European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Improving the Comparability Dimension of European Statistics by Minimizing Unnecessary Variation

E. Baldacci[1], L. Japec[2], I. Stoop[3]

[1] *Eurostat, Luxembourg, Luxembourg; Emanuele.BALDACCI@ec.europa.eu*
[2] *Statistics Sweden, Stockholm, Sweden; lilli.japec@scb.se*
[3] *The Netherlands Institute for Social Research (SCP)/ESAC, Den Haag, The Netherlands; i.stoop@scp.nl*

Abstract

One quality criterion for European statistics is comparability across countries. Often there are trade-offs between this and other quality dimensions, other design aspects, and national good practices. There are several factors that could affect differences across countries due to data collection and processing practices. A lack of comparability seriously diminishes the value and usefulness of statistics. Two main methods are used in the production of European statistics to increase comparability: input and output harmonisation. Although both approaches are used in economic and social statistics, sources of unnecessary variation between countries should be (further) minimized.

For survey data, this can be pursued by focusing on ensuring "conceptual equivalence" in data collection, implementing scientific translation procedures, adapting questions that convey the same meaning across countries, enforcing definitions, using probability sampling, and aiming for high response rates and small measurement errors. To strengthen comparability, surveys such as the European Social Survey, Survey of Health, Ageing and Retirement in Europe (SHARE), and the Programme for the International Assessment of Adult Competencies (PIAAC) could be benchmarked.

Necessary variation will still remain in survey data, owing to differences in sampling frames, response rates, experience of interviewers, internet access, and languages. This should be kept, however, to an acceptable level and requires detailed documentation and dissemination of quality information including non-coverage, response rates, and survey modes impacts.

An additional potential source of unnecessary variation in European statistics stems from the increase in the use of non-survey data sources, such as administrative registers data and new sources such as big data, and the associated new statistical production methods. Non-survey data are already widely used for economic and population statistics and are increasingly collected to measure social phenomena, integrating social surveys. The availability, accessibility, quality and use of non-survey data sources vary greatly across countries. Moreover, new statistical production processes entail a more frequent use of model-based and algorithm-based estimation techniques, which add an additional layer of complexity to ensuring data comparability across countries. Statistical outputs are also evolving reflecting users' needs, with a focus on indicators, visual analytics and dashboards. Harmonisation is critical to ensure that European statistics and related information services produced according to new sources and methods are comparable.

The above results point to the need of strengthened research in methods and practices to ensure cross-country comparability in European statistics in a fast changing production environment. This will be a topical area of work for the ESS in the near future.

Keywords: comparability, variation, benchmark.

1. Introduction

Monitoring progress in society, both on a national and international level, requires different types of statistics (Trewin et al., 2010). Research questions that statistical information should be able to shed some light on include the following (Trewin et al., 2005): has there been any change over time, is there variation across different subgroups, what are the causes of changes, what are the links between indicators and how does a change in one country compare with other countries? Data quality, most notably accuracy, is at the core of these research questions. Producers of data need to make sure that errors are minimised so that users get accurate and reliable statistics. As important decisions are made based on statistics at the international level, statistical organisations need to work together in order to enhance comparability across countries. Bearing in mind that each step in the statistical production process will affect the outcome, comparability between countries becomes a real challenge. This reflects the fact that each country has its own set of data collection and processing conditions. In addition, there are often trade-offs between comparability and other quality dimensions.

Section 2 of this paper provides an illustration of the sources of variability in European cross-country statistics. There are several methods to increase comparability, ranging from input to output harmonisation (Körner and Meyer 2005). In Sections 3 we give examples of these methods and discuss their pros and cons. The ongoing change in paradigm for statistical production, relying on the use of multiple data sources and their integration in statistical processes, affects comparability beyond surveys as described in Section 4. In particular, multi-source statistical production and the use of model-based estimation add an extra layer of complexity in the efforts to minimise unnecessary cross-country variations. We conclude with some thoughts on priority areas for future methodological work.

2. Sources of variability across countries

The European Statistics Code of Practice (Eurostat 2013), consists of the following principles regarding statistical outputs:

- *Relevance:* outputs, i.e. European Statistics meet the needs of users.

- *Accuracy and Reliability:* outputs accurately and reliably portray reality.
- *Timeliness and Punctuality:* outputs are released in a timely and punctual manner.
- *Coherence and Comparability:* outputs are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different sources.
- *Accessibility and Clarity:* outputs are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.

The framework highlights different dimensions that are important in statistical production. Good data quality (accuracy and reliability dimension) can be achieved by applying sound methods and quality assurance in the statistical production process. The other dimensions of quality are also important (e.g., if a user needs data by a certain date and does not get it by then it will be of no consolation to that user that the data is very accurate).

Often there are trade-offs between comparability and other quality dimensions, as well as national good practices. Since each step in the statistical production process can affect the final estimates, the design and execution of these steps at national level could affect the degree of comparability across countries in Europe.

Some steps are more crucial than others in order to gain comparability.

Social and cultural environments – concepts that have to be measured may be of different relevance or may manifest themselves differently in different social and cultural environments. What is considered to be sensitive varies between these environments. Harkness et al. (2003) give an example of the effect of cultural environment for the concept of religiosity. In some religious denominations, attendance at a place of worship is an essential element of religiosity, whereas for other religions rites at home are important. In this example, careful considerations must be taken when designing questions for comparisons between countries if interested in comparing religiosity between countries. A question that asks about the frequency of church attendance, for example, would not alone provide a basis for comparing religiosity.

Data collection modes and mixes of modes – the mode that is used to collect data can affect statistical outputs. There are many examples in the literature of different mode effects, where statistical results can differ substantially depending on whether data collection is based on face to face interviews, telephone interviews, mail or web questionnaires (De Leeuw 2008).

Questions and questionnaires –the way questions are worded and the order in which they appear in the questionnaire are examples of design decisions that can affect statistics. Ensuring conceptual equivalence is important when designing questions for comparisons between countries. This is problematic when some concepts do not exist in all countries or the meaning varies across countries.

Translation – there are different methods for translating questions in international surveys and they can yield different results. Harkness et al. (2003) provide a good overview of these methods. One translation method is word for word translation. This method will definitely not solve the problem of ensuring conceptual equivalence, as a word for word translation will very often not be meaningful in the target language. Sometimes adaptation of questions, i.e., tailoring questions to better fit the target population, is necessary in order to ensure conceptual equivalence. Recently, more elaborate translation methods have been developed to ensure good translation quality. One such method is team translation (see Section 3).

Methodological and financial resources – differences in available skills and financial resources in different countries and organisations limit what is feasible to achieve in terms of survey design. Unnecessary variation between countries due to these methodological decisions should be minimised.

Even when a survey is designed to minimise unnecessary variation between countries, however, necessary variation may remain, e.g. in sampling frames, response rates, experience of interviewers, measurement error structures, prevalence of internet access, and the different languages in which the survey is fielded.

When administrative data is used additional issues arise. Administrative data are collected based on national regulations and the needs of the collecting agency which do not necessarily lead to comparable definitions across countries. Initial quality of administrative data sources may also vary. In the data processing phase, data imputation and integration rely on a variety of assumptions and methods which could lead to additional comparability issues. Finally, data estimation can rely on methods which are not necessarily harmonised across countries. With administrative data sources increasingly used for the production of European statistics and integration of different data sources frequently used to supplement survey data, including for statistical indicators and accounting frameworks, comparability of cross-country statistics is challenging.

3. Harmonisation work

Improving harmonisation of concepts and practices used at the different steps of the statistical production process can foster better comparability of European statistics across countries. Reference literature would typically distinguish two main harmonisation categories: input and output (Körner and Meyer, 2005; Hoffmeyer-Zlotnik & Warner 2013).

Input harmonisation implies the use - at each step of the statistical process - of methods, sources, tools and procedures in order to produce comparable data at European level. This means that not only concepts, definitions and classifications are harmonised, but also that data is collected via the same tool in all countries from the same sources (e.g., a survey with the same questions) and that the survey design, characteristics and methods are precisely defined and possibly even regulated. This approach requires a high degree of co-ordination and upfront harmonisation and could result in inflexibility. Some steps of such a harmonised process or even the whole process could be centrally implemented, which in principle could reduce its cost.

A case where input harmonisation is used in all stages of the production process is the European Social Survey. We describe the process in more details here as it provides an example of the different steps involved. Figure 1 gives an overview of its life cycle.

Figure 1. Survey Life Cycle (Survey Research Centre, 2010)

It was first conceived in the 1990s, first went into the field in 2002 and seven rounds of data collection have been completed in 2016. Fieldwork has now been conducted in more than 30 countries, and 90 thousand individuals have registered as data users. The Core Scientific Team of the European Social Survey is directly involved in all phases of the survey, from specifying the national design (within the general design framework), testing and translating questions and training interviewers, monitoring fieldwork and coding, to delivering data. The Project Specifications of the European Social Survey prescribe in detail how the survey should be conducted.

Questionnaire and translation- The core questionnaire was developed based on proposals from experts in social research and survey methodology, and has remained largely identical over the years. In each round, two rotating modules are fielded. The first test (using the Survey Quality Predictor) looks at the form and structure of questions – its length, the number of complex words it contains, and so on – and then, based on what is known about the quality of items with similar features that have been fielded before, makes a prediction about how well the question will perform (Saris & Gallhofer 2007). A second approach is cognitive interviewing, using human subjects to help predict likely problems with questions (Willis 2005). Finally large-scale pilots are used where whole batches of items are tested on a sample large enough to include meaningful numbers of subgroups. Once the questionnaire has been finalized in English, annotations are added. These annotations do not form part of the final

questionnaire but serve as a guide to translators. In the European Social Survey, the Translation, Review, Adjudication, Pretesting and Documentation (TRAPD) approach has been implemented, developed and promoted by Harkness and her colleagues (Harkness 2003; 2007). TRAPD consists of five interrelated procedures (Harkness, 2003), performed by three key agents: translator, reviewer and adjudicator. Reviewers not only need to have good translation skills but also need to be familiar with the principles of questionnaire design and the particular study design and the topics covered. They attend the review sessions and contribute to and guide the revisions. The adjudicator is the person responsible for the final decisions about which translation options to adopt. Pretesting may again result in modifications before the adjudicator signs off on the version for final fielding. Central in TRAPD is the team approach and the detailed documentation required.

Data collection - The European Social Survey is representative of all persons aged 15 years and over (no upper age limit) resident within private households in each country. The sample is selected using strict random probability methods. Different sampling frames, designs and procedures can be used across countries (see Häder & Lynn 2007). The questionnaire is administered face-to-face. This mode allows the fielding of a rather long questionnaire results in a reasonably high response rate and is feasible in every country. Because of the high costs of face-to-face data collection a range of mixed mode pilots have been conducted. The result of these pilots was that moving to mixed modes would be detrimental to data quality, whereas the savings were very modest because of the relatively small sample size. Detailed contact forms are used to monitor fieldwork and control quality. The data from the contact forms are publicly available for research on nonresponse and data quality. Data from the contact forms can also be used to improve fieldwork. After fieldwork, data are coded and sent for further harmonisation to the European Social Survey Data Archive at NSD (Kolsrud et al., 2007). The archiving process is aimed at releasing the first integrated dataset around nine months after the first country has finished the fieldwork. After this first release data are accessible.

Such a high degree of input harmonisation is rare to find in the European Statistical System (ESS) reflecting cost considerations and national specificities. The early versions of the

European Community Household Panel (ECHP) and the Land Use and Cover Area frame Survey (LUCAS) survey are examples of input harmonisation with the panel being implemented through a common questionnaire, harmonised definitions and sampling requirements and LUCAS being conducted centrally by Eurostat. Another example of input harmonisation are the European Statistical System ICT surveys, which include several elements of comparability at design stage, given that the mode of data collection and related fieldworks to be conducted by NSIs are based on a fully harmonised methodological manual.

In contrast to the previous approach, *output harmonisation* focuses on the final statistical product. Target variables are defined, however due to conceptual arguments the actual questions may vary between countries. Granda (2010) gives an overview of the issues concerning output harmonisation and especially ex-post output harmonisation. In the ESS, pure output harmonisation is as rare as pure input harmonisation (Clemenceau & Museux 2008). For all statistical domains, at least the concepts and definitions of variables as well as the classifications to be used are subject to ESS agreements/legislation. Most European statistics, however, use a blend of input and output harmonisation, where some process steps are standardised and some flexibility is allowed for the others. This mainly reflects the need to take into account the different conditions and varying methodological approaches at national level and reflect the subsidiarity approach adopted by the ESS. The European Union Survey on Income and Living Conditions (EU-SILC) for instance, defines a harmonised list of target variables, definitions and concepts, classification, guidelines and recommendations, while the data collection process, including the selection of the appropriate data sources, is left to the NSIs. Structural Business Statistics are another example in the field of economic statistics where the choice of data sources is left to the member states based on commonly agreed methodologies, definitions and classifications.

Actual harmonisation practices in European statistics could be seen as a continuum of interventions along the statistical production process phases ranging from pure input to pure output harmonisation. When looking beyond surveys as data collection tools, it becomes difficult to classify European statistics harmonisation practices according the input/output

dichotomy; in Figure 2 a framework is proposed mapping the elements of harmonisation against the Generic Statistical Business Process Model (GSBPM).

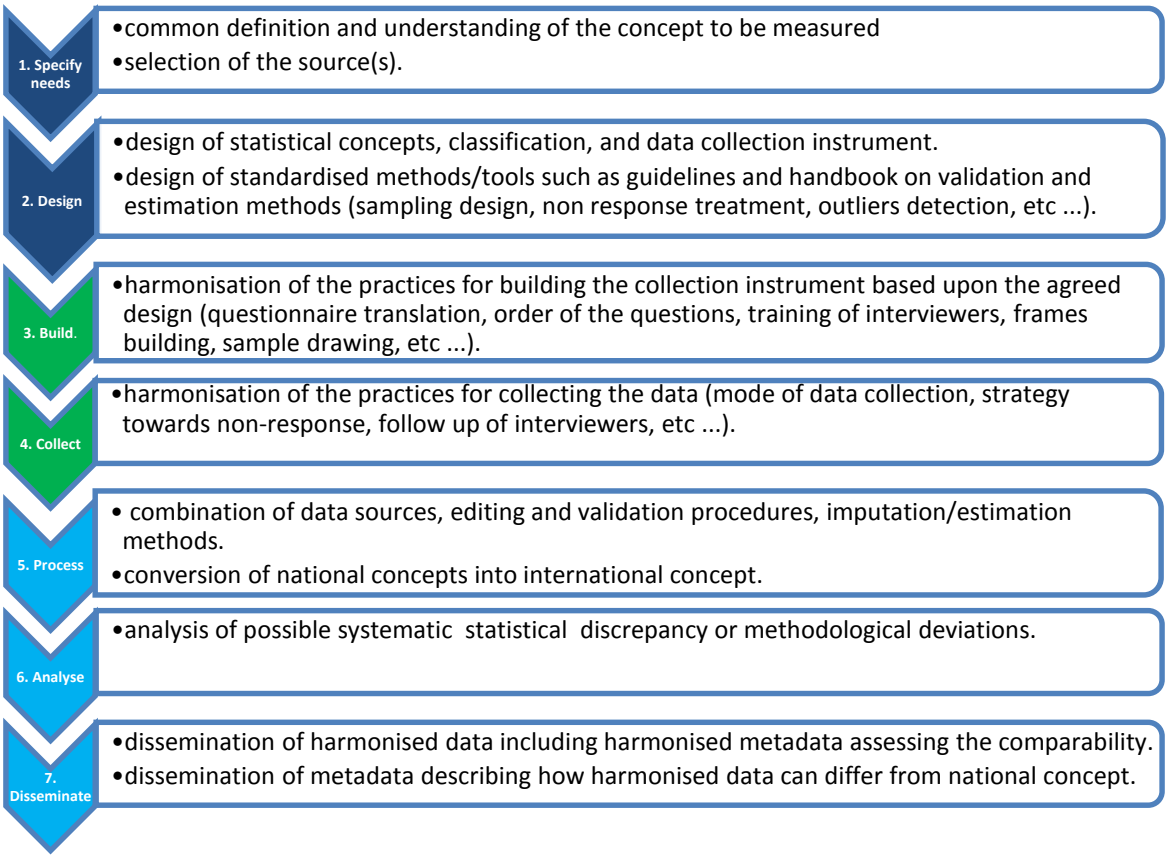

Figure 2 Harmonisation through the different steps of the statistical process

Lynn (2003) distinguishes several strategies for cross-national comparability. It is crucial to establish whether cross-country comparability is the main aim, and the trade-off with achieving the highest (national) quality. For example, in the "consistent quality" approach for survey data, response rates could be set in advance that are feasible in every country. This could result in response rates that are much lower than usual in several countries. To ensure optimal comparability, PAPI data collection could be prescribed and the use of register data be forbidden. This would again result in comparable data, but most likely in poorer national quality than could be achieved had CAPI data collection and the use of registers be allowed. On the other hand, the "maximum quality" approach aims at achieving national data with the best possible quality. This may harm comparability, e.g., when different survey modes are

used. Other strategies usually constrain some quality aspects (e.g., sampling), but leave other design aspects to individual countries (e.g., the survey mode). Still other quality aspects would be formulated as targets rather than constraints (e.g., aim at a response rate of at least 60%). In practice most surveys strike a between these different dimensions.

4. Comparability in a changing statistical production paradigm

Availability of new data sources reflecting more open access to administrative data and the data revolution, budget constraints and the need to lower the response burden have led producers to move to a new paradigm for statistical production (Citro, 2014). This is based on integrated use of multiple data sources (including new sources), adoption of innovative data analysis tools and methods, and a focus on statistical information services to respond to users (ESS Vision 2020). These changes are leading to the following transformations:

- Increasing use of multiple, integrated data sources.
- New modes of data collection (e.g., administrative registries, smart phones and smart meters).
- Model-based and algorithm-based estimates in the production of statistics.
- Move from pure statistical products to a range of on-demand data analysis and information services.

In the current production of demographic, social and economic statistics in Europe, administrative data play an important part, complementing and in part replacing survey data. Data imputation, estimation of information for sub populations, and missing variables are examples of statistical production processes that already rely on multiple sources and their integration through models to estimate statistical information of interest.

These changes imply that traditional harmonisation methods and practices (both on inputs and outputs) may fail to fully account for significant sources of unnecessary variation in European statistics and need to be complemented by additional efforts to achieve comparability. For example, the combination of data sources may add to variability of statistics across countries, as in the case with uneven quality of original sources when administrative data are used.

Differences in method-based estimations could also contribute to increase unnecessary variations across countries, as harmonised approaches may not be followed for the use of such tools. This calls for an augmented quality conceptual framework (building on existing process models such as GSBPM) to assess the impacts on comparability of the multi-source statistical production environment and more research to ensure adequate harmonisation.

However, not all changes in statistical production systems may harm the comparability of European statistics. For example, more integrated data based on multiple sources can help detect and assess cross-country discrepancies. For example, administrative data can complement survey data to better measure the tails of the households' income distribution, thereby lowering the risks that comparisons across countries are biased. Model-based estimations based on larger data sets can lead to more powerful tests of hypothesis and formal model selection processes, leading to better comparability when these tools are used in cross-country statistics. Based on such analysis, methodological recommendations and statistical reconciliation methods (Stone model) could be envisaged for harmonising statistics. Consideration should also be given to the potential of statistical methods that currently are not widely used in official statistics. Bootstrap procedures for example, could lead to better assessment of the measurement errors, in particular when analysing data derived from multiple sources. In the same way, panel data models, where country effects can be identified and estimated, could be used for formal testing of comparability.

5. Way forward: an agenda for future work

Lack of comparability can seriously harm the value and usefulness of European statistics. To strengthen comparability both output and input harmonisation practices have to be improved. For survey data, differences regarding countries' cultural environments and collection mode are critical. For non-survey data, better assessment of quality of data sources and methods for data integration and statistical estimation are important. Methodological skills should also be enhanced to increase knowledge of error sources and mitigation measures. Investment in methodological studies to foster harmonisation of cross-country statistics is needed.

Comparability is hard to achieve even when we design for it. More information should be provided to users of European Statistics on the comparability dimension. For example, the Comparative Survey Design and Implementation (CSDI: www.csdiworkshop.org) network is active in studying methods for enhancing comparability between countries, regions and cultures. The network organises yearly workshops and scientific conferences. The statistical agencies within the European Statistical System would benefit from collaborating with the CSDI network. To strengthen comparability, surveys such as the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE), and the Programme for the International Assessment of Adult Competencies (PIAAC) could also be benchmarked.

At the same time European statistics require looking beyond the design stage. To ensure comparability of data across countries within a new paradigm of statistical production, characterized by mixed modes and multiple sources, more efforts are needed in applied research on quality dimensions, as the ones undertaken in the context of the ESS Vision 2020 (e.g., on administrative and big data sources for statistical production). An increasing emphasis has to be put on the consequences for data comparability of integrating data stemming from multiple sources and using modelling techniques for data analysis and statistical information services. This goes hand in hand with the need to invest further in assessing and communicating to stakeholders the quality of cross-country statistics in Europe.

6. References

Clemenceau, A. & Museux, J.-M., 2008. *Harmonisation of household surveys from an Official Statistics Perspective: experience gained from HBS, ECHP and EU-SILC.* Berlin, Paper presented at the 3MC Conference.

Citro, C., 2014, From multiple modes for surveys to multiple data sources for estimates, Journal of Official Statistics, 30(3, September), 381-442

De Leeuw, E. D. (2008) Choosing the Method of Data Collection. In: E. D. de Leeuw International handbook of Survey Methodology (113-135). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Granda P, Wolf C, Hadorn R. Harmonizing survey data. In: Harkness JA, Braun M, Edwards B, et al., editors. Survey methods in multinational, multicultural and multiregional contexts. Hoboken, NJ: John Wiley & Sons; 2010. pp. 315–32.

Häder, S. & P. Lynn (2007) How representative can a multi-nation survey be? In: R. Jowell, C. Roberts, R. Fitzgerald & G. Eva (eds): Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey (33-52). London, Sage.

Harkness, J. A. (2003). Questionnaire Translation. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph Mohler (Eds.), Cross-cultural survey methods (35–56). New York: Wiley.

Harkness, J.A. (2007) Improving the Comparability of Translations. In: Roger Jowell, Caroline Roberts, Rory Fitzgerald & Gillian Eva (eds) Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey (79-93). London, Sage.

Hoffmeyer-Zlotnik, J. & Warner, U., 2013. *Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research.* s.l.:Springer Netherlands.

Körner, T., and I. Meyer (2005) Harmonising Socio-Demographic Information in Household Surveys of Official Statistics: Experiences from the Federal Statistical Office Germany. In: Jürgen H.P. Hoffmeyer-Zlotnik and Janet A. Harkness (eds) Methodological Aspects in Cross-National Research. ZUMA-Nachrichten Spezial Band 11, pp. 149-162.

Kolsrud, K., K. Kallgraf Skjåk & B. Henrichsen (2007) Free and immediate access to data. In: R. Jowell, C. Roberts: Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey (139-156). London, Sage.

Lynn, P. (2003) Developing Quality Standards for Cross-national Survey Research: Five Approaches. International Journal of Social Research Methodology, Vol. 6, No. 4, pp. 323-336.

Stoop, I., J. Billiet, A. Koch and R. Fitzgerald (2010) Improving Survey Response. Lessons Learned from the European Social Survey. Chichester, John Wiley & Sons. Ltd.

Stoop, I., and E. Harrison (2012) Repeated Cross-sectional Surveys Using F2F. In: Lior Gideon (ed.) Handbook of Survey Methodology for the Social Sciences (249-276). Heidelberg, Springer.