Statistiska centralbyrån
Statistics Sweden

Lilli Japec, PhD
Director R&D Department

August 31, 2015

Quality Assessment of Indicators and Trade-offs Between Different Dimensions of Quality

1. Introduction

NSIs are important players in providing statistical information for decision-making both on the national and the international level. To monitor progress in society (UN 2015) different types of indicators are produced (Trewin et al 2010) e.g., suite-of-indicators, composite indicators and accounting frameworks. Research questions that these indicators should be able to address include (Trewin et al 2005): has there been any change over time, is there a 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, i.e., accuracy, is at the core of these research questions. NSIs need to make sure that errors are minimized so that users get accurate and reliable data. Furthermore we also need to work together in order to enhance comparability across countries. In this paper we will describe the system used at Statistics Sweden in order to assure quality and prioritize activities. We will also discuss trade-offs within and between quality dimensions and the special challenge that comparability poses to the European Statistical System.

2. Quality dimensions in statistics production

Several frameworks for assessing survey quality have been developed over the years e.g., Australian Bureau of Statistics, Eurostat, IMF, OECD, Statistics Canada and Statistics Sweden have developed their own frameworks with different number of dimensions. There are considerable overlaps between all those frameworks. The Eurostat framework (Eurostat 2013) consists of the following dimensions:

- *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 frameworks highlight different dimensions that are important in statistics production. Most of these frameworks, however, have been developed without any active involvement of users. There may very well be other dimensions of quality that are important to users. We can achieve good data quality (accuracy and reliability dimension) by applying best methods, quality assurance and quality control in the statistics production process (section 3). The other dimensions can be seen as restrictions 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.

3. Quality assurance and quality control

Ideally for each estimate e.g., the unemployment rate we would like it to reflect the true state in a country. However, the statistics production process generates errors that jeopardize the accuracy of the estimates e.g., (i) not all people in a country are part of the labour force survey, (ii) interviewers affect the way respondents answer a question, (iii) respondents may not want to participate in the survey or cannot be contacted, and (iv) respondents might not answer truthfully. The first type of error (i) we can easily estimate by calculating the standard deviation and present a confidence interval. The latter types of error, so called nonsampling errors can cause bias and additional variance components and they are much more difficult to estimate. Furthermore the latter types of error are not included in the traditional standard deviations and confidence intervals that we present to users. The result of this is that we underestimate the length of the confidence interval.

Also the point estimate may be biased. Therefore the main idea in statistics production is to minimize nonsampling errors so that the standard deviation and confidence intervals will accurately reflect the uncertainty of the estimate. The main prescription for minimizing nonsampling errors is to standardize steps in the production process that are known to cause large errors e.g., the interviewing and the coding process. Standardizing is usually not enough. We also need to have mechanisms in place to control the quality and make sure that standards are being followed e.g., monitoring of interviewers and coding control and checking that the data collection is conducted as planned by means of paradata analysis. We call these control systems quality assurance and quality control. For instance training of interviewers is a quality assurance measure and interviewer monitoring is a quality control measure.

What agencies typically do is that they handle sampling, nonresponse and coverage errors in a satisfactory way but many other errors are more or less ignored, such as interviewer effects and coding errors resulting in understated margins of error. In addition, users are often not informed about this state of affairs.

What agencies should strive for is to try to minimize the effects of each error source so that the only variation left is due to sampling. Implementing a continuous quality improvement system will directly or indirectly help mitigate the errors that affect the estimates. In the next section we will discuss such a system adopted by Statistics Sweden.

3.1 Example from Statistics Sweden

Statistics Sweden's quality assurance and quality control system for statistics production consists of adherence to ISO 20252 for market, opinion and social research as a standard for process quality and using ASPIRE (A System for Product Improvement, Review and Evaluation).

ISO 20252 contains about 450 requirements mainly on the statistics production process (International Standardization Organization 2012). This process standard has a client focus and transparency and traceability in methods are important requirements in the standard. The use of checklists and templates is also crucial in order to reduce unnecessary process variation within the organization. Validation of results is an important requirement for subprocesses that have a large impact on data quality or costs. Examples of validation requirements include monitoring of interviewers and coding control. Statistics Sweden is certified according to ISO 20252 since 2014 which means that all our statistical products meet the requirements of the standard.

As a result of errors experienced in the critical products Consumer Price Index and National Accounts in 2011, the Ministry of Finance required improvements in Statistics Sweden's products. We needed quantitative and objective measures of product quality. We decided to focus on the accuracy component and prioritize the ten most important statistical products. The products evaluated are surveys, registers and compilations. Paul Biemer, Distinguished Fellow at RTI International and Dennis Trewin, former Australian Statistician, helped Statistics Sweden develop ASPIRE, a management tool with two main goals (Biemer et al 2014). One is to evaluate our products and another is to inspire staff to make important quality improvements in their products. The error sources are slightly different for different types of products and the framework has been adjusted accordingly.

The quality criteria that we use for all products are:

- Knowledge (of the producers of statistics) of the risks affecting data quality for each error source,
- Communication of these risks to the users and suppliers of data and information,
- Available expertise to deal with these risks (in areas such as methodology, measurement and IT),
- Compliance with appropriate standards and best practices relevant to a given error source and,
- Plans and achievements for mitigating the risks.

Using external evaluators is an important feature of ASPIRE. The main reasons are that we want to achieve objectivity, factual as well as perceived. We also believe that an external influence, by highly competent and respected evaluators, will inspire improvement work among staff. This would be much harder to achieve with a self-assessment approach with internal evaluators which has a tendency to be more forgiving.

We have developed guidelines and checklists for the review process to make it as transparent as possible and to minimize the variation between judgments made by evaluators. Each production team starts by making a self-assessment. The assessment and relevant documentation are sent to the evaluators. The next step is a meeting between evaluators and the production team focusing on discussions of changes from the previous year, review of the quality declarations, progress made on previous recommendations followed by preliminary ratings by the evaluators. In the meeting recommendations on improvements are also discussed. There is a control step where the production team receives a draft of the evaluation report. This is an opportunity for the production team to provide feedback to the evaluators and to discuss any disagreements with the evaluators. The rating in terms of scores are then finalized. This process is repeated annually. (When ASPIRE is implemented for the first time there are of course no previous recommendations to discuss and evaluate.)

3.2 Results from ASPIRE

In their final report the evaluators provide examples of types of studies or improvements each product should make. The results for each product are presented in a summary table. In the table below the results for the Labour Force Survey are shown. In the rows we find the different error sources and in the columns the quality criteria are displayed. The scale that is being used ranges from poor to excellent. In our example we can, for instance, see that the available expertise on measurement error in the LFS is very good. We have two red spots and that is for frame and nonresponse error and compliance with standards. This is because the frame covers the registered population and the ILO recommendation is that the resident population should be covered. The other red spot is due to the fact that efforts to mitigate nonresponse appear to be ineffective. Compliance to standards and best practices with data collection has not been kept up to meet the present challenges with increased nonresponse.

Another feature of ASPIRE is that we assign a risk to each error source. The risk will vary between products. For instance, in the LFS the nonresponse and measurement errors are considered to be high risk areas. The risk score is used to calculate the total score for each product. High risk areas have a higher impact on the total score. This is to help the product to focus on important error sources and to set priorities. In our example we can see that the total score is 66.0. Compared to the previous year this is an improvement. The shaded green and pink cells indicate the change from the previous year. In the LFS example work has been done on nonresponse errors and studies have been carried out to estimate them. Additional highly qualified expertise has been assigned to examine the nonresponse bias in 2015. Furthermore an experiment is carried out to see if an external company can achieve higher response rates and plans are in place to evaluate the experiment on a continuing basis. A study of measurement errors has been carried out and documented in a report. This is considered to be an improvement compared to the previous year.

Table 1. Results for the Labour Force Survey, Round 4 in 2014

	Error Source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	Specification error	70	70	◡	◡	◡	◡	◡	L
	Frame error	58	58	◡	◡	◡	◠	○	L
	Non-response error	52	58	○	○	◡	◠	◡	H
	Measurement error	68	70	◡	◡	◡	○	◡	H
	Data processing error	62	62	○	○	◡	◡	◡	M
	Sampling error	80	80	◡	◎	◡	◎	◡	M
	Model/estimation error	64	64	○	○	◡	◡	◡	M
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Total score	64,3	66,0						

Scores					Levels of Risk			Changes from round 2	
●	◠	○	◡	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

We have seen concrete improvements as a result of ASPIRE. Here are some examples. In the first round of ASPIRE we found that all of the evaluated products were weak on measurement errors. A project that looked specifically at methods to study measurement errors was initiated. Methodologists were trained in this area and measurement error studies have been carried out. The quality declarations themselves have improved as well. A special effort was made in this area with hands-on workshops with the specific goal to improve the information and readability of the quality declarations. We have also seen an increased activity in the area of planning studies and improvement projects. We have also redesigned the Survey of Living Conditions with substantial improvements as a result.

The strengths associated with ASPIRE are that it is a comprehensive tool covering all main error sources and that it contains criteria that identify risks for data quality. The evaluator checklists are effective for assigning reliable ratings. The fact that we distinguish between error sources in terms of their impact on the total error is an important feature since we have limited resources and we would like to make sure that we use our resources in the best possible way. It is very inspiring for our staff to have the possibility to discuss their products and possible improvements with very competent evaluators. It is a systematic approach that drives improvements and it is relatively simple and easily understood by managers.

One weakness is that ASPIRE cannot directly measure the true accuracy of a statistical product. Also it relies on skills and experiences of external evaluators, and on that objective information is provided by the product staff.

4. Trade-offs between and within quality dimensions

There is no one-number quality indicator such as a quality index that encompasses all the dimensions described in section 2. Such an index would be very hard to develop and then for users to interpret. Also the different dimensions are not equally important to all users.

In statistics production there are trade-offs to be made both between and within quality dimensions. Perhaps the most obvious trade-off situation between quality dimensions is the one between timeliness and accuracy e.g., improving timeliness may mean that we have to settle with less accurate data due to high nonresponse. With new data sources such as big data the trade-off situation becomes even more complicated for instance big data will most likely improve timeliness but currently we do not know much about accuracy of the data.

There are challenges in terms of trade-offs that have to be made within quality dimensions. ASPIRE, described in the previous section, is a tool for making trade-offs within the accuracy component. For the European Statistical System perhaps the biggest trade-off challenge is within the comparability component. Striking a good balance between national and EU needs is vital for international comparisons. Applying good methods, quality assurance and quality control procedures on a national level will not necessarily produce statistics that are comparable across different countries.

There are two ways to achieve comparability, output and input harmonization. Within the European Statistical System output harmonization with some elements of input harmonization is the most commonly used method to enhance comparability. Output harmonization specifies the deliverables while essential survey conditions are allowed to vary. This is usually not good for comparability. With input harmonization a lot of effort goes into keeping as many essential survey conditions as possible fixed, thereby facilitating country comparisons. An example of a survey where only input harmonization is used is the European Social Survey which is an ERIC (European Research Infrastructure Consortium). Experience shows that international comparability is hard to achieve even if input harmonization is used (Jowell 1998, Harkness et al 2010).

Comparability between countries in the EU is a joint responsibility of the NSIs and Eurostat. It is therefore important to identify and agree on steps in the survey production process where input harmonization is crucial in order to achieve comparability. These decisions should be made by considering both quality and cost aspects and national and EU needs. For instance the use of proxy interviews in the EU-SILC varies a lot between countries ranging from 49% proxy interviews to no or almost no proxy interviews. Based on the literature on proxy interviewing, we may suspect that this will hamper comparability across countries due to bias. There are other steps where we could allow flexibility e.g., the sampling methods used. The important aspect in sampling is to get unbiased estimates and this can be achieved with many different sampling methods.

There are research groups e.g., CSDI, the workshop on comparative survey design and implementation that specializes in research to develop methods to achieve equivalence in multinational, multiregional and multicultural contexts (Harkness et al 2010). The ESS should get more involved in this work.

References

Biemer, P., Trewin, D., Bergdahl, H. and Japec, L. (2014). ”A System of Managing the Quality in Official Statistics”, *Journal of Official Statistics*, 3, pp. 381-573.

Eurostat (2013). The ESS Handbook on Quality Reports.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, P. Ph., Pennell, B-E., and Smith, T. (Eds) (2010). Survey Methods in Multinational, Multiregional, and Multicultural Contexts.

Jowell, R. (1998). How Comparative is Comparative Research. American Behavioural Scientists, 42, 2,168-177.

Trewin, D. and Hall, J. (2005). Measures of Australia’s Progress—A Case Study of a National Report Based on Key Economic, Social and Environment Indicators. OECD.

Trewin, D. and Hall, J. (2010). Developing Societal Progress Indicators: A Practical Guide. OECD.

UN (2015). Transforming our World: the 2030 Agenda for Sustainable Development.