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 2
Fall/Winter 2020

Editors
Irina Tomescu-Dubrow
and
Joshua K. Dubrow
CONSIRT

consirt.osu.edu/newsletter
ISSN 2392-0858

Patience

Patience has been defined as "the capacity to accept or tolerate delay, problems, or suffering without becoming annoyed or anxious." Patience can, in turns, be a virtue and a hindrance. Annoyance and anxiety are beasts to be tamed in order to solve problems through even-keeled deliberation; or they are beasts to be harnessed in order to solve problems through agitation and direct action. Covid 19's second and third waves have forced upon us a time to reflect on what patience means and how it benefits us and those around us. As you reflect, please wear a mask.

In this issue of *Harmonization*, **Marcin W. Zieliński** and **Anna Turner** present their research on harmonizing data from different sources with a focus on Google search data and surveys; **Kazimierz M. Slomczynski** and **Zuzanna Skóra** discuss rating scales in inter-survey harmonization; **Ranjit Singh** returns with news about his on-going GESIS web series on the topic of ex-post harmonization; and we present a **book round-up** featuring recent works in and around survey data harmonization.

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

In This Issue…

Harmonizing Data from Different Sources, p. 2

Rating scales in Inter-survey Harmonization, p. 16

News, p. 29

Book Round-up, p. 30

Support & Copyright Information, p. 32

The editors thank Kazimierz Slomczynski for his assistance with this newsletter.

Articles

To What Extent Do Aggregate Measures of Google Searches Relate to Individual Responses to Survey Items? On Harmonizing Data from Different Sources

by Marcin W. Zieliński and Anna Turner

In this note we reflect on how machine-collected big data can be analyzed together with cross-national survey data in a comparative research framework. Specifically, we use search queries from Google, spanning a period of two years (at monthly intervals, from April 2013 to March 2015) and 25 European countries, to estimate the level of public interest in seven topics (Edward Snowden, Julian Assange, WikiLeaks, NSA, online surveillance, privacy, and data protection) that relate to the Edward Snowden revelations. For the same countries, we use survey data from the Special Eurobarometer 431: Data Protection, conducted in March 2015, to measure peoples' awareness of government surveillance (Special Eurobarometer 2015).[1]

We assume that the Google searches capture personal interest in issues related to privacy of behavior on the Internet, government surveillance, and security breaches. Unlike that reported in surveys, this interest is spontaneous, meaning that it does not depend on a researcher's decision to study a given topic using specific questions. Put differently, Google data are a direct observation of peoples' information-seeking about privacy breaches that the media present in connection with data leakage scandals.

Cross-national survey data can measure knowledge about whether and how government agencies gather personal data for security purposes. These survey data were collected in the last month of the period that the Google data cover. As such, we can expect that at least a fraction of the population who expressed spontaneous interest in general data protection issues should also know about the issues that the survey items ask about.

Our hypothesis is that, although they stem from different sources whose underlying methodology for collecting information is dissimilar, Google search data and survey-derived measures of interest in, and knowledge of, internet privacy, government surveillance, and security breaches, will be positively correlated. The reason is that both data types can be used to measure the same phenomenon.

[1] QB6 "Have you ever heard of recent revelations about government agencies collecting personal data on a large scale for the purpose of national security?" QB7 "Would you say these recent revelations have had an impact on the trust in how your online personal data is used?"

We link our discussion of how and why researchers can join these different types of data to issues of comparability and data harmonization. The primary goal of harmonization is to obtain comparable data. The classic typology of this set of methods involves either ex-ante (i.e. prior to data release) or ex-post procedures (e.g. Granda and Blasczyk 2016). Both types of harmonization are based on a "source to target" methodological framework and are frequently applied to survey data.

In the case of machine-collected big data, which are gathered by a kind of naturalistic observation, measurements are taken in a standardized manner using the same unit regardless of the population. These measurements are standardized so long as the data were collected using the same algorithm and no changes in its way of operating were made by a human. In such cases, harmonization of input values is not needed. Yet, the specificity of the subject of measurement prevents direct comparisons between populations: the probability of becoming a subject is unequal and depends on engagement in the activity that is the subject of observation. Such situations are frequent in the case of observational data even if, theoretically, we can observe the whole population.

In sampling-based studies, we generally have to deal with two situations of subjects eluding observation. The first concerns survey non-response, when the reason for non-observation depends on some property of the observed variable itself. If so, we are dealing with bias, that is, a pattern of "missingness" that is not at random (MNAR) (Little and Rubin, 1987). For MNAR the bias cannot be corrected, because the relation between the fact of not participating in the study and the variable of interest is unknown.

In the second situation of survey non-response, observations are missing at random (MAR). If so, most often we can correct the data through weighting or statistical modelling. In both situations we just discussed, the populations for which the measurement was carried out need harmonization, even though we may have fully harmonized concepts and measures.

We face a similar problem with data collected in the course of 'natural' observation, for example, using Google's Keyword Planner. This tool allows researchers to measure the number of searches of particular words or phrases that people use when they try to find something in Google's search engine. The method of collecting (i.e. measuring) Google searches is standardized, because data are accumulated and aggregated at country level directly from Google's keywords database through the Google Ads API.[2] Thus, the resultant country-level measures are directly comparable: each counts the total number of monthly searches for a given topic, carried out using the Google search engine.

However, these raw data obscure potential behavioral differences of population members. There is variation in who initiates Google searches, likely along systematic characteristics of users themselves, and of the contexts in which they are embedded. People vary in their level of interest in a given topic. More importantly, acting on one's interest, and thus being observed as a Google user

[2] Standardized in the meaning of involving an algorithm for collecting data that makes the unit of observation the same regardless of the time and place it was collected.

searching for terms specified in this article, depends on factors such as access to a computer or any other device that enables Internet use, access to the Internet, and use of Google's search engine. We can expect such systematic variability among users within a given country, and from different countries. Consequently, to analyze computer-generated big data cross-nationally, we need to find ways to harmonize the populations from which these data are collected.

Data Leaks, Data Surveillance, and Tested Hypothesis

In June 2013, Edward Snowden, the computer analyst and whistle blower, publicized information about the top-secret PRISM global data surveillance program of the US National Security Agency (NSA). The world learned that, since 2007 and without the knowledge or consent of their users, companies like Google, Facebook, Microsoft, and Apple provided data such as IP addresses, logins, passwords, sent messages, photos, posts, and videos to NSA. Additional material published later by WikiLeaks further confirmed that the neither the NSA nor major Silicon Valley companies respected the right to Internet privacy.

Snowden's disclosures, widely covered in both traditional and social media, sparked a debate about the global reach of surveillance and increased social science research interest in surveillance and privacy topics. Several studies emerged that measured the extent of citizen awareness of these revelations (CIGI-Ipsos 2014; Pew Research Center 2013). Subsequent studies questioned people's attitudes to Snowden's actions and to government surveillance practices (Friedewald et al. 2017; Murata, Adams, and Lara 2017). Some investigated whether people changed their online behavior or introduced privacy protection practices to secure their online data (Hampton et al. 2014; Penney 2016; Dencik and Cable 2017; Marthews and Tucker 2017). Other studies analyzed the broader context by monitoring traditional and social media response to the NSA exposure and whether, as a result, legal regulations were implemented (Davies 2014).

Our analyses of search queries from Google, and survey data from the March 2015 Special Eurobarometer 431: Data Protection that included two questions dedicated to government surveillance (Special Eurobarometer 2015), contribute to this field.[3] We test the following hypothesis: *Public interest in the topics of Data Leaks and Data Surveillance, manifested in Google searches, will correlate positively with survey-derived measures of countries' level of awareness that government agencies collected personal data on a large scale.*

The strategy of analyzing Google and survey data at country level is relatively new, and although Google data are free to the public, perhaps few use Google data because of low awareness and familiarity with Google tools, but also perhaps due to the lack of a clear methodology for such analyses. Our article aims to help clarify this method.

[3] For specific items, see footnote 1.

Data and Methods

Google Data

We choose Google data for the following reasons: (1) Google provides tools that can help to find out what people around the world search for, and how often, with results broken down into countries and search frequency divided into any time period;[4] (2) no other study has measured the public's interest in the topics of surveillance internationally, which prompted us to look for alternative solutions to surveys; (3) Google data have already been successfully used as an indicator of public interest (Mellon 2014, Zhu, Wang, Qin, Wu 2012, Granka 2010, Scharkow & Vogelgesang 2011), as scholars assume that "the more search requests for a certain issue are made, the more salient is to the public" (Maurer and Holbach 2015); (4) the daily updates to Google data constitute useful statistics that show public interest in real time; (5) because Google data are unobtrusive – queries are private and anonymous – they are less prone to social desirability bias than survey information extracted through the process of questioning by an interviewer; (6) within Europe, even in countries with lower Internet penetration, the reach of Google is very high (with the exception of Russia).

As data source, we use Keyword Planner, a tool that Google has provided since 2004 and that, so far, only marketing and marketing research use widely. In the social sciences up to now, researchers used only Google Trends, despite this platform's very limited data availability and it representing the number of queries by Google's normalized index, not as an integer (Turner, Zielinski, and Slomczynski 2018). Google Keyword Planner accounts for the number of searches in a particular country irrespective of population size and regional context. Thus, raw data from Keyword Planner (in our case, the number of Google queries on each of the seven topics related to the Edward Snowden revelations we present below) do not need to be harmonized as they measure different aspects of public interest in a standardized way across countries. However, to be used in comparative analysis, these measures need to be adjusted to account for differences between populations with regard to size of population, internet coverage, and use of Google's search engine.

Constructing a Harmonized Search Indicator Using Google data

There are at least three environmental dependencies due to which the raw Google search data are not directly comparable across countries, despite standardized measurement. These are: (a) access to the computer devices, (b) access to the Internet, (c) using Google as a search engine. Put differently, these three factors cause differences in the populations whose behavior Google data measurement captures, but not to measures themselves.

[4] In Google Trends, you can divide data per minutes, hours, daily, weekly, bi-weekly, monthly. In Google Keyword Planner – monthly.

For us, the first two factors (computer and internet access) can be combined; one always needs a computer or similar device for internet access. But not all who have access to the internet use Google as a search engine.

Thus, instead of measuring the level of public interest using raw number of searches per given topic, we construct a Search Indicator. The Search Indicator captures the same information as the raw Google searches, but includes a correction for internet coverage and penetration of Google search engine among internet users of a given country.

Internet World Stats collects data on computer/Internet users and Google users independently of Google. These data can be expressed as:

$$C_u \subset \mathrm{I_u} \subset \mathrm{G_u}$$

where:

C_u are computer users, $I_u$are Internet users and G_u are Google users. Which leads to $G_u$as a denominator for creating the Search Indicator.

The Search Indicator we built takes into account search queries from Google Keyword Planner to estimate the level of public interest in seven topics: Edward Snowden, Julian Assange, WikiLeaks, NSA, online surveillance, privacy, and data protection. To create the list of keywords we wanted to analyze within each of these topics, we first defined official languages for all 25 countries, and then cooperated with native speakers to define all synonyms of keywords in the respective languages. We then checked all synonyms in Google Keyword Planner.

For each individual topic and by country, we measured the number of searches at 24 month intervals. Specifically, we took the number of searches of keywords in a certain topic (e.g., all keywords in the topic 'Edward Snowden'), in a particular country, divided by the total number of Google users in that country, and multiplied this figure by 100,000.[5] In this way, we established the number of searches for the topic 'Edward Snowden' per 100,000 Google users (G_U) in this country.

This formula allows us to estimate public interest regardless of differences in population size.

$$\frac{\text{Nr of searches ‚Snowden'}}{\text{Nr of people using Google}} \times 100\,000 = \boxed{\begin{array}{c}\text{Number of searches ‚Snowden'} \\ \text{per 100.000 GU}\end{array}}$$

Next, we increased the resulting indicator by 1, and then applied a natural log linear transformation. The natural log linear transformation normalizes the indicator's distribution, which originally was highly skewed to the right. Adding 1 allowed us to keep countries where the original indicator took meaningful 0 values in a particular occasion, since $\log_N(0)$ does not exist (i.e. without adding 1, we would have lost information in the data).

[5] Since there is no source that provides number of Google Users in the countries/continents, we established this number by multiplying number of Internet users in country by Google market share in this country.

After indexing and summing up the number of searches on a monthly basis, we obtain repeated-measures at fixed occasions, where the number of searches on each topic is measured in each country on 24 occasions in the same periods:

$$SearchIndicator = \log_N\{[(\frac{Z_{(1\ \cdots 24)}}{G_u}) * 100000] + 1\}$$

Where:

($Z_{1...24}$) -- number of searches about each topic in each of the consecutive 24 months (from April 2013 to March 2015)

G_U -- number of Google users in a country

The theoretical minimum number of observations for each topic (Snowden, Assange, Wikileaks, NSA, online surveillance, privacy and data protection) in each country at the particular month is zero. There is no theoretical maximum number of observations. Thus, the Search Indicator can range from 0 to infinity.[6] In practice, observational values range from 0 to 7.24. The number of minimum and maximum searches by topic is presented in Table 1. We used this Search Indicator in further analysis.

Table 1. Number of Searches of Each Topic per 100,000 Google Users Regardless of the Month of Observation

Topic of interest	Minimum number of searches per month	Maximum number of searches per month	Mean value of searches per month	Standard deviation of the mean
Snowden	0	1401.3	69.2	128.7
Assange	0.4	249.1	19.0	17.6
WikiLeaks	1.2	397.1	52.1	34.8
NSA	0	22.1	0.9	2.1
Data privacy	0	195.9	17.0	21.9
Online surveillance	0	13.3	0.5	1.2
Data protection	0	329.8	24.7	44.5

[6] As a ln(0+1)=0

Linking Google Data with Administrative and Survey Data

The Special Eurobarometer 43.1 survey, conducted in March 2015, asked 28,000 EU citizens what they think about the protection of their personal data.[7] For our research, we used the question *QB6: Have you ever heard of recent revelations about government agencies collecting personal data on a large scale for the purpose of national security?*

Possible answers were: Yes, No, and I don't know.

We recoded the original scale values so that No answers are coded 0, Yes are coded 1, and people who answer "don't know" are removed from analyses, and then aggregated respondent-level information into country means. The resulting 'mean value' country-level variable captures, for each of the 25 countries, the percentage of persons who indicated that they heard of recent revelations about government agencies massively collecting personal data.[8] We combined this EB-derived aggregate measure with the country-level data on indexed Google searches. For statistical purposes, we treat the survey-derived measure as the independent variable, and Google-based measures as the dependent variable.

Reducing the Multidimensionality of Dependent Variable(S) into Two Concepts: Data Leaks and Data Surveillance

Regarding public interest in Internet privacy, we theorized that two latent variables exist and they are correlated with each other: Data Leaks (includes the *people* and *institutions* involved most in the revealed global mass surveillance program: Edward Snowden, Julian Assange, WikiLeaks, and the NSA), and Data Surveillance (includes the *issues* brought to public attention by Edward Snowden and the media: the scale of mass surveillance, the importance of securing information privacy, and data protection). We confirmed this assumption statistically by using confirmatory factor analysis (CFA, see Figure 1). The correlation coefficient for Data Leaks and Data Surveillance is 0.62.

[7] https://web.archive.org/web/2019*/https://data.europa.eu/euodp/en/data/dataset/S2075_83_1_431_ENG

[8] If multiplied by 100.

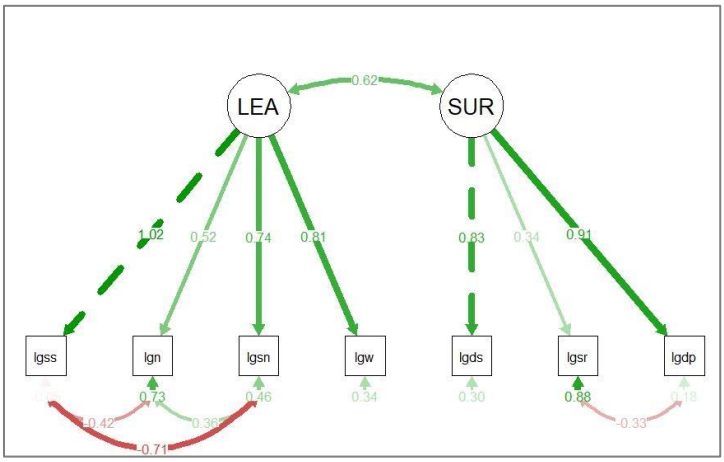

Figure 1. Confirmatory Factor Analysis for Data Leaks and Data Surveillance

Analytical Model

To examine the statistical relation between the Google-based and survey-derived measures of public interest in internet surveillance and security, we estimate Google data with survey data without assuming causality. There are strong reasons to expect that the two types of indicators relate to each other.

The specifics of Google data are crucial in thinking about the nature of this relation. Google Data measure a level of public interest: If someone searches in Google, they are already interested in, and aware of, the topic they are searching for – searches are not random; rather, they are purposeful. For Google to provide information on a topic, the searcher must put in key words. Consequently, we cannot assume that those who positively answered the surveillance question in Eurobarometer had first found out about surveillance in Google. We assume that searchers were already aware of the issue before they started searching. Yet, it is possible that reading more about the issues they searched for via Google reinforced this awareness. In short, while we measure the level of this interest/awareness first with Google data (2013 - 2015) and then with survey data (March 2015), it is reasonable to assume that both phenomena influence each other.

To investigate to what extent Google searches can be associated with individual responses collected in EB surveys, we used a type of multilevel model known as means-as-outcomes model (Raudenbush and Bryk 2001), which can be expressed as:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(MEAN\ SQ_j) + u_{0j}$$

which can be substituted into:

$$Y_{0j} = \gamma_{00} + \gamma_{01}(MEAN\ SQ_j) + u_{0j} + \varepsilon_{ij}$$

The dependent variable -- indexed Google searches -- has two dimensions – Data Leaks and Data Surveillance. Each is measured by the logarithm of number of searches calculated as an average of 24 measured time points in each country separately. The independent variable, constructed via aggregation of Eurobarometer survey data, is the mean value response for the question on knowledge of collecting personal information by governmental agencies (i.e. the proportion of people who have heard about this issue).

Results

Data Leaks

Results presented below show the relationship between variables of interest and lend support to our hypothesis. Public interest in the topic of Data Leaks is positively correlated with level of awareness about government agencies collecting personal data on a large scale.

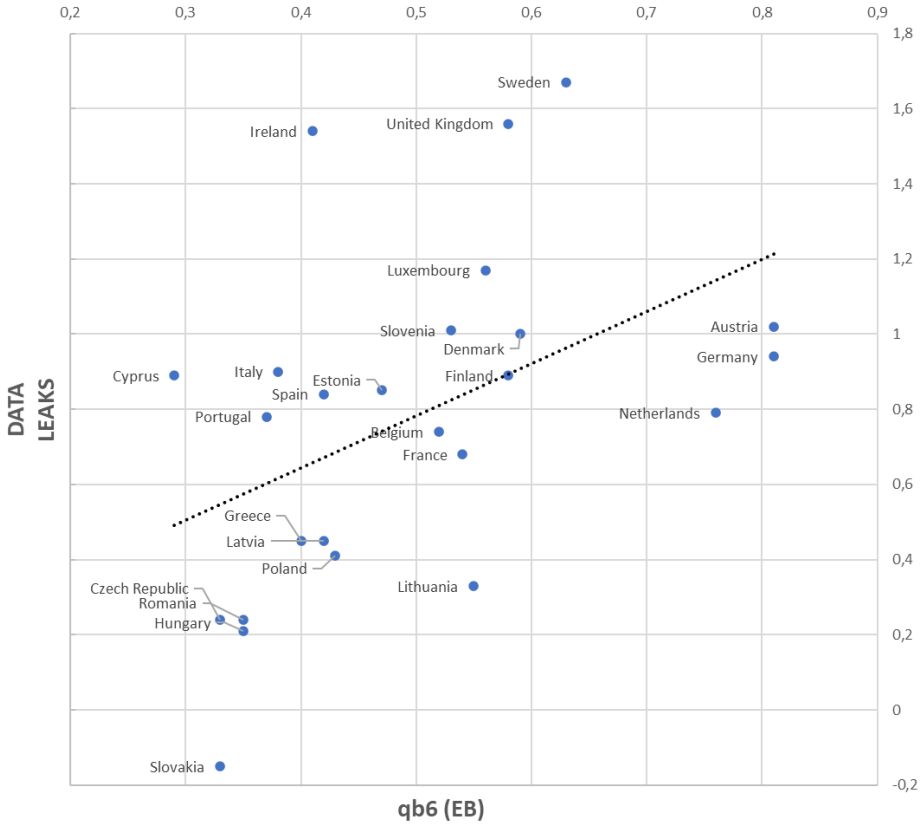

Figure 2. Relationship between Public Interest in Data Leaks Measured with Google Data, and Aggregated Survey Responses from Eurobarometer

As seen in Figure 2, countries differ with regard to the relation between public interest in Data Leaks topics and degree of awareness that government agencies collected personal data on a large scale. For one group of countries, public interest in Data Leaks topics is low and the level of awareness is low - this occurs mainly in post-communist countries: Slovakia, Hungary, Romania, Czech Republic, Poland, and Latvia, and additionally in Greece. In the other group are countries with high level of interest in the subject of Data Leaks and high level of awareness. These are mainly Western European democracies: Sweden, United Kingdom, Luxembourg, Denmark, and also Slovenia.

Data Surveillance

We find a similar relation with regards to Data Surveillance, which we expected given the high correlation (0.62) between Data Leaks and Data Surveillance. Public interest in the topics of Data Surveillance is positively correlated with level of awareness about government agencies collecting personal data on a large scale. However in this case, the relationship is not as strong as in the case of Data Leaks.

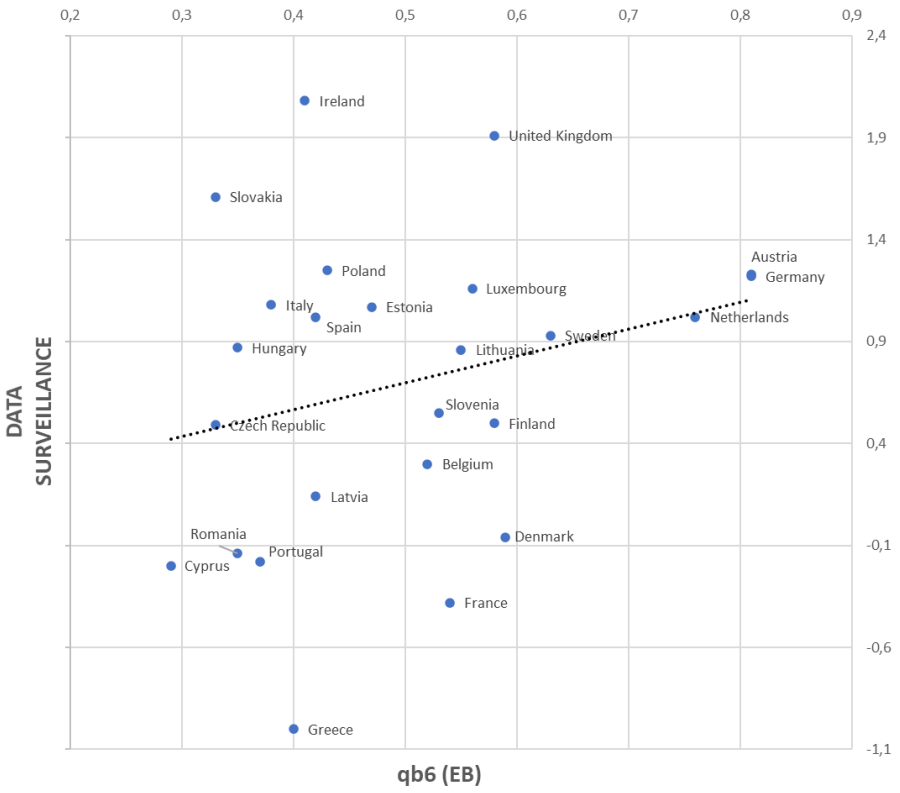

Figure 3. Relationship between Public Interest in Data Surveillance Measured with Google Data and Survey Responses from Eurobarometer

As seen in Figure 3, Greece, Cyprus, Portugal and Romania form a cluster of countries where public interest and awareness in Data Surveillance topics is low. High level of public interest and awareness in the Data Surveillance occurs only in the United Kingdom. Further analysis with country level variables will be necessary to explore if economic, political, social, or perhaps cultural determinants can explain such diversity.

Conclusions

These analyses are a fruitful first step in showing how data from different sources can be combined and how they can be efficiently used in statistical modelling. Google searches were linked with survey data, however administrative data were also used as a basis to harmonize populations from which Google data were collected. Such a step was necessary because of between-country differences in population size, in access to devices that enable Internet use, in access to the Internet, and in use of the Google search engine. This type of harmonization, where its subject is the context (between-country differences) in which the data originated rather than input data themselves, adds a different perspective to ex-ante or ex-post procedures applied to definitions of substantive concepts and their measures.

Google data and survey data capture and measure something that is likely driven by a common factor, which we assume to be Edward Snowden's revelations about the top-secret global surveillance program run by the US National Security Agency. Snowden's disclosure of secret information provoked an international debate in which politicians, journalists, scientists, and ordinary Internet users took part. Attention was directed to the clash between, on the one hand, the fundamental human right to privacy and, on the other hand, the government's obligation to provide security.

Google data and survey data provide evidence that Snowden's revelations became widely known and thus had some effect on societies. One can hypothesize that these two types of indicators, (a) level of public interest as measured by Google searches, and (b) awareness of government surveillance as reflected in aggregate survey data, have a reciprocal, reinforcing effect. People search for issues about internet security to the extent to which they are aware that something is going on, and this in turn leads to greater levels of societal awareness.

Can this relation be untangled? Public interest in a given subject is a broader concept than just the moment at which it is measured. Unfortunately, in our study there is no "time zero" for survey data (i.e. we do not have items in major cross-national surveys of Europe about peoples' awareness of government surveillance before Snowden's revelations). With Google search data, it would be possible to measure the incidence of searches before and after Snowden's disclosure. This opens possibilities for further study of the relationship between concepts of public interest and public awareness.

The research for this paper was supported by a grant from the National Science Center (Preludium, 2017/25/N/HS6/01169, Anna Turner Principal Investigator), with support from the Cross-national Studies: Interdisciplinary Research and Training Program (CONSIRT, consirt.osu.edu), of which the authors are members. CONSIRT is a program of the Polish Academy of Sciences (PAN) and The Ohio State University (OSU), based at the Institute of Philosophy and Sociology PAN and the Department of Sociology at OSU.

Marcin W. Zieliński, Assistant Professor at the Institute of Philosophy and Sociology, Polish Academy of Sciences, sociologist specialized in survey methodology.

Anna Turner received her PhD from the Polish Academy of Sciences. Dr. Turner's research is on the use of Google search data in business and academic research.

References

Bakir, V., Cable, J., Dencik, L, Hintz, A., McStay, A. (2015). Public Feeling on Privacy, Security and Surveillance: A Report by DATA-PSST and DCSS'. Monograph. 2015. https://dcssproject.net/public-feeling/.

Bauman, Z, Bigo, D., Esteves, P. Guild, E., Jabri, V., Lyon, D., Walker, R. B. J. (2014). After Snowden: Rethinking the impact of surveillance. *International Political Sociology*, 8, (2), 121–44. DOI: 10.1111/ips.12048.

Bauman, Z. & Lyon, D. (2012). *Liquid Surveillance: A Conversation*. Hoboken, NJ: Wiley.

Castells, M. (2011). *The Rise of the Network Society*. Hoboken, NJ: Wiley.

CIGI-Ipsos. (2014). Global Survey on Internet Security and Trust. Centre for International Governance Innovation. https://www.cigionline.org/internet-survey-2014.

Davies, S. (2014). A Crisis of Accountability. A Global Analysis of the Impact of the Snowden Revelations. Privacy Surgeon. https://www.academia.edu/7305250/A_Crisis_of_Accountability_A_Global_Analysis_of_the_Impact_of_the_Snowden_Revelations_2014_University_of_Amsterdam_and_Vrie_Universiteit_of_Brussels_

Dencik, L. & Cable, J. (2017). Digital citizenship and surveillance. The advent of surveillance realism: Public opinion and activist responses to the Snowden Leaks. *International Journal of Communication*, 11, (0), 19-32.

Friedewald, M., Burgess, P. J., Cas J., Bellanova, R., Peissl, W. (2017). Surveillance, Privacy and Security: Citizens' Perspectives. New York: Routledge.

Granda, P. & Blasczyk, E. (2016): Data harmonization. In: Cross-cultural Survey Guidelines. https://ccsg.isr.umich.edu/chapters/data-harmonization/ (Accessed 10 Jan 2021

Haggerty, K. D., & Samatas, M. Eds. (2010). *Surveillance and Democracy*. New York: Routledge.

Hampton, K., Rainie, L., Lu, W., Dwyer, M., Shin, I. K., Purcell, K. (2014). Social Media and the "Spiral of Silence". *Pew Research Center: Internet, Science & Tech* (blog). 26 August 2014. https://www.pewresearch.org/internet/2014/08/26/social-media-and-the-spiral-of-silence/.

Lyon, D. (2014). 'Surveillance, Snowden, and Big Data: Capacities, Consequences, Critique'. *Big Data & Society,* 1, (2), 1-13.
DOI: 10.1177/2053951714541861

Marthews, A. & Tucker, C. E. (2017). Government Surveillance and Internet Search Behavior. SSRN Scholarly Paper ID 2412564. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2412564.

Murata, K., Adams, A. A., Lara, P. A. M. (2017). Following Snowden around the world: International comparison of attitudes to Snowden's revelations about the NSA/GCHQ. *Journal of Information, Communication and Ethics in Society,* 15, (3), 183–196.
DOI: 10.1108/JICES-12-2016-0047

Penney, J. (2016). Chilling Effects: Online Surveillance and Wikipedia Use'. SSRN Scholarly Paper ID 2769645. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2769645.

Pew Research Center. (2013). 'Public Split over Impact of NSA Leak, But Most Want Snowden Prosecuted'. *Pew Research Center for the People and the Press* (blog). https://www.people-press.org/2013/06/17/public-split-over-impact-of-nsa-leak-but-most-want-snowden-prosecuted/.

Raudenbush, S. W. & Bryk, A. S. (2001). *Hierarchical Linear Models Applications and Data Analysis Methods*. Thousand Oaks, CA Sage

Special Eurobarometer. (2015). Special Eurobarometer 431: Data protection. http://data.europa.eu/euodp/en/data/dataset/S2075_83_1_431_ENG.

Turner, A., Zielinski, M. W., Slomczynski, K. M. (2018). Google Big Data: charakterystyka i zastosowanie w naukach społecznych. *Studia Socjologiczne*, 4, (231), 49–71. DOI: 10.24425/122482

Watson, H. & Wright, D. (2013). The PRIvacy and Security MirrorS: Towards a European Framework for Integrated Decision Making. Deliverable 7.1: Report on Existing Surveys. https://cordis.europa.eu/project/id/285399

Rating Scales in Inter-Survey Harmonization: What Should Be Controlled? And How?

by Kazimierz M. Slomczynski and Zuzanna Skóra

In the context of survey research, rating scales S are sets of ordered responses (options) to closed-ended questions. The responses refer to the valuation of an object O, concerning an attribute A by a criterion C in format F. For the question, *How much do you trust your government?*, a rating scale S could be expressed on an 11-point scale from (s_0) *very much* to (s_{10}) *not at all*. In this case, the attribute (A) is *trust*, while the *government* is the object (O); criterion C refers to the *strength* by asking *how much*; the format of the question is direct since it refers to the respondent's opinion as such. Let us assume that we want to harmonize several surveys with similar but not identical questions so that our target variable, called *trust in government*, could be used to analyze the combined dataset. What issues of comparability do we face?

In this note, we describe the main issues pertaining to inter-survey harmonization of rating scales that we face in the Survey Data Recycling, SDR, project (Tomescu-Dubrow and Slomczynski 2016; Slomczynski and Tomescu-Dubrow 2018). We refer here to general issues that apply to any ex-post harmonization of survey data containing rating scales. These issues deal with inter-survey methodological variability of the formulation of the question, especially its meaning and the formal properties of the set of answer options from which respondents are supposed to choose. The SDR approach's main idea is to assume a *standard formulation of the question* and *standard scale for answers,* while controlling for the deviation from these standards across given surveys.[1] Thus, the main issues of this note are narrowed to defining control variables. For SDR, control variables of this type are also described in Slomczynski, Tomescu-Dubrow, and Jenkins (2016), and Kołczyńska and Slomczynski (2018).

Control of meaning of the questionnaire items

Objects (O) *and attributes* (A). In generic terms, objects are entities, things, or beings. Attributes are properties assigned to these objects. In survey questionnaires, attributes describe, in subjective sense, characteristics of objects, their functioning or relationships. For example, in the SDR project, the objects of rating scales items are public institutions, such as *the parliament*, *legal system*, and *political parties*, with an attribute *trust*; respondents evaluate the intensity of the attribute to these objects.

[1] In SDR, all properties of scales are assessed using, for the vast majority of source data, their English language documentation. The source documentation includes description of the study, questionnaires, codebooks, and computer files. In this research note most examples of the questionnaire items and scales come from the SDR documents, but some we took from the other sources.

Criteria (C). They refer to the ways in which attributes are evaluated. In a number of questions, criteria are clearly linked to the attributes. Generally, criteria could refer to attributes' strength (e.g., *very much*, *somewhat*, *a little*, *not at all*), quality (e.g., *very poor*, *poor*, *fair*, *good*, *excellent*), frequency (e.g., *very often*, *often*, *rarely*, *very rarely*), likelihood (e.g., *not at all likely*, *somewhat likely*, *extremely likely*), experience (e.g., *very negative*, *somewhat negative*, *neutral*, *somewhat positive*, *very positive*), engagement (e.g., *very much engaged*, *somewhat engaged*, *not engaged at all*), and many other characteristics. Rating scales could also pertain to the degree of agreement with a statement, like in typical Likert's (1932) scale (e.g., *strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, *strongly disagree*). In the literature, item-specific criteria (IS) are considered preferable over agree-disagree ones (AD); see Saris, Revilla, Krosnick, and Shaefer 2010.

Formulation (F). In surveys, the respondent's opinion is often sought directly: *How strongly do you agree or disagree with…?* However, rating scales include also projection questions (*In your opinion, to what extent do most people agree or disagree with the following statement?*) or hypothetical (conditional) questions (*Assume that you won a lottery for one million dollars, how likely is that you would invest a large portion of this sum in your own business?*).

In addition, for some instances:

Qualifiers (Q). These are additional descriptors of the attributes, providing *time-space reference*, or other important characteristics. For example, some questions on *subjective assessment of health* refer to *presently* or *these days* but others to the *last year*. *Interest in local politics* differs from *interest in politics* in general.

Rating scales are built into the questionnaire items in the form of a question that contains *object* O and *attribute* A for valuation according to *criterion* C, in formulation F, possibly with *qualifiers* Q. In the case of inter-survey variability, semantic control can be exemplified as follows:

Object O, coded: 1 – the meaning of the object is extended in comparison to the standard, 2 - the meaning of the object is curtailed in comparison to the standard, 0 – otherwise. For example, when asking about *attending demonstration*, some surveys add the terms *picket or protest meeting* so that the object of the question goes beyond demonstrations. In contrast, in some questions on *trust in people* the object of trust is narrowed to *people that you meet in everyday life,* while the standard question refers to *people* without any condition.

Attribute A, coded: 1 – the meaning of the attribute is extended in comparison to the standard, 2 - the meaning of the attribute is curtailed in comparison to the standard, 0 – otherwise. For items on *interest in politics*, the attribute *interest* sometimes is extended by adding *and concerned about politics*, or *politics matter for you*. In some surveys, questions about subjective assessment of health are narrowed down to physical health.

Criterion C, coded: 1 – agree-disagree question, 0 – otherwise. Usually, rating scales refer directly to respondent's assessment of intensity of the attribute. However, sometimes the intensity is measured through the degree of agreement with a given statement.

Formulation F, coded: 1 – projection question, 2 – hypothetical (conditional) question, 0 – otherwise. Examples include: *Do you think that people like you*…(projection). *If you would be given*…(condition).

For Qualifiers, coding depends on the particular formulation of the standard question.

Coding schema for O, A, C, F and Q can be, and in the SDR practice sometimes is, more complex. To what extent variables O, A, C, F, and Q are needed (i.e. usable in further statistical analyses) depends *on the degree of inter-survey variation* in the questions' formulation. In the SDR project, we use O, A, and C more frequently than F and Q. Valuation of object O is referenced, explicitly or implicitly, concerning time and space, providing an appropriate context.

Control of the Formal Properties of Answer Options

In the SDR database, surveys frequently differ with respect to three important characteristics of rating scales. We capture these differences through separate, control, variables:

Scale length L, coded: Number of elements in the sets of answers presented to respondents. Values range from 3 to 11.

Scale direction D, coded: 1 if the scale is ascending. i.e., responses are ordered from the lowest intensity of the fulfillment of C; 0 otherwise.

Scale polarity P, coded: 1 if the scale is unipolar, i.e., the concept is measured so that the opposition to the highest intensity of the fulfillment of C is low-intensity fulfillment of C (high satisfaction – low satisfaction); 0 for bipolar scales (using antonyms at its poles, like distrust to trust).

For a comprehensive review of the effects of L, D, P, and other scale characteristics, on the quality of data, see Menold and Bogner (2016), DeCastellarnau (2018). In the SDR project we also take into account other properties of scale, based on the literature (e.g. DeCastellarnau 2018).

Expressing Rating Scales on a Common Metric

The goal of harmonizing rating scales across surveys is to maximally condense the information which these scales contain, preserving their substantive and formal properties. Knowing O, A, C, F, and Q for each survey allows us to decide which surveys should be subject to further harmonization and express a target variable in a common metric.

In preparatory work, taking into account information contained in variables L, D, and P, for each set of ordered answers, we assign consecutive numbers k, k = 1,…,*n*, so that answers with the lowest intensity of C achieved 1, and *n* means the highest intensity of C.[2] These numbers constitute a *preparatory scale* that we further use to construct two target scales, using linear transformation and distribution frequencies, respectively.

Common Scores through a Linear Transformation

The first target scale assigns scores in the interval from 0 to 10, taking into account information from source (original) scales of all lengths, and using the transformed consecutive numbers k, k = 1,…,*n*. We apply the following linear transformation:

$$\text{Target\_scale\_(one)}, l(k) = (10 / n*2) + [(k - 1) *(10 / n)]$$

where $l(k)$ is a score for a target variable corresponding to the initial score k, and n is the number of k-values. Table 1 provides assigned scores for source scales that appear in the SDR database.

Our transformation of preparatory scores k into the final target scores ensures that, independently of the source scale length (L), the median and mean values of target scores equal 5, which is a convenient property. As Table 1 shows, the smaller the length of the source scale, the smaller the variance of scores.

[2] In transforming rating scales, the importance of the variable L is obvious. In addition, the transformation of scales so that the direction of intensity is uniform calls for relying on D. However, we must consider the variable P as well. In the case of *bipolar scale,* sometimes coded as -1, -0.5, 0, 0.5, 1, with the ends having the same absolute value but different numeric signs, we must decide that numbers with different signs lay on the same dimension.

Table 1. Transforming Original Rating Scales into Common Metric from 0 to 10, with the Mean and Median Value 5 and Minimized Inter-scale Differences in the Variability

Original (source) scale	Recodes	Median Mean	Average absolute deviation	Variance	Standard deviation
11-points	0,1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0	5.0	2.72	10.00	3.16
10-points	0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5	5.0	2.50	8.25	2.87
9-points	0.6, 1.7, 2.8, 3.9, 5.0, 6.1, 7.2, 8.3, 9.4	5.0	2.44	8.07	2.84
8-points	0.6, 1.9, 3.1, 4.4, 5.6, 6.9, 8.1, 9.4	5.0	2.50	8.23	2.87
7-points	0.7, 2.1, 3.6, 5.0, 6.4, 7.9, 9.3	5.0	2.46	8.25	2.87
6-points	0.8, 2.5, 4.2, 5.8, 7.5, 9.2	5.0	2.50	8.18	2.86
5-points	1.0, 3.0, 5.0, 7.0, 9.0	5.0	2.40	8.00	2.83
4-points	1.25, 3.75, 6.25, 8.75	5.0	2.50	7.81	2.80
3-points	1.66, 5.00, 8.33	5.0	2.22	7.41	2.72
2-points	2.0, 8.0	5.0	3.00	9.00	3.00

The transformation into Target_scale_(one) has an important property: For a given scale-length, the variability of scores is minimized. In other words, for scales with equal intervals, no other score assignments will give a smaller average of differences from the median, or a smaller average of squared differences from the mean.[3] This property is essential and adds to the property of the constant median and mean scores. It makes scales of different lengths as similar as possible, not only concerning the central tendency of scores but also concerning their variabilities.

From an algebraic point of view, the distribution of values of Target_scale_(one) can be compared between samples s, s = 1,…m. However, to assess statistical differences requires to take into account the standard deviation of scores in the compared samples. As Table 1 shows, the standard deviations depend on the original scales' length even if the inter-scale differences in this parameter are subject to minimization. In particular, the standard deviation of scores on the 11-point scale is 3.16, while the corresponding standard deviation for the 3-point scale is 2.72. Although for source scales from 4 to 10 points, the standard deviation of target scores does not differ much (from 2.80 to 2.87), it should still be controlled. Thus, we advocate that in multi-sample analyses with different original scales' lengths, the control variable L will be used.

[3] Counts are performed on scores of items (not on respondents' answers that could correspond to various distributions, different than the uniform distribution).

The logic of the proposed transformation could be extended, to express the target variable on any convenient sets of scores from 0 to r, using the following expression:

$$\text{Target\_scale\_(r), } l(k) = (r / n*2) + [(k - 1) *(r / n)]$$

In the SDR project, we use a 5-point target scale when most source surveys use this scale length, as is the case, for example, with the measure of *interest in politics* (see Table 2 for the transformation).

Table 2. Interest in Politics: Transforming Source Scales into a Common Metric from 0 to 4, with the Median and Mean Values 2 and Minimized Inter-scale Differences in the Variability

Original (source) scale	Recodes	Median Mean	Average absolute deviation	Variance	Standard deviation
5-points	0.0, 1.0, 2.0, 3.0, 4.0	2.0	1.20	2.00	1.41
4-points	0.5, 1.5, 2.5, 3.5	2.0	1.00	1.25	1.12
3-points	0.7, 2.0, 3.3	2.0	0.87	1.13	1.06

Frequency-distribution Metric

The second transformation is more complex, since it takes into account the distribution of scores in a sample.[4] For an *n*-point scale, for values k, k = 1,…,*n*, where X_i is the percent distribution of the variable in sample s, k was recorded to:

$$\text{Target\_scale\_(two), k} = \text{Sum of individuals of scores k-1} + (X_k / 2)$$

Thus, for a given sample *m*, the scale points correspond to the mid-points of the cumulative distribution of scores k. Scores of the scale are percentiles within the sample. The sample distribution is such that the median value equals 50 (or, more precisely, it is around 50), but the variance, like in the case of the distribution of Target_scale_(one), depends on the scale length.

[4] It is worthwhile to note that one of the two methods of assigning numbers to ordered categories originally proposed by Likert (1932; see also Edwards 1957), called the "sigma method," was also based on the distribution of responses. However, Likert favored the "simple method," assigning consecutive integers, 1 to 5, to ordered categories since it correlated strongly with the more complicated sigma method.

Percentiles are an intuitive way to understand where a value falls within a distribution of values. They correspond to the underlying lognormal distribution (Snell 1964, Wu 2007). Table 3 shows an example of the distribution-based transformation of a 5-point source rating scale.

Table 3. Distribution-based Transformation of Rating Scale

Transformed value k	Percentage distribution	Cumulative percentage distribution	Interval	Interval lower bound plus interval midpoint	Target value rounded to integer
1 (low)	10.7	10.7	0 - 10.7	5.35	5
2	22.8	33.5	10.8 - 33.5	22.15	22
3	32.0	65.5	33.6 - 65.5	49.55	50
4	21.7	87.2	65.6 - 87.2	76.40	76
5 (high)	12.8	100.0	87.3 - 100	93.65	94

Using Target Variables with Controls

Any analysis of the distribution of Target_scale_(one) and Target_scale_(two) in a large set of surveys should take into account controls of both kinds: substantive (O, A, C, F and Q) and formal (L, D, and P). For simplicity, we assume that all these control variables are defined on the survey level, and respondents are nested in surveys.[5] In the remaining discussion, we denote the both substantive and formal controls with Z.

Using Rating Scales in an Arbitrary Metric

In the general framework of measurement theory, scale points are just "observations" governed by a given theoretical (latent) variable. For some analyses, using Target_scale_(one) as an approximation of the theoretical variable could be justified on empirical and pragmatic grounds (Bollen and Barb 1981, Leung 2011, Leung and Wu 2017, Carifo and Perla 2007, 2008, Beauducel and Herzberg 2006,

[5] In practice, the controls are defined on the level of the project wave, and in some cases even on the level of the project. Thus, the models discussed in the following sections become multi-level models involving complex data structures. However, the simplified data structure is sufficient to illustrate how the control variables could be used in substantive analyses.

Norman 2010). In particular, if rating scales are unimodal, mesokurtic, and have close to zero skewness, researchers assume that they approximate interval-level measurement. In this context, we can consider how the Target_scale_(one), denoted by Y, depends on some determinants on the individual level, X, and control variables, Z. This is a problem of a two-level model, in which individuals are nested in surveys. In a regression framework, we can write the equation:

$$Y_{ij} = a + b_{10}X + b_{01}Z + b_{11}X*Z + u_{ij}X + u_{0j} + e_{ij}$$

The effect of Z, expressed by b_{10}, is on the survey level. We include the effect of the interaction of Z with X, that is b_{11}. The segment $[a + b_{10}X + b_{01}Z + b_{11}X*Z]$ is a fixed-effect part of the model. The segment $[u_{ij}X + u_{0j} + e_{ij}]$ refers to the random effects. Since the explanatory variable X and the corresponding error term u_j are multiplied, the resulting error term will be different for different values of X.

Using Ratio Scale in a Frequency-Distribution Metric

The Target_scale_(two) provides percentiles that indicate the percentage of respondents in a given survey for whom their answer is at or below a particular value to which this percentile was assigned. Since the mean value is 50 for each survey, it could not be affected by Z measured on the survey level. However, Z may influence the variance of Y.

Incorporating Controls in Statistical Models (Ordinal or Metric)

When we apply metric models to ordinal data, we make an implicit assumption not only about intervals (the same differences between numerical points on the scale are equivalent) but also about probability distribution (normality). These assumptions could be questioned. Even in a symmetrically constructed scale, people can see that the distance for two low ratings is different from the distance between two high ratings. The probability distribution of the scale could be multi-modal, with very thin or thick tails, or heavily skewed. In cross-national research, there is also an assumption that the applied rating scale is invariant among societies studied. However, research shows that, for example, in subjective health assessment, cultural factors influence choosing the option labeled as "fair" as close or far from "poor." In some cultures, "fair" means "average," but in others, it means less than "average," indicating that the distance between "fair" and "poor" is not culturally invariant (Yan and Hu 2019).

Liddel and Kruschke (2018) provide an overview of the consequences of fitting metric models for ordered variables. Amongst them are: (a) increased rate of false alarms, i.e., "detecting" non-existent effects; (b) loss of statistical power, i.e., failures to detect effects; and (c) inversion of effects. Points (a) and (b) are best demonstrated by analyses of simulated data. Analyses of the real-world data on movie ratings are relevant for point (c), an inversion of effect: depending on whether

we used the ordered or metric model, two movies switched their positions on the rating list. The authors argue that because the ordinal model is based on more plausible assumptions and better describes the data, we should rely on ordinal models' predictions.

Many scholars suggest flexible approaches to ordinal data (e.g., Wright and Masters 1982, Agresti 1984, Harwell and Gatti 2001, Cavanagh and Romanoski 2005, Liu and Agresti 2005, Kolen and Brennan 2014, Jonge, Veenhoven, and Kalmijn 2017, Wetzel and Greiff 2018). In particular, rating scales applied to survey samples can be transformed into *latent scores* through models of item response theory (IRT). For finding the latent scores θj, $k = 1, 2, \ldots, m$, we can use a classical formulation of Samejima (1969; see also Andrich 1978) of the two-parameter model for ordinal scales. As it is known, the two-parameter IRT model could be considered as equivalent to the confirmatory factor analysis (CFA); the latent scores (θj) are equal, after some transformation, to the factor scores (η) for a categorical variable. For comparing various IRT and CFA properties relevant for transforming rating scales, see, among others, Takane and De Leeuw 1987, Reise, Widaman, and Pugh 1993, and Muthén and Muthén 2006. For IRT for rating scales in the Bayesian framework, see Wetzel and Greiff 2018. Besides, a variety of approaches for recalibration of a rating scale, including various forms of equating, can be found in Kolen and Brennan 2014, Jonge, Veenhoven, and Kalmijn 2017; see also Sigh 2020. In epidemiology, the cumulative logit modeling for ordinal response variables (Lee 1992) is still popular.

To perform ordinal regression analysis, we recommend the *brms* package (Bürkner and Vuorre, 2019). The package was written for the use in R environment and its syntax resembles the well-known package for frequentist linear regression - *lme4* (Bates et al. 2015). *Brms* is based on Bayesian methods, which provide accuracy and richness of information. Their flexibility enables fitting a multilevel structure, typical of data in harmonized datasets. On top of modeling the shift in thresholds between different groups, thanks to the easily implemented multilevel structure, we can also model this shift while considering differences between countries or projects. For example, when we predict that the option "fair" in health rating could be understood differently in different countries. By introducing separate intercepts for each survey projects we can capture the effect of our predictor X on the outcome variable Y with minimized influence of between-survey differences. The same logic follows for the controls, Z.

Depending on the specific research question, there are several ways in which we could include scale controls in our *brms*-based ordinal model. Here is an example in which we are interested in differences in health assessment (Y) with respect to gender (X). We assume that the outcome variable is expressed on a harmonized 5-point scale, recoded to consecutive integers, and accompanied by a control variable Z.

The simplest model, only accounting for the predictor and our data structure:

$$Y \sim X + (1 \mid \text{survey} / \text{respondent})$$

With this formulation, we express that respondents are nested within surveys and we account for inter-survey differences while assessing the effect of X (varying intercept model). For simplicity, we have not included here the family from which the response distribution Y comes from. For ordinal regression, we would choose the Cumulative Model which assumes that Y originates from the categorization of a latent continuous variable $\tilde{Y}$. The model provides estimates of the placement of thresholds on $\tilde{Y}$. There are at least three possible ways we could include the control variable Z in such a model:

I. Z may be associated with different health assessment but does not affect the main effect ($Y \sim X$). Once we include Z as a predictor, we will acquire the estimation of Y for each level of our control variable Z (impact of Z).

$$Y \sim X + \mathbf{Z} + (1 \mid survey / respondent)$$

II. Z may influence the effect in question ($Y \sim X$). Including Z, on the left hand-side of the grouping structure, will provide us with the estimate of our effect ($Y \sim X$) for all levels of the control variable Z (varying effect model).

$$Y \sim X + \mathbf{Z} + (1 + \mathbf{Z} \mid survey / respondent)$$

III. Z may influence the absolute response pattern. Notice that the formulation resembles our inclusion of survey/respondent grouping (varying intercept model).

$$Y \sim X + (1 \mid survey / respondent) + (1 \mid \mathbf{Z})$$

These model propositions should be treated only as an initial guide on how we could include scale control variables in data analysis. If there are reasons to think that the response distribution Y fulfills the assumptions of a normal probability distribution, such a model can be easily fit in *brms* and requires only the change in the distribution family to Gaussian distribution.

Comparing Different Methods

One of the central questions is about the relationship between results obtained by applying ordered-data models vs. metric models. Do these results converge? If so, could we rely on a simpler way of assigning numbers to ordered categories? Or, what is the cost of using more complex models – how much more do we learn from them? These are the questions that users of harmonized data may try to answer. In doing so, it is necessary to pay attention to harmonization controls.

Summary

When harmonizing survey data ex-post, there are multiple ways to account for inter-survey differences. This research note described certain properties of rating scales that should be accounted for, since they are likely to influence substantive analyses. The control variables developed within the SDR analytic framework and available in the SDR database capture a set of such properties. Researchers can use them, simultaneously, for: (a) survey selection – to narrow down the scope of the research problem; (b) inclusion in statistical models – to analyze as much data as a research problem requires, while accounting for potential variability in responses that inter-survey differences may generate.

This research was funded in part by the National Science Foundation (PTE Federal award 1738502) "Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research", a joint endeavor of the The Ohio State University and the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Kazimierz M. Slomczynski is Emeritus Professor at The Ohio State University and co-PI of the National Science Foundation grant that funded this research.

Zuzanna Skóra is a research assistant in the Survey Data Recycling project and a member of the Consciousness Lab at the Jagiellonian University. She is in the process of finishing her PhD on the topic of measuring consciousness with subjective visibility scales.

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrica*, 43 (4), 561–573.

Bates D, Mächler M, Bolker B, Walker S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, (1), 1–48.

Beauducel, A. & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186–203.

Bollen, K. A. & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, 46, (2), 232–239.

Bürkner, P. C. & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2 (1), 77-101.

Menold, N. & Bogner, K. (2016). Design of Rating Scales in Questionnaires. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences.

Carifio, J. & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Science*, 3, (3), 106-111.

Carifio, J. & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, 42, (12), 1150–1152.

Cavanagh R. F. & Romanoski, J. (2005). Rating scale instruments and measurement. *Learning Environments Research*, 9, (3), 273–289.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: a literature review. *Quality and Quantity*, 52 (4), 1523–1559.

Edwards, A. L. (1957). *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts.
Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in education research. *Review of Educational Research,* 71 (1), 105–131.

Jonge, T. de, Veenhoven, R., Kalmijn, W. (2017). *Diversity in Survey Questions on the Same Topic: Techniques for Improving Comparability*. New York: Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.) New York: Springer.

Kołczyńska, M. & Slomczynski, K. M. (2018). Item metadata as controls for ex post harmonization of international survey projects. In Johnson, T. P., Pennell, B-E., Stoop. I. A. L. & Dorer, B. (Eds.), *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts (3MC).* Hoboken, NJ: Wiley, pp. Pp. 1011-1034.

Lee, J. (1992). Cumulative logit modelling for ordinal response variables: applications to biomedical research. *Bioinformatics*, 8, (6), 555–562.

Leung, S-O. (2011). A Comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research,* 37, (4), 412-421.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328-348.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 (140), 5-55.

Liu, I. & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test* 14, 1–73.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15, (5), 625-632.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

Singh, R. K. (2020). Harmonizing Instruments with Equating. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, 6(1), 11-18.

Saris, W. E., Revilla, M, Krosnick, J., Shaefer, E. M (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods* 4, 1, 61-79.

Snell, E. (1964) A scaling procedure for ordered categorical data, *Biometrics,* 20, 3, 592-607.

Slomczynski, K. M., Tomescu-Dubrow I., Jenkins J. C., with Kołczyńska M., Powałko P., Wysmułek I., Oleksiyenko O., Zieliński M. W, Dubrow J. K. (2016). *Democratic Values and Protest Behavior. Harmonization of Data from International Survey Projects*. Warsaw: IFiS Publishers

Slomczynski, K. M. & Tomescu-Dubrow, I. 2018. Basic principles of Survey Data Recycling. In Johnson, T. P., Pennell, B-E., Stoop. I. A. L. & Dorer, B. (Eds.), *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts (3MC).* Hoboken, NJ: Wiley, pp. 937-962.

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.

Wetzel, E. & Greiff, S. (2018).The world beyond rating scales: Why we should think more carefully about the response format in questionnaires. *European Journal of Psychological Assessment*, 34, (1), 1-5.

Wright, B. & Masters, G. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

Wu, Ch-H. (2007). An empirical study on the transformation of Likert-scale data to numerical Scores. *Applied Mathematical Sciences*, 58, (1), 2851 - 2862.

Wu, H. & Leung, S-O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of Social Service Research,* 43, (4), 527-532.

Yan, T. & Hu, M. (2019). Examining translation and respondents' use of response scales in 3MC surveys. In Johnson, T. P., Pennell, B-E., Stoop. I. A. L. & Dorer, B. (Eds.), *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts (3MC).* Hoboken, NJ: Wiley, pp. 501-518.

News

Methodological Blog Series: Adventures in ex-post Harmonization

by Ranjit K. Singh

Ex-post harmonization is gaining in popularity as it allows us to make use of the treasure trove of existing survey data across the world and unlock its potential to answer novel research questions. At the same time, practitioners of ex-post harmonization still face many daunting methodological challenges. As the field of ex-post harmonization is still developing, there is also a lack of specialized methodological literature. Instead, promising approaches and tools are scattered across many different areas of survey methodology and psychometry.

In a bid to make methodological issues and potential solutions more accessible, I started a series of monthly blogposts here at GESIS. The series discusses the methodological side of ex-post harmonization in easily accessible, bite-sized posts. Topics range from the benefits of ex-post harmonization, over how to decide if variables measure the same construct, to approaches of how to make different variables numerically comparable. Many posts are applicable to a wide range of harmonization issues. In other posts, there is a special focus on harmonizing variables capturing latent constructs (e.g., attitudes, values, interest, or intentions), owing to my current research on the topic.

If you are interested, you can read the introduction to the blog series or jump right to the overview of published and upcoming blog posts.

Aside from the introduction post, there two substantive posts have already been published:

The sum and its parts: The benefits of combining data from different surveys

"Before delving into the "how" of ex-post harmonization, we look into the "why" by exploring the various benefits of and use-cases for ex-post harmonization."

Apples and Oranges: How to find out if two questions measure the same concept?

"Here we delve into an important but difficult matter in harmonization in general and ex-post harmonization in particular: How can we determine if two instruments measure the same concept? This is especially hard if the concepts are latent (i.e., not directly observable)."

The current list of planned post will be completed in April 2021. Afterwards, the series will enter its second season, covering additional topics together with coauthors from GESIS or from ex-post harmonization projects. Topics will then cover new aspects such as harmonizing socio-structural variables, manifest variables such as frequency questions, or linking survey data to other data types (administrative, geospatial, or social media).

To keep posted on new posts in the series (and other resources here at GESIS), you can subscribe to the GESIS Blog (right sidebar) or follow either GESIS (@gesis_org) or me (@_R_K_Singh) on Twitter. Also, the overview of published and upcoming blog posts in the series will be updated with the links to new posts as they get published.

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 survey instruments.

Book Round-up

Survey data harmonization requires methodological knowledge of both surveys and harmonization. Although large-scale harmonization projects involving data collection and data reprocessing have proliferated since the 1980s, social science publications devoted to the methodology of harmonization are scant. We, the newsletter editors, present here a short history of books that provide information that led to the intersection of surveys and data harmonization procedures.

In the last two decades, most survey methods handbooks, while excellent in many ways, do not contain specific chapters on harmonization. These include *The Handbook of Survey Methodology for the Social Sciences* edited by Gideon (2012, Springer) and the *International Handbook of Survey Methodology*, by de Leeuw, Hox, and Dillman (2008, CRC Press). Some, however, include single chapters on a specific type of its methodology, e.g. Hoffmeier-Zlotnik and Warner's chapter, "Harmonization for Cross-National Comparative Social Survey Research: A Case Study Using the 'Private Household' Variable," in *The Palgrave Handbook of Survey Research* (2018 Palgrave Macmillan Cham).

While of great use to survey methodologists and social scientists, and the themes they touch on are relevant for harmonization, most do not explicitly address survey data harmonization as a research process in its own right. Harmonization has specific methodological requirements and challenges – including those posed by documentation and dissemination of harmonized data – that ultimately bear on the extent of data comparability and quality. Notable exceptions are the chapter by Christof Wolf, Silke L. Schneider, Dorothée Behr and Dominique Joye, "Harmonizing Survey Questions Between Cultures and Over Time," in the *SAGE Handbook of Survey Methodology* (2016), and Granda, Wolf and Hadorn's "Harmonizing Survey Data" in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (Harkness, J.A. et al 2010). Moreover, these books do not feature the harmonization experience gained in other disciplines, notably from medicine and health research (e.g. Fortier et al. 2011).

The few books or book sections on harmonization methodology are topic-specific. For example, the multi-authored book *Democratic Values and Protest Behavior: Harmonization of Data from International Survey Projects* (Slomczynski, Tomescu-Dubrow, Jenkins et al. 2016), details the logic of,

and methodology for, creating, via ex-post harmonization, a multi-year multi-country database for cross-national research of political protest. Hoffmeyer-Zlotnik and Warner's book *Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research* (2014) deals with ex-ante output harmonization of socio-demographic indicators. The volume *Advances in Cross-national Comparison: a European Working Book for Demographic and Socio-economic Variables* that Hoffmeyer-Zlotnik and Wolf edited (2003) focuses on ex-ante harmonization of demographic and socio-economic variables for cross-national surveys in Europe. The recently edited volume *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts* (3MC) (Johnson, T.P., Pennell, B., Stoop, I.A.L., Dorer, B., Eds., 2018) features a full section on harmonization, but its five chapters, led by Slomczynski and Tomescu-Dubrow deals with cross-national survey data reprocessing via ex-post harmonization.

To create datasets that facilitate comparative analysis, scholars use survey data harmonization, whether ex-ante or ex-post, to combine survey methods, statistical techniques, and substantive theories. Over the decades, a rapidly growing community of scholars, institutions, and government agencies has used public opinion, health, census, and panel surveys to conduct fascinating research on survey data harmonization. Our newsletter has been a part of that process.

Yet, there remains a knowledge gap between, on one side, the projects that harmonized survey data and, on the other side, a methodological literature on how harmonization occurs. The consequence is a divided field in which harmonized survey data are readily available but harmonization theories, methodological approaches, and best practice recommendations are not widely shared.

In response to this strong imbalance between the relevance and practice of harmonization methods and shared knowledge about harmonization methodology, Irina Tomescu-Dubrow, Christof Wolf, Kazimierz M. Slomczynski, and J. Craig Jenkins will edit the volume *Survey Data Harmonization in the Social Sciences* (under contract at Wiley Publishers). The book is designed to integrate the discussion of concepts and methodology developed around harmonization with practical knowledge, including challenges and best practices, accumulated in the process of building longitudinal and cross-national datasets for comparative research. It takes a multi-disciplinary perspective, where fields such as demography and public health, that have long experience with survey data harmonization, directly communicate with sociology, political science, economics, and survey methodology.

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 dubrow.4@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