Pushing Forward with ASPIRE[1]

Heather BERGDAHL
Quality Coordinator, Statistics Sweden

Dr. Paul BIEMER
Distinguished Fellow, RTI International

Dennis TREWIN
Former Australian Statistician

0. Abstract

Statistics Sweden has gained valuable experience with ASPIRE during the two years and three rounds of evaluations we have been through for ten important statistical products. It has not only given the agency an additional tool for meeting the European Statistics Code of Practice in a fuller way but also measurable, objective and credible results to communicate to stakeholders, users and staff on quality changes in statistics. ASPIRE gives a comprehensive picture of the state of quality in a statistical product which not only provides the necessary overview for improvements but also gives priorities and direction for such work.

In this paper we further describe ASPIRE and how it can be used to set clear measureable goals for product quality. Also we present some results from product evaluations along with associated recommendations for quality improvement.

Finally we summarize lessons learned and present our plans to act upon these in order to make further progress on the road towards higher quality.

1. Introduction

Statistics Sweden submitted a first paper describing the ASPIRE assessment approach to the European Conference on Quality in Official Statistics (Q2012) in Athens titled, A Tool for Managing Product Quality [1]. In this context, the tool was referred to as the SCB model of Quality Indicators. With further refinement and subsequent assessments, the need arose to come up with an official name for the system which we now call ASPIRE (A System for Product Improvement, Review and Evaluation). As the name of this paper conveys, Statistics Sweden is pushing forward with ASPIRE and finds it to be a very useful system with which to manage quality improvements for a range of different statistical products. It is also a great source of inspiration for the staff whose task it is to carry out improvement efforts with the ten of the agency's important products currently engaged in the work.[2]

[1] A System for Product Improvement, Review and Evaluation

[2] The statistical products involved are the Labour Force Survey, Consumer Price Index, Gross Domestic Product – both quarterly and annual (production side only), Foreign Trade of Goods Survey, Annual Municipal Accounts, Structural Business Statistics Survey, Living Conditions Survey, Business Register and Total Population Register

In the present paper we will further describe ASPIRE, how it is used, some results and lessons learned as well as plans for pushing even further. The description of ASPIRE will be subsequently be briefer in the present paper and focus more on refinements to the system in our descriptions. We advise readers who wish to see a more thorough and detailed description of ASPIRE to study the earlier paper [1] which is readily available on the Q2012 website. For additional detail, readers are referred to the latest Biemer and Trewin report [3] available at Statistics Sweden, on which the present paper relies heavily and to an upcoming article in the Journal of Official Statistics [2]. In order to facilitate comparisons with the Q2012 paper [1] we will follow the same structure in the present paper as much as possible.

2. The ASPIRE Model

The ASPIRE model has since the start focused on the Accuracy dimension of quality. Although the model can be extended to all dimensions of quality[3] as tested in Biemer and Trewin (2013) [4], the Statistics Sweden management has chosen to focus the work on Accuracy for the time being. Also the scope of the evaluations is limited to ten important products with the agency.

2.1 Error Sources

For Accuracy, current risks to accuracy are assessed separately for each error source that may affect product quality. Error sources are not the same for all products so they are allowed to differ by type of product in the evaluation. For example, sampling error does not apply to products that do not employ sampling. Or if revised estimates are not issued for a product, then there would be no risk of revision error. As shown in Exhibit 1, three sets of error sources have been identified for the ten products considered in this evaluation. Note that the error sources associated with the two registers – Business and Total Population – are somewhat different than the error sources for the other products. Likewise, the error sources associated with GDP are somewhat different from those for the survey products.

Exhibit 1. Sources of Error Considered by Product

Product	**Error Sources**
Survey Products Foreign Trade of Goods Survey (FTG) Labour Force Survey (LFS) Annual Municipal Accounts (RS) Structural Business Survey (SBS) Consumer Price Index (CPI) Living Conditions Survey (ULF/SILC)	Specification error Frame error Nonresponse error Measurement error Data processing error Sampling error Model/estimation error Revision error
Registers Business Register (BR) Total Population Register (TPR)	Specification error Frame: Overcoverage Undercoverage Duplication Missing Data Content Error
Compilations Quarterly Gross Domestic Product (GDP) Annual GDP	Input data error Compilation error Data processing error Modelling error Balancing error Revision error

[3] The other dimensions of quality comprise Contents/Relevance, Timeliness and Punctuality, Comparability and Coherence, and Accessibility and Clarity.

2.2 Risk assessment

Each error source is also assigned a risk rating depending upon its potential impact on the data 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.

As an example, a product may have very little risk of nonresponse bias as a result of current efforts to maintain high response rates and ensure representativity in the achieved sample. Therefore, its residual risk is considered to be Low. However, should all of these efforts be eliminated, nonresponse bias could then have an important impact on the TSE and the risk to data quality would be high. As a result, the inherent risk is considered to be high although the current, residual risk is low.

Residual risk can change over time depending upon changes in activities of the product to mitigate error risks or when those activities no longer mitigate risk in the same way due to changes in inherent risks. However, inherent risks typically do not change all else being equal. Changes in the survey taking environment that alter the potential for error in the absence of risk mitigation can alter inherent risks, but such environmental changes occur infrequently. For example, the residual risk of nonresponse bias may be reduced if response rates for a survey increase substantially with no change in inherent risk. However, the inherent risk may increase if the target population is becoming increasingly unavailable or uncooperative, even if response rates remain the same due to additional efforts made to maintain them.

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 has not played a very active role in the evaluation and is generally not noted in the evaluation. Rather, its primary purpose is to clarify the meaning and facilitate the assessment of inherent risk. However, changes in residual risk can indicate the success of mitigation efforts and if so should be reflected in the improvement ratings.

2.3 Quality Criteria

ASPIRE involves the rating of quality efforts for the products according to the following quality criteria:

- 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 or IT),
- Compliance with appropriate standards and best practices relevant to the given error source, and,
- Plans and achievements for mitigating the risks.

One significant change in the latest round of ASPIRE, round 3, was the addition of "communication with suppliers" of data and information under the Communication criteria. Prior rounds only assessed "communication with users" regarding the error sources for a product.

2.4 Ratings according to Quality Guidelines

The explicit guidelines developed for each criterion to aid the assessment of current quality and quality improvements have also been further enhanced and improved with successive rounds of evaluation. The application of these guidelines is now facilitated by the use of checklists as illustrated in Exhibit 2. See Biemer and Trewin 2014 [3] for the most recent checklists.

Exhibit 2 Example of Quality Guidelines and conversion to checklist items

Guidelines for the Criterion of Knowledge of Risks regarding "Good" and Very Good"	**Conversion of guideline to checklist item**
"Good": Some work has been done to assess the potential impact of the error source on data quality. **But**: Evaluations have only considered proxy measures (example, error rates) of the impact with no evaluations of MSE components.	Reports exist that gauge the impact of the source of error on data quality using proxy measure (e.g. error rates, missing data rates, qualitative measure of error, etc.) – Yes or No. *Yes, to achieve the level of "Good".*
"Very Good": Studies have estimated relevant bias and variance components associated with the error source and are well-documented. **But**: Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses.	At least one component of the total MSE (bias and variance) of key estimates that is most relevant for the error source has been estimated and is documented – Yes or No. *Yes, to achieve the level of "Good".*

The checklists are generic in that the same checklist could be applied to each relevant error source. Moreover, we believe the simple "yes/no" format used for the checklists eliminates much of the subjectivity and inter-rater variability associated with the quality assessments. In addition, the checklists incorporate an implied rating feature so that upon completing the checklist for a criterion, the rating for that criterion is largely pre-determined based upon the last "yes"-checked item in the list.

A product's *error-level score* is then simply the sum of its ratings (on a scale of 1 to 10) for an error source across the five criteria in section 3.3 divided by the highest score attainable (which is 50 for most products) and then expressed as a percentage. A product's overall score, also expressed as a percentage, is computed by following formula:

$$\text{Overall Score} = \sum_{\text{all error sources}} \frac{(\text{error-level score}) \times (\text{error source weight})}{50 \times (\text{weight sum})}$$

where the "weight" is either 1, 2, or 3 corresponding to an error source's risk; i.e., low, medium, or high, respectively, and "weight sum" is the sum of these weights over all the product's error sources.

3. Application to the Products

The application of the model to the ten products in Exhibit 1 followed a multistep process as follows:

a) Pre-interview activities include two primary activities. The product staff perform a self-evaluation by completing the criteria checklist for each error source. The evaluators then review these checklists along with quality declarations and any additional relevant materials for each product.

b) Each quality interview takes approximately four hours to conduct. The meetings were organized into essentially five parts:

- discussion of any notable changes that have occurred during the preceding 12 months that may have some effect on data quality,
- review of the quality declarations focusing on clarifications of the processes associated with product design, data collection, data processing, estimation, and reporting and emphasizing changes occurring within the past year,
- progress that was made on the recommendations from the previous round
- assignment of preliminary ratings for each criterion by error source using the quality checklists, and
- review of all assigned ratings with a discussion of the results and recommendations for improvement.

Detailed minutes are kept of all the interviews which provide a record of the proceedings and are used extensively in refining the ratings.

c) Control is then made by the experts for consistency in the ratings within and across products. Feedback is then given to the product staff on their preliminary ratings and comments to them. They are asked to correct any inaccurate or misleading information. Thereafter the ratings and recommendations are finalized by the experts.

d) The plan is to repeat this evaluation process annually to monitor quality improvements or deteriorations and to provide feedback – both positive and negative – regarding were future improvement efforts should be directed.

4. Strengths and limitations to the approach

ASPIRE, as any model or method for evaluating the quality of processes as complex as those associated with these ten products is subject to limitations. It does not, in fact, measure the true accuracy of a statistic, which is virtually impossible for many of these products because the data is not available for such calculations. Besides, even if the data were available for bias and variance calculations, we would have to take into account that even these would have limitations. ASPIRE 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 limitation 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 them.

There are, however, three important strengths of ASPIRE:

a) The approach is comprehensive in that it covers all the important error sources of a product and examines criteria that cover important risks to product quality.
b) The checklists used to assign ratings are quite effective at identifying and assessing both manifest and hidden risks to data quality. The process seems therefore capable of assigning reliable ratings that reflect true data quality risks provided that the information provided to the evaluators is accurate and complete.
c) ASPIRE identifies areas where improvements are needed ranked in terms of priority among competing risk areas e.g. high risk areas with lower ratings should be prioritized, all other factors remaining equal.

Given these strengths and limitations we are convinced that the ASPIRE approach is capable of achieving the following goals:

- identify the current, most important threats or risks to the quality of a product,
- apply a structured, comprehensive approach for assessing efforts aimed at reducing these risks,
- identify areas where future efforts are needed to continually improve process and product quality.

We believe that meeting these goals is an important prerequisite for Accuracy to improve, a process which of course also depends on the efforts made with product staff to achieve these goals.

5. Findings based on the Assessments

The evaluations result in a large number of single assessments within each criteria pertaining to each error source and for each statistical product. For each of these assessments a score is assigned. We will present here findings for two products as well as some general recommendations that the expert assessors have offered to Statistics Sweden as a result of the assessments.

5.1 Specific findings

Exhibit 3 Quality evaluation for the Labour Force Survey (LFS)

	Error Source	Average score round 2	Average score round 3	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	52						**H**
	Measurement error	56	68						**H**
	Data processing error	62	62						**M**
	Sampling error	78	80						**M**
	Model/estimation error	60	64						**M**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total score	60,9	64,3						

Scores					**Levels of Risk**			**Changes from round 2**	
					H	**M**	**L**		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

The presentation in Exhibit 3 gives an easily-grasped picture of the quality efforts that the LFS staff and their partners are investing in order to gain control over the relevant error sources for the LFS. Many ratings show a level of "Very good". The level of "Excellent" has also been awarded within Sampling Error. There are two error sources that have been identified with high risk within this product, nonresponse error and measurement error. It is quite evident in the exhibit that work is progressing on measurement error. The staff's knowledge, communication and expertise have improved between rounds 2 and 3 of ASPIRE due to an excellent study regarding measurement error in the LFS [6]. Merely looking at the assessments for the area of non-response would give the impression that work is not being invested in this high-risk area. On the contrary, many activities are in progress at Statistics Sweden with the objectives of counteracting the downward trend in response rates and to gain control over this error source that can potentially bias the LFS estimates. However, according to the experts' assessment, necessary control has not been achieved with the level and focus of the agency's present activities. Several recommendations have been given to Statistics Sweden by the experts on how to focus activities in this area to achieve better results. See Biemer and Trewin 2014 [3] for more details.

In total, the LFS have improved their total score by 3.3 percentage points. Although there are many plans with some indication of progress to mitigate the risks to data quality as indicated in the column on the right for Plans or Achievement towards mitigation of risks, it remains to be seen if these plans will progress to see more substantial improvements for Accuracy with the LFS. This would be indicated by an "Excellent" score in this column.

Exhibit 4 Quality evaluation for the Structural Business Survey (SBS)

	Error Source	Average Score round 2	Average Score round 3	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 over error sources)	**Specification error**	54	58	○	○	◡	○	◡	**M**
	Frame error	64	60	◡	◡	◡	○	◠	**M**
	Non-response error	70	70	◡	○	◡	◡	◡	**M**
	Measurement error	52	56	○	○	○	○	◡	**H**
	Data processing error	60	60	○	○	◡	○	◡	**H**
	Sampling error	84	86	◎	◡	◡	◎	◎	**M**
	Model/estimation error	56	48	○	○	○	◠	◠	**H**
	Revision error	56	54	○	○	◡	◠	◠	**H**
	Total score	60,8	60,1						

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

The presentation in Exhibit 4 also gives a good view of quality efforts that the SBS staff and their partners make to control and minimize the risks to data quality within the relevant error sources for this product. The conditions for the two surveys in Exhibits 3 and 4 are totally different from one another with the SBS collecting economic data from enterprises instead of individuals. The assigned ratings are within a range between Fair and Excellent the latter of which is very clearly in the error source of sampling. The SBS are dealing with four areas that pose high risk to the quality of the statistics they produce. These are measurement and data processing error as well as model/estimation and revision error. Some improvements have obviously been made between rounds 2 and 3 but what is concerning in Exhibit 4 compared to Exhibit 3 is the number of deteriorations noted and the effect this has on the total score, causing a slight decrease between the rounds of evaluation. The reasons for the deteriorations are also apparently in the areas of expertise and planning and reflect shifts in priorities within the area of economic statistics as a whole, which affects the conditions for the SBS to provide statistics of high quality. Improvements in areas such as business profiling, over-coverage in the Business Register, and the storage of metadata have become of less priority in light of development needs for new IT-systems.

Again, several recommendations have been given to Statistics Sweden by the experts on how to focus activities for the SBS in order to achieve higher quality. See Biemer and Trewin 2014 [3] for more details.

5.2 General findings and recommendations

With the three rounds of assessment with these selected products, the external evaluators Biemer and Trewin, have gained useful insights into more cross-cutting issues at Statistics Sweden. Biemer and

Trewin have for this reason regularly offered a list of general recommendations to Statistics Sweden for management to consider. Examples of areas where such recommendations are offered are:

- the need for integration of economic statistics
- the need for additional evaluation studies
- stabilising nonresponse rates in household surveys and managing the potential nonresponse bias
- development of quality declarations
- the need for a systematic approach for archival and retrieval of manuscripts and reports that document quality improvement projects
- the need for an annual process for planning and monitoring projects that address recommendations in the annual ASPIRE reports

As evident, the effects of improvement efforts in cross-cutting issues such as these can have much more far reaching effects in the agency as a whole than any product specific recommendations can have. The last recommendation in the listing above is therefore probably quite crucial for Statistics Sweden to consider for the success of the agency's present product improvement journey.

6. Evaluation and Future work with ASPIRE

The ASPIRE assessment system has been the object for review by the Scientific Board of Statistics Sweden in 2012. The Board was generally pleased with the system and recommended that Statistics Sweden prioritize the quality improvement that is advocated by the ASPIRE guidelines and experts, i.e. to focus on facilitating evaluation studies that increase the knowledge and understanding of error sources that pose risks for poor data quality. The studies should have the ultimate objective of reducing or gaining control over these risks. Such efforts should likely result in statistics with higher accuracy/quality. The Board also had some minor suggestions for refinements to ASPIRE which have been taken into consideration in successive rounds.

Another way of evaluating ASPIRE is to look at the achieved results after the three assessment rounds bearing in mind that the first two rounds were largely test rounds. The following results have been noted thus far:

1. Given the fact that measurement error was identified as a high risk area for most products in the evaluations, several studies were launched at Statistics Sweden to explore approaches with the objective of increasing the level of knowledge at the agency within the area of measurement error. The results give survey managers at Statistics Sweden valuable tools to do studies in this area. One specific study that was carried out is a latent class analysis for the Labour Force Survey [7].

2. An initiative was taken during 2012 with the products involved with ASPIRE to improve the quality of the information provided in the products' quality declarations in response to one of the general recommendations given by the experts. This work resulted in clearer quantitative and qualitative information that is made available to users of the statistics and therefore also gave rise to improved scores for selected areas and products in round 2 of ASPIRE within the area of Communication.

3. Round 3 of ASPIRE showed many improvements in the area of planning for studies and improvement projects which comprised over 40 percent of the total number of improvements compared to just over 20 percent in round 2.

4. The Living Conditions Survey, evaluated for the first time in round 2, has received the necessary confirmation from the experts along with valuable recommendations for a major redesign of the survey which has also been approved. The recommendations have also been promptly acted upon which is seen in the substantially improved overall score for the product in round 3 of ASPIRE with 9 points.

5. An interesting observation with the assessment results is that products which historically and systematically make use of methodological staff are generally receiving higher scores than products that do not such as the two registers and National Accounts (GDP). However, we are seeing a tendency for these latter products to realize the need and see possibilities with more methodological work in conjunction with their product. We are also presently trying to increase the support for these products in this area.

In summary, we have not yet seen substantial improvements to Accuracy in these products with the introduction of ASPIRE. Most of the improvements we are seeing are, however, yielding better conditions for real improvements within Accuracy i.e. with greater compliance with standards and best practices, and in planning for studies and improvement projects. We see in this respect that ASPIRE is not only a system capable of assessing changes to quality yielding quantitative and objective measures and but also one that assesses the degree of maturity in the agency's quality efforts and gives structure and inspiration to the improvement work. The ASPIRE process is actually not expected to produce substantial improvements quickly. Rather we can see that it aims to fundamentally change the core processes and the organisational culture that lead to substantial improvements. In this regard, we must admit that we have launched on a longer journey toward "evolutionary" improvement that is real and lasting. Indeed, we have further steps to take, both large and small, in order to achieve higher levels of Accuracy. This fits in well with Statistics Sweden's long term strategy and goal to provide users with statistics of high quality based upon scientific grounds, which adhere to national and international quality guidelines and standards and are continuously developed to meet user needs.

7. References

[1] Biemer, P., Trewin, D., Bergdahl, H., Japec L., and Pettersson Å. (2012). A Tool for Managing Product Quality, Paper submitted to the European Conference on Quality in Official Statistics (Q2012) in Athens. http://www.q2012.gr/articlefiles/sessions/3.2_Biemer_Bergdahl_Tool%20for%20Managing%20Product%20Quality.pdf

[2] Biemer, P., Trewin, D., Bergdahl, H. (2014). A System for Managing the Quality of Official Statistics, Journal of Official Statistics. (*to be published in 2014*)

[3] Biemer, P., Trewin, D. (2014). A Third Application of ASPIRE for Statistics Sweden.

[4] Biemer, P., Trewin, D. (2013). A Second Application of the "ASPIRE" Quality Evaluation System for Statistics Sweden.

[5] Biemer, P., Trewin, D. (2012). Development of Quality Indicators at Statistics Sweden, Report to the Director General of Statistics Sweden.

[6] Näsén, K., et al. (2014). Measurement errors in the Swedish Labour Force Surveys, Background Facts, Labour and Education Statistics 2014:2, Statistics Sweden. (*in Swedish only*) http://www.scb.se/Statistik/_Publikationer/AM0401_2014A01_BR_AM76BR1402.pdf

[7] Karlsson, P., (2013). Markov Latent Class Analysis of Measurement Errors in Labour Force Status, Statistics Sweden. (*unpublished paper)*