Qual Quant
DOI 10.1007/s11135-015-0215-z

The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field

Joshua Kjerulf Dubrow[1,2] · **Irina Tomescu-Dubrow**[1,2]

Abstract Cross-national survey data harmonization combines surveys conducted in multiple countries and across many time periods into a single, coherent dataset. Methodologically, ex post survey data harmonization is especially complex because it combines projects that were not specifically designed to be comparable. We examine the institutional and intellectual history of nine large scale ex post survey data harmonization (SDH) projects in the social sciences from the 1980s to the 2010s. An interdisciplinary methodological field of SDH slowly emerges, facilitated in part by a partnership between academia and government and from the coordinated contributions of social scientists, survey methodologists and computer scientists. While there has been a learning process, it is in terms of accumulated practicalities, and not with the coordination or institutional apparatus one would expect from a 30 year effort.

Keywords Harmonization · Surveys · Methodology · Cross-national · Interdisciplinary

1 Introduction

Cross-national survey data harmonization combines surveys conducted in multiple countries and across many time periods into a single, coherent dataset. It is a generic term for procedures that aim to achieve, or at least improve, the comparability of surveys over time and of surveys from different countries (Granda and Blasczyk 2010; Granda et al. 2010). *Ex post* survey data harmonization is an especially complex process, because it combines

✉ Joshua Kjerulf Dubrow
dubrow.2@osu.edu

[1] Cross-national Studies: Interdisciplinary Research and Training program (CONSIRT), The Ohio State University, Columbus, USA

[2] Institute of Philosophy and Sociology, Polish Academy of Sciences, 72 Nowy Swiat, Room 211, 00-330 Warsaw, Poland

Published online: 09 May 2015

projects that were not specifically designed to be comparable. Though fraught with a daunting methodological complexity, *ex post* harmonization can tap the great wealth of cross-national surveys produced by the international social science community in ways that influence our substantive and methodological knowledge. Yet, the history of cross-national survey data harmonization in the social sciences presents a puzzle. Although harmonization projects have existed for decades, conducted by mix of social scientists, survey specialists, data archivists and computer scientists, harmonization has not become a distinct interdisciplinary methodological field.

We examine the history of survey data harmonization (hereafter, SDH) projects in the social sciences, focusing on *ex post* cross-national survey harmonization. We incorporate knowledge from recent works, such as Burkhauser and Lillard's (2005, p. 9–18) very useful 25 year overview of *ex ante* and *ex post* harmonization attempts, with a focus on its uses for the development of social policy. The history by Granda and his colleagues (2010), while more limited because their paper was designed to present principles of SDH in general—is also insightful.

Unlike interdisciplinary scientific advances elsewhere, SDH projects in the social sciences have proceeded without a coordinated effort (for the example of coordinated biomedical advance, see Marcum (2008). Our review of SDH projects since the 1980s reveals an observable learning process, in terms of accumulated practicalities primarily. In the last 15 years there have also been attempts at theorizing survey data harmonization and developing appropriate methodology, with such exemplary work as that of Hoffmeyer-Zlotnik and Wolf (2003), Minkel (2004), Granda and Blasczyk (2010), Granda et al. (2010). However, a set of general theoretically and methodologically-motivated rules for *ex post* SDH has yet to be established. In the social sciences, *ex post* SDH occurs without an institutionalized or disciplinary apparatus: no journal, no professional association, no academic department, and no research center; survey data harmonization does not even have a handbook. Thus, a methodological field emerges, but without a long-standing coordinated effort to build a comprehensive theoretical and methodological base that one would expect of a 30 year effort.

2 Parameters of our study

SDH in the social sciences has many projects; a complete history is outside the scope of this article. As such, we set the parameters of our study of SDH projects.

2.1 Type of SDH

Depending on whether scholars intend to design a study to collect comparable data, or use existing data not designed a priori as comparative, the literature distinguishes between input and *ex-ante* output harmonization, and ex-post output, or, just *ex-post*, harmonization (Ehling et al. 2006, pp. 1–2, Granda et al. 2010). We focus on the latter. We do not include projects that are primarily *ex ante*, such as the Survey of Health and Retirement in Europe (SHARE, see Börsch-Supan et al. 2013) and the Eurobarometer (for a critical history of that project, see Nissen 2014). There is one exception: We analyze the European Community Household Panel (ECHP), a large scale and expensive *ex ante* SDH project whose premature end resulted in major advances in SDH methodology and provided lessons for existing and future *ex post* SDH projects.

2.2 Scope

We concentrate on large scale projects designed to produce data on a range of research topics with open research questions. They involve multiple institutions—including governments, and especially their financing—and large numbers of researchers and assistants. Typically, these projects produce harmonized data and user manuals for the data, harmonized data and corresponding user manuals, as well as publications on the use of these data for addressing substantive issues. We do not include in our study the second type of SDH projects—those designed by small research teams to answer specific pre-determined research questions, where harmonization is limited to the variables needed to answer those questions.

2.3 Institutions and origins

We focus on the institutional factors behind *ex post* SDH projects and on the intellectual processes that guided their respective harmonization methods. We present the origins of nine studies that, to the best of our knowledge, constitute the population of large scale *ex post* SDH projects. While our focus is on origins, for three of them—ECHP, the Consortium of Household Panels for European Socio-economic Research (CHER) and CHINTEX (see below)—we also examine their full lifespan. Though their lifespan short, they had a long-term impact on subsequent harmonization endeavors.

A key institutional factor in the development of SDH into an interdisciplinary methodological field is the partnership between academia and government. Academic research programs and national statistical institutes may have different motives and goals, but over the decades both have clearly come together to take on the challenge of SDH. Governments, of course, provide much of the funding for large scale academic projects. In their *ex ante* work, national statistical institutes have a long history of working together to provide macro-level measures of cross-nationally equivalent economic and social phenomena.[1] National statistical institutes, and the governments of which they are a part, contribute heavily to survey data harmonization—ECHP, CHER, European Union Statistics on Income and Living Conditions (EU-SILC), CHINTEX and Harmonized European Time Use Study (HETUS) are all large scale projects of this kind. Academics are also essential, as they provide much of the creativity and expertise that further the SDH goals of government statistical agencies. Though this partnership can be rough, overall, SDH could not have proceeded without it.

2.4 Interdisciplinarity

The potential for interdisciplinarity—the integration of two or more disciplines that produces emergent, original, interesting and useful knowledge—comes to fruition when the relationship of the integrated disciplines is strong and symmetrical (Wagner et al. 2011; National Academies 2004; Dubrow 2011). The strength and symmetry of disciplinary relationships depends on the relative similarity in objects of study, methods of inquiry, the size of the profession and its disciplinary apparatus, and its history (Dubrow 2011; Dubrow and Kolczynska 2015; see also Jacobs 2014).

The social sciences contain many instances of unrealized potential, as is the case of sociology and American Studies (Dubrow 2011; Jacobs 2014) and in political sociology

[1] We are thankful for the anonymous reviewer who provided the above insights.

(Dubrow and Kolczynska 2015). By contrast, SDH realizes this potential for interdisciplinarity, it creates a synthesis of knowledge from economists, sociologists, political scientists, survey specialists and data archivists, from academia and from government. The successful collaboration between types of actors coming from different disciplines and spanning decades has yet to produce a coherent institutional apparatus around SDH.

As this is also story of interdisciplinarity, we note the subject orientation of large scale *ex post* SDH projects and the academic disciplines of their intellectual leaders. The interest and respect for cross-national survey methodology unites discipline-diverse scientific teams within projects and is a major force that has thus far allowed for the emergence of an interdisciplinary methodological field.

2.5 Methods and presentation

We present the history of cross-national, *ex post* SDH in the social sciences chronologically, to show how SDH evolved, and how earlier projects influenced later ones. As with other histories of methodologies and international projects (e.g. Platt 1983; Friendly and Denis 2005; Nissen 2014), ours is based on archival research: in addition to journal articles, much of the archives we examine are available on the Internet. As such we use websites and reports of research projects and workshops reported there. Insomuch as history is designed to be learned from, in this article we attempt to draw lessons from the choices made by SDH projects for those who wish to initiate their own, large scale SDH project.

3 Methodological challenges of SDH

SDH dates back to the early 1980s, but it is safe to say that most social scientists are unfamiliar with the literature on it. As Van de Vijver and Matsumoto (2011, p. 6) put it, "Understanding why any study is to be conducted in the first place leads to questions about how to conduct it, which is a discussion in the realm of research methodology." Thus, we begin with a brief outline of the basic research problem all cross-national survey research faces, and how SDH attempts to deal with it. Then, we go in-depth on the definition of SDH, and outline the methodological problems in actually conducting an SDH project.

To analyze cross-nationally the relationship between any two major concepts requires data with individual- and country-level measures, and substantial between-country variation. If we examine contemporary societies with the European Social Survey, for example, we can say little about the influence of the political system as measured by level of democracy, since the majority of countries in ESS have the same high scores on democracy. To increase between-country differentiation, we typically turn to different international survey projects. Some are regionally focused, such as the Afrobarometer or the Latino Barometer. Others are worldwide, but include few countries, such as the International Social Survey Programme. The World Values Survey (WVS), the broadest project in terms of country inclusion, has its own shortcomings, pertaining, among others, to undercoverage of historically marginalized regions, such as Sub-Saharan Africa, Southeast Asia and the Middle East and limits on the measurement of major concepts.

Cross-national SDH proposes a different solution to overcoming limitations inherent in any single cross-national project: rather than analyzing separately several surveys, it combines, *ex post*, information from different projects into an integrated dataset with

comparable measures. Harmonization has many benefits, and chief among them is the potential to increase cross-national variation, and thus lead to new substantive insights into human attitudes and behaviors.

Data harmonization is a generic term for procedures that aim to achieve, or at least improve, the comparability of surveys over time and of surveys from different countries (Granda and Blasczyk 2010; Granda et al. (2010). We can generically say of *ex post* SDH that it is a process (a) in which different survey datasets that were not specifically designed to be compared are pooled and adjusted (i.e. recoded, rescaled, transformed) to create a new integrated dataset that could be analyzed as a typical single-source dataset; and (b) that is based on clear criteria that specify which datasets are included into the new dataset, and clear methods for how variables in the new dataset are created.

To identify all of the methodological challenges inherent in SDH, its best to should start with the overarching methodological challenge in data comparability, and then recognize that there are numerous methodological problems and room for error at each step of the harmonization process (see also Smith 2011). In harmonization, this means moving from source variables—the original variables in the datasets of particular surveys—to target variables, i.e. the harmonized, common variable produced from the source variables. The challenge of cross-national SDH is to produce meaningful data that have accounted for all of the error produced in what Granda and Blasczyk (2010) refer to as the data lifecycle. This lifecycle begins at the initial data source (e.g. each country involved in the international survey research project) to the harmonization decisions undertaken by the SDH project (creation of the target variables), to data cleaning of the final master file (the harmonized data). Thus, not only do SDH projects inherit the errors of the initial data source, but they may create their own in the harmonization process. In short, SDH is an especially daunting endeavor.

4 The beginning of large scale SDH: the 1980s

One of the earliest attempts to integrate data from different extant surveys, and perhaps the most successful, is the Luxembourg Income Study, now simply called LIS. The idea of LIS was generated by a conference on the topic of poverty in cross-national perspective, held in Luxembourg in 1982 (for a detailed history, see Smeeding et al. 1985, p. 2–4).[2] Smeeding et al. (1985, p. 2) describe how the idea of LIS took shape:

> Some participants in that conference were highly experienced in the microanalysis of income distribution data sets for their own countries. It became apparent during the conference deliberations that it would be possible to pool the knowledge and experience in these various countries to create internally and externally consistent data sets for comparative studies which are far superior to those currently in existence.

The government of Luxembourg agreed, and from their sponsorship, LIS was born. From 1982 until 2006, Smeeding, a PhD in Economics from the University of Wisconsin-Madison, directed the project.[3] At the outset, LIS intended to include social indicators other than income, and to be useful in both basic and applied "social economic research" (Smeeding et al. 1985, p. 2–3). At that point, LIS was in its "experimental stage," and

[2] http://www.lisdatacenter.org/wps/liswps/1.pdf. Accessed February 7, 2014.

[3] http://www.lafollette.wisc.edu/facultystaff/smeeding/smeeding-timothy-cv.pdf. Accessed February 7, 2014.

envisioned to update and expand its data over time (Smeeding et al. 1985, p. 4). The initial budget did not allow the research team to cover their wider ambitions: "There was only enough money to conduct a feasibility study on the seven datasets currently included" (Smeeding et al. 1985, p. 5). Two years later, LIS claims to have "moved beyond the initial experimental stage" (Rainwater and Smeeding 1987, p. 5).[4]

It was 1983, and LIS attempted to codify the procedures designed "to produce the final harmonized LIS data set" (Smeeding et al. 1985, p. 4). For each dataset under consideration, LIS identified an expert who knew that data well and asked them about the strengths and weaknesses of that data "for the type of comparable social policy analyses" they hoped to do (Smeeding et al. 1985, p. 4). As Smeeding et al. (1985, p. 4–5) describe it, LIS tried to obtain a representative sample of suitable datasets, they developed criteria for inclusion: timeliness,[5] data quality,[6] sample size, income accounting unit,[7] and geographical location. These data were collected by government agencies or in close collaboration with them and their national statistical agencies. Some governments were suspicious of the LIS project, and refused to share some of their datasets (Smeeding et al. 1985, p. 5). Datasets that were available but did not meet the criteria were excluded. The first version of the LIS database contained data from a handful of Western countries—USA, Norway, Canada, UK, Germany, and Sweden—and Israel.

Ex post harmonization was the goal, and nothing was more troubling than income, their main variable of interest. "It should be stressed," Smeeding et al. (1985, p. 11) write of income harmonization, "that international comparability and consistency rather than perfection is our goal." The point was to make a generally comparable dataset that also allows individual researchers to decide on what to do with certain anomalies.

Being aware of imperfections of existing datasets, the LIS team put a lot of effort in understanding ways in which "relative data quality" (p. 18), as they called it, could affect the comparability of income measures between countries. This included not only constructing a conceptual model of levels of income reporting, but also an analysis of the quality of administrative data—that was used for comparison—to make necessary adjustments for income definitions and populations.

At the outset of the project, with almost no methodological literature on SDH to draw upon, it was difficult to anticipate all the problems that seemingly simple variables would pose to harmonization. In the early years, LIS spent a lot of time thinking how to get at comparability, and considered many elements of harmonization. As the data developed over the decades, their harmonization procedures have apparently satisfied its many users. In the early 1980s, though, LIS did not look as deeply at the problems in harmonizing

[4] http://www.lisdatacenter.org/wps/liswps/12.pdf. Accessed February 7, 2014. Harmonization speeds forward via technological advance. Since its inception, LIS faced a major problem in supplying its data to interested researchers. By 1987, LIS began to solve the problem of data access through the then new system called BITNET, described as "an electronic mail and file transfer network" that linked about 400 academic and research institutions around the world (Rainwater and Smeeding 1987, p. 9). It was the early internet.

[5] Data had to be available for 1979, the baseline for all datasets. They transgressed that rule by including 1981 data from Germany and Canada, which they justified on practical grounds (see Smeeding et al. 1985: Footnote 3, p. F-1).

[6] "…(as measured by response rates and other indicators of nonsampling error)" (Smeeding et al. 1985, p. 4). They begin to discuss data quality in detail on p. 18.

[7] Household, family or both. This problem required some harmonization of what "household" versus "family" means.

demographics as they had with income. Of demographics, LIS claimed then that "most of these variables are self-explanatory" (Smeeding et al. 1985, p. 16). The Appendix in which they listed these demographics did not contain much more information. Education, for example, has the following comment: "If not coded in years, recoded as minimum years needed to attain given level of education" (Smeeding et al. 1985, p. 40). It is not clear why, at that time, they did not use the International Standard Classification of Education (ISCED), which was created by UNESCO in 1976.[8] The lesson to draw, here, is that in SDH, the best approach is to assume that all variables will pose major difficulties.[9]

While LIS was getting off the ground, scholars interested in the concept of "time use" started to consider how to compare all of the Time Use Studies (TUS) conducted in various countries, past and present.[10] The resulting project, named the Multinational Time Use Study (MTUS), has its roots in the 1970s, but only took shape as a harmonized time use study in the 1980s (for a detailed history, see MTUS User's Guide 2013, Chap. 2). Jonathan Gershuny, Ph.D. in History and Social Studies of Science from the University of Sussex, and now Professor of Sociology at Oxford University, led MTUS from the beginning. The European Foundation for the Improvement of Living and Working Conditions (EFILWC), an agency of the European Union, paid for the initial release of MTUS; the collaboration between MTUS researchers and the EU led to the Harmonized European Time Use Study, or HETUS (see below).

MTUS is based on time use diaries, though they did collect sociodemographics, as well. As MTUS (MTUS User's Guide 2013, Chap. 1: 5) explains:

> The time use diary is a narrative account and not a series of quantitative questionnaire answers. People can give a full account of their day without necessarily completing all columns of the diary for all potential time slots…

The early MTUS lacked some basic methodological rigors—"in early versions, limited or no efforts were made to clean data," and "some diaries originally included as good quality diaries now are classified as low quality diaries not suitable for more analysis" (MTUS User's Guide 2013, p. 25)—that MTUS now does thoroughly. They follow their own harmonization rules which they consider as appropriate to TUS. For example, MTUS refuses to impute data: "We also do not impute data based on estimations from what similar people do on similar days. All adjustments to diaries work from information that the participant supplies" (MTUS User's Guide Chapter 1: 6). This sets the TUS studies apart from other survey harmonization projects.

[8] Classifications Newsletter, Number 27, August 2011, p. 5. ISCED was not updated again until 1997, and then again in 2011. http://www.uis.unesco.org/Education/Documents/UNSD_newsletter_27e_ISCED.pdf. Accessed February 7, 2014. Sometime afterward, LIS began to also use ISCED http://www.lisdatacenter.org/wp-content/uploads/standardisation-of-education-levels.pdf. Accessed February 7, 2014.

[9] The LIS harmonization process is described here: http://www.lisdatacenter.org/wp-content/uploads/our-lis-documentation-harmonisation-guidelines.pdf. According to the LIS Guidelines (p. 12): "Socio-demographic variables… are all individual level variables and report the major socio-demographic characteristics of the household members." They are: Living arrangements, demographics, immigration, health, and education. LIS' wealth data also has harmonization guidelines, but as of this writing, they are not available on the website. Instead, they have a document on "behavioral variable mapping" http://www.lisdatacenter.org/wp-content/uploads/2011/02/behavioural-variable-mapping-2011-03.pdf.

[10] MTUS has time use data from the 1960s to the present.

5 SDH becomes popular: the 1990s

There were three significant SDH projects of the 1990s. One was the Cross-national Equivalent File (CNEF) (http://cnef.ehe.osu.edu/). The CNEF project, currently directed by Dean Lillard, PhD in Economics from the University of Chicago, was initially directed by Richard Burkhauser, a PhD in Economics from the University of Chicago, and Gert G. Wagner, a PhD in Economics from Berlin University of Technology. CNEF is simultaneously based on the successful LIS model[11] and designed to overcome some of the problems of LIS. As Frick et al. (2007, p. 628) write,

> While the standardized LIS data are impressive, they cannot meet some goals of the cross-national research community. For example, the LIS allows researchers only indirect access to the underlying confidential microdata which in several cases is official data. Further, researchers cannot easily get access to the original data sources. This limitation means that most researchers must accept the LIS standardization rules. Finally, and perhaps most importantly, the LIS data are cross-sectional, and so do not serve researchers interested in longitudinal analyses.

CNEF harmonizes household panel studies (such as the US Panel Study of Income Dynamics PSID and British Household Panel Study BHPS). It began in 1990 as a project funded by the U.S. National Institute on Aging and administered by Cornell University. The idea was to create a cross-national harmonized panel dataset dating back to 1980.[12] The US academics behind the development CNEF (and LIS, too) had first to convince the creators of the government funded country data sets to be partners, either for free or even at some additional cost to them. This was done to secure additional funding to *ex post* harmonize their data and to collectively increase the value of each of their country datasets for cross-national research. The academics also had to convince a funding agency that the additional cost had sufficient scientific value that it should be funded by that agency. At first, without much of a track record of *ex post* SDH data other than LIS, this was difficult to do. Thus, initially, CNEF included only Germany and the U.S. By 2007, with CNEF building an impressive track record of ISI publications, it expanded to six countries. In the early years, variables were limited to income and demographics.[13]

Unlike LIS, CNEF was designed to be developed and enhanced by its user community. As Frick et al. (2007, p. 629) write,

> Equivalently defined variables are added when researchers develop cross-nationally comparable measures as part of a particular research project. … Consequently, the harmonized data included in the CNEF are an amalgam of the knowledge of many researchers answering a diverse set of questions. Just as importantly, the CNEF continuously evolves as researchers refine and add to the set of harmonized variables.

CNEF can be called a bottom–up approach, with users having strong say in the direction of CNEF's target variables, as opposed to LIS' top–down approach. When it comes to top–

[11] Burkhauser and Lillard (2005, p. 10).

[12] It has since been dated back to 1970 http://www.human.cornell.edu/pam/research/centers-programs/german-panel/cnef.cfm. Accessed February 11, 2014.

[13] Frick et al (2007, p. 630) write: "The original core variables to be harmonized were income and demographic characteristics of respondents to the PSID and the German Socio-Economic Panel SOEP, and reflects the objectives of the original project that motivated the creation of the CNEF to compare and understand income-based inequality and income mobility in the US and Germany."

down or bottom-up in SDH, there are no ideal types, as LIS uses its working papers to understand how users use the data (with the assumption that LIS makes adjustments to its dataset based on use patterns). According to Burkhauser and Lillard (2005, p. 12), successful SDH of panel studies have been driven and implemented by active researchers, not by government bureaucrats[14], and they do so by identifying theories and research questions before they harmonize.[15] All of these efforts, however, are dependent on government funds.

6 Lessons from a troubled project: the European community household panel (ECHP)

The long-term success of SDH projects depends on sustained (a) access to high quality country data that continue to be updated, (b) funding from external sources for the harmonization of those data and the research time to work on it, (c) research use of that data, and (d) evidence of the scientific value of that research, such as articles in highly regarded peer reviewed journals. All these elements are interrelated and a shortcoming in any is likely to result in the end of the project.[16]

LIS and CNEF followed this model. Their intellectual cores—Rainwater (sociology) and Smeeding (economics) for LIS and Burkhauser (economics) and Wagner (economics) for CNEF—were leading researchers in their respective disciplines prior to taking on these projects. After they obtained funding, they published heavily in ISI journals to show the value of their data for cross-national analysis, and for raising additional funds for outreach efforts to potential new users, especially PhD students. From the outset, these scholars sought to harmonize the data in a way that would facilitate comparisons of the US with European countries. They judged, wisely, that this was the best way to interest American funding agencies and American researchers in the data. (It also helped that other major US scholars in the field published well-regarded articles based on the data). Their approach made it likely that research from this work would be published in US based journals, furthering the virtuous cycle of scientific value.

Initially an idea from the National Institutes of Aging, CNEF's continuation was built also on an explicit desire to inform public policy. Burkhauser and colleagues published in ISI journals as an example of the power of cross-nationally harmonized data to answer public policy questions. They then secured a National Institutes of Health grant from the National Institute of Aging to study transitions out of the workforce of older workers in Germany and the USA. They also promised to create a cross-national data set—CNEF—as a product of that research.[17]

The European Community Household Panel (ECHP), a large scale SDH project run by the European Union's Eurostat beginning with the mid-1990s, came to a premature end in

[14] "The effort to harmonize existing panel studies share one significant organizational feature: active researchers conceived, planned, and carried out how the data would be harmonized. While data managers, some in government statistical agencies, were often involved in the process, it was researchers who decided how to define equivalently the variables of interest."

[15] Burkhauser and Lillard (2005, p. 12): "Researchers guided by theory and concepts flowing from the research pertinent to the object of their studies are best able to make the assumptions necessary to harmonize data across countries."

[16] We are thankful to one of the reviewers who formulated these four principles of sustained SDH success.

[17] Again, we are thankful for the reviewer who pointed this out to us.

2001. Though an *ex-ante* project, ECHP's demise had far reaching and long lasting implications, one that provides lessons for present and future SDH projects.[18] As Burkhauser and Lillard (2005, p. 14) explain,

> Led by Eurostat, the ECHP attempted, by using a common survey instrument, to create a set of country based data sets that were comparable across countries. The ECHP goal was to create comparable panel data for all European Union … countries.

Burkhauser and Lillard clearly see ECHP as a failure, and describe the reasons they think it failed (2005, p. 14):

> The ECHP was plagued by problems from the outset. In part these problems may have arisen because the ECHP was developed by Eurostat and implemented by each country's statistical agency with little or no consultation with the research community. … end users played a minor role in the creation and implementation of the survey instrument. Most troubling, the ECHP project failed to utilize the long experience of researchers who were running mature panel surveys in EU countries.

The listed ECHP's main failures:

> Long delays in processing, Problems with initial responses, Problems with attrition rates, Non-uniform implementation, Lack of input from the research community in design and response to users over time, Initial failure to take advantage of existing panels, Poor dissemination strategy to get the data to the international research community and High costs of use for individual researchers (14–15).

CHER, an EU funded project (see below), also had many complaints about ECHP data, mostly about the lack of comparability with existing datasets. As CHER (2003: 7) described it:

> The lack of longitudinal data that are at the same time comparable, well documented and closely related to relevant macro and meso information and of user-friendly access, has truly regrettable consequences. The potential for a cross-national database to compare the situation in one country with those in other countries is not sufficiently used, and comparative analysis of European issues is still underdeveloped.

Though maligned by subsequent projects, Eurostat—who was in charge of ECHP—has a different spin on how ECHP ended. Eurostat (2012: 11) sees ECHP in somewhat heroic terms, retired only because the nature of its agreement had ended (in a gentlemanly manner) and because of the changes to the EU's political circumstances. They explain:

> The EU commitment to fighting social exclusion was confirmed with the formation of 'The European Community Household Panel (ECHP)' a pioneering data collection instrument. Launched on a gentleman's agreement basis in 1994 it expired in 2001. However the political scene has changed, notably with the introduction of open method of coordination in the fields of social inclusion and pensions reform. Other important changes included enlargement of the EU from 15 to 25 states (and demands for coverage of other neighbouring countries), and the publication by the United Nations expert group on household income statistics of a detailed report and

[18] For more discussion on the shortcomings of ECHP, see Burkhauser and Lillard (2005, pp. 14–15); CHER 2014, pp. 6–7; and undated CHINTEX document, p. 5; for more information on ECHP (2011), see the EuroPanel Users Network. http://epunet.essex.ac.uk/echp.php.html. Accessed February 13, 2014.

> recommendations. In recognition of these changes, the ECHP is being progressively replaced with data collection under the EU-SILC regulations…

The end of ECHP caused difficulties for Eurostat and the users of ECHP data. The demise of ECHP was also the end of EPAG, the European Panel Analysis Group built to use ECHP data, and the EuroPanel Users' Network (EPUNet), created by EPAG to disseminate information and coordinate analyses of ECHP data.

ECHP's end also caused an existential crisis for the EU's harmonization efforts, *ex ante* in particular. Is *ex ante* possible? Is it worth the expense?

To answer these questions, the EU created CHINTEX. CHINTEX is an improbable acronym that somehow stands for, "The Change from Input Harmonization to Ex-post Harmonization in Samples of ECHP—Implications on Data Quality." CHINTEX began in 2000, was funded for three and a half years by the European Commission, and was coordinated by Germany's Federal Statistical Office.[19] It arose out of a problem posed by the end of ECHP, when a few countries pulled out of the *ex-ante* collaboration on which ECHP depended. In short, the point of CHINTEX was to examine whether *ex ante* efforts were worth the time and expense. According to a document that summarizes CHINTEX (date unknown, p. 1):

> It is the overall objective of CHINTEX by means of this unique data situation to clarify if it is necessary to have centralised, standardised survey instruments to achieve harmonisation and comparability or if this objective can also be achieved by ex-post harmonisation, by which independent national sources are satisfactorily converted to common concepts, definitions, survey questions etc.
>
> Secondarily, they sought to better understand the issue of data quality.[20]

In their final conference in May 2003, CHINTEX attempted to explain what they learned from studying ECHP's methods and how this knowledge can be applied to its successor project, EU-SILC (and, presumably, other such projects).[21] CHINTEX's main contribution was to offer a wide range of methodological guidelines for SDH, on topics such as total survey error, panel effects, weighting and imputation, in the form of working papers and PowerPoint slides. CHINTEX cost 1.3 million Euros.[22]

CHINTEX's "post-mortem" methodological analysis of ECHP is the rare trove of methodological literature on SDH. In analyzing the feasibility of ECHP years after it was created, CHINTEX (2003) contributed to what is becoming a theory of data harmonization.[23] Separately, Granda et al. (2010) also contribute to this on-going effort.

[19] https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/FinalConference/Downloads/Hahlen.pdf;jsessionid=7E7753BEE731A012B3D321FA73D189F0.cae3?__blob=publicationFile. Accessed February 13, 2014.

[20] "Furthermore, the project investigates important hypotheses about the data quality of panel surveys (non-response, reporting errors and panel effects) which are of general interest for survey statisticians." (CHINEX data unknown, p. 1).

[21] https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/FinalConference/Einfuehrung.html. Accessed February 14, 2014.

[22] See CROS, a website maintained by the European Commission. http://www.cros-portal.eu/content/chintex. "The CROS Portal is a content management system based on Drupal and stands for "Portal on **C**ollaboration in **R**esearch and Methodology for **O**fficial **S**tatistics". http://www.cros-portal.eu/page/about-cros-portal. "The European Commission maintains this website to enhance public access to information about its initiatives and European Union policies in general." http://www.cros-portal.eu/page/legal-notice.

[23] "One objective of this chapter is to establish a simple framework of harmonisation which is useful to pinpoint the research issues of CHINTEX and to highlight the differences to other issues of harmonisation".

6.1 Was ECHP a failure?

The claim that ECHP was a failure is debatable. Those suggesting or outright declaring it as such—CNEF, CHER, and CHINEX—present a litany of complaints about its administration. Further evidence is that two countries pulled out of ECHP prematurely, which likely encouraged its end. Eurostat counters by arguing that ECHP would have ended, anyway, because the gentleman's agreement was due to expire at that time. Of course, since ECHP was allegedly the product of a gentleman's agreement, there can be no definitive paper-trail that would prove it was to end when it did. Perhaps it was, as Eurostat seems to suggest, a pilot study designed to create its successor project (EU-SILC), but we do not know of documentation from ECHP's beginning that suggests that it was just a big pilot study.

A counter argument—that ECHP was useful, though limited—can be based on how many academic products it resulted in, i.e. its scientific value. Setting aside the issue of whether the monetary costs of ECHP were worth it, to establish this argument we would need comparable information on academic products of each major SDH project, e.g. the number of articles that are based on LIS, CNEF and the like. Such data do not exist, mainly and unhappily because there is no requirement of data users to actually cite the data that they are using. SDH projects encourage users to cite the data, but there is no enforcement. Still, we can make a rough sketch of popularity or, at least, name recognition, based on mentions of the SDH data in major research article databases.

We did this for ECHP. Using the ISI Citation database, under "topic", there were 245 mentions in articles of the search term, "European Community Household Panel." This number is similar when searching in EBSCO and SCOPUS. It would be unreasonable to assume that these are all negative citations of the data, considering that 13 of these articles were published in 2013 alone, and all of them are empirical articles that faithfully use the dataset. To call ECHP a failure, then, is to ignore a very important aspect of any SDH project: That scholars use the dataset to produce scientific knowledge. Clearly, ECHP has been successful in this regard.

7 The EU builds HETUS

While ECHP was signaling the administrative difficulties in *ex ante* harmonization across national statistical agencies, another European project, HETUS, the Harmonized European Time Use Surveys, embraced it. HETUS, however, did not go all in on *ex ante*, but rather combined *ex ante* and *ex post* harmonization. HETUS began in the 1990s when "the need for increased comparability became recognised" (Eurostat HETUS 2000: Preface). In 1996–1997, Eurostat funded a series of pilot-level time use studies in 18 countries, designed to test the feasibility of harmonization. In 1998, the Statistical Programme Committee (SPC) recommended that "harmonisation of time use data was feasible despite recognised national differences" (Eurostat HETUS 2000: Preface). In 2000, around the time that the ECHP project was drawing to a close, Eurostat drew up harmonization guidelines for the Time Use Studies (TUS). Of TUS harmonization, Eurostat lamented its cost: "The chosen survey design is rather expensive, and in some cases it might be somewhat more expensive than a non-harmonised national design would have been" (Eurostat HETUS 2000: 8). To date, HETUS has 15 "comparable countries."[24]

[24] https://www.h2.scb.se/tus/tus/introduction1.html. Accessed February 14, 2014.

8 SDH booms: the early 2000s

The early 2000s saw the maturation of LIS, CNEF and HETUS, the end of ECHP, and the creation of new SDH projects. An early project was the Consortium of Household Panels for European Socio-economic Research (CHER). CHER was initially funded by the European Commission for over one million Euros between 2000 and 2003, and coordinated by CEPS (2014), a research bureau in Luxembourg.[25] CHER is substantively similar in its harmonization aims as CNEF, namely the harmonization of already collected panel data.

Unlike the beginning of CNEF, the aims of CHER at the outset were much more ambitious. In 2003, CHER reported to the European Commission that it aimed to collect and harmonize panel data on "demography, health, education and training, employment and activity, income and expenditure, housing and household durables, subjective information and social relations" (CHER 2003, p. 6). CHER grew out of frustration with the ECHP data, and was built to overcome its shortcomings. By 2003, CHER expanded into many more countries than CNEF and had data dating back to the 1980s. CHER had eclipsed CNEF in terms of country and topic coverage, but the project ended in 2003, and was not updated.[26]

Although CHER was designed to overcome the problems of ECHP, EU-SILC has been named (by Eurostat) as its true successor.[27] Like CNEF and CHER, EU-SILC is *ex-post* harmonized panel data. It was formally created in 2004 and run by Eurostat[28] with data dating back to 2005. Unlike any other harmonization project, however, EU-SILC "is organised under a framework regulation and is thus compulsory for all EU Member States" (Eurostat 2009, p. 13), and as such is expressly designed to inform European Union economic and social policy. Thus, it has data for all EU countries, plus others in or near the European continent.

By the 2000s, perhaps the largest *ex post* cross-national SDH projects run by political scientists and sociologists is the International Stratification and Mobility File (ISMF).[29]

[25] http://ec.europa.eu/research/social-sciences/projects/010_en.html. Accessed February 13, 2014. According to CEPS' website, "CEPS/INSTEAD is a centre of reference for research in the social and economic sciences in the Grand Duchy of Luxembourg, a public institution under the jurisdiction of the Ministry of Higher Education and Research." http://www.ceps.lu/?type=module&id=53. Accessed February 13, 2014.

[26] There is no official information about the fate of CHER. From existing documents, it seems to have ended in 2003 when they submitted their final report to the European Commission. To get more information, and since CHER was a CEPS project, I emailed a colleague to ask when CHER ended. They wrote, "The project has been on for a while as an experiment to test whether various national level panel studies could be homogeneized and merged in one big dataset. I don't know exactly the date when the CHER was closed. Probably it was 2003, but the data it contains should be older, probably dating to 1999. As a matter of fact, it was a nice data-set but I've never used it because it was too old. Not by chance only 16 papers have been produced with CHER." Anonymous, Personal Communication, February 13, 2014.

[27] As of April 2006, EPUnet "is now finished," according to its website. "Jean-Marc Museux from EUROSTAT has produced three papers on EU-SILC, the successor to the ECHP". http://epunet.essex.ac.uk/view_news.php%3FID=36.html. Accessed February 13, 2014.

[28] "The EU-SILC project was launched in 2003 on the basis of a "gentlemen's agreement" in six Member States (Belgium, Denmark, Greece, Ireland, Luxembourg and Austria) and Norway. The start of the EU-SILC instrument was in 2004 for the EU-15 (except Germany, the Netherlands, the United Kingdom) and Estonia, Norway and Iceland." http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc. Accessed February 13, 2014.

[29] Most SDH is about economics and was run by economists. Only the TUS, ISMF, GBS and now, the Harmonization Project, are not explicitly about economics; they are run by sociologists and political scientists.

The goal of ISMF is to be able to compare social stratification and social mobility patterns across countries and time. By 2009, the data expanded to over 250 surveys from 52 nations, with some surveys dating back to the 1940s. Its focus is on educational and occupational status of both respondents and their parents, and has harmonized demographics, education, employment status, occupation and income.

The ISMF is the rare attempt to create a harmonized datafile of cross-national, cross-sectional surveys stretching far back in time. ISMF is run by Harry. B. Ganzeboom, Donald J. Treiman, and Elizabeth Stephenson of University of California-Los Angeles. Ganzeboom and Treiman have been at the forefront of measurement standardization of stratification variables—occupation, and now education[30]—for decades. The ISMF does not have a public official history and it is not clear when it began. Most current information on ISMF is from the personal website of its co-creator, Ganzeboom, PhD in Social Sciences from the University of Utrecht (Netherlands).

The ISMF seems to have grown out of the needs of the social stratification and mobility research community, particularly those from the International Sociological Association's Research Committee (ISA RC) 28, "Social Stratification," for comparable cross-national and cross-time data. In 1990, Treiman and Ganzeboom complained about the poor attention given to comparability issues between empirical studies of status attainment based on surveys conducted in different nations (pp. 109, 116–117). In a sharply worded rebuke, Treiman and Ganzeboom (1990) wrote of the consequence of poor comparability: "Results obtained by analysts who have not troubled to standardize their samples, their measurement instruments, and their models must be dismissed as simply meaningless" (116). In 1990 no such data existed, and their solution was to standardize analyses with common variables and methodological approaches. "Because no one analyst will command the data or the resources to carry out the definitive cross-national comparison of status-attainment processes," they wrote, "our only alternative … is to generate the necessary information collectively" (116). Sometime between 1990 and 1993, they created the ISMF.[31]

From publicly available information it is not clear whether ISMF emerged as an ad hoc way of addressing the data comparability problem via SDH or was, at the outset, a large scale project. It is also not clear what harmonization procedures its researchers employed, or the extent to which they drew on the growing methodological research on SDH. They do not provide any document that discusses their theory of SDH. In fact, most of what they say is this[32]: "…we do provide researchers with some important tools for harmonization efforts, in particular maps to recode occupation and education categories into internationally and historically comparable metrics." After, they provide a long list of surveys with links to syntax for recoding in statistical software packages.

It does not seem as if ISMF has been publicized in the same degree as LIS and the other SDH studies. JSTOR lists 11 articles containing the phrase, "International Stratification and Mobility File". It is not clear whether ISMF is well publicized even within RC28. In 2006, Hout and diPrete (2006) offer a history of accomplishments of the International Sociological Association's RC28 committee, in which they only briefly mention that the

[30] Schröder and Ganzeboom (2013), "Measuring and Modeling Levels of Education in European Societies". European Sociological Review.

[31] Of the founding date, the only evidence we've found is that, in 1993, Ganzeboom and Treiman published a paper in which they reference the ISMF (Ganzeboom and Treiman 1993, p. 470). In a book chapter in the *International Handbook of Sociology* (2000) Treiman and Ganzeboom describe the history of social stratification and mobility research. They presented a clear trend toward cross-national surveys and attempts at standardizing measurement of stratification variables (see also Ganzeboom and Treiman 2000).

[32] http://www.harryganzeboom.nl/ISMF/index.htm. Accessed February 14, 2014.

ISMF confirms a finding on the topic of welfare states, labor markets and occupational mobility (12). Hout and DiPrete cite a paper by Ganzeboom and Treiman from 2000.[33] The data seem a treasure, but it is not clear why this treasure is not more widely recognized.

9 SDH expands: the 2010s

The 2010s have seen the continuation of CNEF, EU-SILC, and ISMF, as well as MTUS and HETUS. As of 2013, the Survey Data Recycling (SDR) project, which started as the Harmonization Project (see Tomescu-Dubrow and Slomczynski 2014), joined the group of large-scale SDH projects. SDR is led by sociologist Kazimierz M. Slomczynski of the Polish Academy of Sciences and The Ohio State University. At its core, it seeks to harmonize individual-level variables relevant to basic models explaining political protest, while also keeping the possibility of harmonizing variables relevant to other topics open. Equally important to the project are contextual variables that influence individual political attitudes and behavior, as the quality of the source data, and of the harmonization procedures (see Slomczynski and Tomescu-Dubrow 2015). This is the only current harmonization project to focus on these issues.

All SDH projects are ambitious, by definition. SDR's ambition is in the size of its hoped-for datafile, as well as in developing a new approach to dealing with errors and biases that likely accompany the source and the harmonized data. From selected 22 international survey projects, including the WVS, ESS, ISSP, the SDR team pooled 1721 surveys covering ca. 2.3 million respondents from 142 countries/territories measured at various points of time from 1966 to the first quarter of 2014. The objective is to apply formal SDR procedures to create a harmonized database with comparable individual-level measures of political behavior, social attitudes, and demographics, and that contains controls for survey data quality and for quality of harmonization processes. After appending it with country characteristics from non-survey sources, the SDR team plans to publicly share the new data.

10 Some lessons of SDH history

In this article we traced the emergence of the interdisciplinary methodological field of cross-national survey data harmonization in the social sciences. Since the 1980s, there has been substantial thought, money and time spent on SDH in the social sciences. In this section we try to draw some lessons of this history.

10.1 Why does SDH exist?

Methods in the social sciences develop because they are a problem-solving enterprise and they diffuse because other social scientists come to see their usefulness in solving their own problems, such as with focus group interviews (Lee 2010), the participant observation method (Platt 1983), or the invention of the scatterplot graph (2005). Development takes time, and in some cases decades of time, and not all methods survive the development stage (such as the life history approach of the 1930s, chronicled by Thomas 1978). It seems that the idea of SDH

[33] Ganzeboom and Treiman (2000) Ascription and achievement after career entry. Paper presented at the 2000 meeting of ISA's RC 28.

grew as the technological, statistical and other knowledge developed to a point where (a) there came to be a wealth of cross-national survey projects and (b) people think it would be useful if these data could be compared but wonder whether these disparate datasets are, in reality, comparable. Its recent popularity is likely owed to the fact that (c) we finally have the computing technologies to start making such an idea a reality.

SDH integrates disciplines and produces emergent knowledge. However, as of now it has not developed an institutionalized or disciplinary apparatus, and as such SDH is not a supra-discipline (Balsiger 2004, p. 410), or even a new discipline. SDH's main characteristic—that of an unusually complex method—means that it can be considered an emerging methodological field, which has become interdisciplinary out of the necessity of dealing with the many methodological problems that accompany it. These problems are so specific, dire and acute that solving them calls for innovative methodological advances. Despite of—or perhaps because of—this, there is no coherent set of guidelines on how best to solve the severe challenges SDH imposes.

Yet, just because we can harmonize does not necessarily mean that it is useful for the social sciences. On the benefits to science in general, Doiron et al. (2012, p. 1) stated it quite well:

> The benefits of harmonizing and pooling research databases are numerous. Integrating harmonized data from different populations allows achieving sample sizes that could not be obtained with individual studies, improves the generalizability of results, helps ensure the validity of comparative research, encourages more efficient secondary usage of existing data, and provides opportunities for collaborative and multi-centre research.

These benefits seem to reasonably apply to SDH in the social sciences. SDH is akin to interdisciplinarity, the combining of different things that were not designed to be combined in order to produce emergent, original and interesting knowledge. The institutional rationale of SDH is, *prima facie*, the quest for this knowledge. By its nature, SDH is ambitious. Millions of Euros have been spent on it. But there is too little research that answers the questions: To what extent does SDH provide knowledge that could not be gained otherwise? How do harmonized data perform, from methodological and substantive points of view, compared to non-harmonized data? Is *ex ante* harmonization preferable over *ex post* harmonization, is a combined approach most fruitful?

10.2 Where is the methodological literature on SDH in the social sciences?

In compiling this history, it is clear that there is now a large methodological literature on SDH in the social sciences but there is no coherence across projects. The problem is that this literature exists in various documents scattered over place and time, and little of it has been synthesized into a manageable and accessible format. In short, there is a need to pull this literature together, to create a handbook of SDH in the social sciences based on the many existing efforts. We can envision this handbook to start with a theory of SDH and general principles, and to continue with chapters on specific aspects of SDH—dealing with the quality of the source data, weighting, data imputation, assessing the reliability and validity of the target variables, and so on—that synthesize what we know.

There is also a clear need for definitions of SDH and its processes, but such a glossary is only now in the making. Some have written guidelines for data harmonization.[34] An

[34] http://ccsg.isr.umich.edu/harmonization.cfm. Accessed February 7, 2014.

attempt at a glossary of survey data harmonization is available from Granda and Blasczyk (2010).[35] The chapter by Granda et al. (2010) is also relevant. The term "clear" is in development, and existing SDH projects are not so transparent in explaining how they did what they did.[36] While there is a call—and an effort at heeding this call—for better documentation Granda et al. 2010, p. 326), there remain few comprehensive guidelines of the SDH process, especially ones that include details on each step, including the ideas that were abandoned. Granda et al. (2010, p. 326–328) call for the development of software that standardizes the documentation process. We seem to have a long way to go in this regard.

Acknowledgements This work is supported by the project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective", funded by the National Science Centre, Poland, under the grant number 2012/06/M/HS6/00322. We thank Kazimierz M. Slomczynski and Marta Kolczynska for their comments on an earlier draft, and Dean Lillard for advice on the citation of harmonized data.

References

Balsiger, P.W.: Supradisciplinary research practices: history, objectives and rationale. Futures **36**, 407–421 (2004)

Börsch-Supan, A., Martina, B., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S., On Behalf of the SHARE Central Coordination Team: Data resource profile: the survey of health, ageing and retirement in Europe (SHARE). Int. J. Epidemiol. (2013). doi:10.1093/ije/dyt088

Burkhauser, R.V., Lillard, D.R.: The contribution and potential of data harmonization for cross-national comparative research, DIW Diskussionspapiere, No. 486. http://www.econstor.eu/bitstream/10419/18337/1/dp486.pdf (2005). Accessed 12 Feb 2014

CEPS/INSTEAD: http://www.ceps.lu/?type=module&id=53 (2014). Accessed 13 Feb 2014

CHINTEX: Final conference: harmonisation of surveys and data quality. https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/FinalConference/Einfuehrung.html (2003). Accessed 14 Feb 2014

Classifications Newsletter: United Nations Statistics Division (UNSD) Number 27 (August). http://www.uis.unesco.org/Education/Documents/UNSD_newsletter_27e_ISCED.pdf (2011). Accessed 7 Feb 2014

Cornell University User Package for the Cross-National Equivalent File (CNEF): http://www.human.cornell.edu/pam/research/centers-programs/german-panel/cnef.cfm (1970–2009). Accessed 11 Feb 2014

Cross-National Equivalent File (CNEF): http://cnef.ehe.osu.edu/ (2014). Accessed 22 Feb 2014

ČSDA: Program of Workshop on Harmonisation of Social Survey Data for Cross-National Comparison will be held in Prague on Tuesday 19th October http://archiv.soc.cas.cz/articles/cz/84/workshop.html (2010). Accessed 22 Feb 2014

Doiron, D., Raina, P., Ferretti, V., L'Heureux, F., Fortier, I.: Facilitating collaborative research—implementing a platform supporting data harmonization and pooling. Norsk Epidemiologi **21**(2), 221–224 (2012)

Dubrow, J.K.: Sociology and American Studies: A Case Study in the Limits of Interdisciplinarity. Am. Sociol. **42**(4), 303–315 (2011)

Dubrow, J.K., Kolczynska, M.: A quem pertence o estudo da democracia? Sociologia, ciência política e a promessa da interdisciplinaridade na Sociologia política desde 1945 (Who Owns the Study of Democracy? Sociology, Political Science, and the Interdisciplinary Promise of Political Sociology since 1945). Sociologias **17**(38), 92–120 (2015)

Ehling, M., Rendtel, U., et al.: Synopsis. Researh Results of Chintex - Summary and Conclusions. http://destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/Downloads/Synopsis.html (2006). Retrieved 15 Feb 2015

[35] http://ccsg.isr.umich.edu/pdf/13DataHarmonizationNov2010.pdf. Accessed February 7, 2014.

[36] This is a problem with social science in general: Social researchers rarely keep good records on the research process and are reluctant—for whatever reasons—to share enough the ups-and-downs of their scientific pursuit.

Elias, P.: 'Big Data' and the social sciences—a perspective from the ESRC. http://www2.warwick.ac.uk/fac/soc/economics/research/centres/cage/events/conferences/peuk/peter_elias_big_data_and_the_social_sciences_pe_final.pdf (2014). Accessed 6 Feb 2014
EPUNET: http://epunet.essex.ac.uk/view_news.php%3FID=36.html (2014). Accessed 13 Feb 2014
EuroPanel Users Network: http://epunet.essex.ac.uk/echp.php.html (2014). Accessed 13 Feb 2014
European Commission website: Consortium of household panels for European socio-economic research (CHER) http://ec.europa.eu/research/social-sciences/projects/010_en.html (2014). Accessed 13 Feb 2014
European Community Household Panel (ECHP): http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/echp (2011). Accessed 18 Aug 2011
European Union Statistics on Income and Living Conditions (EU-SILC): http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc (2014). Accessed 13 Feb 2014
EUROSTAT: Harmonised European time use surveys. Methodologies and working papers. http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-RA-08-014 (2009). Accessed 22 Feb 2014
Fisher, K., Gershuny. J.: Multinational time use study. User's guide and documentation. Version 6, 16 July (2013)
Frick, J.R., Jenkins, S.P., Lillard, D.R., Lipps, O., Wooden, M.: The cross-national equivalent file (CNEF) and its member country household panel studies In: Schmollers Jahrbuch **127**(4), 627–654. http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.77260.de/schmoller_frick_etal_2007.pdf (2007). Accessed 22 Feb 2014
Friendly, M., Denis, D.: The early origins and development of the scatterplot. J. Hist. Behav. Sci. **41**(2), 103–130 (2005)
Ganzeboom, H.B.G., Treiman, D.J.: Ascription and achievement after career entry. Paper presented at the 2000 meeting of the RC28 (2000)
Ganzeboom, H.B.G., Treiman, D.J.: Three internationally standardised measures for comparative research on occupational status. In: Hoffmeyer-Zlotnik, J.H.P., Wolf, C. (eds.) Advances in cross-national comparison. A European working book for demographic and socio-economic variables, pp. 159–193. Kluwer Academic Press, New York (2003)
Gartner: Definition of big data. http://www.gartner.com/it-glossary/big-data/ (2014). Accessed 22 Feb 2014
GESIS: An illustrated user guide to the CharmCats database for classifications and conversions. http://www.cessda.org/project/doc/D12.2c_An_illustrated_short_Usermanual_CharmCats_0.5.pdf (2009). Accessed 22 Feb 2014
Granda, P., Blasczyk, E.: Data harmonization. In: Cross-cultural survey guidelines. http://ccsg.isr.umich.edu/pdf/13DataHarmonizationNov2010.pdf (2010). Accessed 7 Feb 2014
Granda, P., Wolf, C., Hadorn, R.: Harmonizing survey data. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Pennell, B.-E., Smith, T.W. (eds.) Survey Methods in Multinational, Multiregional, and Multicultural Contexts, pp. 315–334. Wiley, New York (2010)
Harmonised European time use surveys (HETUS): Introduction. https://www.h2.scb.se/tus/tus/introduction1.html (2014). Accessed 14 Feb 2014
Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis. http://www.ncbi.nlm.nih.gov/books/NBK132553/pdf/TOC.pdf (2014). Accessed 22 Feb 2014
Hoffmeyer-Zlotnik, J.H.P., Wolf, C.: Comparing demographic and socio-economic variables across nations: synthesis and recommendations. In: Hoffmeyer-Zlotnik, J.H.P., Wolf, C. (eds.) Advances in Cross-national Comparison: A European Working Book for Demographic and Socio-Economic Variables, pp. 389–406. Springer, New York (2003)
Hout, M., DiPrete, T.A.: What we have learned: RC28's contributions to knowledge about social stratification. Res. Soc. Stratif. Mobil. **24**(2006), 1–20 (2006)
International Stratification and Mobility File (ISMF): http://www.harryganzeboom.nl/ISMF/index.htm (2014). Accessed 22 Feb 2014
Jacobs, J.A.: In Defense of Disciplines: Interdisciplinarity and Specialization in the Research University. University of Chicago Press, Chicago (2014)
Lee, R.M.: The secret life of focus groups: Robert Merton and the diffusion of a research method. Am. Sociol. **41**, 115–141 (2010)
Nissen, Sylke: The eurobarometer and the process of european integration: methodological foundations and weaknesses of the largest european survey. Qual. Quant. **48**, 713–727 (2014)
LIS Behavioral variable mapping. Undated. http://www.lisdatacenter.org/wp-content/uploads/2011/02/behavioural-variable-mapping-2011-03.pdf. Accessed 7 Feb 2014
LIS Database: Guidelines. Undated. http://www.lisdatacenter.org/wp-content/uploads/our-lis-documentation-harmonisation-guidelines.pdf. Accessed 7 Feb 2014

LIS standardisation routines for highest level of completed education. Undated. http://www.lisdatacenter.org/wp-content/uploads/standardisation-of-education-levels.pdf. Accessed 7 Feb 2014

Marcum, J.A.: Instituting science: discovery or construction of scientific knowledge? Int. Stud. Phil. Sci. **22**(2), 185–210 (2008)

Minkel, H.: Report on data conversion methodology of the change from input harmonization to ex-post harmonization in national samples of the European Community Household Panel—Implications on data quality. CHINTEX Working Paper 20. https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/Downloads/WorkingPaper20.pdf?__blob=publicationFile (2004). Retrieved 15 Feb 2015

National Academies.: Facilitating Interdisciplinary Research. National Academies Press, Washington (2004)

Olenski, J.: SSDIS. Global Standard for Harmonization of Social Statistics with special reference to transition and globalization processes. United Nations Statistics Division. ESA/STAT/AC.88/10 (April 7, 2003). http://unstats.un.org/unsd/demographic/meetings/egm/Socialstat_0503/docs/no_10.pdf (2003). Accessed 22 Feb 2014

Platt, J.: The development of the 'Participant Observation' method in sociology: origin myth and history. J. Hist. Behav. Sci. **19**, 379–393 (1983)

Portal on Collaboration in Research and Methodology for Official Statistics (CROS): http://www.cros-portal.eu/page/legal-notice (2014). Accessed 14 Feb 2014

Quandt, M.: Data harmonisation as a cumulative effort a platform designed to foster the cumulation of knowledge. Paper presented at the workshop on harmonisation of social survey data for cross-national comparison. Prague, 19th October. http://archiv.soc.cas.cz/download/860/06_Quandt.pdf (2010). Accessed 22 Feb 2014

Rainwater, L., Smeeding, T.: The luxembourg income study: The use of telecommunications in the social sciences. Luxembourg income study working paper series. Working paper No. 12 (May). http://www.lisdatacenter.org/wps/liswps/12.pdf (1987). Accessed 7 Feb 2014

Schröder, H., Ganzeboom, H.B.G.: Measuring and modeling levels of education in european societies. Eur Sociol Rev **30**, 119–136 (2013)

Slomczynski, K.M., Tomescu-Dubrow, I.: Survey data recycling: towards a formalized approach to ex-post harmonization of international projects. In: Harmonization: newsletter on survey data harmonization in the social sciences, vol. 1, pp. 10–13 (2015)

Smeeding, T.M.., Schmaus, G., Allegrezza, S.: An introduction to LIS. Luxembourg income study working paper series. Working Paper No. 1 (June). http://www.lisdatacenter.org/wps/liswps/1.pdf (1985). Accessed 7 Feb 2014

Smeeting, T.M.: CV. http://www.lafollette.wisc.edu/facultystaff/smeeding/smeeding-timothy-cv.pdf (2013). Accessed 7 Feb 2014

Smith, T.W.: Refining the total survey error perspective. Int. J. Pub. Opin. Res. **23**(4), 464–484 (2011)

Thomas, E.A.: Herbert Blumer's critique of the polish peasant: a post mortem on the life history approach in sociology. J. Hist. Behav. Sci. **14**, 124–131 (1978)

Treiman, D.J., Ganzeboom, H.B.G.: Cross-National comparative status attainment research. Res. Soc. Stratif. Mobil. (9), 105–127 (1990)

Treiman, D.J., Ganzeboom, H.B.G.: The fourth generation of comparative stratification research. In: Quah, S.R., Sales, A. (eds.) The International Handbook of Sociology, pp. 122–150. Sage, London (2000)

Tomescu-Dubrow, I., Slomczynski, K.M.: Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Leve Modeling in Cross-National Perspective. Ask. Res. Methods. **23**(1), 103–114 (2014)

UK's Economic and Social Research Council: http://www.esrc.ac.uk/ (2014). Accessed 22 Feb 2014

Van de Vijver, F.J.R., Matsumoto, D.: Introduction to the methodological issues associated with cross-cultural research. In Matsumoto, D., van de Vijver, F.J.R. (eds.) Cross-cultural research methods in psychology, pp. 1–16. Cambridge University Press, New York (2011)

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I., Börner, K.: Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. J. Infometrics. **165**, 14–26 (2011)

Workshop "Harmonization Strategies for Behavioral, Social Science, and Genetic Research" organized by the US Department of Health and Human Services, National Institutes of Health, National Institute on Aging and Division of Behavioral and Social Research, Bethesda (MD), November 29–30, 2011. http://www.nia.nih.gov/sites/default/files/nia_bssg_harmonization_summary_version_2-5-20122.pdf (2014). Accessed 22 Feb 2014