The Effects of Data Harmonization on the Survey Research Process

Conference on Building Multi-Source Databases for Comparative Analyses

Peter Granda
University of Michigan

16-17 December 2019
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw

A Bit of History

- Standardization/Harmonization
- Similar developments in data harmonization over time in both the physical and social sciences
- Key role of government statistical agencies (biologic and social)
- Some early references: NIDA desire to create operational definitions in socio-behavioral drug use research because of " a constant struggle in attempting to compare the results of one research effort with another".(1975)
- Eurostat call for greater comparability in statistical data through the fulfillment of "harmonization principles" developed in the post-war period by the UN and OECD (1991)
  - producing standards with worldwide validity
  - drawing up at regional level, e.g. the European Community, standards compatible with world standards, better adapted to the regional specificities
  - linking the methods used for related or complementary fields to each other

Contributions/Developments (among many) over the last 25 years or so

- Smeeding (1996) and Burkhauser/Lillard (2005) articles describing the overall benefits of harmonized and comparative cross-national data files
- Pioneering work of Ruggles on US (and then international) census public-use files
- Clear definitions in the literature of the different types of data harmonization possible:
  - Input
  - Ex-ante output
  - Ex-post output
- Specific practical strategies and techniques recommended for considering and implementing data harmonization protocols as more projects got underway

<-- previous variable next variable -->

CPES V00233 | NLAAS SC8_1 | NSAL C8 †

† For the individual studies, this variable exists in different sections from CPES.

Variable Label: Physical health rating

Total | English | Spanish | Vietnamese | Tagalog | Chinese

SC8_1

En general, ¿diría usted que su salud física es excelente, muy buena, buena, regular, o pobre?

- ○ 1 EXCELENTE
- ○ 2 MUY BUENA
- ○ 3 BUENA
- ○ 4 REGULAR
- ○ 5 POBRE

A discrepancy exists between the question text in the surveys. You may wish to compare question text to determine how it affects the variable results.

View Universe

Value	Label	Frequency	Valid Percent	Total Percent
1	EXCELLENT	263	17.52%	17.52%
2	VERY GOOD	304	20.25%	20.25%
3	GOOD	433	28.85%	28.85%
4	FAIR	417	27.78%	27.78%
5	POOR	84	05.60%	05.60%
.	Missing	-	-	00.00%

Disclaimer: Frequencies displayed above are not weighted.

Missing Data Codes: (Missing)

Some recent initiatives in this space

- Outreach from government agencies responsible for the collection and dissemination of health data to employ data harmonization strategies to improve the analytical power of their datasets (NIH and Canadian Institutes of Health Research)
- Publication of comprehensive guidelines which recommend best practices to create, disseminate, and preserve harmonized data (CCSG and Malestrom)
- New task force report, soon to be completed, on 3MC surveys for submission to AAPOR/WAPOR

Survey Research Lifecycle

Effects of (data) harmonization "MINDSET" on all aspects of the survey process NOT ONLY 3MC SURVEYS

- Initial planning to conduct a survey
- Data collection/field operations
  - More carefully managed fieldwork
  - More overall project coordination
- Data processing, e.g., editing, missing data, imputation, weighting
- Creation of public-use data files and documentation
- Preservation
- Sharing
- Replication

Effect of Data Harmonization Practices on the Archival Community

- A case where data repositories have learned significant lessons from researchers and data producers
- Greater emphasis on standard procedures for ingesting new data collections
- Employing more rigorous processing techniques in preparing data files for public-use
- Defining a common core of materials to fully document data collections and, in particular, WHAT IS REQUIRED FOR HARMONIZED DATA COLLECTIONS
- Production of ex-post harmonized data files as additional "public resources" for meta-analyses and other secondary analytical investigations

Ongoing Importance of Input and Output Harmonization Techniques on the Survey Research Process

- CSDI workshops for more than 15 years (2002)
- High costs of output harmonization encourages more projects to consider implementing input strategies whenever possible and feasible
- Recognized importance of both input and output harmonization strategies on the same project
- New ex post output harmonization projects done by researchers and archives and what might be called “SUPER HARMONIZATION” efforts
- Very strong interest in ex post strategies and creation of recommended guidelines among health researchers
- Strategies to impute item missing data and to specify weights for new ex post harmonized datasets

-1 – Illogical value
0 – Not imputed
1 – Imputed in original study
2 – Logically imputed
4 – Age known, month imputed
5 – Originally imputed as June, recoded to random month
6 – Month imputed
7 – Multiple regression imputed: Not ascertained
8 – Multiple regression imputed: Refused
9 – Multiple regression imputed: Don’t know
10 – Midpoint of a given range
11 – Answer comes from a follow-up question
12 – Midpoint of a range given as a follow-up
13 – Answer is rounded up
14 – Answer is rounded down or is a lower bound
15 – Reported months and weeks, weeks<=4
16 – Reported months=trunc(weeks/4.33)
17 – Reported months=weeks and weeks>4
18 – Reported months!=trunc(weeks/4.33), months!=weeks, and weeks>4
19 – Months coded NA/RF/DK
20 – Weeks coded NA/RF/DK
21 – Impossible to determine which method used last
22 – Computed from other harmonized variables

Where are Data Harmonization Strategies and Techniques Moving in the Future?

- Practical experiences with output harmonization leads to better input harmonization?
- Increasing importance of automation tools, machine learning and AI in reducing the time and expense to complete data harmonization tasks (i.e., "data wrangling") on very large, complex data inputs
- Realization that data harmonization is only the first important foundational step – need for robust analytics platforms and tools to successfully mine large datasets

Data → Harmonization → Analysis → Better Outcomes

Where are Data Harmonization Strategies and Techniques Moving in the Future in Other Research Spaces?

- Health and Medical Data:
  - Large datasets available on similar medical conditions from diverse sources
  - Established rules and guidelines for ex post harmonization to create large datasets on specific health issues to do meta-analyses
  - Large overlap with the experiences of social science researchers
  - Addressing increasing demand from national funding agencies to make better use of existing data to improve health outcomes
  - Greater challenges for this community with attempting to harmonize data on more esoteric research agendas
    - Shift from fee-for-service health models to value-based care
    - Virtual autopsies

Where are Data Harmonization Strategies and Techniques Moving in the Future in the Social Sciences?

- Further experimentation and development of new methods to create more ex post harmonization data resources
- New types of data of great interest to social science researchers but also present great challenges to use such data effectively
- Further integration into the Total Survey Error framework

Total Survey Error (TSE)

- Data processing (editing, data entry, coding, weights, tabulation) identified as one of the chief factors in measuring total survey error.
- Usability/Interpretability, Accessibility recognized dimensions of a survey quality framework.
- Assessment of the magnitude of these errors often neglected in the study of nonsampling error.
- Groves, Lyberg (2010): future research agenda in TSE should include "the role of standards for measuring and reporting sources of error".
- Standard documentation (including both description (codebooks) and assessment (quality profiles) essential to measure TSE.
- DOCUMENTATION: AN INTEGRAL PART OF TSE

SOCIAL MEDIA DATA

- Attractions:
  - Unique data on social/economic issues
  - Unique perspective since not collected directly by researchers
  - Huge number of cases possible
  - Potential for use with survey data to expand research questions and analytical results
  - Immediacy, availability, quantity
- Distractions:
  - Representativeness issues
  - Often incomplete original documentation
  - Obtained in some cases from private data producers
  - Unsure data availability in future and no tradition of preservation
  - DATA NOT PRODUCED FOR RESEARCH PURPOSES

SOCIAL MEDIA DATA HARMONIZATION?

- Urban Systems Example – Location Based Social Networks (LBSN)
  - Retrieved data from different social networks which have different purposes raising comparability issues
  - Millions of potential users
  - Ease of data collection
  - Volunteered geographic information
  - Higher incidence of data from higher income locations
  - Need for data harmonization prior to visualization or analysis
  - Challenges of verification with such large datasets
  - Validation, selection, filtering and interpretation of data based on the source and the research topic under investigation

SOCIAL MEDIA DATA HARMONIZATION?

- Urban Systems Example – Location Based Social Networks

Table 4

Example research topics that can be addressed by combining LBSNs data variables.

	FOURSQUARE	TWITTER	GOOGLE PLACES	INSTAGRAM	AIRBNB
Research topic	Identification of the most visited/checked-in venues. [1]	Spatiotemporal patterns of people presence, activities and languages. [3]	Quantity and diversity of economic activities in an area. [2]	Identify relevant spatial features/ character related to the user experience and perception. [4]	location and clusters of accommodation typology –single family house, multifamily/ apartment building– [5]
Variables selected	UGDAT, ID, TEMP	LOC, TEMP	LOC, CAT, ID	LOC, UGDAT (pho)	LOC, CAT

SOCIAL MEDIA DATA HARMONIZATION?

- Further discussion emanating from BigSurv18 (Big Data Meets Survey Science) and, in particular, session on "Social Science Infrastructure for Big Data", chaired by Christof Wolf, and featuring among others presentations from FORS, GESIS, and ICPSR
- SOMAR at ICPSR: A new archive of curated datasets, workflows, and code for use by social science researchers for the empirical analysis of social media platforms, content, and user behavior.
- ONGOING CHALLENGE: Then, how to combine this new harmonized data source and join it with more traditional social science survey sources, either harmonized themselves or not?

SOMAR Open Questions

- What about content that's integral but not native to the social media post (e.g., links, images, videos)?
- What are the right metadata enhancements?
- How should SOMAR fit/model data management practices?
- How should we connect to existing collections and tools?
- How should we sustain the enterprise?

QUALITY

- Lots of data documentation available, growing more comprehensive every day, but still not nearly enough about the QUALITY of the data being described.
- Even if we have complete descriptions of all variables and good summaries of sampling, weighting, fieldwork, and, in particular, harmonization decisions, etc., we still do not know about the quality of the collected information unless the data producers provide an equally comprehensive assessment.
- Specific need for set of best practices to assess, measure, and report the quality of data harmonization efforts
- Need to include quality indicators as variables in the data files
- Concept of quality profiles as necessary addition to study documentation

Additional Final Thoughts

- Dubrow & Tomescu-Dubrow (2016) describe the emergence of a new methodological field but " without a coordinated effort to build a comprehensive theoretical and methodological base" since "no institutionalized apparatus" existed to provide the means to develop a complete theory of survey data harmonization

- Yet, despite the lack of coordination, the effects of several decades of practical (some successful and some not) data harmonization practices and projects have directly affected almost all aspects of the survey research process

Some Outstanding Issues

- How can we successfully measure the result of ex-post harmonization efforts?
- Can we create a template/set of rules/agreed standards to judge when harmonization efforts are possible?
- How does the increasing number of ex-post harmonization projects affect future input harmonization efforts?
- What to do about the lag between practical harmonization work and agreed community procedures to guide such work?
- WHAT IS THE ROLE OF DATA HARMONIZATION AS RESEARCHERS WANT TO INCREASINGLY COMBINE "DESIGNED" DATA WITH "ORGANIC" DATA?

An Alternate View

In case you thought data harmonization a complicated process with uncertain results, the business world might beg to differ. I quote:

- Simply put, data harmonization is all about creating a "single source of truth." It does this by taking data from disparate sources, clearing away any misleading or inaccurate items, and presenting it as a whole. This means you get a single window view of everything and anything that supports ongoing decision-making, including financial information and business performance.
- Data is coming at you from many different angles. But once it's harmonized, it's been cleaned, sorted, and aggregated to provide a complete picture. And everyone sees the same data. So it's easier to get people on board and easier to steer your business in the right direction.