U/KVAL
Heather Bergdahl

Tema 4 Kommunikasjon av resultat og kvalitet til brukeren (Statistiske produkter)

A Tool for Managing Product Quality

Heather Bergdahl[1], Lilli Japec[2], Åke Pettersson[3]

[1] Quality coordinator, R&D, Statistics Sweden, heather.bergdahl@scb.se
[2] Department Head, R&D, Statistics Sweden, lilli.japec@scb.se
[3] Quality Manager, R&D, Statistics Sweden, ake.pettersson@scb.se

Abstract

In 2011 Statistics Sweden was presented with a challenge by its main stakeholder – the Ministry of Finance – to develop indicators that could show developments in product quality. There are numerous quality frameworks that address different dimensions, such as organizational, process and product quality. The challenge, however, is to measure and monitor changes in product quality in a comprehensive and systematic way and to clearly and concisely present progress on total survey quality to stakeholders. A tool, known as ASPIRE, was developed and has been tested in 2011 and 2012 for the Accuracy Dimension of quality on ten key products including Gross Domestic Product (GDP) and the Labor Force Survey (LFS). In this paper we describe this tool and how it can be used to set clear measureable goals for product quality. Also, some results from the product evaluation and the associated recommendations for quality improvement are presented.

1. Background

The government of Sweden has during recent years tried to monitor quality improvements in official statistics for which Statistics Sweden is responsible. In this context the government has requested a report in the form of specific indicators that signify any quality improvements that are occurring in pre-specified programs.

Until 2008, Statistics Sweden monitored the quality of statistical programs via self-assessments, the results of which were reported publicly. However, due to the inherent bias in self-assessments, the process did not yield the informative and accurate measures needed for effective, continual quality improvement. The self-assessment process was thus discontinued and progress on product quality has not been quantified since 2009.

Therefore in 2011, Statistics Sweden's R&D Department took steps to develop a model that will capture quality changes in the agency's statistical programs. Review and evaluation of accuracy of eight important products was done in November/December 2011 using this approach referred to as ASPIRE (A System for Product Improvement, Review, and Evaluation). A baseline for these products (Round 1) is reported in Biemer and Trewin (2012). The most recent work in November/December 2012 (Round 2) with ten important products is reported in Biemer and Trewin (2013).

2. The ASPIRE Model

The ASPIRE approach reported in Biemer and Trewin (2013) is general in that it can be applied to a range of products produced by a data collection program, a frame or register, or a compilation of a number of statistical inputs such as the system of National Accounts. ASPIRE is also comprehensive

in that it considers the errors arising from all major error sources from the design of the data collection to final publication or data release.

2.1 Sources of error

The ASPIRE model assesses product quality by first decomposing the total error for a product into major error components: sampling error, frame error, nonresponse error, measurement error, data processing error, modelling/estimation error, revision error and specification error.

ASPIRE can also be customized so that it considers only those error sources that pertain to a specific statistical product. For example, sampling error would not apply to products that do not employ sampling. The model also accommodates the risk variation across error sources so that a product's overall quality depends more on error sources that pose greater error risks. For example, in the Municipal Accounts, revision error is of low risk because preliminary and final estimates seldom differ appreciably and data users are not affected appreciably by revisions. On the other hand, data processing error is of high risk due to the amount of editing data receive and its potential to affect the final estimates.

2.2 Risk assessment

Each error source is also assigned a risk rating depending upon its potential impact on the quality for a specific product. In this regard, it is important to distinguish between two types of risk referred to as "residual" (or "current") risk and "inherent" (or "potential") risk. *Residual risk* reflects the likelihood that a serious, impactful error might occur from the source *despite* the current efforts that are in place to reduce the risk. *Inherent* risk is the likelihood of such an error *in the absence of* current efforts toward risk mitigation. In other words, inherent reflects the risk of error from the error source if efforts to maintain current, residual error were to be suspended.

Inherent risk is an important component of a product's overall score because it determines the weight attributed to an error source in computing a product's average rating. Residual risk does not play an active role in the evaluation and is seldom noted in the evaluation. Rather, its primary purpose is to clarify the meaning and facilitate the assessment of inherent risk.

2.3 Ratings

A two-step rating process is used to assign a rating from 1-10 for each criterion. First, a criterion is graded on a five point qualitative scale corresponding to Poor, Fair, Good, Very Good, and Excellent. These ratings are later refined by choosing between low or high numerical point ratings within each of the five categories. For example, if an error source is assigned a rating of "Good" in step 1 of the evaluation, a numerical rating of either 5 or 6 is later assigned in step 2 to refine this rating.

The model provides a set of quality guidelines for each rating and each criterion to aid the evaluators in their assessment. These guidelines give the model and the subsequent assessment the necessary conditions to achieve consistency and objectivity, even though it is difficult to totally rule out all subjectivity in the final analysis. The guidelines also give clarity and transparency to the assessment process and allow the product staff to clearly note what is required to attain higher levels of ratings.

2.4 Quality Guidelines

In addition to decomposing total error for a product into its component sources and identification of the risks associated with each source, the ASPIRE model evaluates the potential for these error sources to affect data quality according to five quality criteria, viz., Knowledge of Risks, Communication with

Users, Available Expertise, Compliance with Standards and Best Practices, and Achievement Towards Risk Mitigation or Improvement Plans.

In short, the quality guidelines explain what is specifically required for each criterion in order to receive a specific rating. For the first criterion of *Knowledge of the risks*, the range goes from little acknowledgement of an error source being a potential factor for data quality, which would be given a rating of Poor, to high requirements for a rating of Excellent, where there exists an ongoing program to evaluate bias and variance components associated with the error sources and their implications for data analysis. Likewise, for the second criterion of *Communication of these risks with users*, the model requires the producers of statistics to be thorough, cogent and clear to receive the highest rating of Excellence.

The third criterion, *Available Expertise* in order to be able to deal with the risks, rewards products who access expertise familiar with the adequate techniques required to address the risk factors for a particular error source. The guidelines go from the state where no staff are familiar with these techniques (Poor), to the available expertise being more than adequate to achieve the highest ratings across all quality criteria. These are actively addressing the errors from the source as well as keeping up to date with and contributing to developments in the area of their expertise.

Along similar lines, the fourth criterion, *Compliance with appropriate standards and best practices*, ranges from unawareness and non-compliance (Poor) to full awareness and compliance with standards and best practices. To attain the highest rating, the relevant staff is even contributing to the latest standards and best practices within the particular error source.

For the fifth, and last criterion, *Achievement and/or improvement plans for mitigating the risks*, the guidelines give direction for evaluators to rate what is actually being done to mitigate the risks, which assumes that plans exist as a basis for good results. On the lower end of the scale, very little planning has been done, while on the upper end of the scale, plans are in place and mitigation work has made excellent progress, signifying that the error source is being maintained at an acceptable level given the primary uses of the data.

3. A Case Study – Foreign Trade of Goods (FTG)

ASPIRE has been tested on the following statistical products: Annual Municipal Accounts (RS), Consumer Price Index (CPI), Foreign Trade of Goods Survey (FTG), Labour Force Survey (LFS), Structural Business Survey (SBS), Business Register (BR), Total Population Register (TPR), Survey of Living Conditions (LCS) and Gross Domestic Product (quarterly and annual). In the first round of evaluations, reported in Biemer and Trewin (2012), FTG's evaluation score was among the highest. The FTG continued this high level of performance in Round 2 of ASPIRE, reported in Biemer and Trewin (2013). The following are noteworthy quality improvement activities that occurred in 2012:

- Communication with users regarding survey error generally improved as a result of improvements to the QD.
- Three important studies were completed and documented in reports providing more information on survey error.
- Swedish Customs adopted Statistics Sweden's editing system which demonstrates that it is a state of the art system.
- Plans are in place to better understand the causes of revision error, its impact on important users such as the NA, and some effective means for reducing it over time.
- An asymmetry study with Finland was completed which focused on the effects of coding error on the trade statistics.

- Work is underway to replace the current Excel-based macro-editing software with much improved and flexible software written by IT professionals.
- Use of the agency's "Standardized Toolbox" increased, meaning improved practices.
- A new survey of statistical value is scheduled for 2013.

The current and previous round's ratings are shown below in graphical form.

FTG Accuracy Ratings for 2012

Error source	Score round 1	Score round 2	Knowledge of Risks	Communication to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Specification error	58	58	○	○	◡	◡	○	M
Frame error	58	58	○	○	◡	○	◡	L
Non-response error	62	66	◡	◡	◡	○	◡	M
Measurement error	54	62	◡	○	◡	◡	○	H
Data processing	46	60	◡	◡	◡	◠	○	H
Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Model/estimation	66	80	◡	◡	◎	◎	◡	M
Revision error	62	76	◡	◡	◡	◎	◡	H
Total score	**57,3**	**65,8**						

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

The external evaluators, Biemer and Trewin, also offered a number of recommendations to the FTG staff for their improvement plans for 2013 and beyond. These recommendation included suggestions on (1) reducing the size of their revisions and understand the impact of these on different uses including the National Accounts (2) improving the information provided in the quality declaration on the size of the revision error and making comparison across EU-countries of revision errors and (3) researching to find better ways of estimating the trade below the cut-off limit for Intrastat in order to reassure staff and users that it is insignificant.

4. Limitations of ASPIRE

There are three important strengths of ASPIRE. First, the approach is comprehensive in that it (a) covers all the important sources of error for a product and (b) uses criteria that span all the important risks to product quality. Second, the extent to which the documentation and other information shared during the ASPIRE process is both accurate and complete, the current approach can be used to assign reliable ratings that reflect true data quality risks. Third, ASPIRE identifies areas where improvements are needed ranked in terms of their priority among competing risk areas. For example, priority should be given to areas having highest risk and lowest ratings, assuming other factors being equal.

One weakness of the model is that it is, at best, a proxy measure for product quality. ASPIRE does not provide a direct measure of the total error of a variable, estimate, or product. It relies on the assumption that reducing the risks of poor data quality and improving process quality will lead to real improvements in data quality. Another weakness of the approach is that it is somewhat subjective in that it relies heavily on the knowledge, skill, and impartiality of the evaluators as well as the accuracy and completeness of the information available to the evaluators.

5. References

Biemer, P. and Trewin, D. (2013). A Second Application of the "ASPIRE" Quality Evaluation System for Statistics Sweden, Report to Statistics Sweden.

Biemer, P. and Trewin, D. (2012). Development of Quality Indicators at Statistics Sweden, Report to Statistics Sweden.