

XII. Harmonization of Survey and Statistical Data

Introduction

Harmonization in cross-national and cross-cultural contexts occurs when producers of [survey](#) or [statistical data](#) create common measures of key economic, political, social, and health indicators. The goal is to present data that permit a degree of comparability over time or space.

When collecting data on the core characteristics of their populations, national statistical agencies strive in their harmonization efforts to create measures of such important concepts as income, poverty levels, gross domestic product (GDP), etc., which are intended to have the same basic conceptual meaning no matter when or how they were collected. Producers of longitudinal survey data often ask respondents the same question over multiple waves of a single data collection and/or plan the entire survey collection process so that concepts are closely comparable between waves.

Even those who conduct cross-sectional surveys may consider harmonization after the fact ([ex-post](#)) if they decide the subject matter is important enough that users interested in a given topic would benefit from comparing concepts and responses over multiple surveys. Secondary users (for example, individual researchers not directly connected to the original data collection effort, or social science data archives) may undertake harmonization projects of their own to create new data resources.

Various harmonization strategies exist, both [ex-ante](#)/input oriented and ex post/output oriented. Each has advantages and disadvantages.

Guidelines

Goal: To ensure that survey and statistical research teams follow accepted standards when creating harmonized data and documentation files, and use a harmonization strategy that best fits their basic source materials and the objectives they wish to achieve.

- 1. Decide what type of harmonization strategy to employ, taking into account that all harmonization efforts will require some combination of strategies.**
 - Consider 'input' harmonization when the survey collection process derives from a central authority.**

Rationale

'Input' harmonization, usually applied in a multi-national context, seeks to impose strict guidelines by which each country follows both the same survey procedures and a common questionnaire. The strategy permits a high degree of comparability, which allows analysts maximum flexibility in studying the information collected from diverse populations. Effective input implementation of common standards across countries increases the chances of collecting high-quality data in the field. However, this does come with significant cost implications, since it requires a high degree of planning at the very outset of the project, as well as continuous monitoring throughout the process.

Procedural steps

- Create an overall monitoring team that coordinates the work of data collection agencies.
- Produce a comprehensive planning document that provides detailed specifications and procedures for collecting and producing national data files. An example is the Data Protocol of the European Social Survey [\[3\]](#).
- Publish the details of the plan and provide a schedule for the release of public-use files to the user community.

Lessons learned

- This approach may be the most costly because it involves adherence to strict methodologies throughout the data collection life cycle. For example, the ESS seeks to collect data every other year, uses face-to-face interviews, creates and distributes detailed sampling and fieldwork procedures, develops translation protocols in all participating countries well in advance, draws comparable national samples, aims to achieve high response rates, adopts consistent coding procedures, and creates and distributes well-documented datasets in a timely fashion. All of these procedures require greater organizational capabilities and resources throughout the planning and data collection stages, but can result in more valuable public-use data files at the end.
- Not all harmonization projects will be able to follow such procedures, so it is important to decide which methods are best, given the resources at their disposal. In addition, the creation of such common standards and implementation at the local level requires considerable expertise. This also may not be feasible in all cultural contexts.

- It is sometimes difficult to have common standards applied once the survey is in the field. The World Mental Health Survey Initiative created such standards in the planning stage, but was unable to implement all of them, as some were not relevant in each survey country, and some countries' current survey collection practices could not support the recommended standards.
- **Consider 'output' harmonization when the survey collection process is determined at the level of individual countries or cultures and there is no strong central authority.**

Rationale

This type of harmonization is implemented through two main strategies, one "ex-ante" and the other "ex-post." When harmonization has already been considered during survey planning with regard to the development of common goals, measurements and understanding of concepts, the [ex-ante](#) strategy ensures that specific targets are established for the collection of data on key variables. However, the questions used to collect these data may vary from country to country.

The second variant is an [ex-post](#) strategy, by which statistical or survey data are made comparable through a conversion procedure after completion of the data collection process. Ex-post strategies can be used in situations where intensive early planning is not possible because of financial or policy constraints. The goal of the harmonization procedures could, for example, be to present in a comparative fashion data on a given topic that has been collected in separate, cross-sectional surveys.

Procedural steps

- Use an ex-ante strategy whenever possible. This enhances comparability since harmonization is addressed at the planning stage of each national data collection, as well as at the end of the process when creating harmonized data files.
- Use an ex-post strategy if no consideration regarding harmonization has been given by data collectors at the start of data collection(s), but researchers/data publishers later believe that a harmonized data file can be produced through a [conversion process](#) to create comparable variables or statistics.
- Adopt a detailed 'conversion' process that includes descriptions of how the producer(s) of the harmonized data dealt with the following:
 - Differences across studies with regard to what is to be measured (e.g., definitions of population, concepts, variables)

- Differences in how to measure (e.g., scale of measurement, wording/routing of questions, respondents)
 - Data [editing](#)
 - Procedures used to create and define harmonized variables
 - Construction of recoded variables
 - Sample weights and [sample design](#) variables
 - [Imputation](#)
 - Differences in how estimates are generated (weighting, non-response adjustments)
- Record all decisions about the ‘conversion’ process systematically. One option is to use two separate databases to record all work: a production database which stores the original and harmonized materials, and a user’s database which provides the analysts access to the overall process.

Lessons learned

- In a working paper, Roland Gunther describes in detail the harmonization efforts surrounding the European Community Household Panel (ECHP) [\[6\]](#). This survey began as a major example of input harmonization, with its design of uniform questionnaires as well as detailed definitions, rules, procedures, and models to ensure comparability across nations. After the first phase of the project, a few countries decided to cease collecting national samples for the ECHP, and instead to conduct their own national surveys, resulting in the need to do ex-post harmonization. Those doing the harmonization work learned that this kind of ex-post harmonization was resource-intensive and required staff experienced in both the original source and target formats of the ECHP framework. They also had to know in detail how their national questionnaires differed. Common problems included concepts heavily affected by national contexts, as well as differences in scales of measurement, variable [coding](#) schemes, and definitions of these concepts. Solutions to such problems were often found through ad hoc decisions about recoding, combining, or collapsing variables, and almost never through estimation techniques.
- These harmonization strategies are almost never applied exclusively on any single statistical or survey data collection. Depending on specific cultural and national characteristics, data producers should consider strategies that will enable them to collect their data in the most efficient manner. In some situations, they may want to combine strategies. For example, data producers may start with an input harmonization plan, but should be prepared to do some ex-post output harmonization to account for differences across cultures.

2. Create harmonized variables whenever it makes sense to do so, and always follow a systematic and adequately documented approach.

Rationale

Following a systematic approach from the beginning of the harmonization process allows data producers to document all of their decisions at the time they are made. When documentation is produced at the end of the process, it is often incomplete because producers might not remember the rationale for some of the decisions they made.

Procedural steps

- Identify and become familiar with software tools that facilitate a comparison of variables from different surveys, in order to determine if and how these could be harmonized. Such tools often work from a common database that stores all the information about each variable.
- Establish partnerships if possible with producers of harmonization tools.
- In making harmonization decisions, consult with substantive and methodological experts.
- Form an advisory committee of researchers knowledgeable about the subject matter at the beginning of the harmonization process, if possible, and consult with them regularly.
- Before fieldwork, consult with experts or an advisory committee on a systematic design process, and with methodology groups to investigate comparability issues.
- Show the group analytic results at different points in the process to allow for possible changes in rules used to create new variables.
- Consider establishing a testing group of users knowledgeable about the subject matter but not about the harmonization process, who provide feedback on the analytic usefulness of the data before they are released publicly.

Lessons learned

- Good decision-making about the harmonization process will benefit from the use of software tools, as well as input from a diverse group of survey researchers who can offer advice on various procedures and techniques to use when producing harmonized files. The ISSP Data

Wizard [\[5\]](#) is used by the International Social Survey Programme (ISSP). The Data Wizard supports procedures that were previously performed manually to harmonize data at the national level. The tool offers rule-based checks, automation of partial steps, and the visualization of certain conditions, to make the harmonization process more efficient, easier, and less susceptible to mistakes.

- The European Values Study (EVS) formed a number of work groups, both before and after fieldwork. The aim on one hand was to set standards at an early stage, and on the other to consolidate and merge data which had been cleaned by participating national survey teams. This project produced an integrated source questionnaire and a set of equivalency tables to assist secondary researchers. The project web site makes all of this information easily accessible [\[4\]](#).
- Realize that not all concepts measured in the survey process are equally susceptible to harmonization efforts. For example, cross-national harmonization of the number of births and marriages is a far easier task than comparisons of divorce rates where local laws, customs, and data collection methods may differ substantially. Other concepts, such as international population migration, may not, due to a lack of precise definition and great variety in measurement criteria, lend themselves to harmonization at all, or only at the most basic level.
- Establishing partnerships with producers of harmonization tools may be more beneficial than creating new tools, which may require costly programming efforts.

3. Focus on both the variable and file levels in the harmonization process.

Rationale

Harmonization efforts usually concentrate on comparing and integrating information involving specific variables across data files. However, it is equally important to consider the overall characteristics of the surveys that make them good candidates for harmonization, and to report the decisions involving this process to end users.

Procedural steps

- Recognize the different aspects involved in converting [source variables](#) into [target variables](#), such as “differences in the definitions of underlying concepts or in the definitions of the variables, deviations in the scales of measurements and so on” [\[9, p. 4.\]](#).

- Describe similarities and differences in how individual variables change, including discussion of [universe](#), question wording, [coding](#) schemes, and missing data definitions.
- Consider file-level attributes when creating the harmonized data file, including how the understanding of [survey weights](#), [imputation](#) procedures, variance estimation, and key substantive and demographic concepts will change in the process.

Lessons learned

- Data producers must recognize the degrees of individual item or variable persistency when creating questionnaires and collecting data. Item persistency over time is very important in generating harmonized data files. There are considerable differences, for example, between an ‘absolute’ persistent variable, such as “country of birth,” and a less persistent variable, such as “country of citizenship.” The concept might mean different things in different countries, is subject to change, and could be reported validly for multiple countries by some respondents [\[9, p. 16\]](#).
 - The Collaborative Psychiatric Epidemiology Surveys (CPES) [\[1\]](#) created a harmonized data file from three comparable surveys on mental health. Data producers created a pooled weight for the harmonized file, based on race/ancestry groupings and on the geographic domains of the [sampling frames](#) of each individual survey. Understanding the specific characteristics of each input file was an essential part of creating a harmonized output file [\[7\]](#).
- 4. Develop criteria for measuring the quality of the harmonization process by testing it with users knowledgeable about the characteristics of the underlying surveys, the meaning of source variables and also their transformation into target variables.**

Rationale

Researchers may analyze harmonized files in new and unexpected ways. It is crucial to provide them sufficient information about the concepts and definitions presented, and the assumptions underlying the decisions made in their construction.

Procedural steps

- Devise ways to judge the quality of the harmonization outputs.

- The Statistical Office of the European Communities proposed the following set of quality criteria when evaluating harmonization outputs:
 - Relevance of the statistical concepts
 - Accuracy of the estimates
 - Topicality and timeliness of the dissemination of results
 - Accessibility and clarity of the information
 - Comparability of the statistical data
 - Coherence
 - Completeness

Strictly speaking, these traits apply to [statistical data](#). However, many of them would apply equally to [survey data](#), particularly those regarding the comparability of social, economic, and demographic concepts cross-nationally or cross-culturally.

- Anticipate the need to modify and update harmonized datasets after public release, based on comments from the research community.
- Prepare presentations at social science research conferences that describe the harmonization process to potential users.

Lessons learned

- The usefulness of well-harmonized data is clearly recognized by many international organizations. For example, the United Nations Economic and Social Council indicated in a recent report that it “was working towards the harmonization of relevant environmental data-collection activities with concepts and definitions of environmental accounts. Such harmonization would result in substantial benefits in the quality of the data because it would introduce consistency checks to the environmental data and would also provide additional analytical value. The dissemination of national accounts, complemented with environment statistics information, was a very powerful analytical tool for the derivation of consistent and coherent indicators, such as resource efficiency indicators and resource use as percentage of value added. It would also allow for more in-depth analysis through scenario-modeling using input-output techniques” [\[10\]](#).

5. Provide the widest range of data and documentation products about the complete process.

Rationale

Regardless of whether researchers adopt input or output harmonization as a strategy, all aspects of the survey planning, collection, and dissemination process should be considered when producing harmonized

data files or creating accompanying documentation. Users should have access not only to the harmonized end result, but also to detailed information about all steps taken by the producers, in order for them to fully understand what decisions were made during the entire process.

Procedural steps

- Start the documentation process as soon as possible, in order to ensure that all decisions are captured even before a definite plan to produce a public-use data file exists.
- Document each target variable with information from all source variables, [transformation algorithms](#), and any deviations from the intended harmonized approach, if known.
- If possible, provide users with access to the original data files used in producing the harmonized file.
- Provide users with the code or syntax used in creating new variables for the harmonized file.
- Provide users with complete documentation, including [crosswalks](#), which describe all the relationships between variables in individual data files with their counterparts in the harmonized file. An interactive, web-based documentation tool is often the best way to present such documentation.
 - Include original questionnaires and information about the data collection process whenever possible.
- Report on as many of the following elements of the data life cycle as apply to the particular harmonization process.
 - Project planning
 - [Sampling frame](#)
 - Sample size
 - [Sample design](#) (See [Survey Instrument Design](#), [Questionnaire Design](#), [Sample Design](#))
 - Extent of the field period
 - Instrument construction and design
 - Translation and [adaptation](#)
 - Interview method
 - Respondent follow-up if panel survey
 - Data collection
 - [Editing](#)
 - [Item non-response](#)
 - Any special treatment given to demographic and country-specific variables

- Sample Weights
- Variance estimation
- Data production, including both long-planned and ad-hoc decisions implemented during variable conversion
- Documentation production
- Dissemination

This list is based on documentation provided in the Integrated Health Interview Series (IHIS). The IHIS is an effort to provide an assortment of variables from the core household and person level files from the National Center for Health Statistics' seminal data collection effort on the health conditions for the US population from 1969 to the present. It provides extensive user notes and FAQ pages to describe how their harmonization project coped with several of these components [\[8\]](#).

- Consider archiving the harmonized data with a social science data archive to ensure continued availability of all data and documentation files.

Lessons learned

- The Eurobarometer Survey Series, in operation since 1970, now includes several dozen cross-sectional surveys, all of which have been harmonized before being made available to researchers. These surveys are released initially with basic information about each study and the characteristics of all variables, and are then further processed by the social science data archives, led by the Gesis-Zentralarchiv, to include variable frequencies, more complete documentation, and online analysis services for researchers [\[2\]](#). Such partnerships between data producer and social science data archives encourage long-term preservation, enhance access, and make it possible to continually improve services to the research community.

Glossary

Adaptation	Changing existing materials (e.g., management plans, contracts, training manuals, questionnaires, etc.) by deliberately altering some content or design component to make the resulting materials more suitable for another sociocultural context or a particular population.
Coding	Translating nonnumeric data into numeric fields.
Conversion process	Data processing procedures used to create harmonized variables from original input variables.
Crosswalks	A description, usually presented in tabular format, of all the relationships between variables in individual data files and their counterparts in the harmonized file.
Editing	Altering data recorded by the interviewer or respondent to improve the quality of the data (e.g., checking consistency, correcting mistakes, following up on suspicious values, deleting duplicates, etc.). Sometimes this term also includes coding and imputation , the placement of a number into a field where data were missing.
Ex-ante	The process of creating harmonized variables at the outset of data collection, based on using the same questionnaire or agreed definitions in the harmonization process.
Ex-post	The process of creating harmonized variables from data that already exists.
Imputation	Computational methods that assign one or more estimated answers for each item that previously had missing, incomplete or implausible data.
Item non-response	The lack of information on individual data items for a sample element where other data items were successfully obtained.
Sample design	Information on the target and final sample sizes, strata definitions and the sample selection methodology.

Sampling frame	Lists or materials used to identify all elements (e.g., persons, households, establishments) of a survey population from which the sample will be selected. These lists or materials can include maps of areas in which the elements can be found, lists of members of a professional association and registries of addresses or persons.
Source variables	Original variables chosen as part of the harmonization process.
Statistical data	Data from a survey or administrative source used to produce statistics.
Survey data	Information collected by researchers, which encompasses any measurement procedures that involve asking questions of respondents.
Target variables	Variables created during the harmonization process.
Transformation algorithms	Changing all the values of a variable by using some mathematical operation.
Universe	Another term for population. A group of persons (or institutions, events, or other subjects of study) that one wishes to describe or about which one wishes to generalize. To generalize about a population, one often studies a sample that is meant to be representative of the population.
Weighting	A post-survey adjustment that may account for differential coverage, sampling, and/or nonresponse processes.

References

- [1] Collaborative Psychiatric Epidemiology Surveys (CPES). Retrieved Sept. 15, 2008 from <http://www.icpsr.umich.edu/CPES>
- [2] Eurobarometer Survey Series. Retrieved Sept. 15, 2008 from http://www.gesis.org/en/data_service/eurobarometer
- [3] European Social Survey. Retrieved Sept. 15, 2008 from <http://europeansocialsurvey.org/>
- [4] European Values Study. Retrieved Sept. 15, 2008 from <http://www.europeanvalues.nl/>
- [5] German Social Science Infrastructure Services (GESIS). *ISSP DataWizard*. Retrieved Sept. 15, 2008 from http://www.gesis.org/en/research/information_technology/ISSPWizard.htm
- [6] Gunther, R. (2003). Working Paper #19, Report on compiled information of the change from input harmonization to ex-post harmonization in national samples of the European Community Household Panel – Implications on data quality (CHINTEX). Retrieved May 23, 2008 from http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/Projekt/Downloads/WorkingPaper1__092003,property=file.pdf
- [7] Heeringa, S., & Berglund, P. National Institutes of Mental Health (NIMH), Collaborative Psychiatric Epidemiology Survey Program (CPES) Data Set: Integrated Weights and Sampling Error Codes for Design-based Analysis. Retrieved Sept. 15, 2008 from <http://www.icpsr.umich.edu/cocoon/cpes/using.xml?section=Weighting#l.++Introduction>
- [8] Integrated Health Interview Series (IHIS). Retrieved Sept. 15, 2008 from <http://www.ihis.us/ihis/>
- [9] Minkel, H. (2004). Working Paper #20, Report on data conversion methodology of the change from input harmonization to ex-post harmonization in national samples of the European Community Household Panel – Implications on data quality (CHINTEX). Retrieved May 23, 2008, from http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/Projekt/Downloads/WorkingPaper2__012004,property=file.pdf

- [10] United Nations Economic and Social Council. Environmental-economic accounting. E/CN.3/2005/15. Retrieved December 20, 2005, from <http://unstats.un.org/unsd/statcom/doc05/2005-15e.pdf>

Further Reading

- Bauer, G., Jungblut, J., Muller, W., Pollak, R., Weiss, F., & Wirth, H. (2006). *Issues in the comparative measurement of the supervisory function*. Unpublished manuscript. Retrieved May 23, 2008, from http://www.mzes.uni-mannheim.de/publications/papers/Supervisor_Function.pdf
- Bilgen, I., & Scholz, E. (2007). *Cross-national harmonisation of socio-demographic variables in the International Social Survey Programme (ISSP)*. Anaheim, CA: American Association of Public Opinion Research.
- Burkhauser, R. V., & Lillard, D. R. (2005). The contribution and potential of data harmonization for cross-national comparative research. *Journal of Comparative Policy Analysis*, 7(4), 313-330.
- Carlson, R. O. (1958). To talk with kings. *Public Opinion Quarterly*, 22(3), 224.
- Desrosieres, A. (2000). Measurement and its uses: Harmonization and quality in social statistics. *International Statistical Review/Revue Internationale de Statistique*, 68(2), 173-187.
- Ehling, M. Harmonising data in official statistics.(2003). In J. H. P. Hoffmeyer-Zlotnik & C. Wolf, *Advances in cross-national comparison: A European working book for demographic and socio-economic variables* (pp. 17-31). New York: Kluwer Academic / Plenum Publishers.
- Esteve, A., & Sobek, M. (2003). Historical methods; challenges and methods of international census harmonization. *Historical Methods*, 36(2), 66-79.
- Gandek, B., Alacoque, J., V, U., Andrew-Hobbs, M., & Davis, K. (2003). Translating the short-form headache impact test (HIT-6) in 27 countries: Methodological and conceptual issues. *Quality of Life Research*, 12, 975-979.
- Gil Alonso, F. (2006). Toward a European statistics system: Sources of harmonized data for population and households in Europe. Paper presented at the International Data Session of the EAPS European Population Conference.
- Hantrais, L., & Mangen, S. (1996). *Cross-national research methods in the social sciences*. New York: Pinter.
- Hoffmeyer-Zlotnik, J. H. P. (2004). Data harmonisation. Roundtable at a conference for the Network of Economic and Social Infrastructures, Luxembourg.

- Kennett, P. A. (2001). *Comparative social policy: Theory and research*. Buckingham: Open University Press.
- Korner, T., & Meyer, I. (2005). Harmonising socio-demographic information in household surveys of official statistics: Experiences from the Federal Statistical Office Germany. In J. H.P. Hoffmeyer-Zlotnik, & J. A. Harkness (Eds.), *Methodological aspects in cross-national research* (pp. 149-162). Mannheim: ZUMA Nachrichten Spezial, Band 11.
- Niero, M., Martin, M., Finger, T., Lucas, R., Mear, I., Wild, D., et al. (2002). A new approach to multicultural item generation in the development of two obesity-specific measures: The obesity and weight loss quality of life (OWLQOL) questionnaire and the weight related symptom measure (WRSM). *Clinical Therapeutics*, 24(4), 690-700.
- Olenski, J. (2003). *SSDIS: Global standard for harmonization of Social statistics*. Unpublished manuscript. Retrieved May 23, 2008, from http://unstats.un.org/UNSD/demographic/meetings/egm/Socialstat_0503/docs/no_10.pdf
- Pennell, B. E. (2006). Survey design and management. From a course at the Summer Institute, Survey Research Center, University of Michigan.
- U.K. Office for National Statistics. *National statistics harmonization*. Retrieved May 23, 2008, from <http://www.statistics.gov.uk/about/data/harmonisation/default.asp>