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

Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences

Volume 6, Issue 1
Spring/Summer 2020

Editors
Irina Tomescu-Dubrow
and
Joshua K. Dubrow
CONSIRT

consirt.osu.edu/newsletter
ISSN 2392-0858

Protect Others, Protect Yourself, Wear a Mask

The Covid 19 "Coronavirus" Pandemic is still here and so are changes in how we live, including where we work, how we educate our children, and how we navigate public spaces. The WHO, CDC, and most other health authorities advise wearing a mask as part of a larger strategy to prevent the spread of the virus. WHO sums up their advice as follows:

> "Does WHO advise the use of non-medical, fabric masks in public?
> ...If there is widespread community transmission, and especially in settings where physical distancing cannot be maintained, governments should encourage the general public to wear a fabric mask."[1]

The responsibility to protect others and ourselves from the virus while gathering the comparative administrative and survey data that are key resources used to shape collective action on the pandemic significantly changes regular processes of data collection (see recent WAPOR webinars on this topic). The fast pace of such changes and their degree of unpredictability add complexity to harmonization methods, both at the stages of data collection design and implementation, and ex-post, after data release.

In this issue of *Harmonization*, **Joshua K. Dubrow** surveys the landscape of administrative data sources used to understand Covid 19 cases and mortality; **Ranjit Singh** discusses the use of harmonizing instruments with equating; **Joonghyun Kwak** explores inter-survey methodological variability in the SDR framework; and **Julie de Jong** & **Kristen Cibelli Hibben** consider survey quality in the context of the EQLS 4. Plus, we bring news from the **Maelstrom** harmonization project, a December 2019 harmonization event, and survey sources to examine people's Covid 19 attitudes and behaviors.

As with every issue of *Harmonization*, we welcome your articles and news. Please send them to the newsletter editors.

----

[1] World Health Organization "Advice on the use of masks in the context of COVID-19; Interim guidance 5 June 2020"

[2] World Health Organization "Q&A: Masks and COVID-19"

In This Issue…

The editors thank Ilona Wysmulek and Kazimierz Slomczynski for their assistance with this newsletter.

Articles

Administrative Data Sources to Understand the Covid 19 Pandemic

by Joshua K. Dubrow

The Covid 19 pandemic is a modern-day disaster whose severity the media broadcasts, often via tables and graphs, the growth, stability, and decline in cases and mortality across nations and time. Media broadcasts are generally based on administrative data – cases of infected persons and deaths – that various people and organizations collect, harmonize, and aggregate. Many data providers have made their data available in machine-readable form for public scrutiny. An intention is to guide policy-makers and educate the public.

This note focuses on cross-national administrative Covid 19 data that are a main empirical foundation for the innumerable present and future comparative analyses on the virus' impact, and for possible harmonization projects. I ask two basic questions: What does the landscape of major distributors of cross-national Covid 19 counts look like? And what are some major sources of error[1] in these data?

The Landscape of Administrative Data on Covid 19 Cases and Mortality

In this information age, if data can be collected, they will be collected. This is true for Covid 19 – there are a great many sources of the numbers of infected persons (Covid 19 cases) and mortality (Covid 19 deaths), or what can be called "counts," which are widely available for modelers and media. Many organizations, through their websites, distribute count data. These counts are not raw numbers (especially at the individual or hospital level for obvious ethical and privacy reasons). Organizations distribute aggregates at various levels of administration that they produce from raw numbers, or from already aggregated information that other actors (people, organizations) gave them. Many, but not all organizations provide their data in GitHub, a data and computer program repository.[2]

To survey this landscape of data sources, I ask two basic questions: First, who are some of the major distributors of cross-national Covid 19 count data? Second, do their data overlap? I define

[1] Error can be defined as "a measure of the estimated difference between the observed or calculated value of a quantity and its true value" (Google). Here, I am interested in the difference between observed values of Covid 19 cases and mortality as collected by people and organizations and the true counts across nations and time. Deviations from the true counts are errors. Reasons for the deviations can be called "the sources of error."

[2] e.g. HealthMap provides a link to GitHub, but Reuters Graphics does not.

overlap as the situation in which an organization relies, in part or in whole, on another source for the data it distributes. The answers to these questions will tell us a bit about the field of data providers that governments, academics, and others, must choose from.

To identify highly visible organizations that report Covid 19 counts across nations and time, in April 2020, at what arguably was the height of the pandemic, I Googled the term "covid 19 data."[1] I chose a group of 14 organizations[2] such as one would find referenced in news reports: (in alphabetical order) 1point3acres, Bing/covid, European Centre for Disease Prevention and Control (ECDC)[3], Google's webpage for "Covid 19", Healthdata.org, HealthMap, Humanitarian Data Exchange (HDX), Johns Hopkins University (JHU) COVID Tracker, Reuters Graphics, *The New York Times*, Statista, Wikipedia, World Health Organization (WHO), and Worldometer. Figure 1 displays these organizations and the data sharing relationships between them.

On the 14 data sources' websites, I searched for information on where they get their Covid 19 count data. A first observation is that, in their textual description, data providers can be vague about where exactly their data come from and, perhaps to assure the reader that the data are thorough, some boast about how many sources they have. The European Union's ECDC claims that "a team of epidemiologists screens up to 500 relevant sources to collect the latest figures."[4] Worldometer writes that they "validate the data from an ever-growing list of over 5,000 sources." The initially popular 1point3acres reports in their FAQ:

> "Where does your data come from? We have a mix of scripts, API calls, and user submissions. Our data team then jumps in and fact check, cross referencing difference sources, de-dup and make sure we present most accurate data from credible sources… we use data from COVID Tracking and JHU (Rest of the World)."

---

[1] I focus here on cross-national counts and thus exclude US-only sources, e.g. CDC, The COVID Tracking Project, and Wunderground (an IBM company).

[2] The list is not exhaustive, but it does cover many of the websites available on the first few pages of Google search results.

[3] ECDC is "An agency of the European Union" https://web.archive.org/web/*/https://www.ecdc.europa.eu/en/Covid 19-pandemic

[4] They do mention the other sources, but as topics. Specifically, they write:

> "This includes websites of ministries of health (43% of the total number of sources), websites of public health institutes (9%), websites from other national authorities (ministries of social services and welfare, governments, prime minister cabinets, cabinets of ministries, websites on health statistics and official response teams) (6%), WHO websites and WHO situation reports (2%), and official dashboards and interactive maps from national and international institutions (10%). In addition, ECDC screens social media accounts maintained by national authorities, for example Twitter, Facebook, YouTube or Telegram accounts run by ministries of health (28%) and other official sources (e.g. official media outlets) (2%). Several media and social media sources are screened to gather additional information which can be validated with the official sources previously mentioned. Only cases and deaths reported by the national and regional competent authorities from the countries and territories listed are aggregated in our database." https://www.ecdc.europa.eu/en/Covid 19/data-collection

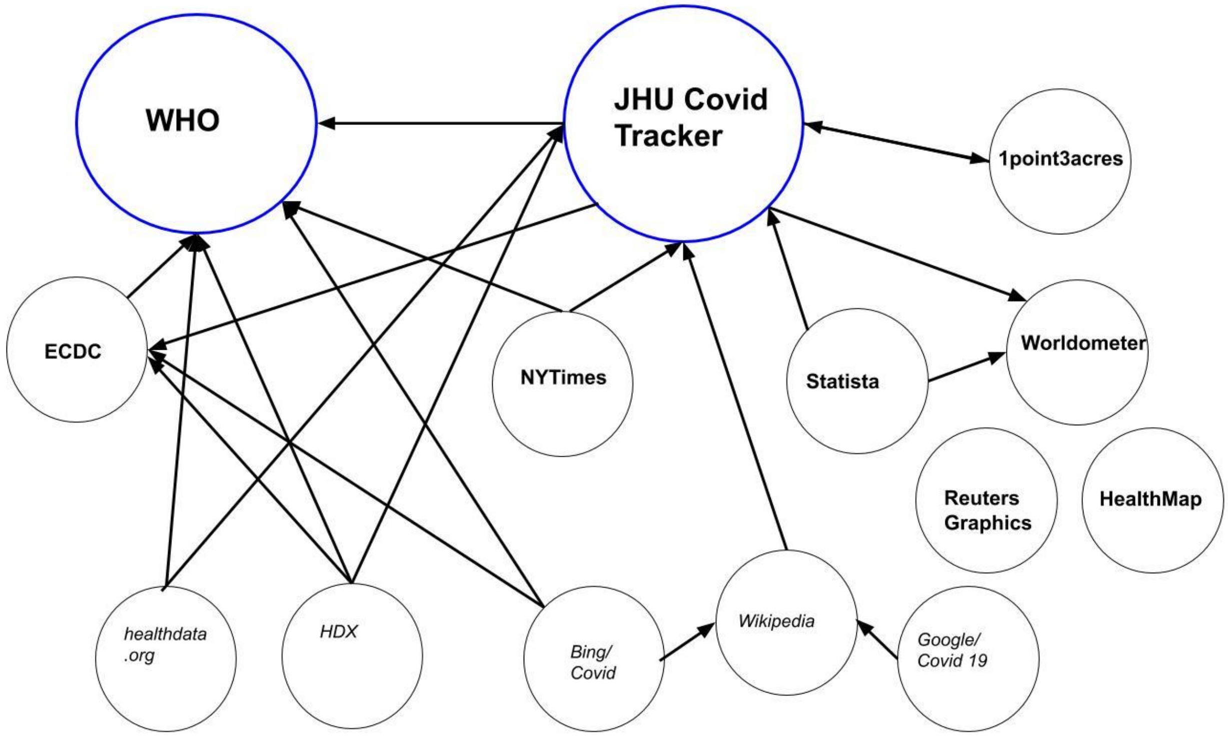


Figure 1. Graph of 14 "Covid 19" Data Sources of Cases and Mortality

Note: Data hubs are marked in large print and **blue** circles. Organizations that claim to collect their own data are in **bold**. Organizations that do not claim to collect their own data, but instead rely on data from others in this network (and also outside of this network), are in *italics*. Arrows indicate that organizations share Covid 19 data. The tip of the arrow points to where, in this group, an organization takes the data from (e.g. healthdata.org takes data from WHO and JHU; arrows point to WHO, but WHO does not publicly claim to use data from anyone in this network).

Of the 14 data sources, nine claim to collect their own data (Figure 1, names in **bold),** whereas five do not make such statements (names in *italics*). Among the organizations that report data collection, two – WHO and JHU Covid Tracker (Figure 1, circle in **blue**) – constitute what I call main data hubs. Covid 19 counts that WHO collects are used by six other organizations, and counts by JHU Covid Tracker are used by eight other organizations (including Bing and Google).

Most organizations (11 out of 14) explicitly indicate on their websites that the Covid 19 count data they distribute come, at least in part, from other organizations. Most of these (9 of the 11) rely on at least two other sources (in Figure 1, see nodes with at least two outgoing arrows). Within this network, JHU Covid Tracker is the distributor with the most heterogeneous source base; it relies on four other organizations: WHO, ECDC, 1point3acres, and Worldometer. Bing/Covid depends on three sources: WHO, ECDC, and Wikipedia and thus, without explicit mention, Bing/Covid depends on JHU Covid Tracker.

This map is useful to understand the not-explicitly-stated basis of Covid 19 count data. Consider the following data chain. For cross-national counts, Google/Covid 19 depends solely on

Wikipedia. Wikipedia depends solely on JHU Covid Tracker. Google and Wikipedia thus depend on the same source – JHU.

Of the 14 organizations' websites I examined, four do not explicitly mention that they take Covid 19 counts from other organizations shown in Figure 1: WHO, Worldometer, Healthmap, and Reuter Graphics. Yet, the absence of such a statement does not necessarily mean that there is no data overlap. For example, while Worldometer does not list any of the organizations in Figure 1 as a data source, they do claim to use 5000+ sources. Since I do not have the resources (e.g. patience) to go through each of them, I listed Worldometer as non-overlap. Healthmap's textual description of sources is too vague to allow for an assessment.[1] Reuters Graphics lists only local and national health authorities and themselves as sources of their data; to get their data, they write on the website that users must contact Reuters Graphics directly.

Of course, data providers within this group of 14 rely on outside sources. These sources include public authorities (various national health authorities and various subnational health authorities, including their press conferences and social media presence), reports in the mainstream media, social media (Twitter, Facebook, Telegram), specialty media sources such as BNO News and 24/7 Wall St., and what they call "user submissions," meaning that anybody in the world can contact them to report some information that could, perhaps, be included in their dataset.

Sources of Error

These Data Providers Depend on Upstream Reporting

Data collection of Covid 19 counts is difficult. At root, organizations depend on information provided by various national and subnational data sources that, in turn, received it from hospitals, labs, and other health organizations and medical authorities, which in turn depend on professionals within those organizations to report on Covid 19 cases and mortality. We have limited descriptions of upstream reporting from the USA via the CDC. Descriptions within other nations, in English, that share details of this upstream data collection process are difficult to find.

There are attempts to standardize Covid 19 count data collection in order to compare counts across nations and within nations, among lower administrative units, and over time. For example, WHO provides guidelines for case reporting. They do so through the "Revised Case Report Form for Confirmed Novel Coronavirus COVID 19 (report to WHO within 48 hours of case identification)"[2] Standardization requires that the data from a variety of sources and at different

---

[1] https://www.healthmap.org/Covid 19/# reports on data source: "All data used to produce this map are exclusively collected from publicly available sources including government reports and news media." This is their textual description, contained in a pop-up.

[2] They are protective of this publicly available resource. At the bottom of this document they wrote:

> "This is a draft. The content of this document is not final, and the text may be subject to revisions before publication. The document may not be reviewed, abstracted, quoted, reproduced, transmitted, distributed, translated or adapted, in part or in whole, in any form or by any means without the permission of the World Health Organization."

levels of aggregation are harmonized ex-post. Two short excerpts from WHO and *The New York Times*, respectively, illustrate this need well. According to WHO:

> "Due to differences in reporting methods, retrospective data consolidation, and reporting delays, the number of new cases may not always reflect the exact difference between yesterday's and today's totals."[1]

The New York Times effort to track Covid 19 wrote:

> "In tracking the cases, the reporting process is labor-intensive but straightforward much of the time. But with dozens of states and hundreds of local health departments using their own reporting methods — and sometimes moving patients from county to county or state to state with no explanation — judgment calls have sometimes been required."[2]

Ex-post harmonization of Covid 19 data counts is difficult and data providers (including those who aggregate the data) do not often explicitly state these difficulties. I mention some of the difficulties in this brief note. First, we can imagine the difficulties that the initial reporting agencies – i.e. the tens of thousands of hospitals with unequal economic development (as Covid Tracking Project hints at with their State Data Quality Grade) – are likely to have. Belgium stated well the problem of standardizing information when data sources are so disparate. In reporting the prevalence of Covid 19 in Belgium, "The Health, Food Chain Safety and Environment, a Federal Public Service of Belgium," notes: "In practice, we collect the data reported to us by: the national reference lab; the hospitals; the residential care centres; the General Practitioners (GPs); and the network of sentinel GPs and hospitals for the monitoring of flu-like syndrome."[3] They go on to write that: "The various sources do not always report the same type of data by any means, and the manner and frequency of reporting can also vary."[4]

Second, there will be discrepancies in the quality of data reporting. Sometimes, political reasons lead to inaccuracy on Covid 19 cases and deaths. For example, in May 2020, *The New Yorker* reported about Iran:

---

[1] From "Coronavirus disease 2019 (COVID 19) Situation Report – 96." https://web.archive.org/web/20200503174111/https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200425-sitrep-96-Covid 19.pdf?sfvrsn=a33836bb_4

[2] "We're Sharing Coronavirus Case Data for Every U.S. County" by *The New York Times* March 28, 2020.

[3] "The COVID 19 figures: collection, verification and publication," April 14, 2020. https://web.archive.org/web/*/https://www.info-coronavirus.be/en/news/collection-data/

[4] "The COVID 19 figures: collection, verification and publication," April 14, 2020. https://web.archive.org/web/*/https://www.info-coronavirus.be/en/news/collection-data/

> "Soon, Iran became a global center of the coronavirus, with nearly seventy thousand reported cases and four thousand deaths. But the government maintained tight control over information; according to a leaked official document, the Revolutionary Guard ordered hospitals to hand over death tallies before releasing them to the public."[1]

Also in May 2020, Russia experienced a huge upswing in cases, but reported low mortality. In the *Bloomberg News* article, "Experts Question Russian Data on Covid 19 Death Toll"[2]

> "Russian Deputy Prime Minister Tatyana Golikova Tuesday rejected suggestions Russia was understating the death rate. 'That's what it is and we never manipulate official data,' she said."

Yet, critics of Russia have questioned this general stance:

> "Russian authorities detained the leader of an independent doctors' union, an outspoken critic of the Kremlin who has dismissed as 'lies' the country's low official numbers for coronavirus infections."[3]

In response to the criticism, Russia in June retrospectively doubled the number of Moscow's mortalities that they had reported in April.[4]

Recently, as Brazil became one of the world leaders in Covid 19 cases, for a weekend in June the country simply stopped the daily reporting of cases and removed all previous information about Covid-19 tracking.[5]

We find unequal reporting between and within countries. In the US, the Center for Disease Control and Prevention (CDC) admitted that, in their harmonization and aggregation of data, they combined serology tests for antibodies with diagnostic tests of active viral infection, a data situation that may have led to a slight over-count of the number of Americans tested for Covid 19. In a *New York Times* article, blame for this mishap was attributed to too much pressure on too much work in

---

[1] "The Twilight of the Iranian Revolution" by Dexter Filkins in *The New Yorker*, May 25, 2020. https://www.newyorker.com/magazine/2020/05/25/the-twilight-of-the-iranian-revolution

[2] "Experts Question Russian Data on Covid 19 Death Toll," by Henry Meyer in *Bloomberg News*, May 13, 2020. https://web.archive.org/web/*/https://www.bloomberg.com/news/articles/2020-05-13/experts-question-russian-data-on-Covid 19-death-toll

[3] "Russian Doctor Detained After Challenging Virus Figures" by Andrew Higgins, *The New York Times*, April 3, 2020 Updated April 10, 2020. https://web.archive.org/web/20200527132124/https://www.nytimes.com/2020/04/03/world/europe/russian-virus-doctor-detained.html?action=click&module=RelatedLinks&pgtype=Article

[4] "Moscow more than doubles city's Covid 19 death toll," *BBC News*, May 29, 2020. https://web.archive.org/web/20200529002935/https://www.bbc.com/news/world-europe-52843976

[5] "Brazil stops releasing Covid 19 death toll and wipes data from official site" by Dom Phillips, *The Guardian*, June 7, 2020. https://web.archive.org/web/20200608015743/https://www.theguardian.com/world/2020/jun/07/brazil-stops-releasing-Covid 19-death-toll-and-wipes-data-from-official-site

too short span of time, an understandable situation that unfortunately led to poor decision making:

> "Epidemiologists, state health officials and a spokeswoman for the C.D.C. said there was no ill intent; they attributed the flawed reporting system to confusion and fatigue in overworked state and local health departments that typically track infections — not tests — during outbreaks. The C.D.C. relies on states to report their data."[1]

The CDC, and many state health officials, acknowledged the error and vowed to separate these counts in their future counts, *The New York Times* reported.

There have also been recent complaints that states have found errors in their Covid 19 counts. An article by *NBC News* from May 25, 2020, "'I'm looking for the truth': States face criticism for COVID 19 data cover-ups" summarizes some of the headlines since the beginning of May. Georgia apologized for a "processing error" that led to the erroneous presentation that counts were decreasing, rather than increasing. And in Florida

> "… officials last month stopped releasing the list of coronavirus deaths being compiled by the state's medical examiners, which had at times shown a higher death toll than the total being published by the state. State officials said that list needed to be reviewed as a result of the discrepancy."[2]

Indeed, widespread is the notion that counts are *under*-reported. Although the Coronavirus Conspiracists argue that there is a case and mortality over-count, there is no logic or evidence for that argument. In the second half of April, *The New York Times* reported "63,000 Missing Deaths: Tracking the True Toll of the Coronavirus Outbreak"[3] and *The Economist* reported that "Official Covid 19 death tolls still under-count the true number of fatalities."[4] "The data is limited," writes *The New York Times*, "and, if anything, excess deaths are underestimated because not all deaths have been reported."[5] In June, the director of the CDC said, "Our best estimate right now is that for

---

[1] "C.D.C. Test Counting Error Leaves Epidemiologists 'Really Baffled'" by Sheryl Gay Stolberg, Sheila Kaplan and Sarah Mervosh, *The New York Times*, May 22, 2020.
https://web.archive.org/web/20200526230234/https://www.nytimes.com/2020/05/22/us/politics/coronavirus-tests-cdc.html

[2] "'I'm looking for the truth': States face criticism for COVID 19 data cover-ups" by Allan Smith, *NBC News*, May 25, 2020.
https://web.archive.org/web/20200525164215/https://www.nbcnews.com/politics/politics-news/i-m-looking-truth-states-face-criticism-Covid 19-data-n1202086

[3] https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html

[4] https://web.archive.org/web/*/https://www.economist.com/graphic-detail/2020/04/16/tracking-Covid 19-excess-deaths-across-countries

[5] https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html. Note that they consider "data" to be singular, not plural.

On death counts, see also an article in *The Independent*, who writes: "One reason was for the time-lag in cases being reported, and how information about the disease included in coroners' reports was not always complete. Factors such as a person dying because they were every case that was reported, there actually were 10 other infections."[1]

A third difficulty for ex-post harmonization is that collecting Covid 19 counts, like any other data collection process, introduces errors related both to representation (e.g. who gets tested) and measurement (e.g. inadequate testing instruments and data processing errors). Regarding the processing errors, one reason for changes in the numbers of Covid 19 cases and mortality is the continual re-definition of "what is a case" and what counts as a fatality due to Covid 19.

Systematic errors can come from humans, from machines, or from some combination of the two, especially since Covid 19 data collection and reporting is not well standardized within or between nations, and in the midst of a pandemic, people and systems are severely stressed. Many of these errors may be extremely difficult to identify, let alone to correct ex-post, which in turn can affect the accuracy of statistics derived from Covid 19 counts.

The COVID Tracking Project for the US reports that the data situation has improved: "Reporting on even basic testing data was very patchy when we first began collecting data in early March, but is mostly now well reported."[2] Still, they have "State data quality grades," ranging from A+ (best, such as Iowa) to F (worst, such as Arkansas)[3].

The CDC seems to also make great efforts to standardize reporting. To improve the timeliness of reporting, and apparently, in the beginning of the pandemic there had been quite a bit of manual entry of forms, the CDC has worked to improve their electronic case reporting (eCR) system:

> "According to the CDC, electronic case reporting (eCR) is defined as the automated generation and sending of EHR[4] case reports to public health officials. eCR allows for automatic, complete, and accurate data to be reported in real-time. In return, it lessens burden for providers by improving the timeliness and accuracy of case reports… In an effort to reduce the healthcare system's burden of manually completing the COVID 19 reporting forms, the CDC will make these forms available electronically."[5]

---

too afraid to go hospital would likely not be included."

https://web.archive.org/web/*/https://www.independent.co.uk/news/world/americas/us-coronavirus-real-death-toll-covid-29-cases-a9504911.html

[1] "As Virus Surges, Younger People Account for 'Disturbing' Number of Cases." By Julie Bosman and Sarah Mervosh, *The New York Times*, June 26, 2020, Section A, Page 1. https://www.nytimes.com/2020/06/25/us/coronavirus-cases-young-people.html

[2] https://covidtracking.com/about-data

[3] As of May 2020.

[4] Electronic Health Reports (EHR)

[5] "CDC Unveils FHIR-Based COVID 19 EHR Reporting Application" by Christopher Jason April 20, 2020.

https://ehrintelligence.com/news/cdc-unveils-fhir-based-Covid 19-ehr-reporting-application

This does not describe how individual states conduct upstream reporting. I found one county that described it, in part. According to the official website of Pinal County, Arizona: "How are COVID 19 cases reported? All infectious diseases, including COVID 19 cases, are reported to local Public Health departments through an electronic platform known as MEDSIS (Medical Electronic Disease

The speed required in the pandemic to standardize across all parts of the reporting system whose infrastructure cannot handle the load can cause problems in the production of timely and accurate data. The CDC is a case in point. A recent *New York Times* article reported that the CDC's data infrastructure contains

> "antiquated data systems, many of which rely on information assembled by or shared with local health officials through phone calls, faxes and thousands of spreadsheets attached to emails. The data is not integrated, comprehensive or robust enough, with some exceptions, to depend on in real time."

To cope, "The agency rushed to hire extra workers to process incoming emails from hospitals," but the White House had turned toward JHU for timely accounts.[1]

Improvements in collecting and reporting Covid 19 counts likely occur in other countries, too, as organizations at different levels of administration get more experience with the process, and learn from previous mistakes. Yet, it is important to remember that, in the pursuit of data, errors occur. Identifying error sources in publicly released datasets of Covid 19 counts and understanding to what extent errors can be accounted for is an important part of ex-post harmonization decisions that will inform the future use of these data for social scientists.

Discussion

The importance of Covid 19 administrative data cannot be overstated. Data on Covid 19 cases and deaths have led to a host of institutional decisions that potentially impacted the lives of billions of people. These data are the basis of social distancing policies, economic redistribution, and how to conduct elections, among other things. This brief note explored some of what we know, and some of what we do not know, about the landscape of data sources that provide crucial information for governments, academics, and the public.

The pandemic has given policy-makers and academics plenty of data, and for their projects present and future, they will need to choose which data sources to use. In a limited analysis presented here, we see that, for cross-national Covid 19 counts, the data landscape is dominated by WHO and JHU. There is considerable overlap. Many data sources use others' sources. I explored 14

Surveillance Intelligence System). Our epidemiologist team manually checks each of the data entered to ensure the data is accurate." https://www.govserv.org/US/Florence/726925130750461/Pinal-County-Board-of-Supervisors

[1] "Built for This, C.D.C. Shows Flaws in Crisis." By Eric Lipton, Abby Goodnough, Michael D. Shear, Megan Twohey, Apoorva Mandavilli, Sheri Fink and Mark Walker in *The New York Times*, June 3, 2020. https://web.archive.org/web/20200607180433/https://www.nytimes.com/2020/06/03/us/cdc-coronavirus.html

> "As the number of suspected cases — and deaths — mounted, the C.D.C. struggled to record them accurately. Still, many officials turned to Johns Hopkins University, which became the primary source for up-to-date counts. Even the White House cited its numbers instead of the C.D.C.'s lagging tallies."

different organizations, but I predict that WHO, JHU, and Worldometer will emerge as the core cross-national data sources that academics and governments will use. WHO is the sole official source of cross-national data. Yet, JHU quickly established itself as a premier data provider early in the process and has held on to that status. Worldometer has been well-known to social scientists, and thus may be used because their data can easily be merged with the social, economic, and political data that they already provide.

I do not assess the validity and reliability of these data, but I do mention some of the sources of error that may lead to discrepancies across nations and time. A main source is the frequent redefinitions or cases and mortality, which are due in part to changes in knowledge about Covid 19. As countries update their knowledge, it is not clear whether they will expend the effort to retrospectively change their data to reflect the new knowledge. Other errors occur in any large-scale data collection process, such as processing errors.

A source of error that has yet to gain much attention is the difficulty in harmonizing and aggregating data from multiple local sources that report upstream to national organizations. These discrepancies may be due to unequal infrastructures and the unequal resources of hospitals, labs, and other organizations staffed with time and social pressured people who, due to systemic problems and simple fatigue, make mistakes that can introduce a series of minor errors in the data that they report upstream. The upstream reporting problem may not matter much, or it may matter a lot. We don't know. Upstream reporting is a black box that we should open.

Joshua K. Dubrow is co-editor of Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences. *This material is based upon work supported by the National Science Foundation under Grant No. (PTE Federal award 1738502) and by the National Science Centre, Poland (2016/23/B/HS6/03916).*

Harmonizing Instruments with Equating

by Ranjit K. Singh

This is a brief introduction to equating, which is a promising approach to harmonising survey instruments that measure latent constructs such as attitudes, values, intentions, or other individuals' attributes that are not directly observable. The focus is on measurement instruments with only one question (in contrast to multi-item questionnaires). The article is intended to inspire ex-post harmonization practitioners who struggle to harmonize such variables into a homogenous target variable. Some researchers might find that the equating approach is directly applicable to their work. However, even if the specific approach is not a good fit for a particular project, the basic idea of equating could still be a helpful way of thinking about instrument harmonization in general.

The Goal of Instrument Harmonization

A common challenge in ex-post harmonization is how to combine data on a concept measured with different instruments. This is especially hard if the concept is a latent construct; i.e., a construct that cannot be directly observed, such as attitudes, values, or intentions. The central problem here is that latent constructs have no natural units. There is, for example, no self-evident way to compare how much "strongly interested" is on one scale of political interest as compared to "somewhat agree" on another scale measuring political interest.

To get a better understanding of the problem, it helps to clarify the ideal result of an ex-post harmonization process: Taking data from different surveys that were not intended to be combined and that use different instruments, we want to create a seamless dataset. Seamless here implies that once the dataset has been harmonized, it should no longer matter which instrument was used for a particular case in the dataset. To that end, we need to harmonize scores measured with different instruments so that the same score always "means" the same, regardless of the source instrument.

While seemingly self-evident, it is necessary to take a closer look what "meaning the same" implies. To understand harmonization, we should first remind ourselves that the output of measurement (the observed scores on an instrument) is not reality itself, but only something that is related to this reality (Raykov & Marcoulides 2011). What this entails is best explained with the following example.

Consider a latent construct such as political interest. If we measure political interest, we implicitly assume that respondents have a certain attribute strength that governs the extent to which they direct their attention towards or away from sources of political information. In other words, they have a theoretical *true score* that directly reflects their real interest. When we ask respondents to answer a standardized question about their political interest, we assume that respondents will choose one of the offered response categories based on their true political interest. In other words, a measurement instrument projects the true score of respondents onto an arbitrary numerical scale: People within a certain range of political interest will likely chose a "1" on the instrument, people in a somewhat higher range of political interest will likely chose a "2," and so on for each score up to "5" on the five-point scale. The crux is that different instruments project the same reality — true scores — differently. The "3" on one scale does not automatically represent the same range of true scores as the "3" on another instrument. This is true even if the two instruments have the same number of response options but differ with respect to other features, such as wording or layout.

The relationship between true scores and measurement scores in different instruments has two important implications: (1) The *observed scores* we have in our source data represent a mix of truth and measurement; and (2) Different instruments change the measurement component and are likely to result in different observed scores for people with the same true scores (Raykov & Marcoulides, 2011). With this in mind, we can now formalize what instrument harmonization should do: Respondents who are the same with regard to a construct should get the same harmonized score, no matter which instrument was used.

Linking and Equating

Fortunately, this goal of harmonization is shared by psychometric performance and aptitude testing, where there is a need to make comparable different tests for the same construct. This has resulted in an extensive literature dating back to the 1970s on what is today called score linking (Dorans & Puhan, 2017). Equating, meanwhile, is a subfield of score linking that directly addresses our problem, i.e., making comparable scores from instruments that measure the same construct.[1] As we will see, both the logic of equating and its formulas can be of great use.

At the heart of equating is its equity property. Respondents with a certain true score should, on average, get the same converted (i.e., equated) score in a source instrument than they would get on the target instrument (Kolen & Brennan, 2014). The equity property contains the qualification "on average" to reflect random error in measurement and in equating itself. If we achieve such a matching of converted source scores to target scores, we have corrected differences in measurement without eliminating or biasing real differences.

The obstacle we now face is, of course, that we do not have the true scores — we only have observed scores. Psychometry tackles that problem with multi-item instruments such as personality questionnaires. True score estimations are extracted from the interplay of different measurements of the same construct for each respondent, often in factor models (Raykov & Marcoulides, 2011). In the social sciences, large scale survey programs often cannot accommodate multiple questions for all their constructs of interest. Fortunately, not all forms of equating rely on multiple items. Observed score equating relies only on the observed scores (what we have in the dataset) and not on true score estimations. Without multiple measurements, that is, multiple data-points for each person, we cannot disentangle measurement and reality on the respondent level. We can, however, disentangle measurement and reality on the aggregate level.

Observed Score Equating

The basic idea of observed score equating is that if we cannot isolate the effect of different instruments, we control for population differences in true scores via random group designs. This is done by taking two random samples of the same population. In one, respondents answer the source instrument, and in the other, there is the target instrument. Since both samples have similar true scores distributions (they randomly sample the same population), differences in the observed score distributions are due to the instrument differences. Next, we apply a mathematical transformation to

[1] A short note on terminology: In the formal literature, equating also denotes a very strict quality standard for test comparability in psychometric diagnostics, such as professional aptitude testing. This formal standard is only attainable with test forms constructed to very similar test specifications such as length or reliability (Kolen & Brennan, 2014). For harmonization purposes, equating can still be done even if the instruments for example differ in reliability. The standard only cautions us that equating makes units of measurement comparable but does not correct the limitations of the equated instruments. For an extensive overview of the terminology and its history, see Kolen and Brennan (2014) or Dorans and Puhan (2017).

scores of the source instrument so that the distribution of transformed source scores is similar in shape to the distribution of target scores (Kolen & Brennan, 2014). In other words, observed score equating basically matches scores in the two instruments based on their position along the frequency distribution. Average respondents get the same scores regardless of instrument used, and the same is true for below average or above average respondents.

Equating does not mean that the survey data we want to harmonize has to be drawn from the same population. We can use different datasets to perform the equating that results in a transformation table. This table can then be used to harmonize the data we intend to harmonize. Equating is also always symmetrical, meaning that we can transform scores from one instrument to the other and *vice versa*.

Next, we take a closer look at two ways to transform distribution shapes in observed score equating: (1) Linear equating for approximately normally distributed source and target scores and (2) equipercentile equating if the distribution of one or both instrument scores are non-normal (e.g., strongly skewed or even bimodal).

Linear equating

Linear equating assumes that both score distributions are approximately normally distributed, which implies that the two instrument score distributions only differ in two parameters: The mean and the standard deviation. In linear equating, scores of the source instrument are linearly transformed so that the transformed source score mean and standard deviation become equal to the target score mean and standard deviation (Kolen & Brennan, 2014). Respondents now have very similar scores on the transformed source instrument and the target instrument depending on their position along the normal distribution. Respondents with the same z-score have the same harmonized score but scaled to the format of the target scale.

To avoid confusion, I add two clarifications. First, linear equating is distinct from a mere z-transformation. The mathematical transformation that is used to align the distribution shapes is indeed similar to a z-transformation. However, at the heart of linear equating are the two instrument samples drawn from the same population. By setting the population as equal, we can isolate and eliminate the measurement differences. The resulting translation table can then be used in instances where the two instruments were used on non-equal populations. The result is a harmonization of the measurement while preserving true population differences. A mere z-transformation, in contrast, would indiscriminately destroy true population differences because for each sample we will have mean = 0, and SD = 1.

Second, linear equating is also quite different from the frequently used harmonization approach, which is the linear stretch method. Linear stretch applies a linear transformation to instrument scores solely based on differences in scale points. Consider a five-point source scale and a seven-point target scale. With linear stretch, we would assign minimum score to minimum score (a source score of 1 would remain a 1) and maximum score to maximum score (a source score of 5 would become a 7). All points in between are stretched so that they fit in the space between 1 and the new maximum score with equal distances (Jonge, Veenhoven, & Kalmijn, 2017). The source scale 1, 2, 3, 4, 5 would become the transformed scale 1, 2.5, 4.0, 5.5, 7.

Linear equating, in contrast, harmonizes scales based on the distribution of responses for the same population. Linear stretch only takes the number of response scale points into account. The difference becomes apparent if we apply both methods to two instruments with different question and response option wording, but with the same number of scale points. Even in the same population we would expect different response frequency distribution for both instruments because both question wording and response option wording change the response options that respondents choose. Linear stretch would ignore that and assign each score in one instrument to exactly the same score in the other instrument because the scale points are the same (Jonge et al., 2017). Linear equating, meanwhile, would transform scores so that the distributions become aligned. Hence, respondents at the same position along the construct distribution (e.g., average respondents) may well get different harmonized scores with linear stretch, but with very similar scores with linear equating.

Equipercentile Equating

Equipercentile equating, meanwhile, drops the assumption of normally distributed instrument scores. Instead, we transform scores so that the distribution of transformed source scores fluidly matches the shape of the distribution of target scores. And as a reminder: Just like with linear equating, this is performed on two random samples from the same population. Equipercentile equating operates like this: We take a score from the source instrument and based on the frequency distribution we calculate the percentile rank of that score (i.e., the position of that score along the distribution of the construct in the population). Then we look up which score in the target instrument corresponds to that percentile rank. After that, we transform the source score with a certain percentile rank into that target score with the same percentile rank (Kolen & Brennan, 2014). Consequently, all scores now "mean the same" in the sense that each transformed source score and target score point to the same specific place in the population distribution.

One remaining challenge is that response scores are ordinal and not continuous. This has two implications that the equipercentile equating formulas solve with linear interpolation (Kolen & Brennan, 2014). (1) A score does not represent an exact percentile rank along the continuous distribution of the construct. Instead, each score represents a range of respondents (e.g., if the first response option is chosen by 20% of respondents, then it represents percentile ranks from the 0th to the 20th). Equipercentile equating solves this with linear interpolation and simply assigns the middle (e.g., 10%). (2) If we have a percentile rank for a source score, we likely have no target score at exactly that percentile rank. Again, we linearly interpolate and assign a transformed score between the two target scores. If the two applicable target scores 1 and 2 represent the 10th and 40th percentile rank, and if the percentile rank of the source score is 20, then we would assign the transformed score 1.25. This is because 20% is 33% along the distance from 10% to 40% and 1.33 is 33% along the distance from score 1 to score 2. The outcome of equipercentile equating is then, just as with linear equating, a transformation table that translates scores of one instrument into the format of the other instrument.

Observed Score Equating in Practice

At this point I was hopefully able to pique your interest in observed score equating. Yet the hurdle of requiring two samples from the same population, one for each instrument, poses a challenge. In the following, I will lay out some arguments and ideas how observed score equating can be applied in practice.

First, to stress again, the data used to equate two instruments does not have to be from the dataset we want to harmonize. Any point in time where the two instruments are used in a probabilistic sample of the same population is enough. This might also mean that the matching is done with a completely different survey program that just so happens to have copied one of the instruments. Finding such serendipitous scale-population-time matches is facilitated by using the databases of large harmonization projects, such as the Survey Data Recycling project (and previous project: see Tomescu-Dubrow and Slomczynski 2016), which may well have the data you need conveniently searchable, bundled, and cleaned.

Second, if you have probabilistic samples of the same population, but at different points in time, then this is only a problem if the population changed substantially regarding the construct in the meantime. Chances are that another survey program has a time series of the construct. If no change occurred, equating can be done as is, and if a change occurred, one can apply linear equating but correct the transformation for the change over time.

Third, observed score equating is preferably done with samples from a population that is similar to the population that the harmonization project is interested in. Yet, if no such data exist, equating can also be done with a split-half experiment in a non-probabilistic sample (e.g., an online access panel). With a single survey, several instruments can be equated at the same time. Please note that comparability issues can occur if the experiment sample and the target population of your project are very different and if your measurement instrument is not measurement invariant across those populations (i.e., if it is not interpreted the same way across populations). Dorans and Holland (2000) discuss the problem, the potential ways of estimating the extent of the problem, and ways to mitigate it.

Fourth, with regard to harmonizing cross-cultural data, equating often cannot be done directly with existing data. This is often the case if data from national survey programs is to be harmonized, meaning that the same construct is measured with different instruments in different countries. However, equating can perhaps still be used via chained equating (Kolen & Brennan, 2014). Consider harmonizing instruments from two probabilistic national surveys from Country A and Country B. Now assume that the relevant construct has also been included in a cross-national survey program including Countries A and B. We can then equate the instruments from the national survey A to the cross-cultural survey. And then equate the cross-cultural instrument further to the national survey B. However, this chained equating cannot be more comparable across cultures than is the instrument in the cross-national survey.

In sum, equating is worth considering. It is not a panacea, but if the preconditions are met, it is a rigorous method that will increase the quality of the harmonized dataset considerably. If suitable data are available, then equating is easily done due to many existing specialized programs and packages for statistical applications. A package specialized in observed score equating is *equate* (Albano, 2016). Packages that also include other techniques are *kequate* (Andersson, Branberg, and Wiberg 2013) and *SNSequate* (González, 2014). For an overview and guidance on applying these R packages see González & Wiberg (2017). There are also stand-alone programs for equating, such as *RAGE-RGEQUATE* for observed score equating (Zeng, Kolen, Hanson, Cui, & Chien, 2005) which can be downloaded at Brennan (2020) who also curates a comprehensive list of other programs.

Yet even if a direct application is not possible, the idea of equating is helpful in rethinking harmonization. It also guides us to potential pitfalls in analysing harmonized data where equating has not been performed. If you would like to learn more about what equating can do for your project or what alternative approaches exist, feel free to contact me. At GESIS, I offer consultation on how to make survey instruments comparable in ex-post harmonization (see below).

Ranjit K. Singh is a post-doctoral scholar at GESIS, the Leibniz Institute for the Social Sciences, where he practices and researches the harmonization of substantive instruments in surveys. At GESIS he now also offers consulting on the ex-post harmonization of substantive instruments for researchers who either combine data from different survey programs or who want to change an instrument in their ongoing survey program. Consultation topics include assessing instrument comparability, weighing consequences of comparability issues, and strategies for harmonizing substantive instruments. For more information, contact ranjit.singh@gesis.org or visit: https://www.gesis.org/en/services/data-analysis/data-harmonization/harmonizing-substantive-instruments

References

Albano, A. D. (2016). equate: An R Package for Observed-Score Linking and Equating. *Journal of Statistical Software*, *74*(8). https://doi.org/10.18637/jss.v074.i08

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software*, 55(6). https://doi.org/10.18637/jss.v055.i06

Brennan, R. L. (2020). *Computer Programs*. Retrieved June 24, 2020, from https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Dorans, N. J., & Holland, P. W. (2000). Population Invariance and the Equatability of Tests: Basic

Theory and The Linear Case. *Journal of Educational Measurement*, *37*(4), 281–306.

Dorans, N. J., & Puhan, G. (2017). Contributions to Score Linking Theory and Practice. In R. E. Bennett & M. von Davier (Eds.), *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS* (pp. 79–132). https://doi.org/10.1007/978-3-319-58689-2_4

González, J. (2014). SNSequate : Standard and Nonstandard Statistical Models and Methods for Test Equating. *Journal of Statistical Software*, *59*(7). https://doi.org/10.18637/jss.v059.i07

Jonge, T. de, Veenhoven, R., & Kalmijn, W. (2017). *Diversity in Survey Questions on the Same Topic: Techniques for Improving Comparability*. https://doi.org/10.1007/978-3-319-53261-5_1

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.). https://doi.org/10.1007/978-1-4939-0317-7

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

Tomescu-Dubrow, I., & Slomczynski, K. M. (2016). Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling. *International Journal of Sociology*, *46*(1), 58–72. https://doi.org/10.1080/00207659.2016.1130424

Zeng, L., Kolen, M. J., Hanson, B.A., Cui, Z., & Chien, Y. (2005). *RAGE-RGEQUATE [computer program]*. Iowa City, IA: The University of Iowa.

Inter-Survey Methodological Variability in Institutional Trust from the Survey Data Recycling Framework

By Joonghyun Kwak

A critical problem in survey data harmonization is methodological inter-survey variability. Methodological differences between surveys have been treated as unmeasured errors that might be inherent in survey question properties or might occur during fieldwork and data processing (Slomczynski and Tomescu-Dubrow 2018). The Survey Data Recycling (SDR) framework offers a solution to this problem by creating harmonization and survey quality control variables that measure potential sources of inter-survey variability. This research note examines the extent to which the harmonization and survey quality control variables adjust for inter-survey variability, using trust measures in parliament, legal system, and political parties from the SDR database v1.1 (Slomczynski et al. 2017).

Survey project (Abb.)	Survey Project	Trust in Parliament			Trust in Legal System			Trust in Political Parties		
		Time Span	Countries	Compared Projects	Time Span	Countries	Compared Projects	Time Span	Countries	Compared Projects
ABS	Asian Barometer	2006-2010	4	LITS, WVS	2006-2010	4	LITS, WVS	2006-2010	4	LITS, WVS
AMB	Americas Barometer	2004-2010	20	ISSP, LB, WVS	2004-2010	20	ISSP, LB, WVS	2004-2010	20	LB, WVS
ASES	Asia Europe Survey	2000	8	EB, WVS	2000	8	EB	2000	8	EB, WVS
CNEP	Comparative National Elections Project	2005	1	LB	2005	1	LB	2005	1	LB
EB	Eurobarometer	2000-2012	14	ASES, EQLS, ESS, ISSP, LITS, WVS	2000	8	ASES	2000-2012	14	ASES, ESS, LITS, WVS
EQLS	European Quality of Life Survey	2007-2012	11	EB, ESS, EVS, LB, WVS	2007-2012	11	ESS, EVS, LB, WVS	2007-2008	9	ESS, EVS, LB, WVS
ESS	European Social Survey	2004-2012	22	EB, EQLS, EVS, ISSP, LB, LITS, NBB, WVS	2004-2012	22	EQLS, EVS, ISSP, LB, LITS, NBB, WVS	2004-2012	22	EB, EQLS, EVS, LB, LITS, NBB, WVS
EVS	European Values Study	2008-2009	13	ESS, ISSP, LB	2008-2009	13	ESS, ISSP, LB	2008-2009	13	ESS, LB
ISSP	International Social Survey Programme	2006-2010	15	AMB, EB, ESS, EVS, LB, LITS, WVS	2008-2010	13	AMB, ESS, EVS, LB, LITS	_	_	_
LB	Latinobarometro	1995-2010	19	AMB, CNEP, EQLS, ESS, EVS, ISSP, WVS	1995-2010	19	AMB, CNEP, EQLS, ESS, EVS, ISSP, WVS	1995-2010	19	AMB, CNEP, EQLS, ESS, EVS, WVS
LITS	Life in Transition Survey	2006-2010	15	ABS, EB, ESS, ISSP, WVS	2006-2010	15	ABS, ESS, ISSP, WVS	2006-2010	15	ABS, EB, ESS, WVS
NBB	New Baltic Barometer	1996-2004	2	ESS, WVS	1996-2004	2	ESS, WVS	1996-2004	2	ESS, WVS
WVS	World Values Survey	1995-2007	28	ABS, AMB, ASES, EB, EQLS, ESS, ISSP, LB, LITS, NBB	1995-2007	28	ABS, AMB, EQLS, ESS, LB, LITS, NBB	1995-2007	28	ABS, AMB, ASES, EB, EQLS, ESS, LB, LITS

Table 1. International Survey Projects in the Analysis, 13 Projects for 137 Country-Years

Data and Variables

SDR data has a three-level hierarchical structure with respondents who are clustered within national surveys nested within country-years. Theoretically, the mean value of institutional trust should be identical between national surveys collected in the same year (assuming no major event severely and quickly upsets degrees of trust); therefore, an actual difference in the mean values between surveys is the result of inter-survey variability. To focus on the effects of harmonization and survey quality control variables on inter-survey variability, I restrict the sample to countries that have at least two surveys in the same year. To eliminate selection bias that would hinder comparison across models, I also confine the sample to country-years that cover all three trust measures. Ultimately, the final country-year level sample size is 137 country-years, covering 53 countries between 1995 and 2012. The regional and temporal data coverage includes 13 international survey projects with different sample sizes for national surveys across models: 305 surveys for trust in parliament, 293 surveys for trust in legal system, and 288 surveys for trust in political parties. Table 1 provides a description of the international survey projects in terms of time spans, number of countries, and compared projects in the same country-years for each institutional trust measure. The list of country-years is provided in the Appendix.

The dependent variables measure respondents' trust in three political institutions: *trust in parliament*, *trust in legal system*, and *trust in political parties*. I use an 11-point scale of trust measures ranging from 0 (lowest intensity) to 10 (highest intensity), which were re-scaled from various response scales in the source surveys by linear transformation (Slomczynski et al., 2016: 56).

The institutional trust variables in the SDR data are accompanied by three harmonization control variables that capture question properties in the source survey (Slomczynski et al. 2016). *Scale length* indicates the response scales in the original question ranging from 2 to 11. *Scale direction* measures whether the original scale is ordered from lowest to highest trust or from highest to lowest. *Scale polarity* indicates whether the response values are defined by one dimension—from no trust to strong trust—or if the response values are measured by two dimensions—from distrust to trust.

An important source of inter-survey variability is variation in data quality across surveys. The SDR data set offers three indices of survey quality that can be used as control variables (Kwak and Slomczynski 2019; Tomescu-Dubrow et al 2017). *Computer file quality index* measures errors in computer data files by constructing an additive scale in three dichotomous variables that capture whether the survey has: (1) duplicate cases; (2) over 5% of missing data on either age or gender variable; and (3) errors in respondent ID. *Survey documentation index* measures survey quality as reflected in the documentation of the source data (Kołczyńska and Schoene 2018). This index is also created as an additive scale in five dichotomous variables that measure whether the survey documentation has information on: (1) sampling; (2) response rate; (3) control of the quality of the questionnaire translation; (4) questionnaire pretesting; and (5) fieldwork control. *Processing error index* measures processing errors that indicate a contradiction between data file and survey documentation by counting the number of errors in seven selected variables (gender, age, birth year, education level, schooling year, trust in parliament, and participation in a demonstration) and dividing it by the number of variables in the survey (Oleksiyenko, Wysmulek, and Vangeli 2018). This index captures the number of processing errors adjusted by the total number of variables for which these errors were checked in a given survey. In this paper, for each index, higher values indicate poorer quality.

Methods

I examine variance components in the three-level multilevel model. Variance components represent the unexplained group effects that indicate inter-group variance. The three-level multilevel model for the constrained model takes the following form:

$$y_{ijk} = b_0 + b_{1\ldots3}h_{1\ldots3jk} + b_{4\ldots6}q_{4\ldots6jk} + u_k + u_{jk} + e_{ijk}$$

where y_{ijk} is trust in a given political institution for individual *i* nested within survey *j* in country-year *k*; b_0 is the intercept; $h_{1\ldots3jk}$ represent three harmonization control variables in survey *j* in country-year *k*; $q_{4\ldots6jk}$ represent three quality control variables in survey *j* in country-year k; u_k is country-year random effect; u_{jk} is survey-level random effect; e_{ijk} is individual-level random effect.

To facilitate interpretation of the variance components, u_k, u_{jk}, e_{ijk}, I calculate intraclass correlation coefficients (ICCs) that represent the degree of similarity between individuals of the same grouping factor and the amount of variation explained by the grouping factor (Pais 2010). Inter-survey variability in a three-level multilevel structure is estimated by (1) *country-year-level ICC,* which represents the degree of homogeneity among individuals within the same country-year, but different surveys and (2) *survey-level ICC,* which represents the homogeneity among individuals in the same survey—and, therefore, the same country-year—as well as how much individuals' variation in institutional trust is attributed to differences between surveys in the same country-year. I calculate country-year-level and survey-level ICCs in the following manner:

$$ICC_{county-year} = \frac{u_k}{u_k + u_{jk} + e_{ijk}}$$

$$ICC_{survey} = \frac{u_k + u_{jk}}{u_k + u_{jk} + e_{ijk}}$$

Given that the ICCs in an unconstrained model capture the total variance at the levels, the impact of the harmonization and survey quality control variables on inter-survey variability can be identified by a change in ICCs between an unconstrained model and a constrained model that adjust for the effect of the controls. An increase in country-year-level ICC from unconstrained model to constrained model indicates that the control variables increase the degree of similarity among individuals in different surveys within the same country-year. This suggests that inter-survey variation decreases. A decline in survey-level ICC from unconstrained to constrained model suggests that the control variables reduce the difference of institutional trust between surveys. Therefore, an increase in country-year-level ICC and a decline in survey-level ICC reflect the decline in inter-survey variability that is attributed to the harmonization and survey quality control variables.

Analysis

Table 2 presents the results of three-level multilevel analyses for trust in parliament. Model 1 is an unconstrained model that does not include any covariates. The ICC in Model 1 is 0.083 for the country-year level and 0.130 for the survey level. These indicate that 8.3% of variance of trust in parliament is explained by country-year-level variation and 13.0% of variance is attributed to the survey-level variation. This suggests that survey factors influence trust in parliament to a greater degree than country-year context.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Scale length		0.042**						-0.176***
		(0.015)						(0.045)
Scale direction			0.408***					1.206***
			(0.071)					(0.217)
Scale polarity				-0.138†				0.203
				(0.071)				(0.134)
Computer file quality index					0.178†			0.084
					(0.100)			(0.090)
Data documentation index						-0.028		-0.075*
						(0.027)		(0.033)
Processing error index							0.379***	-0.499**
							(0.097)	(0.177)
Constant	4.306***	4.068***	4.124***	4.428***	4.280***	4.385***	4.220***	4.907***
	(0.068)	(0.093)	(0.066)	(0.087)	(0.069)	(0.118)	(0.068)	(0.199)
Variance Components								
Country-year level	0.492	0.489	0.524	0.496	0.494	0.476	0.508	0.530
Survey level	0.277	0.264	0.218	0.274	0.273	0.281	0.260	0.190
Individual level	5.168	5.168	5.168	5.168	5.168	5.168	5.168	5.168
Intraclass Correlation Coefficients								
Country-year level	0.083	0.083	0.089	0.083	0.083	0.080	0.086	0.090
Survey level	0.130	0.127	0.126	0.130	0.129	0.128	0.129	0.122
Model Fit Statistics								
AIC	1799532.0	1799524.3	1799495.5	1799532.5	1799531.2	1799533.0	1799525.4	1799481.4
BIC	1799575.6	1799578.8	1799550.0	1799587.1	1799585.7	1799587.5	1799579.9	1799590.4
-2log likelihood	1799524.0	1799514.3	1799485.5	1799522.5	1799521.2	1799523.0	1799515.4	1799461.4

*—p < 0.05; **—p < 0.01; ***—p < 0.001; †—p < 0.05 (two-tailed tests).
a—Unstandardized regression coefficients (standard errors in parentheses).

Table 2. Three-level Multilevel Analysis for Trust in Parliament (N=401,355 Individuals Nested Within 305 Surveys in 137 Country-Years)[a]

To investigate bivariate relationships between trust in parliament and the control variables, in Models 2 through 7, I add harmonization and survey quality control variables one at a time. Among harmonization control variables, scale length and scale direction have positive and significant effects on trust in parliament, whereas scale polarity is negatively associated with trust in parliament at p <0.10. Among the survey quality control variables, the processing error index has a positive and significant effect. The effect of the computer files quality index is also positive, but significant only at p<0.10. The effect for survey documentation index is negative and not statistically significant. The results suggest that most properties and qualities of the source survey identified by the SDR framework are systematically related to trust in parliament.

The country-year-level ICCs show that models with scale direction and processing error index have higher ICCs than Model 1 by 7.2% (=[0.089-0.083]/0.083) and 3.6% (=[0.086-0.083]/0.083) respectively, whereas models with scale length, scale direction, and computer file quality index have the same levels of ICCs as in Model 1. The decline in country-level ICC by survey documentation index is in the unexpected direction. Survey-level ICCs in all but one model are consistently lower than Model 1, which indicates that the control variables cause a decline in inter-survey variability. Only scale polarity has no effect on survey-level variance. Overall, the results of changes in country-year-level and survey-level ICCs suggest that, with the exception of scale polarity, harmonization and survey quality control variables contribute to reducing inter-survey variability in trust in parliament.

In Model 8, which is fully constrained by harmonization and survey quality control variables, I find that the directions of coefficients for scale length, scale polarity, and processing error index are reversed compared to those in the bivariate regression models. I suspect that these reversed effects are led by multi-collinearity between control variables. Bivariate correlation between scale length and scale direction is 0.792, and that between scale direction and processing error index is 0.457. Additionally, the variance inflation factors of scale length and scale direction are greater than 10, which is commonly regarded as a sign of severe multi-collinearity (O'Brien 2007). Therefore, the effects for the control variables should be cautiously interpreted, considering the serious multi-collinearity problem.

Changes in ICCs from Model 1 to Model 8 is an 8.43% increase (=[0.090-0.083]/0.083) for country-year level and a 6.15% decline (=[0.122-0.130]/0.130) for survey level. This suggests that the harmonization and survey quality controls increase the similarity of trust in parliament among individuals in different surveys within the same country-year and reduce unexplained variance of protest potential between surveys in the same country-year. The changes in the ICCs indicate that both the properties of the question in the source survey and the qualities of the survey play a role in increasing homogeneity of trust in parliament between surveys—that is, they reduce inter-survey variability.

Table 3 presents results from three-level multilevel models for trust in legal system and political parties. Although the strengths of the coefficients in these two models differ from the model for trust in parliament, except for processing error index in the model for trust in political parties, the directions of the coefficients are the same as in Table 1. For trust in the legal system, ICC in the constrained model increases 6.32% (=[0.101-0.095]/0.095) for country-year level and reduces 6.92% (=[0.121-0.130]/0.121) for survey level compared to the unconstrained model. For trust in political parties, ICC in the constrained model is higher 1.59% (=[0.064-0.063]/0.064) for country-year level and is lower 5.58% (=[0.088-0.093]/0.088) for survey level than the unconstrained model. Although the extent to which harmonization and survey quality control variables account for this varies, the results suggest that the variables consistently contribute to reducing variation between surveys for institutional trust.

	Trust in Legal System[b]		Trust in Political Parties[c]	
	Model 1	**Model 2**	**Model 1**	**Model 2**
Scale length		-0.220 ***		-0.105 ***
		(0.032)		(0.027)
Scale direction		1.350 ***		0.560 ***
		(0.135)		(0.126)
Scale polarity		0.284 †		0.303 **
		(0.146)		(0.111)
Computer file quality index		0.068		0.129
		(0.059)		(0.078)
Data documentation index		-0.114 ***		-0.057 †
		(0.023)		(0.029)
Processing error index		-0.714 ***		0.111
		(0.122)		(0.115)
Constant	4.628 ***	5.498 ***	3.567 ***	3.752 ***
	(0.071)	(0.169)	(0.054)	(0.174)
Variance Components				
Country-year level	0.579	0.607	0.326	0.327
Survey level	0.208	0.119	0.156	0.125
Individual level	5.283	5.283	4.680	4.680
Intraclass Correlation Coefficients				
Country-year level	0.095	0.101	0.063	0.064
Survey level	0.130	0.121	0.093	0.088
Model Fit Statistics				
AIC	1773454.5	1773378.9	1679504.4	1679479.8
BIC	1773498.0	1773487.8	1679547.8	1679588.4
-2log likelihood	1773446.5	1773358.9	1679496.4	1679459.8

*—$p < 0.05$; **—$p < 0.01$; ***—$p < 0.001$; †—$p < 0.05$ (two-tailed tests).
a—Unstandardized regression coefficients (standard errors in parentheses).
b—N=393,430 individuals nested within 293 surveys in 137 country-years
c—N=382,984 individuals nested within 288 surveys in 137 country-years

Table 3. Three-level Multilevel Analysis for Trust in the Legal System and Political Parties[a]

Conclusion

In this research note, I tested the idea that harmonization and survey quality control variables reduce inter-survey variability in harmonized institutional trust measures. I used data from country-years that have multiple surveys in the SDR database. In three-level multilevel analyses, I accounted for the different properties of questions in the source survey and variation in survey qualities throughout the models for three measures of institutional trust and found an increase in country-year-level ICC and a decline in survey-level ICC. This finding suggests that the control variables designed by the SDR framework can help account for inter-survey variability by improving the homogeneity of institutional trust between surveys in the same country-year, that is, reducing unexplained variance of institutional trust between surveys.

It is important to note that a large portion of variance between surveys remains unexplained. This research focused only on a methodological component of inter-survey variability; therefore, the unexplained variance might be attributed to country-year level substantive variables that affect institutional trust. Future research should examine how the methodological control variables are associated with the effects of other substantive covariates on inter-survey variability in institutional trust. In addition, I invite researchers to develop new potential sources of methodological inter-survey variability, such as weights (Zieliński, Pował, and Kołczyńska 2018) and fieldwork period (Voicu 2019), and construct more elaborate harmonization and survey quality control variables to enhance survey data harmonization.

Joonghyun Kwak is a post-doctoral scholar at The Ohio State University for the Survey Data Recycling (SDR) project, funded by the National Science Foundation. This material is based upon work supported by the National Science Foundation under Grant No. (PTE Federal award 1738502)

References

Kołczyńska, Marta and Matthew Schoene. 2018. "Survey Data Harmonization and the Quality of Data Documentation in Cross-National Surveys." Pp. 963–84 in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, edited by T. P. Johnson, B-E. Pennell, I. A. L. Stoop, and B. Dorer. Hoboken, NJ: Wiley.

Kwak, Joonghyun and Kazimierz M. Slomczynski. 2019. "Aggregating Survey Data on the National Level for Indexing Trust in Public Institutions: On the Effects of Lagged Variables, Data Harmonization Controls, and Data Quality." *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences* 5(1):1–21.

O'Brien, Robert M. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality and Quantity* 41:673–90.

Oleksiyenko, Olena, Ilona Wysmulek, and Anastas Vangeli. 2018. "Identification of Processing Errors in Cross-National Surveys." Pp. 985–1010 in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, edited by T. P. Johnson, B-E. Pennell, I. A. L. Stoop, and B. Dorer. Hoboken, NJ: Wiley.

Pais, Jeremy. 2010. "Competing Sources of Earnings Inequality: A Comparison of Variance Components." *Research in Social Stratification and Mobility* 28:359–73.

Slomczynski, Kazimierz M., J. Craig Jenkins, Irina Tomescu-Dubrow, Marta Kołczyńska, Ilona Wysmułek, Olena Oleksiyenko, Przemek Powałko, and Marcin W. Zieliński. 2017. SDR Master Box (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ VWGF5Q).

Slomczynski, Kazimierz M. and Irina Tomescu-Dubrow. 2018. "Basic Principles of Survey Data Recycling." Pp. 937–62 in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, edited by T. P. Johnson, B-E. Pennell, I. A. L. Stoop, and B. Dorer. Hoboken, NJ: Wiley.

Slomczynski, Kazimierz M., Irina Tomesu-Dubrow, J. Craig Jenkins (with Przemek Powałko, Marta Kołczyńska, Ilona Wysmułek, Olena Oleksiyenko, Marcin W. Zieliński, and Joshua Dubrow). 2016. *Democratic Values and Protest Behavior: Harmonization of Data from International Survey Projects*. Poland: IFiS PAN Publishers.

Tomescu-Dubrow, Irina, Kazimierz M. Slomczynski and Marta Kolczynska. 2017. "Quality Controls and Their Application to Substantive Analyses of Data from International Survey Projects." Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences 3(1):9–13

Voicu, Bogdan. 2019. "Do Differences in Fieldwork Period Affect Survey Comparability? Examining World Values Survey and European Values Study in Romania, 2017-2018." *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences* 5(2):20–27.

Zieliński, Marcin W., Przemek Pował, and Marta Kołczyńska. 2018. "The Past, Present and Future of Statistical Weights in International Survey Projects: Implications for Survey Data Harmonization." Pp. 1035–52 in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, edited by T. P. Johnson, B-E. Pennell, I. A. L. Stoop, and B. Dorer. Hoboken, NJ: Wiley.

Appendix: List of Country-Years in the Analysis: 137 Country-Years in 53 countries, 1995-2012

Country	Year	Country	Year	Country	Year
Argentina	1995, 2006, 2007	Great Britain	2000	Peru	1996, 2006, 2008, 2010
Austria	2007	Greece	2000	Poland	2006, 2008, 2010
Bulgaria	2006	Guatemala	2004, 2006, 2008, 2010	Portugal	2000, 2007
Bolivia	2004, 2006, 2008	Honduras	2004, 2006, 2008, 2010	Paraguay	2006, 2008, 2010
Brazil	1997, 2006, 2007, 2008, 2010	Hungary	2006, 2010	Serbia	2006
Canada	2006	Indonesia	2006	Russian Federation	2006, 2008
Switzerland	2008	Ireland	2000, 2007	Sweden	2000, 2006, 2010
Chile	1996, 2006, 2007	Iceland	2012	Slovenia	2006, 2008, 2010
Colombia	1997, 1998, 2004, 2005, 2006, 2008, 2010	Italy	2000, 2010	Slovakia	2007, 2008, 2010
Costa Rica	2004, 2006, 2008, 2010	Latvia	1996	El Salvador	1999, 2004, 2006, 2008, 2010
Cyprus	2006, 2008	Moldova	2006	Turkey	2006, 2007, 2009
Germany	2010	Macedonia	2008	Taiwan	2006
Denmark	2008	Mongolia	2006, 2010	Ukraine	2006
Dominican Republic	2006, 2008, 2010	Mexico	1996, 2004, 2005, 2006, 2008, 2010	United States	2006
Ecuador	2006, 2008, 2010	Nicaragua	2004, 2006, 2008, 2010	Uruguay	1996, 2005, 2006, 2007, 2008, 2010
Estonia	1996, 2004, 2007	Netherlands	2006, 2008, 2012	Venezuela	1996, 2007, 2008
Spain	2000, 2006, 2007, 2008	Norway	2007, 2008	Viet Nam	2006
France	2000, 2006, 2008, 2010	Panama	2004, 2006, 2008, 2010		

Assessing the Quality of Survey Processes and Outputs: The 4th European Quality of Life Survey

By Julie de Jong & Kristen Cibelli Hibben

The European Quality of Life Survey (EQLS), carried out by European Union (EU) agency Eurofound, collects data every four to five years through face-to-face surveys with nationally representative samples across member and candidate states on current life conditions and attitudes on a wide range of topics. Given the potential impact of the data to inform EU policy and social research in Europe more broadly, it is critical for Eurofound to ensure rigorous and comparable data.

The most recent wave of the survey – the 4th EQLS – was conducted in 2016 and early 2017 in 33 EU and EU Candidate countries. Upon completion of fieldwork, a number of organizations received an invitation to submit a tender for an external quality assessment of survey outputs and processes of the 4th EQLS and recommendations for improving quality of future surveys. As survey methodologists at the Survey Research Center International Unit at the University of Michigan, specializing in comparative survey methodology, we bid on and subsequently were awarded the contract to complete the assessment. The following is a brief description of the assessment process that we led, with additional support from a third methodologist, a sampling statistician, and a linguist specializing in translation studies.

Before we began our assessment, we reviewed existing quality frameworks to ascertain the approach best suited to such an evaluation. There are several frameworks used when assessing survey quality, including: 1) total survey error (TSE); 2) fitness for intended use; and 3) monitoring survey production process quality (Biemer & Lyberg, 2003; Groves et al., 2009; Lyberg & Biemer, 2008; Lyberg & Stukel, 2010).

A first attempt at integrating these three approaches was done in the Harmonization project (Slomczynski, Tomescu-Dubrow & Jenkins 2016), where survey quality is defined along three main survey dimensions – the survey documentation, data processing, and data records in computer files (e.g. Tomescu-Dubrow & Slomczynski 2015). For each dimension, the project developed indicators that correspond to components of TSE, fitness for intended use, and survey process quality monitoring, respectively (e.g. Slomczynski, Tomescu-Dubrow & Jenkins 2015). These indicators were applied to 81 publicly available 3MC data files stemming from 22 international survey projects, to measure variability in survey quality (e.g. Kołczyńska and Schoene 2015; Slomczynski, Powałko & Krauze 2015; Wysmułek & Oleksiyenko 2015; Zielinski and Powałko 2015).

In the 4th EQLS, as in other multinational, multicultural, or multiregional (3MC) surveys, success hinges on the comparability or equivalence of data across many cultures and countries. Yet the challenges of documentation, survey quality assessment procedures and criteria are far more complex in the 3MC context, necessitating, in our view, these approaches for the evaluation of processes and outputs, integrated into the framework shown in Figure 1. Each of these guided our assessment of one or more areas of interest and, in the case of TSE, our recommendations for future surveys.

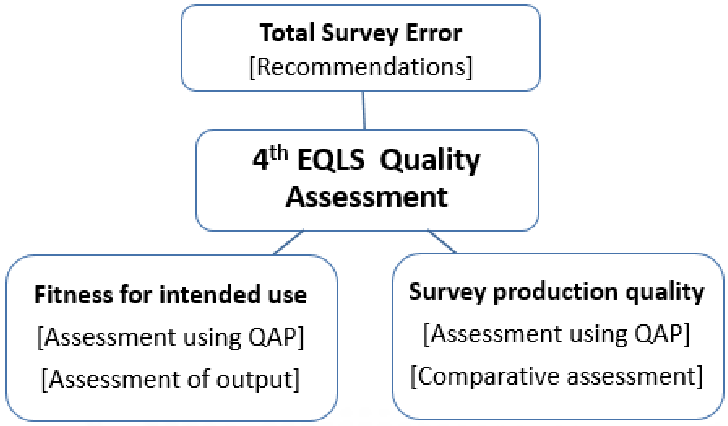


Figure 1. Integrated Quality Approach

Step 1

The first step in our assessment was to evaluate compliance with the Quality Assurance Plan (QAP), which was developed by Eurofound to accompany the Terms of Reference for the design and implementation of the 4th EQLS. The QAP consisted of a detailed set of quality indicators and associated targets, grouped by twelve themes associated with the stages of the survey lifecycle. Each quality indicator is linked to one of the following quality dimensions: Relevance and Timeliness, Accuracy, Punctuality, Accessibility, and Coherence and Comparability. These criteria are widely used within the European Statistical System as a framework to assess quality (Eurostat, 2015). The QAP integrated two approaches to survey quality—fitness for intended use and monitoring survey production process quality—by monitoring and assessing quality according to each of the fitness for intended use quality dimensions.

While the QAP separated the survey lifecycle into twelve themes, we reorganized the indicators into the following four sections:

- Sampling frame development;
- Questionnaire development & advance translation, cognitive testing, and translation;
- Fieldwork (implementation, monitoring, contact procedures, nonresponse, and paradata); and
- Weighting.

Indicators and the associated targets were also prioritized, with many indicators "absolutely required" (in red in Figure 2), and others categorized as "real world" (yellow) and "ideal world" (green). The research and consultancy firm, Kantar Public, was awarded the data collection contract for the 4th EQLS and served as the coordinating center. Following fieldwork, Kantar Public produced a Quality Assurance Report (QAR). Using this report, along with other documentation from Eurofound, we assigned compliance scores (target met, target not met, outcome unknown, target no longer applicable) to each quality indicator. Figure 2 illustrates several of these quality indicators.

Theme	Sub-theme	Key quality dimension addressed	Indicator	Target	UM Assessment
Sampling	Register vs. enumeration	Accuracy	Percentage of countries where a high quality register is used (up-to-date and approximating full coverage)	100%	Not met
	Sampling frame (country)	Accuracy	Percentage of countries where the sampling frame covers at least 95% of the populations	100%	Met
	Reference statistics (overall)	Accessibility	Percentage of countries for which the characteristics of the reference statistics are documented in complete accordance with the template	100%	Met
Translation	Initial translation	Punctuality	Initial translation delivered at agreed date	Y	Met
Fieldforce traininig	Construction of interviewer training materials	Accuracy	Training materials cover strategies for convincing reluctant respondents	Y	Met

Figure 2. Example of Quality Indicators

Although the QAP indicators span five quality dimensions, our assessment focused primarily on those indicators related to the dimension of accuracy, "considered fundamental to product quality" (Biemer et al., 2014, p. 381). Indeed, roughly half (73) of the 145 total indicators were related to accuracy. In general, we found a high level of compliance across all stages. While there are several exceptions, we assessed non-compliance as generally having a minimal effect on data quality. We also evaluated the QAP itself more generally as a framework for assessing quality. Eurofound has made important strides in advancing survey quality through the development and implementation of its QAP. However, its use to assess quality revealed some limitations in the framework itself. We found the QAP to be a useful tool for monitoring survey process quality as initially envisioned. However, as the external assessors, we also found several ways in which the QAP could be improved and determined that it alone could not provide a comprehensive review of survey quality.

Step 2

We turned next from an assessment focused primarily on processes to a review of several different types of survey outputs, including the English-language source questionnaire, nonresponse statistics, sample composition statistics, and several other types of statistical output. This exercise both facilitated an evaluation of the quality of the data and was another method used to further assess the processes used to collect the data within the framework of fitness for intended use.

About half of the questionnaire items in the 4th EQLS were fielded in previous wave(s), while the remaining half were developed specifically for this wave. We conducted an expert review of the source questionnaire, focusing exclusively on the new items. To assess the quality of the final source questionnaire, we drew upon the comprehensive set of guidelines for writing good questions developed by Sudman & Bradburn (1982) summarized and updated by Groves et al. (2009), as well as principles regarding questionnaire design drawn from Saris and Gallhofer (2014), to develop criteria to guide development of survey questions, particularly as they relate to 3MC surveys such as the EQLS. We noted several areas where questionnaire design decisions could both increase measurement error and have implications for comparability between countries. Our item-by-item assessment also uncovered a number of items for which a survey methodologist trained in 3MC questionnaire design may have suggested revisions. We also noted certain items where quality was difficult to assess because the relevant research objectives and analysis intentions were not clear. The Recommendations section of the report included the suggestion that the development of new items, and their placement in the questionnaire, be reviewed by an expert in 3MC questionnaire design in future waves of the survey.

We then examined nonresponse statistics from the 4th EQLS including overall nonresponse rates, and rates of refusal, contact, and cooperation. We first compared nonresponse statistics from the 4th EQLS and the 3rd EQLS, and then with the last two rounds of the European Social Surveys (Round 7 and 8). We note that while the ESS shares neither the organizational structure nor number or breadth of countries as the EQLS, a recent external ESS impact report notes its synergy with other 3MC surveys in the European context and the extent to which it is considered a benchmark in the industry (Kolarz et al., 2017). Given the dearth of comparable surveys, the extensive reach, impact, and use of the ESS data, and the relative value for both stakeholders and data users in understanding how the two surveys compare to each other, we elected to use it here and elsewhere in our assessment as a point of reference to the EQLS. To compare nonresponse, refusal, cooperation, and contact rates, we applied the standard definitions and formulas of the American Association for Public Opinion Research (AAPOR).

Overall, the trend in response statistics between the 3rd and 4th EQLS painted a mixed picture. Response rates increased in 12 countries and refusal rates went down or held constant in approximately half of the study countries. This may be attributable to improved field procedures and increased contact rates, which largely increased or stayed nearly constant (within one percent) in 25 countries. Comparisons between the ESS and EQLS also presented a mixed but perhaps slightly more positive picture. Response rates fell in fewer countries in the ESS than for the EQLS and where response rates did go down, the decline was not as sharp, and overall, we saw less extreme variation in each type of response statistic for the ESS. We noted, however, that the EQLS covers a much larger and more diverse set of countries including a number of countries with less established survey research tradition. These results underscore the importance of continued focus on response rates for the EQLS and the implementation of efforts to better understand the potential effects of nonresponse bias and to further improve response rates in future waves.

We also compared the EQLS data with Eurostat sociodemographic data used to calculate the calibration weights from Eurostat and survey data from the ESS to assess representativeness of the target population. These comparisons suggest that the sampling and fieldwork processes are in line with other major cross-national data sources. Analysis of both the coefficient of variation and the design effects also provided evidence of increased efficiency, which lessens the impact of subsequent weighting calculations on estimates of statistical precision, while increasing comparability.

Step 3

In the final step of the evaluation stage, we assessed the processes of the 4th EQLS against best practices in the survey research industry, particularly as applied to 3MC contexts. We first defined for each of the four phases of the survey lifecycle a set of 3MC survey best practice guidelines, considering both the processes of other major 3MC surveys in the European context, including ESS, Survey of Health, Ageing, and Retirement in Europe (SHARE), and the *Programme for the International Assessment of Adult Competencies (*PIAAC), as well as the survey methodology literature, to support the inclusion of each specific standard in our suggested framework.

We then considered the process followed in relation to each guideline by the ESS and the EQLS, with discussion of each guideline concluding with a statement about overall compliance with the best practice standard in the 4th EQLS. For example, within the section on fieldwork implementation, one such guideline stated "A standard CAPI instrument should be developed centrally and then thoroughly evaluated in all participating countries", noting that the objective of comparability in a 3MC survey necessitates that the design implementation does not contribute to measurement error. In the assessment process, documentation specified that (at that time) the ESS did not have a standardized instrument (European Social Survey, 2017). While the 4th EQLS developed and deployed a standardized CAPI instrument for data collection, documentation on the development process and subsequent testing was sparse and may be related to issues of competition among data collection firms and reluctance to share possible industry advantages.

We found that, overall, a high level of standardization and quality was achieved in sampling frame development and sampling procedures for the 4th EQLS, although there are areas for continuous process improvement. The questionnaire development process incorporated many best practices including consultation with subject matter experts, a translatability assessment, team translation, and a sizable investment in pretesting. However, the process would have benefited from additional expert review and pretesting, as well as more thorough documentation. Fieldwork was a particular area of strength, with key fieldwork procedures highly standardized including the respondent selection process at the household level, the use of a standardized CAPI instrument for both sample management and questionnaire administration, and the pilot test protocol. While we note a few areas where documentation could be improved, we found overall that weighting procedures were in line with best practices.

Step 4

We concluded our assessment with a discussion of our general findings, highlighting the strengths of the EQLS, and providing a series of recommendations for quality assurance and assessment for future Eurofound surveys and 3MC surveys more broadly. We drew on the principles of the SWOT framework, which considers (s)trengths, (w)eaknesses, (o)pportunities, and (t)hreats, to prioritize recommendations by cost and impact, in each of the four main stages of the survey lifecycle, while simultaneously considering the impact in terms of the source of TSE addressed. Some recommendations were guided by the comparisons between best practices, ESS processes, and EQLS processes, while others were a result of considering a broader scope of design decisions faced by the EQLS and what opportunities for improvement might be considered in Eurofound's future surveys.

Using this framework, we considered those items listed under the 'threat' and 'opportunity' sections of each task in the SWOT assessment and prioritized them, noting impacts on specific components within the TSE framework in the 3MC context. Recommendations were organized by the survey lifecycle and listed in order of priority. We identified as high priority recommendations in areas where the threat and opportunity nexus is associated with a greater impact on data quality and associated error, and can potentially be addressed with a relatively low cost solution. Considered within the framework of *fitness for intended use*, the majority of recommendations categorized as *high impact* speak to issues of accuracy. Figure 3 provides an example of recommendations pertaining to sampling frame development, which in the report was followed by justification and suggestions for implementation if applicable.

	Low cost		**High cost**		
	High impact	**Low impact**	**High impact**	**Low impact**	**Source of error addressed**
Consider alternate respondent selection methods	X				Nonresponse error
Calculate effective sample size	X				Sampling error
Clearly define target the population		X			Coverage error, nonresponse error
Thoroughly document sampling frame sources		X			Comparison error
Consider alternatives to enumeration methods			X		Sampling error

Figure 3. Example of Recommendations for Sampling Frame Development Activities

We also included recommendations for Eurofound's reporting of output statistics, data dissemination, and data disclosure, and how the QAP could be enhanced in the future.

Future Application of the Quality Assessment Process

While it can be challenging to monitor quality in international and comparative studies, several other 3MC surveys have also performed internal quality audits and/or have commissioned external quality assessments (Börsch-Supan et al., 2008; Gallup Europe, 2010; Wuyts & Loosveldt, 2019). However, the focus tends to differ in the specific elements of the survey lifecycle between the topic or main field of study of the survey. For example, generally speaking, assessment surveys (e.g., PISA, PIAAC) devote more attention to the psychometric qualities of questions and assessment instruments, official statistics tend to focus on sampling, coverage and nonresponse, health surveys emphasize validated measurement instruments, academic surveys acknowledge the importance of questionnaire testing, and surveys from the market research world advertise timeliness as an important asset. Indeed, assessments of the 2nd and 3rd EQLS (Petrakos et al., 2010; Vila et al., 2013) included extensive analysis of standard errors and design effects for a wide range of variables for each participating country, with limited interpretive discussion on the implications for comparability of the data. In surveys where there are just a few key variables of interest, an analysis of these data can inform interpretation (e.g., a survey focused on a specific health outcome, with several key variables relating to prevalence of a condition). However, in an omnibus survey such as the EQLS and many other studies, such a broad analysis is statistically likely to produce a number of outliers or extreme values, many of which are apt to be spurious and due to statistical chance, rather than due to any issue in the survey process. Evaluators with methods training, who are not experts in the subject area of the survey are less likely to be able to determine when extreme values reflect actual phenomenon versus a spurious result. In contrast, we argue that the type of holistic approach employed in the assessment of the 4th EQLS offers a more comprehensive picture of quality and should be applied to more 3MC survey quality assessments in the future.

Indeed, the best practices for comparative survey research defined for this assessment have been adapted for the forthcoming AAPOR/WAPOR Joint Task Force on Quality in Comparative Surveys. The Task Force also notes the important contribution of external quality assessments and their recommendations to the work of survey research organizations. We hope to have the opportunity to apply the approach developed for the 4th EQLS to other 3MC surveys in the future.

The complete European Quality of Life Survey 2016: Quality Assessment (de Jong and Cibelli Hibben, 2018), can be found at:
https://www.eurofound.europa.eu/sites/default/files/wpef18059.pdf

Julie de Jong is a survey methodologist in the International Unit within the Survey Research Center at the University of Michigan and specializes in design and management of cross-national and single-country surveys, with a focus on increasing data quality and promoting survey research best practices, particularly in countries with limited survey research capacity and infrastructure.

Kristen Cibelli Hibben, previously a survey methodologist in the International Unit within the Survey Research Center at the University of Michigan, is currently at the Collaborating Center for Questionnaire Design and

Evaluation Research (CCQDER) which conducts question evaluation and development studies at the US National Center for Health Statistics and where she continues to work on issues of cross-cultural and cross-national comparability.

References

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley and Sons.

Biemer, P., Trewin, D., Bergdahl, H., & Japec, L. (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics*, *30*(3), 381–415. https://doi.org/10.2478/jos-2014-0022

Börsch-Supan, A., Brugiavini, A., Jürges, H., Kapteyn, A., Mackenbach, J., Siegrist, J., & Weber, G. (2008). First results from the Survey of Health, Ageing and Retirement in Europe (2004-2007). *Starting the Longitudinal Dimension. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).*

European Social Survey. (2017). *ESS Round 8 (2017/2017) technical report*. ESS ERIC.

Eurostat. (2015). *Quality assurance framework of the European Statistical System, Version 1.2*. http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646

Gallup Europe. (2010). *Quality Assessment of the 5th European Working Conditions Survey*. Eurofound.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons Inc.

Kolarz, P., Angelis, J., Krcál, A., Simmonds, P., Traag, V., & Wain, M. (2017). *Comparative impact study of the European Social Survey (ESS) ERIC*. United Kingdom: Technopolis Group.

Kołczyńska M. & Schoene M. (2015) " Quality Control of Variables Harmonized Ex-Post: An Analysis of 22 International Survey Projects."

csdiworkshop.org/wp-content/uploads/2020/04/KolczynskaSchoeneCSDI2015.pdf

Lyberg, L. E., & Biemer, P. P. (2008). Quality assurance and quality control in surveys. In *International handbook of survey methodology*. Lawrence Erlbaum Associates.

Lyberg, L. E., & Stukel, D. M. (2010). Quality assurance and quality control in cross-national comparative studies. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 225–249). John Wiley & Sons Inc.

Petrakos, M., Kleideri, M., & Ieromnimon, A. (2010). *Quality Assessment of the 2nd European Quality of Life Survey*. Eurofound.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.

Slomczynski, K.M, Tomescu-Dubrow I., & Jenkins J.C. (2015) "Re-Thinking Ex Post Harmonization for Comparative Research: Data Recycling of 22 International Survey Projects." Paper presented at the CSDI International Workshop, Mmarch 26-28, London, England

csdiworkshop.org/wp-content/uploads/2020/04/CSDI_2015_SLOMCZYNSKI_TOMESCU_DUBROW.pdf

Slomczynski, K.M., Powalko P, Krauze T. 2015. "The Large Number of Duplicate Records in International Survey Projects: The Need for Data Quality Control." *CONSIRT Working Papers Series* 8 *(CONSIRT Labs: Methodology of Survey Data Harmonization)* at consirt.osu.edu.

Słomczyński, Kazimierz M., Irina Tomescu-Dubrow, J. Craig Jenkins, with Marta Kołczyńska, Przemek Powałko, Ilona Wysmułek, Olena Oleksiyenko, Marcin Zielińsk and Joshua K. Dubrow. 2016. *Democratic Values and Protest Behavior: Harmonization of Data from International Survey Projects*. Warsaw: IFiS Publishers (asc.ohio-state.edu/dataharmonization/wp-content/uploads/2016/11/slomczynski-democratic-1_282-archiw.pdf)

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. Jossey-Bass.

Tomescu-Dubrow I. and Slomczynski K.M. (2015) "Assessment of survey quality. Overview" Lecture delivered at the International Workshop "Cross-national Survey Harmonization and Analysis: Weights, Data Quality and Multi-level

Modeling", May 11-16, 2015, Columbus, OH, USA.

Vila, J., Cervera, J., & Carausu, F. (2013). *Quality assessment of the third European Quality of Life Survey*. Eurofound.

Wuyts, C., & Loosveldt, G. (2019). *Quality matrix for the European Social Survey, Round 8*. European Social Survey.

Wysmułek, I. & Oleksiyenko, O. (2015) "Documentation Quality Assessment in Ex-Post Survey Harmonization: Implications for Comparative Research." csdiworkshop.org/wp-content/uploads/2020/04/CSD2015QC.pdf

Zielinski M & Powalko, P. (2015) "Weighting Data in Ex-Post Harmonization Process: The Consequences of Different Practices of Survey Organizations for Comparative Studies." csdiworkshop.org/wp-content/uploads/2020/04/WeightingData.pdf

News

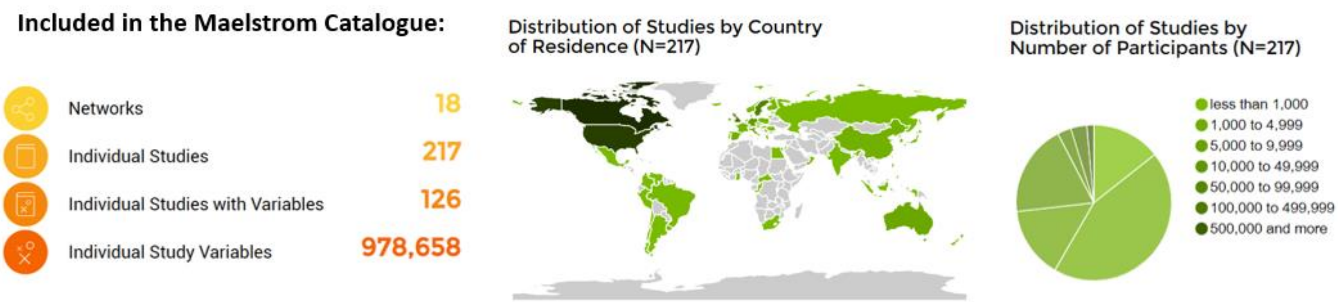
Maelstrom Research: Supporting Collaborative Epidemiological Research through Rigorous Data Documentation, Harmonization, and Co-Analysis

by Tina W. Wey and Isabel Fortier

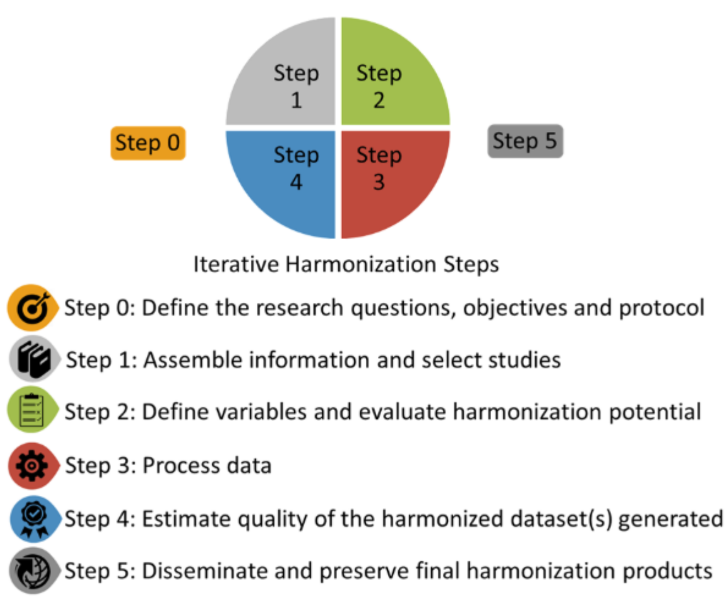
Cohort studies are valuable resources for better understanding the relationship between risk factors and long-term health outcomes and advancing knowledge for public health. Conducting such studies requires enormous investments of time and resources from researchers, participants, and institutions alike, yet the data products often remain under-utilized. Collaborative research programs to leverage existing databases and harmonize data from different studies have increased. However, significant challenges exist due to the quantity and complexity of information collected by individual studies, the heterogeneity of study designs, and the variability of the ethical, legal, social, and cultural contexts affecting data access.

The Maelstrom Research (maelstrom-research.org) initiative aims to leverage the use of existing data resources by providing tools and expertise to promote data discovery, harmonization, co-analysis, and sharing across epidemiological studies. Hosted at the Research Institute of the McGill University Health Centre (Montreal, Canada), Maelstrom Research partners with Canadian and international networks to implement collaborative programs. Maelstrom engages in diverse activities include leading cataloguing and harmonization initiatives, providing expertise and support for others to implement programs, and further developing tools and methods.

To facilitate data discovery, Maelstrom Research has implemented an interactive, publicly accessible online catalogue, which documents and allows exploration of metadata at the level of research networks, individual studies, and study variables. The catalogue serves as the documentation and dissemination platform for 18 partner research networks and five harmonization projects. It currently includes 217 individual studies (comprising over 6,500,000 participants from 60 countries), with comprehensive information on the list variables collected for 126 studies (over 970,000 variables). A search tool facilitates the identification of relevant information for specific research questions and allows users to explore harmonization potential of variables across studies or data collection events. The catalogue was implemented with a cataloguing toolkit[1] developed at Maelstrom and available for use by other initiatives.



To encourage rigorous and effective approaches to data harmonization, Maelstrom Research developed systematic but generic guidelines for rigorous retrospective harmonization[2] based on assessments of existing harmonization projects, expert feedback, and iterative testing of earlier versions. The guidelines delineate and provide the rationale for essential iterative harmonization steps to: define the research question, objectives, and protocol; assemble existing information; define the targeted variables and evaluate harmonization potential; process data; estimate the quality of the harmonized dataset; and disseminate and preserve final products. Fundamental to the guidelines are emphasis on a collaborative framework, expert input, validation of study data inputs, validation of harmonized data outputs, rigorous and transparent documentation, and respect for stakeholders.



To support the implementation of such activities, the Maelstrom Research team utilizes a suite of integrated open-source software[3], including Opal software developed for a secure database repository and management system which interfaces with R or RStudio environments for data processing and analysis and with Mica software designed to catalogue and publish metadata, create data portals, and manage data access.

Maelstrom Research continually seeks to develop further tools and collaborations to meet the needs of the research community. The importance of systematic and adaptable methodologies and tools for collaborative health research are exemplified in the context of the current COVID-19 pandemic. As organizations and researchers rapidly mobilize to gather data to conduct research and inform policy, major challenges will be posed by heterogeneous approaches, and implementing thoughtful harmonization approaches will greatly facilitate the future interoperability and maximize research impact of data being collected now.

We encourage you to visit our website (maelstrom-research.org) or contact us through the website or at info@maelstrom-research.org.

Tina W. Wey, PhD, is a data analyst at Maelstrom Research.

Isabel Fortier, PhD, is the director of Maelstrom Research and a researcher at the Research Institute of the McGill University Health Centre (RI-MUHC).

1. Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. Beiki O, ed. PLOS ONE. 2018; 13(7):e0200926.

2. Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. Published online June 6, 2016.

3. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. Int J Epidemiol. 2017; 46(5):1372-1378.

Building Multi-Source Databases for Comparative Analyses: Event Report

In December 2019, Cross-national Studies: Interdisciplinary Research and Training program (CONSIRT) of The Ohio State University and the Polish Academy of Sciences (PAN), with the support of the Institute of Philosophy and Sociology IFiS PAN, organized the international event, *Building Multi-Source Databases for Comparative Analyses*. The event (December 16-20) comprised a two day conference on survey data harmonization in the social sciences followed by a 3-day workshop on ex-post survey data harmonization methodology. Both the conference and the workshop took place at IFiS PAN in Warsaw, Poland. They were jointly set within the Survey Data Recycling (SDR) Project (NSF 1738502) and the Political Voice and Economic Inequality across Nations and Time (POLINQ) Project (NCN 2016/23/B/HS6/03916). Additional funding was provided by PAN.

The Program of the event, including presentations, are available here on the SDR Project website: asc.ohio-state.edu/dataharmonization/about/events/building-multi-source-databases-december-2019/

About the Conference

The Conference facilitated discussions on methodology of survey data harmonization, and collaboration on the co-edited volume *Survey Data Harmonization in the Social Sciences* that Christof Wolf (University of Mannheim, and GESIS) and the PIs of the Survey Data Recycling (SDR) Project, Kazimierz M. Slomczynski, Irina Tomescu-Dubrow, and J. Craig Jenkins are preparing (under contract with Wiley Publishers). To garner insights from discipline-specific and interdisciplinary views on the challenges inherent to harmonization and how these challenges are met, the conference joined contributions from sociology, political science, demography, economics, and health and medicine.

Presenters included (in alphabetical order): **Piotr Cichocki** (U. Poznan), **Claire Durand** (U. Montreal), **Peter Granda** (U. Michigan), **Ewa Jarosz** (IFiS PAN), **Andrew Klassen** (the HUMAN Surveys Project), **Dean Lillard** (OSU), **Paolo Segatti** (U Milan), **Matthew Sobek** (U. Minnesota), **Jennifer Oser** (Ben-Gurion University of the Negev), **Markus Quandt** (GESIS), **Irina Tomescu-Dubrow** (IFiS PAN and CONSIRT), and **Tina Wey** (The Maelstrom Project, Research Institute of McGill University Health Centre).

About the Workshop

The Workshop highlighted theoretical and practical considerations that researchers have when reprocessing cross-national survey data to build multi-source databases for comparative analyses. It featured a keynote lecture by **Filomena Maggino** (Sapienza Università di Roma and Editor of *Social Indicators Research*). Experiences within the SDR and the POLINQ projects, and the SDR database as a key empirical resource, informed the Workshop.

Day 1, led by members of the **SDR Team**, was devoted to discussing survey data recycling (SDR) as a framework for reprocessing extant cross-national survey data and ex-post harmonization, the structure of the SDR database, and conceptual and practical issues of constructing datasets stemming from the SDR database.

Day 2 focused on missing data imputation. **Stef van Buuren**, professor of Statistical Analysis of Incomplete Data, University of Utrecht and statistician at the Netherlands Organisation for Applied Scientific Research TNO in Leiden, delivered two lectures on missing data imputation for survey datasets with a multi-level structure, focusing on comparability problems. Dr. Michał Kotnarowski (IFiS PAN) led the computer lab sessions.

Day 3 focused on the potential and challenge in the construction and analysis ofmulti-source databases for comparative research, including the issue of using individual-level data from cross-national surveys to construct measures of characteristics of countries in given years (macro-level). Recent experiences of the SDR and POLINQ teams informed discussion, via presentations by (in alphabetical order): **Joshua Dubrow** (IFiS PAN and CONSIRT), **Craig Jenkins** (OSU), **Joonghyun Kwak** (OSU), and **Olga Zelinska** (IFiS PAN). After Maggino's keynote lecture, *Complexity in Society: From Indicators Construction to their Synthesis,* a roundtable discussion led by **Claire Durand, Dean Lillard, Filomena Maggino, Kazimierz M. Slomczynski and Stef van Buuren**, and involving participants closed the Workshop.

Some Survey Data Sources on Covid 19

From the beginning of the Covid 19 Pandemic, survey organizations from around the world are collecting data on how everyday people react to it and to businesses and government policies. These surveys provide the sources of information for comparative research, including future efforts to harmonize ex-post cross-national survey data on Covid 19. To this end, an inventory of available surveys and their documentation would be a valuable resource. Attempts to create such a repository exist.

A prominent initiative is the Societal Experts Action Network (SEAN). SEAN describes itself as an "expert group convened by the US National Academies of Sciences, Engineering, and Medicine, in collaboration with the National Science Foundation, to connect policymakers, researchers and the public with critical social, behavioral and economic inquiry relating to the pandemic."[1] SEAN is a product of a private company called Langer Associates. They have a "COVID-19 Survey Archive", in which they list their survey sources, such as PEW and IPSOS,

[1] https://covid-19.parc.us.com/client/index.html#/

among many others. Their weekly reports group survey topics on how the pandemic is related to the economy, education, politics, social cohesion, and public health, among others.[1]

Another international compilation is Gilani's Gallopedia, which they write was in support for WAPOR, the World Association of Public Opinion Research. It collected polls related to the coronavirus from January to the end of April 2020. Their topic list: "General awareness about the virus, perceived threat, impact on daily life, satisfaction with government responses to the situation, and financial impact of Coronavirus." gilani-gallopedia.org/01/Gallopedia633-01-Item.htm As of April, they had a list of surveys from 39 countries and several cross-national surveys, for a total of 219 surveys. In May, the list was updated and moved to WAPOR's website wapor.org/resources/covid-19-public-opinion-research. We note a survey conducted by Kantar for 21 European countries in the last week of April for the European Parliament (europarl.europa.eu/at-your-service/en/be-heard/eurobarometer/public-opinion-in-the-eu-in-time-of-coronavirus-crisis).

[1] From the May 22, 2020 report: "Reopening America," "Contact and Concern," "Daily Life," "Impact on Education," "Economic Impacts," "Health Impact," "Healthcare and Public Health Attitudes," "Potential Mitigation Strategies," "Government Response," "Voting in November," "Coming Together," "State Polls," and "International Results."

Harmonization would like to hear from you!

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

1. Send us content!

Send us your announcements (100 words max.), conference and workshop summaries (500 words max.), and new publications (250 words max.) that center on survey data harmonization in the social sciences; send us your short research notes and articles (500-1000 words) on survey data harmonization in the social sciences. We are especially interested in advancing the methodology of survey data harmonization. Send it to the co-editors, Irina Tomescu-Dubrow tomescu.1@osu.edu and Joshua K. Dubrow, dubrow.2@osu.edu.

2. Tell your colleagues!

To help build a community, this *Newsletter* is open access. We encourage you to share it in an email, blog, or social media.

Support

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

Copyright Information