A SIXTH APPLICATION OF ASPIRE FOR STATISTICS SWEDEN

Paul Biemer, Dennis Trewin, Dan Kasprzyk and Jesper Hansson

July 15, 2016

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

In 2011, the Ministry of Finance directed Statistics Sweden to develop a system of quality indicators for a number of key statistical products. This system was to include metrics that reflect current data quality as well as capture any changes in quality that occur over time. With the help of external consultants, Statistics Sweden developed a quality evaluation approach that is referred to as ASPIRE: A System for Product Improvement Review and Evaluation or ASPIRE (see Biemer and Trewin, 2013 and Biemer, Trewin, Japec and Bergdahl, 2014). The review process has been conducted annually since 2011 for essentially the same core set of statistical products.

This report summarizes the results from the sixth annual review (Round 6) of ASPIRE which was conducted in May/June 2016 by the ASPIRE team (viz., Biemer, Trewin, Kasprzyk and Hansson). Because of the need to propose recommendations prior to the annual planning process of Statistics Sweden, this round of ASPIRE took place earlier in the year and only 9 months since Round 5. The Round 6 report covers the following 10 products which were also reviewed in the previous round: Annual Municipal Accounts (RS), Consumer Price Index (CPI), Foreign Trade of Goods Survey (FTG), Labour Force Survey (LFS), Survey of Living Conditions (LCS/SILC), Structural Business Survey (SBS), Business Register (BR), Total Population Register (TPR), Producer and Import Price Index (PPI), and the GDP component of the quarterly National Accounts, (GDP(Q)).

Although not part of the data product review process, the ASPIRE team participated in a number of additional discussions and presentations including the integration of economic statistics, a review of the EVRY outsourcing experiment for the LFS, several projects that are part of the broader nonresponse project, the sensitivity analysis in economic statistics project, the work on quality indicators for base registers project, and the research on the BR survey feedback project. Comments on these discussions are found in this report. The ASPIRE team presented two seminars on trade-offs among quality dimensions – one in Stockholm and one in Örebro – and led a breakfast seminar/discussion on coherence in official statistics. The team also presented a seminar on the ASPIRE process as it related to the CPI to the CPI Board.

The practice begun in Round 5 of having four external reviewers participate in ASPIRE evaluations continued in this round but the protocol was changed. Two external reviewers, rather than four, were assigned to each data product evaluation. This change allowed two product interviews to be conducted simultaneously, thus increasing the efficiency of the interview process. As in the prior rounds, the evaluation for each product involved a self-assessment, reviews of relevant documentation, presentations by staff, interviews of key staff, and a staff review of the preliminary evaluation results with feedback.

As in previous rounds, each product was scored (on a 10-point scale) against criteria that were standardized across error sources. In Round 6, we continued the use of six criteria that we began in Round 5. Four of the six criteria were the same in Rounds 1-4. Last year, for Round 5, one prior criterion on "planning and achievements towards the mitigation of risks was split into two criteria: one for "planning toward risk mitigation" and the second for the "effectiveness of the risk mitigation activities." This change was needed because relatively too much emphasis was being placed on planning and not enough on the implementation and evaluation of these plans. The second criterion places greater emphasis on the effectiveness of the planning and risk mitigation activities. The use of quality criteria guidelines and checklists greatly facilitate the application of the criteria and, we believe, provide more consistent ratings. Overall scores were tallied as a weighted average of the scores for each error source where the weights were 1, 2, or 3

corresponding respectively to low, medium, or high intrinsic risks associated with each error source. More details on the ASPIRE process may be found in Biemer, et al (2014).

To highlight some of the key results from this round, all but two of the ten survey and register products showed an improvement in ratings. With a maximum possible score of 100 percent (indicating perfect quality), the product scores ranged from 48.2 percent (for LCS/SILC) to 64.6 percent (for FTG) with an average rating of 57.9 percent. (Exhibits 2a and 2b provide the scores for each product by error source.) Although not in this report, we prepared a 'Change Matrix' for each product that provides explanations for any changes in ratings since the previous round. They are available from Heather Bergdahl on request.

Changes in scores from Round 5 to Round 6 ranged from 2.4 (PPI) to -1.1 (LCS/SILC) with an average of 0.8 percentage points. While an average increase of 0.8 is not a substantial improvement, it does indicate that overall, the quality of data products continue to steadily improve.

Some additional findings from the data product reviews found in Section 2 include the following:

- As in prior rounds, model/estimation has the lowest mean rating although it was tied with frame error in this round.
- Model/estimation error scores improved by 2 percentage points on average. This error source is medium to high risk for all survey products and high risk for the GDP(Q) so this is a positive development.
- As in prior rounds, the error source with the highest quality score is sampling error.

In addition to the product reviews, the ASPIRE team had the opportunity to hear and discuss a number of current crosscutting projects and bring our perspective to these projects:

- *Statistical Coherence (Section 4.1).* Statistical Coherence can be regarded as a comprehensive indicator of Accuracy. Previous round ASPIRE recommendations identified a need for more Coherence across Statistics Sweden's statistical products. In Round 6, recommendations are made for better internal communication to address coherence issues, as well as better communication with core external users, to better understand the technical and communication issues related to Coherence from their perspective.
- *Integration and Coordination of Economic Statistics (Section 4.2).* This topic has been discussed in previous ASPIRE reports and is of major importance. Identifying where Statistics Sweden would like to be in five years in terms of an integrated economic statistics system is key to addressing the recommendations made by the ASPIRE team with regards to the BR, profiling of large enterprises, harmonization of business units, standard classifications, consistent methodological decisions, and the rationalization of collections.
- *The Outsourcing Project with the LFS (Section 4.3)* We note that a preliminary analysis of the LFS comparing data collected by the external partner EVRY and data collected by Statistics Sweden showed some differences in data quality indicators and differences in cost metrics. Before extending the extent of outsourcing, additional research is needed on several dimensions of the results to better understand the reasons for the data quality and cost differences.
- *Nonresponse in Household Surveys (Section 4.4).* Nonresponse continues to be a major problem in household surveys. Significant resources, staff time, and research project effort continue to be expended to mitigate this problem. Progress has been noted through increased knowledge of the causes of the problem. Some mitigation efforts have been effective, most

notably the use of mixed mode data collection and the development of models to reduce both callbacks and noncontacts. Much more remains to be done in these areas including the use of paper and pencil interviewing (PAPI) to mitigate web and telephone nonresponse and a greater use of propensity models to direct the data collection efforts.

- *The Role of Methodologists (Section 4.5).* The ASPIRE team had several opportunities to speak with methodologists who work on the various products reviewed as part of ASPIRE as well as group discussions. Data quality improvement projects do not always use methodologists, when, in fact, they bring a strong technical perspective to a project. We note the need for a greater involvement of methodologists in the planning, implementation, and evaluation of data quality improvement projects.
- *Sensitivity Analysis in Economic Statistics (Section 4.6).* During recent years, many innovative and important projects have been carried out as part of the sensitivity analysis project of GDP. These projects have had the purpose of building knowledge about how sensitive estimates in the National Accounts are to uncertainties in the input data. They have led to several important changes. Statistics Sweden should continue to provide resources for the sensitivity analysis in economic statistics project.
- *Enhancing Evaluation (Section 4.7.1).* In prior rounds, we noted that Statistics Sweden did not have a strong evaluation culture even though outstanding quantitative skills exist within Statistics Sweden. Evaluation work is undertaken but there is considerable scope for improvement, particularly in documentation and follow-up action. Staff would benefit from identifying the most relevant evaluation methods used at Statistics Sweden together with their strengths and weaknesses, followed up with supplementary support and training arrangements. These methods might be documented in a Manual or Guidelines of Best Practice on Evaluation Methods. There is also a need to address cultural issues. Once the Manual is available, there should be planned socialisation activities with the product areas. Key issues to address in this socialisation are the importance of evaluation (regarding it as an investment rather than a cost), the need to plan and fund evaluation activities, the importance of engaging methodological support, and the need for documentation.
- *Quality Indicators for Base Registers (Section 4.7.2).* During the past year, staff conducted a pre-study that identified a selection of quality indicators for registers, then implemented, and reported on a number of such indicators. Staff focused on quality indicators that used variables related to coverage, linkage and classification. The pre-study also identified contact information variables for study. The registers and the variables they contain are critical for sampling and statistical operations. This useful project should continue, if possible.

In Section 5, we have identified 11 recommendations that we consider highest priority for improving the quality of Statistics Sweden's data products. Priorities were assessed on the basis of impact and viability with cost being an important aspect of viability. They are listed below in no particular order and further discussed in Section 5.

1. Commission an external expert to conduct a comprehensive review of Statistics Sweden interviewing facilities and operations.
2. Conduct additional evaluations aimed at better understanding the costs and quality differences between Statistics Sweden and EVRY data collection for the LFS.
3. Pursue continued research on the use of mixed mode interviews in household surveys, incorporating PAPI.

4. Consider increased use of call monitoring for quality improvement with more frequent interviewer performance feedback.
5. Investigate alternative approaches to estimating household consumption, particularly through non-survey sources, given the difficulties with the Household Budget Survey.
6. Convene a technical advisory committee (TAC) to provide guidance on the redesign of the LCS/SILC/Children's Survey trilogy and answer fundamental questions about consolidation and simplification.
7. Implement additional improvements to the Business Register.
8. Develop a more integrated approach to Economic Statistics data products and rationalize collections where the uses do not justify the costs to Statistics Sweden and respondents.
9. Improve staff knowledge of evaluation methods and the techniques for assessing mitigation effectiveness.
10. Encourage greater use of methodologists in all improvement projects for Statistics Sweden's data products.
11. Rethink modalities for producer-user communication and feedback.

2 BACKGROUND AND INTRODUCTION

This is the sixth round of ASPIRE. In Round 4, we decided to shorten the reporting process because the background and technical details of ASPIRE have been well documented in prior reports as well as in the JOS journal article by Biemer, Trewin, Bergdahl and Japec (2014). This report conforms to this shortened format. As with the previous rounds, the focus of this ASPIRE round is on the Accuracy quality dimension.

This year, the same ten products as last year were reviewed. The ten products that comprise the scope of our review are listed in Exhibit 1.

As in prior follow up rounds, one objective of Round 6 was to identify areas where clear improvements (or deteriorations) had been made since the previous evaluation. Another objective was to follow up on previous year's recommendations. For all products, our report identifies the highest priority areas for improvement at the product level. Furthermore, some general recommendations are made for high priority crosscutting issues.

The ASPIRE process, error sources and evaluation criteria, that were applied in this review are essentially the same as in Round 5 (see Biemer, Trewin, Kasprzyk and Hansson (2015). The general ASPIRE process is described in greater detail in Biemer, et al (2014).

Exhibit 1. Sources of Error Considered by Product

Product	**Error Sources**
Survey Products 1. Labour Force Survey (LFS) 2. Living Conditions Survey (LCS/SILC) 3. Consumer Price Index (CPI) 4. Producer and Import Price Index, (PPI) 5. Annual Municipal Accounts (RS) 6. Foreign Trade of Goods (FTG) 7. Structural Business Statistics (SBS)	Specification error Frame error Nonresponse error Measurement error Data processing error Sampling error Model/estimation error Revision error
Registers 8. Business Register (BR) 9. Total Population Register (TPR)	Specification error Frame: Overcoverage Undercoverage Duplication Missing Data Content Error
Compilations 10. Quarterly Gross Domestic Product (GDP(Q))	Input data error Compilation error Modelling error Data processing error Deflation/Reflation error Balancing error Revision error

2.1 CHANGES TO ASPIRE IN ROUND 6

This sixth round of ASPIRE was conducted in May/June 2016, only nine months since the last round. The review was conducted several months earlier in the calendar year to facilitate Statistics Sweden's annual planning process that should be completed by August. A similar change was made last year when the process was moved from November 2014 in round 4 to August 2015 in Round 5. The shorter review period has, of course, implications on the progress one could expect from products. The next round of ASPIRE is tentatively planned for May 2017, which will mean a return to the 12-month review period used for the first four rounds of ASPIRE.

Statistics Sweden has developed a web interface that facilitates the product completions of checklists. All entries in the checklists are stored in a database. After some minor introductory problems, the system worked well. Next year, when the product staff become accustomed to the interface, we expect even greater efficiencies in completing and reviewing the checklists. The system will also allow us to follow ratings and comments to individual error sources over time in a more efficient way.

Beginning in Round 5, the number of external reviewers participating in the ASPIRE process was increased from two to four. In Round 5, all four reviewers participated in all product interviews. In this round, interviews were conducted in parallel with two reviewers participating in each interview. The reviewers met both before and after the interviews in order to exchange views and address issues arising from the ratings. The team also met at the end of the ASPIRE process to reconcile scores and impose greater consistency of ratings among products. All four reviewers are jointly responsible for the conclusions in this report.

In previous rounds of ASPIRE there have often been discussions of how to consistently and logically assign the various types of error risks to the error sources. It is not always obvious under which error source heading (for example, Specification Error, Measurement Error, Modelling/Estimation Error, and so on) a particular error risk should be rated. In order to clarify how we allocate error sources we have started to develop short documents describing the delineation of error sources for each individual product. The delineations identify product specific error risks that are key for each respective error source. These documents provide examples of relevant error sources that have been identified in previous rounds and thereby assist staff in allocating comments in the checklists consistently to specific error sources. Our goal is to complete the delineation documentation project before the commencement of Round 7 and to make these documents available to products areas to facilitate the development of their ASPIRE self-evaluations.

2.2 SCOPE OF THE REVIEW

On the top panel of Exhibit 1 are the seven survey products that are included in the ASPIRE review in this review round. The error sources associated with these products are shown to the right of these products. Likewise, the middle panel shows the two registers included in this review and their error sources that were reviewed in all prior rounds. In this, as well as the previous, round of ASPIRE we review quarterly GDP. The error sources associated with the quarterly GDP (which are discussed below) are shown on the right panel under the heading GDP(Q).

In addition to the ten product reviews conducted in this round, the ASPIRE team followed up on a number of general, more specific issues that pertain to the ASPIRE evaluations. These include:

- issues associated with statistical Coherence at Statistics Sweden,
- the integration of economic statistics,
- the experiment to outsource data collection for the LFS,
- nonresponse in household surveys,
- the critical role of methodologists at Statistics Sweden,
- on-going sensitivity analysis research in economic statistics,
- methods and tools for evaluating the effectiveness of error mitigation plans, and
- quality indicators for base registers

We have included separate sections in the report providing our thoughts and recommendations on these topics.

The individual product recommendations resulting from the ASPIRE review are important for improving the quality of the specific programs. Furthermore, the product reviews and staff discussion enabled the ASPIRE team to observe crosscutting issues that the organization should review carefully as they affect multiple programs and are important to quality improvements throughout Statistics Sweden. Our crosscutting (or general) recommendations are discussed in Section 4.

Reviewing ten data products resulted in numerous individual recommendations. The discussion of crosscutting issues also identified a number of recommendations to improve the quality of Statistics Sweden's data products. As a result, ASPIRE team was asked by senior management to identify high priority recommendations for their consideration. These recommendations are found in Section 5.

The next section summarises the results of the quality evaluations and presents recommendations for quality improvements for each of the ten products reviewed.

3 PRODUCT REVIEWS

Exhibit 2a provides the overall scores for the seven survey products and two registers for each product's relevant error sources. Exhibit 2b provides the overall scores for GDP (Q). The error sources are shown in first column of these tables while the other columns refer to the products being evaluated. For each product, a dark shaded cell corresponds to a "High Risk" error source for that product. Medium shaded cells correspond to "Medium Risk" error sources and unshaded cells correspond to "Low Risk" error sources for a product.

As discussed in previous ASPIRE reports (see, for example, Biemer and Trewin, 2014), the interpretation of the error sources and criteria may vary between surveys, registers and compilations. For example, for a survey, it may be appropriate to consider measures such as bias and variance because the products of surveys are estimates. For registers, the concepts of bias and variance do not apply because they are data sets, not estimates. Instead, it may be more appropriate to consider the validity and reliability of the register data because these quality concepts are more appropriate for data sets and values.

Likewise, Exhibit 2b provides the scores for the quarterly GDP. As discussed in Biemer and Trewin (2014), the error structure used in the evaluation of this product has been customized to reflect the unique operations associated with compiling the data and generating quarterly estimates of GDP. For that reason, the accuracy of GDP is treated separately from the other nine products.

There are a number of other differences among the error models used for surveys, registers and compilations which are explained much greater detail in Biemer, et al (2014).

Exhibit 2c summarizes the total scores for all ten products over all six ASPIRE rounds in the form of a histogram. These exhibits will be discussed in greater detail in the next section.

3.1 GENERAL OBSERVATIONS

Before discussing each product's detailed ratings, some general observations regarding the results in Exhibits 2a, 2b and 2c as well as a few caveats can be stated.

First, there is a natural tendency to compare the overall scores across the products or to rank the products by their total score. However, the ASPIRE model was not developed to facilitate such inter-product comparisons and there are some risks associated with ranking products in this manner. For one, the total score for a product reflects a weighting of the error sources by the risk levels, which can vary considerably across products. Products with many high risk error sources, such as GDP, may be at somewhat of a disadvantage in such comparisons because they must perform well in many high risk areas in order to achieve a high score.

Second, the assessment of low, medium, or high risk is done within a product, not across products. Thus, it is possible that a high risk error source for one product could be of less importance to Statistics Sweden than a medium risk error source for another product if the latter product carries greater importance to Statistics Sweden or for official statistics. If resources devoted to quality improvements are greater for one product than another, this could also explain why some products are able to show greater improvements than others. Further, although we have attempted to achieve some degree of consistency in ratings among products, some inconsistencies surely remain.

Finally, the scores assigned to a particular error source for a product have an unknown level of uncertainty due to some element of subjectivity in the assignment of ratings as well as other imperfections in the rating process. A difference of 2 or 3 points in the overall product scores may

not be meaningful because a reassessment of the product could reasonably produce an overall score that differs from the assigned score by that margin. Thus, any ranking of products would need to acknowledge these inevitable and unknown uncertainties in the ratings.

A more appropriate use of the product scores is to compare scores for the same product across review rounds as a way of assessing progress toward improvements. As noted in Biemer et al (2014), the ASPIRE review process focuses on process changes, new knowledge gained or communicated, and new research conducted or planned since the prior round that could alter the error risks and justify changes in the quality ratings. We believe this process assures a high level of reliability in the round-to-round changes scores for each product.

Before discussing the results in Exhibits 2a and 2b, it should be noted that a number of corrections to the ratings and risk levels were made in Round 6 than may affect the Round 5 to Round 6 comparisons, viz.

- for the PPI, the risk level was raised for measurement error from Medium to High,
- for the RS, the risk level for Specification Error was changed from N/A to Low, and
- for the TPR and BR, ratings for Frame Error (Duplication) for the criteria, Planning for Risk Mitigation and Effectiveness of Risk Mitigation Activities were changed from "not applicable (N/A)" to applicable and were assigned a numerical value.

Close inspection of scores in Exhibits 2a and 2b yield the following observations:

- The last row of Exhibits 2a and 2b shows the Round 5 to Round 6 changes in the overall quality ratings by product. Ratings increased for eight out of ten products participating in both this and the previous round; seven increased by at least 1 point. For Exhibit 3a, the average increase is 0.8 points for surveys and registers and 1.1 points for GDP.
- The largest improvement in Exhibit 3a is the PPI (2.4). However, recall that Round 5 was the first time this product was evaluated. As shown in Exhibit 3c, most products show considerable improvement the second time they are evaluated. This may be due in part to the "low-hanging fruit effect" which means that initial quality improvements tend to address the areas requiring the least effort to raise ratings.
- The LCS/SILC had the largest rating drop – 1.1 points. As noted in our review, this product suffers from a number of high risk quality concerns which, for various reasons, are not being adequately addressed. The TPR also experienced a small drop in ratings. However, this is less of a concern given their high ratings in Round 4 which were sustained in Round 5.
- Two error sources in Exhibit 2a are tied for the lowest mean rating: Frame Error and Model/Estimation Error at 55. The former fell from a rating of 56 in Round 5 while the latter increased from 53. Looking back across all six rounds, Model/Estimation Error tends to be consistently the lowest rated error source with Frame Error only slightly higher.
- Measurement/Content Error is of high risk for seven out of nine products in Exhibit 2a – the most of any error source. It has also seen great improvement across the six rounds; its rating has increased by more than 1 point per round on average.
- Likewise, Data Processing Error is high risk for five out of six products and it, too, has increased by more than 1 point per round on average.
- Not surprisingly, the error source with the highest quality score, and by a wide margin, is sampling error. This was also true in all the prior rounds.

To increase their ratings, products could concentrate on high risk areas having below average ratings provided that viable mitigation strategies can be identified. However, these areas may not be the highest priority areas for quality improvement as other factors need to be taken into account. For example, one should consider the feasibility and costs of the improvements, the needs and relative priorities of other products and improvement activities, the importance of improving Accuracy compared with other dimensions of quality and initiatives that are promoted by Eurostat and other external groups.

Exhibit 2a. Product Error-Level, Overall Level, and Error Source-Level Ratings with Risk-Levels Highlighted and Comparisons to Round 5 Overall Ratings

Error Source/Product	LFS	LCS/SILC	CPI	PPI	RS	FTG	SBS	BR	TPR	Mean rating
Specification error	62	52	68	48	48	58	57	58	52	56
Frame error	57	38	62	53	50	57	60	56	59	55
overcoverage								58	65	62
undercoverage								52	57	54
duplication								57	57	57
Nonresponse error /Missing data	58	40	55	57	58	62	70	47	60	56
Measurement error/Content	65	50	67	48	53	65	55	55	55	57
Data processing error	55	47	67	58	60	65	58	N/A	N/A	59
Sampling error	78	60	67	63	N/A	N/A	85	N/A	N/A	71
Model/estimation error	63	52	52	48	48	73	48	N/A	N/A	55
Revision error	N/A	N/A	N/A	N/A	60	67	55	N/A	N/A	61
Round 6 mean rating	63,0	48,2	63,2	53,7	54,9	64,6	59,7	55,0	58,7	57,9
Round 5 mean rating	61,9	49,4	61,8	51,3	53,6	64,1	58,7	53,8	59,0	57,1
Change (improvement/deterioration)	▲ 1,1	▼ -1,1	▲ 1,5	▲ 2,4	▲ 1,2	▲ 0,5	▲ 1,0	▲ 1,2	▼ -0,3	0,8

HIGH RISK
MEDIUM RISK
LOW RISK
N/A= Not Applicable

Exhibit 2b. Product Error-Level, Overall Level, and Error Source-Level Rating with Risk-Levels Highlighted and Comparisons to Round 5 for the National Accounts

Error Source/Product	GDP(Q)
Input data sources Production Side (Average)	58
Index of Service Production (ISP)	62
Index of Industrial Productions (IIP)	62
Merchanting (including royalites, licensing, R&D)	50
Input data sources Expenditure Side (Average)	54
Turnover	55
Government	57
Investments	52
Inventories	55
Net Exports in Goods and Services	53
Compilation error (modelling)	47
Compilation error (data processing)	50
Deflation error	60
Balancing error	52
Revision error	57
Round 6 mean rating	54,0
Round 5 mean rating	52,9
Change (improvement/deterioration)	▲ 1,1

HIGH RISK
MEDIUM RISK
LOW RISK
N/A= Not Applicable

Exhibit 2c shows the overall ratings by product for the six evaluation rounds. Recall that, in Round 5, a sixth criterion (viz., Effectiveness of Mitigation Measures) was introduced which substantially reduced the ratings for most products and error sources. This criterion was also applied in the current round with essentially the same results although there were a few improvements. Some reasons for the low ratings for this criterion as well as well as some remediation measures were discussed in the Round 5 report. As a result of this new criterion, the ratings for the current round and Round 5 are not directly comparable to the ratings for Rounds 1-4 which did not use the sixth criterion. For Round 5, we reported ratings under both the new (with Effectiveness) and old (without Effectiveness) criteria to bridge the Round 5 ratings with the prior rounds' ratings. For Exhibit 2c, we continue to use both sets of scores for Round 5; however, for the current round, only the scores under the new, revised criteria are shown.

As previously noted, the LCS/SILC is consistently the lowest rated product in the ASPIRE process despite experiencing substantial improvement since Round 2 when it was first evaluated. Hopefully, this product's drop in ratings in the current round can be reversed in Round 7. Our recommendations (in Section 3.2.2) provide some guidance in that regard. The FTG tends to be the highest rated product although its ratings have somewhat stabilised over the last two rounds.

It is somewhat disappointing to observe that the magnitude of the average increase for last three rounds has been somewhat smaller than in earlier rounds, as can be seen from the "Mean" bars, the last set of bars in Exhibit 2c. However, note that the previous rounds of ASPIRE were conducted at 12 month intervals whereas, the last two rounds were only 9 months apart. Thus, the time available to implement improvements was about 25 percent less.

Some additional possible explanations for the small average increase in ratings were also noted in our Round 5 report. One is that the so-called "low hanging fruit" of quality improvement (i.e., improvements that can be more readily accomplished with low budgets and minimal activity) was picked up in early rounds. The achievement of further improvements will require a greater commitment of resources, personnel and innovative thinking.

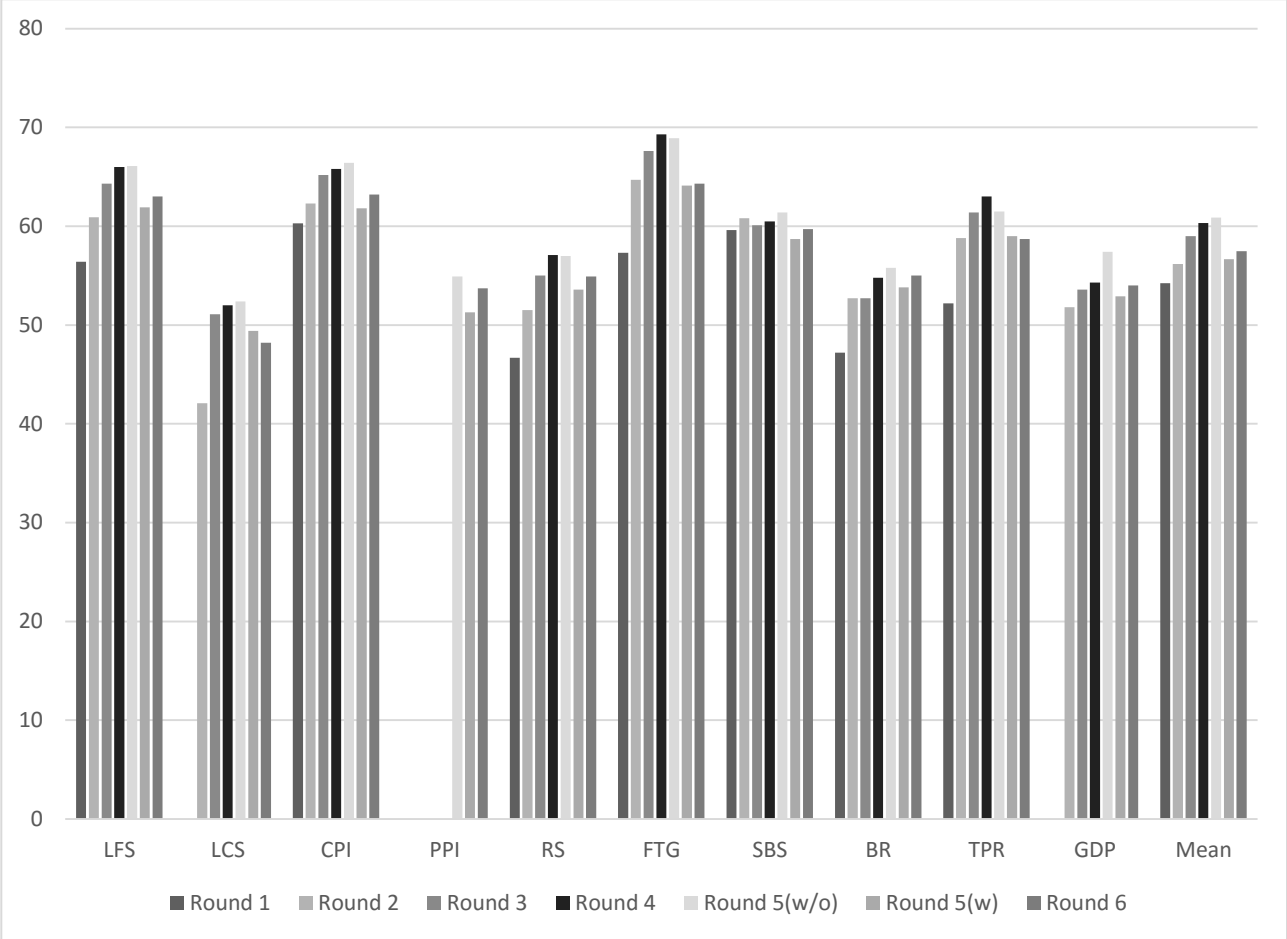
In addition, we have noticed that some products, citing budgetary constraints and production demands, do not assign sufficiently high priority to continuous quality improvements. This can happen when management's attention is so focused on the routine production work that the objectives of continual quality improvement are given lower priority.

Still, even products that attribute a higher priority to quality improvements are finding it to be quite difficult to maintain ratings of "Very Good" or "Excellent" for planning for mitigation for some error sources as they attend to quality improvements for other, needier error sources. Just like there is a limit to the number of "balls in the air" a juggler can handle, so too as mitigation effort with one error source is improved, another may reduce from the relative lack of attention. This should be expected because there are seldom enough resources, personnel and time to do all that is needed to plan quality improvement efforts in all areas simultaneously. The challenge for the product manager is to ensure the focus is on the highest priority quality improvements at a particular point of time.

Finally, as we have cautioned in our prior reports, the results in Exhibit 2c do not necessarily represent the pace of data quality improvements for these 10 products. Although data quality improvement is the ultimate goal of ASPIRE, an improvement in ASPIRE ratings means that products have improved relative to the six ASPIRE criteria. We can only say that data quality has been improved to the extent the six criteria reflect actual reductions in the risks of product error. As an example, products may increase their ratings by developing plans designed to reduce the error. However, actual error reduction may not be realized until these plans have been implemented. This

is one of the reasons the sixth criterion (viz., for Risk Mitigation Effectiveness) was introduced in Round 5.

Exhibit 2c. Overall Quality Ratings for All Products by Round including the Two Ratings for Round 5 Corresponding to "with (w)" and "without (w/o)" the Separate Assessments of Mitigation Effectiveness



Notes: [1]*LCS/SILC was not evaluated in Round 1.*
[2]*PPI was evaluated for the first time in Round 5.*
[3]*The GDP(P) component of GDP(Q) was evaluated in Rounds 2-4. Thus, ratings for GDP(P/Q) reflect GDP(P) for Rounds 2-4 and GDP(Q) (i.e., GDP(P) and GDP(E)) for Round 5 and 6.*
[4]*New criterion added for Effectiveness of Mitigation Activities in Round 5*

3.2 ASPIRE REVIEWS

In this section, we review the progress over the past 9 months for the ten products shown in Exhibit 1 that were also reviewed in Round 5 using the checklists that appear in Annex 1. Customized versions of the checklists were used for the quarterly GDP to take into account the unique error structure of the national accounts (see Biemer, et al, 2014). The ratings for each of the six criteria and applicable error sources are updated to reflect this progress. Then, we conclude the review of each product with our recommendations for the coming year.

3.2.1 LABOUR FORCE SURVEY (LFS)

CONTEXT

The LFS staff have made substantial progress during for the current review period to address a number of error sources. There have been studies of coverage and nonresponse error using comparisons between the TPR and the LFS sample. Progress has been made to improve the employment questions for youth. There has also been progress to improve the variance estimators for change estimators and the models for post-survey adjustments. In addition, although response rates for the LFS, which are now approximately 58 percent, have continued their downward trend, progress has been made on several fronts for reducing the risks of nonresponse bias and achieving greater sample representativity. These include mixed mode data collection; outsourcing of a substantial part of the LFS sample; and the use of stopping rules to maximize callback efficiency and reduce nonresponse bias due to noncontacts. Overall, this is a very good review round for the LFS.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *EVRY Experiment*. The EVRY experiment was conducted to compare costs and quality of the LFS data collection carried out by Statistics Sweden and EVRY and to provide recommendations to Statistics Sweden on how to proceed regarding the outsourcing of data collection. The experiment found a number of important benefits using EVRY including somewhat higher response rates.
- *Analysis of Post-survey Adjustment Models*. To evaluate coverage error, LFS staff is comparing TPR and weighted LFS totals for the full sample by overall and by panel. One result so far is that people born abroad, particularly recent migrants, are less represented in the sample due to the time lag from sampling only once a year. Thus, persons immigrating to Sweden during the sample year have no chance of being selected for that year's sample. There are plans to repeat this analysis for the selection weighted respondent sample as well.
- *Web Data Collection*. LFS is preparing to use web/telephone mixed mode data collection in panels 2-7 for the permanently employed only. A test is planned for 2017 and 2018. The LFS realize that this may not have much impact on the nonresponse rate but could increase representativity and reduce data collection costs.
- *Analysis of Noncontacts*. Work has begun on simulating a selective callback strategy that would curtail callbacks on cases deemed to be non-productive in favour of increasing callbacks on cases deemed to be productive. There are plans to apply this methodology to live LFS cases in the coming year.

- *Continuing Work Related to Measurement Error*. Based upon results from the reinterview survey that was conducted two years ago, questions regarding the concepts of permanently and temporarily employed as well as full and part-time work are being revised for 15-24 year olds. A small test will be carried out (30 persons) in the near future.

Other Accomplishments

- Staff are engaged with the data collection department to better plan the utilization of interviewing resources. They have used "gap" analysis (see discussion below) to identify where interviewer staffing could be improved to achieve greater efficiency and productivity in data collection.
- Eurostat wants to harmonise LFS questions across all member countries and, in that regard, has developed a model for questionnaire organization and wordings. LFS staff participated in a preliminary test that suggests the Eurostat's current approach is quite flawed.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Dependent and Proxy Interviewing*. As far as we know, the practices of proxy and dependent interviewing have never been evaluated for the LFS, yet they could have a substantial effect on data quality. For example, there is evidence from the literature that dependent interviewing results in fewer employment transitions (see, for example, Biemer and Lyberg, 2003). Whether this implies better data quality is debatable, particularly if interviewers are not following standardized questioning protocols. One way to investigate how the interviewers are handling these situations is to use call monitoring with behaviour coding. Some questions that could be answered through analysis and interviewer observation are the following:
    a. Are interviewers following standardized procedures in asking about changes in labour forces status via dependent interviewing?
    b. What proportion of the interviews is conducted with a proxy-respondent?
    c. What are the characteristics of sample members whose information is provided by proxy-respondents versus self-respondents? What is the relationship between the proxy informant and the sample member?
    d. To what extent do interviewers attempt to obtain self-response (for example, how many callbacks are made to interview to the sample member before taking a proxy)? How do the interviewers handle the process for identifying and interviewing a qualified proxy?
    e. How do interviewers handle the process of referring to prior labour force status responses in obtaining the current labour force status? How often do they lead the respondent to a response? How is this handled when the prior interviewer was conducted by proxy and the current one is self-response (or vice-versa)?
    f. How do labour force transitions compare for proxy and self-respondents?
2. *Post-survey Adjustment Modelling*. We encourage the current research being conducted to evaluate nonresponse and coverage bias and believe this work should continue with high priority. Comparisons of the TPR with the full sample and the respondent sample have been very revealing so far regarding the representativity of both samples. Likewise, the work examining the sensitivity of nonresponse adjustments to the exclusion/inclusion of auxiliary

variables has been very valuable for increasing Statistics Sweden's knowledge of the nonresponse bias and how to adjust for it. We believe this work should be augmented by taking into account the new "resident propensity" variable that is being computed by TPR staff for both nonresponse weighting and response rate computation.

3. *Questionnaire Design*. The reinterview study suggested that 15 to 24 year olds have difficulty providing correct responses to the permanent/temporary as well as the full-time/part-time employment questions. It is important to understand why, which is the purpose of the cognitive interviewing research that is being conducted for these questions. In that way, alternative methods for collecting this information from this age group can be developed. In addition, before new questions can be introduced in the LFS, tests of the questions in the production setting should be conducted. Call monitoring with behaviour coding should be used to study the ability of the interviewers to ask the new questions as well as the ability of the respondents to understand and answer them.
4. *Mixed-Mode Research.* As we noted in the last round, the research on mixed mode data collection is potentially quite important for all household surveys, including the LFS. However, we are concerned that the protocols being considered do nothing to increase the Wave 1 response rates, which will have a significant influence on the level of response for all subsequent interview waves. Some experimentation on protocol designs that allow for mixed mode response at wave 1 should be considered.
5. *Analysis of Noncontacts*. Further analysis of the paradata for contacts, noncontacts, interviews and refusals is needed to better understand the causes of nonresponse, particularly with regard to noncontacts. We encourage the LFS staff to continue their work in this area with particular emphasis on optimal calling strategies for noncontact cases.

Other Areas for Consideration

1. *Updating the quality declaration*. We appreciate the recent revision of the quality declaration. However, two areas where the quality declaration could be improved are (a) some discussion of indirect (proxy) interviewing and its effects on measurement error and nonresponse, (b) a discussion and examination of the effects of dependent interviewing on response quality – both advantages and disadvantages, and (c) the addition of quantitative information as it becomes available from the ongoing studies. These elements should be considered in the next revision of the quality declaration.
2. *Gap Analysis*. The LFS methodologists should continue to work with the data collection department regarding interviewer staffing and call scheduling to meet the demands of LFS data collection. In the interview, we discussed the idea of a "gap" analysis; i.e., a report that shows the number of interviewing hours requested by the LFS staff (by timeslot) (denoted by R_s) compared with the number of interviewing hours provided by the data collection department (P_s). We see at least two advantages of such a report:
    a. The gap, i.e., $R_s - P_s$, is an indicator of the shortfall or surplus of interviewing hours for some time slot *s* compared to the hours needed to meeting the rigorous production schedule of the LFS. As such, it can be used as an indication of whether data collection is on track to meet the schedule because a shortfall of interviewing hours dedicated to the LFS suggests that production may be falling behind schedule. It also reflects interviewing efficiency because a surplus of interviewer hours dedicated to the LFS indicates that calling may be inefficient.

b. Hopefully, that the gap is small. If it is large, the objective is to close the gap over time. Thus, the gap should be analysed longitudinally to assess progress by the data collection department in becoming more efficient and productive.

We recommend that the gap analysis be performed at least monthly and that the results be shared with the LFS and data collection department management. At that point, these two groups should meet to discuss the results and the progress being made toward more efficient and productive interviewing.

3. *Industry and Occupation Coding*. The LFS staff is collaborating with the staff that does industry and occupation coding and that work has produced positive results. One area that should be explored is whether the information now being collected by interviewers on occupation is sufficient for the coders to assign accurate codes. A study conducted by Biemer and Caspar (1994) showed that if certain key elements of an occupation are not recorded by interviewers, occupations are coded much less accurately.

Exhibit 3. Labour Force Survey (LFS), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	Specification error	63	62	◡	○	◡	◡	◡	◠	L
	Frame error	52	57	◡	◡	◡	○	○	◠	L
	Non-response /Missing data	55	58	◡	○	◡	○	○	○	H
	Measurement /Content	63	65	◡	◡	◡	◡	◡	◠	H
	Data processing error	57	55	○	○	◡	◡	○	◠	M
	Sampling error	78	78	◡	◎	◡	◎	◡	◡	M
	Model/estimation error	63	63	○	○	◡	◡	◡	○	M
	Total score	61,9	63,0							

Scores					Levels of Risk			Changes from round 5	
●	◠	○	◡	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.2 LIVING CONDITIONS SURVEY (LCS/SILC)

CONTEXT

The data products program that is referred to as the Living Conditions Survey is complex and multi-faceted, producing multiple data products, both cross-section and longitudinal. The LCS/SILC program has a strong and longstanding user base and is subject to a variety of Eurostat data requirements, some of which can be difficult to implement. The subject matter unit, working with the cognitive laboratory, has a regular, ongoing program of questionnaire development for its topical content. Because this is an ongoing program, staff are very engaged in an operational cycle of activity – revising content as necessary, working with the data collection staff as the survey is conducted, processing the survey data, developing essential products for its user community, and planning for the next round. The ongoing production aspects of the LCS/SILC are significant. The program would not have its user constituency if the operational aspects were not addressed well and in a timely fashion. The ASPIRE team acknowledges the need to place a priority on operational and user concerns. We do believe, however, that somewhat more emphasis on measuring the quality of LCS/SILC data and a renewed emphasis on measuring the effectiveness of processes and procedures to improve the program is essential.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Implications of Eurostat's Delivery Schedule*. To study the implications of the new data delivery schedule proposed by Eurostat, staff studied possible seasonal effects in the estimates produced with a six-month data collection period and a report is available in Swedish.
- *Survey Integration*. Following our recommendation, staff initiated a project to consider design options for better integrating the three surveys that constitute the LCS/SILC.
- *Children's Survey Precision Requirements*. Following our recommendation to review precision requirements for the children's survey based on primary uses of the data, the sample size for the children's study was increased.
- *Gini Coefficient Evaluation*. Staff will initiate a project this fall (which has been approved by management) to evaluate the two estimates of the Gini Coefficient based on Statistics Sweden data and understand and document the differences of these estimates.

Other Accomplishments

- Improved timeliness of the SILC data deliveries to Eurostat.
- Development of improved production programs, resulting in a faster production and publication process for the national LCS.
- Working with the cognitive laboratory, development of revised questionnaires for "housing conditions "and "material deprivation".
- Initiating a project to develop new and improved weights using calibration methods.
- Revising the Children's Survey questionnaire and reducing the number of questions.

Recommendations for Coming Year

1. *LCS/SILC Technical Advisory Committee*. The data collection requirements identified recently by Eurostat present an opportunity to change the status quo of the LCS/SILC survey operations. The Eurostat requirements provide an opportunity to think innovatively and creatively about how to reduce the complexity of the LCS/SILC data collection system. An external technical advisory committee should be commissioned to consider a radical redesign of the system that meets the requirements of both Eurostat and Swedish users. Questions that the committee should address include:

   a) What minimum data requirements on living conditions and quality of life will meet the needs of both national and the international users?

   b) How will the recommendations on quality of life measures prepared by external reviewer, Professor Emeritus Robert Erikson (http://www.regeringen.se/contentassets/dbb4c911287747b3943b4f61cf2b344f/far-vi-det-battre-om-matt-pa-livskvalitet-.pf) be addressed in the redesign?

   c) Despite successful reductions in respondent burden over the last few years, can staff continue to address respondent burden and nonresponse reducing the LCS content and still maintain relevance?

   d) Should the Children's Survey continue as a supplement or should it be independent of the LCS?

   e) Noting the reductions made a few years ago, can the LCS/SILC be consolidated into one SILC survey with supplemental questions as needed?

   In general, how can the old survey system be redesigned into a new system that meets both national and international requirements while maintaining essential components of the old system? There are many questions to address and possible redesign approaches to consider. We recommend Statistics Sweden convene a technical advisory committee to provide guidance and support for a redesign of the LCS/SILC/Children's Survey. The committee should have representatives from the product area, survey methodology, and the user community with a chair independent of these areas. The committee should develop and evaluate alternative designs satisfying the data requirements.

2. *Communication with the nonresponse project*. LCS/SILC staff should collaborate with methodologists who have been involved with the broader nonresponse project to understand the nonresponse research conducted on the LFS and how the LCS can benefit. It is important that LCS/SILC stay current with these developments and strategize on applying this research to LCS/SILC.

3. *Communication with TPR staff*. LCS/SILC staff should have regular conversations with the TPR staff on the potential uses of the overcoverage/ resident propensity variable and the household register and assess how LCS/SILC can take advantage of the TPR staff work.

4. *Improved and more complete documentation*. The complexity of the survey requires continued vigilance at documenting processes, procedures, and decisions. Some effort and priority should be given to technical documentation to ensure data users understand the survey's design and estimation procedures. The quality declaration should be regularly updated and translated into English, say at three years intervals. This would serve the needs of the external EU community as well as the ASPIRE review process.

5. *Nonresponse bias in the Children's Survey*. The Children's Survey has a very high nonresponse rate and would benefit from research on nonresponse bias in the survey's key estimates.

Exhibit 4. Living Conditions Survey (LCS/SILC), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	**Specification error**	52	52	○	○	◒	○	○	◓	M
	Frame error	40	38	◓	◓	◒	◓	◓	◓	M
	Non-response /Missing data	45	40	◓	◓	○	○	◓	◓	H
	Measurement /Content	50	50	○	◓	◒	○	○	◓	H
	Data processing error	47	47	○	◓	◒	○	○	◓	L
	Sampling error	60	60	◒	◒	◒	○	○	○	M
	Model/estimation error	52	52	○	◓	◒	○	○	◓	H
	Total score	49,4	48,2							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.3 CONSUMER PRICE INDEX (CPI)

CONTEXT

The Swedish CPI continues to be of a very high standard especially when compared to those of other countries. Nevertheless, there continues to be a range of initiatives that result in continuous improvements to the accuracy of the CPI taking advantage of new technology and new data sources.

We were given a presentation of the planning system that is used for organising and managing CPI development and maintenance activities. It has a 5 year time horizon and includes proposals arising from the CPI Board and ASPIRE. This will help to ensure that the continuous improvement approach will continue in the future.

In previous reviews, we have commented on the Household Budget Survey (HBS) and the limitations on its use for the CPI. The 2016 HBS was discontinued after a few months because of data quality concerns and will likely not be conducted again for some years. There are now other data sources (e.g. scanner data and other data sources that might be used for compiling weights for the CPI (and the National Accounts). Some effort should be devoted to understanding how they might be used for these important users.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *CPI Error Study.* There are plans to undertake a study this September of the most important Total Survey Error components of the CPI. It will be based somewhat on the 1999 quantitative study of error sources in the CPI.
- *Extended Use of Scanner and Internet Data.* Scanner data use has been extended to include prices for alcohol and pharmaceutical goods. Web scraping has been extended to car rentals, car inspections and mobile phones. Consequently, most of the price data collection is now being undertaken centrally where quality is easier to manage.
- *Quality Adjustments.* A very innovative Implicit Quality Index diagnostic tool was introduced last year. This enables the impact of quality adjustments to be assessed and is an important macro-editing tool. It has identified some possible problems with high tech goods for example. It might also be used to analyse differences in interpretation between the central collection staff and the field staff. It will be introduced into the standard production environment.
- *Hand-held Computers.* Updated tablet computers have been successfully introduced and paradata is being collected. There are plans to analyse this paradata.
- *Sample Redesign.* A sample redesign was introduced in 2016. Because of the greater clustering of the sample, the cost of data collection was reduced. Because the intra-cluster correlation is quite small, this increase in clustering was shown to have little impact on sampling errors. In addition, the sample redesign provided an opportunity to reduce the number of price collectors from around 100 to about 40 increasing the overall skill levels of the price collectors while reducing costs.

Other Accomplishments

- *Other Quality Improvements.* There have been a range of important improvements such as updating the list of products included in the CPI basket; an upgraded sample design of rental apartments; commencing a study of quality adjustments in the field; investigating international

internet purchases; and discussions with experts on the determinants of quality for different products (e.g. clothing).

RECOMMENDATIONS

1. *CPI Error Study.* There are plans to estimate the main error components of the CPI based somewhat on the 1999 CPI error study. When the data become available, they should be used to support a Total Survey Error approach to improving the accuracy of the CPI because this would provide the evidence base for deciding where to best place the research and error mitigation efforts. It would also provide information in support of the EU grant to optimise resources across the whole CPI. Previous studies have shown that quality change has been the most important source of error.

2. *Extended Use of Scanner and Internet Data.* Continue to broaden the use of scanner data and 'web scraping' to reduce sampling errors in the relevant components but, perhaps more importantly, to reduce the measurement errors, especially those associated with assessing discounts. It might also provide data that can be used for hedonic models. Extend this approach to problematic areas like international internet purchases.

3. *Quality Adjustments.* Research into methods for estimating quality adjustments, including quality adjustments in the field, should continue as this may be one of the most important sources of error in the CPI. The Implicit Quality Index diagnostic tool should be very useful in assessing the merits of the different methods.

4. *Monitoring the Work of Price Collectors.* There is a lot of dependency on the work of the price collectors and their work should be routinely monitored. The newly introduced tablet technology has been used to collect 'paradata' as well as price data. This capability should be used to better monitor and evaluate the quality and effectiveness of the work of the price collectors. There would be merit in researching how other countries have used paradata in their CPI collections.

5. *CPI Weights.* Because it is unlikely that reliable HBS data will become available in the near future, there should be some investigations into data sources (e.g. scanner data), other than the HBS, that could be used for updating weights at the upper and lower levels of the CPI. The focus should be on those products where the price movements might be quite different to the rest of the CPI.

6. *Owner Occupied Housing Costs.* The treatment of owner occupied housing costs has been under debate for several decades. This effort should be kept in proportion given that the huge amount of effort in Sweden and elsewhere does not seem to have provided an entirely satisfactory solution. The present treatment is somewhat irregular in an international perspective, which creates problems for the analysis of inflation and communication of monetary policy. Harmonization within the EU could be a way forward and should be investigated.

Exhibit 5. Consumer Price Index (CPI), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	**Specification error**	68	68							H
	Frame error	60	62							M
	Non-response /Missing data	52	55							L
	Measurement /Content	65	67							H
	Data processing error	67	67							H
	Sampling error	65	67							H
	Model/estimation error	48	52							H
	Total score	61,8	63,2							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

CONTEXT

Evaluation of the PPI, both for goods and services, was introduced in the ASPIRE process last year. The PPI, among other things, provides very important input to the National Accounts when calculating GDP in constant prices. Our understanding of the ongoing work related to quality improvements for these products has increased significantly since the previous round of ASPIRE. This has resulted in some changes in the ratings but most changes are due to actual improvements.

We have also adjusted the assessed level of risk for Measurement Error from Medium to High, which is consistent with the CPI. This change was also made retroactively to the Round 5 scores thereby neutralizing the effect of this change when comparing the scores from Round 5 to Round 6.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Change in Data Collection Process*: The Data Collection department has taken over data collection, beginning in January this year. During the first few months, the response rate fell slightly but is now at par with last year. There is work going on to further improve the response rate, for instance by reminding enterprises earlier in the month. Further improvements are expected since the Data Collection department can apply their experience and resources.
- *Communication with Core Users*: Discussions are held regularly with the National Accounts and different units in the Department of Economic Statistics. This communication has given important insights to both producers and users of the PPI.
- *Sensitivity Analysis.* Participation in the project "Sensitivity Analysis for GDP" has continued since last year and several papers are going to be presented at conferences this year. An example is the ambitious attempt to assess the size of errors that has resulted in the report "Uncertainties in the Swedish PPI and SPPI", which is to be presented at the Q2016 conference. This has generated increased knowledge, both at the Price Statistics Unit and at other units/departments at Statistics Sweden, about the importance of the PPI in the calculation of GDP in constant prices.

Other Accomplishments

- *New Product Groups.* Two new product groups, SPIN 74 Professional, scientific and technical activities and SPIN 82 Office administrative, office support and other business support activities were introduced in 2016. A pilot survey for four new product groups (TV-broadcasting, radio, travel agency and licences) is being conducted this year. The results for these products will probably be published in 2018 after the quality has been evaluated. Including these products will increase coverage of the services prices, excluding health care and schools where no market prices exist, from about 80 to 86 per cent.
- *Comparison with Other Countries.* A tool for comparing the indices of some important products with several other European countries has been developed. This could be used both as an input to macro-editing and to detect products where other, potentially better, methods may be available.
- *Assessment of Nonsampling Errors.* A new tool with an informative cobweb diagram summarising information showing assessed levels of nonsampling error sources have been

developed. These assessments are supported by a much more structured gathering of information about a number of quality indicators, e.g. the amount of imputation, nonresponse rates and so forth. So far, five product groups have been analysed. With the infrastructure now in place, at least five other product groups could be added each quarter.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Relative Risk Assessment Tool.* Implement the new tool for assessing relative risk for error-by-error source in different product groups in regular production mode. Increase the number of products groups covered so that the most important ones are on board during next year.
2. *Benchmark with Another Country.* Conduct a benchmark study with another country that is considered to be at the frontier regarding best practice (e.g. Denmark or Germany). Focus primarily on the nonsampling error sources (measurement, quality adjustments, coverage of products) where we think the scope for improvements are most likely.
3. *Measure the Price of Trade Margins.* Start planning for the introduction of measuring prices of trade margins. A promising pre-study already exists that could be used as a base for a field study with experimental price collection.
4. *Improve User Communication.* Sustain regular communication with core users of PPI, the most important being the National Accounts. Consider creating a user group with representatives from both inside and outside Statistics Sweden. An alternative could be to extend the responsibilities of the CPI board to cover the PPI as well. In particular, the input from users is needed in order to decide on priorities for future development work.
5. *Monitor Quality Adjustments.* Develop a measure comparable to the Implicit Quality Index of the CPI for the PPI. This measure could be a valuable tool for both keeping track of quality adjustments and as an illustration of their importance to key users (e.g. the NA). It may also provide measures of the effectiveness of the quality adjustment processes.

Other Areas for Consideration

- *Increase Knowledge in Best Practices of Quality Adjustment.* There is a vast ongoing research into methods for measuring complex service prices and quality adjustments. This may have a substantial influence on methods used to improve accuracy in the future. Statistics Sweden should follow this research closely to be able to improve the PPI.
- *Consider the Costs when Increasing Coverage.* The extension of the PPI to cover more products (services) has been done largely without increasing resources. Given resources, the accuracy in other product groups may deteriorate since less time can be devoted to monitor each individual product group. When planning for further extensions of coverage, this issue has to be taken into account and additional resources allocated for this work if necessary.

Exhibit 6. Producer and Import Price Index (PPI), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	**Specification error**	40	48	○	○	○	○	○	◓	H
	Frame error	53	53	○	○	○	◒	○	◓	M
	Non-response /Missing data	52	57	○	○	◒	○	○	◓	M
	Measurement /Content	47	48	○	◓	○	○	○	◓	H
	Data processing error	58	58	◒	○	◒	◒	○	◓	H
	Sampling error	63	63	◒	○	◒	◒	◒	○	H
	Model/estimation error	47	48	○	◓	○	◒	○	◓	H
	Total score	51,3	53,7							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.4 ANNUAL MUNICIPAL ACCOUNTS (RS)

Since the Round 5 ASPIRE review, the RS data collection staff were reorganized into a more consolidated team to allow greater collaboration among the staff members. Priorities within this consolidated team can more easily be identified and acted upon now that the budgets of several data collections are merged, resulting in a shared budget. The team had several important administrative matters to address including developing accurate job descriptions for individual staff members and this took considerable effort. A retirement of the IT staff resulted in several new staff and their work on the project broadened the knowledge base. In spite of these administrative obligations and staff changes, the RS staff were able to make data quality improvements.

We should also note that, in Round 6, the N/A (not applicable) designations previously assigned to Specification Error and to Planning and Effectiveness under Frame Error were reactivated. These changes were also made retroactively to the Round 5 scores thereby neutralizing the effect of this change when comparing the scores from Round 5 to Round 6.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Developing Flow Charts*. In response our recommendation, the RS staff developed flowcharts of the editing process that illustrated possible unwarranted looping and redundancies in the process. The exercise proved helpful in possibly identifying editing system inefficiencies and ideas for improving the effectiveness of the editing process. Additional research is needed, particularly in quantifying the magnitude and extent of edit changes at various editing decision points in the process.
- *Allocating Common Costs Model.* As we had previously recommended, the RS staff performed an experiment whereby the allocation keys used to disaggregate common costs to various sub-activities be evaluated by using the Statistics Sweden allocation model on the municipalities that use their own model (about one fifth of all municipalities) and comparing the two allocations. The results showed that the macro-level differences were not large and were inconsequential to the National Accounts. Differences existed at the micro-level, however, that are being further investigated. In addition, this analysis revealed an important incidental finding that, over the last several years, common costs has been increasing significantly more than gross costs – an inexplicable result that heretofore was undetected by macro-editing.
- *Disability Care Estimates*. Also, as noted in prior ASPIRE rounds, the RS staff should monitor the disability care estimates in the coming year and consider whether their current procedures for mitigating this risk can be improved. In response, the RS staff reviewed the editing rules for the estimates, established maximum differences from average values and maximum differences from the previous year to identify outlier values.

Other Accomplishments

- *Transmission Interface*. The transmission interface for respondents put into service several years ago allows a review of more of the data elements and allows the municipalities to comment on the data they submit, resulting in fewer follow-up contacts with municipalities by staff.

- *Revision Policy*. The Revision Policy adopted in 2015 has resulted in municipalities ensuring the data they submit are correct earlier in the submission cycle. No municipality submitted after the final revision date. Next year, some indicators will be developed to gauge the degree to which data quality has improved as a result of the new revision policy
- *Improving Timeliness.* RS staff worked with late responding municipalities on their data submissions. Staff visited eight municipalities and called another 12 based on their being late in sending data for several years. Interest by the municipalities indicated a desire for providing high quality data, but no measurement tools are in place to quantify changes in data quality.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Increase in Common Costs*. As noted above, research with the common costs model found that common costs has been increasing significantly more than gross costs. Staff need to follow up to establish that this is a correct finding, why it is happening and why this anomaly was not detected in the current macro-editing process. Are the current edits and data reviews insufficient?
2. *Analysis of Changes at Decision Points.* The macro-level flowcharts developed this past year are very helpful to understand the data submission and editing process. RS staff should continue this work especially to identify areas within the system that are inefficient or produce little change in estimates. Statistics, such as edit failure rates and the number and type of changes at the decision points in the flow charts should be produced. The effect of the edits on selected estimates should also be evaluated.
3. *Quantifying Effectiveness.* Processes in the RS program have changed over the last several years – often for what appears to be a positive change. Unless, a data product initiates a program that develops statistics and other measures to quantify the effectiveness of new processes, the user will never know the effects of changed procedures. RS staff should develop measures of the effectiveness of their edits and processes, including statistics that show the sensitivity of the estimates to such changes. Quantitative evidence of the effectiveness of the revision policy should be produced.

Other Areas for Consideration

- *Manual Editing.* The flow charts identified an important aspect of the RS edit and submission system – the manual editing phase. Plans should be made to understand the key aspects of the manual edits – productive and unproductive edits – and ensure edits are applied with a consistent and uniform set of principles. A detailed flow chart of the manual edits is desirable.
- *Analysis of Comments*. RS staff should work with respondents to develop realistic substantive categories of comments in order to monitor the comments through the IT Interface, thus allowing the RS staff to understand or explain the data submitted.
- *Specification Error*. Municipalities respond to data requests from Statistics Sweden in a variety of ways, often times providing data generated through their accounting system, and not necessarily in conformance with the data element specified by Statistics Sweden. An understanding of this type of error, specification error, is desirable, particularly as it affects the National Accounts.

- *Missing Data*. RS staff have a good understanding of the data collected. The user community is not nearly as versed in the quality of the data, especially with respect to missing data. In this case, the RS staff should document the extent of missing data and include this information in an updated quality declaration.

Exhibit 7. Annual Municipal Accounts (RS), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	**Specification error**	48	48	○	○	○	○	◓	◓	L
	Frame error	50	50	○	○	◒	◒	◓	◓	L
	Non-response /Missing data	57	58	○	○	◒	○	○	◓	M
	Measurement /Content	57	53	○	○	◒	◒	○	◓	M
	Data processing error	60	60	◒	○	◒	◒	○	◓	H
	Model/estimation error	40	48	○	◓	◒	○	○	◓	M
	Revision error	58	60	○	○	◒	◒	○	○	L
	Total score	53,6	54,9							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.6 FOREIGN TRADE OF GOODS (FTG)

CONTEXT

For the past nine months, much effort has been devoted to IT-related issues that have absorbed both resources and attention to other possible quality improvements. Nevertheless, important progress in some areas has been made.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Statistical Value Survey***.** FTG staff have also been investigating options for the design of the quinquennial Survey of Statistical Value. In 2017, the staff are hoping to conduct a Eurostat-funded study of the design options.
- *Communication with NA Staff***.** Quarterly meetings with National Accounts staff have continued which has increased cooperation and communication between the two units.

Other Accomplishments

- *Data Editing.* The Eurostat funded project to evaluate components of the FTG data editing system was completed and a report written titled "Evaluating and Improving the Validation Process in the Swedish FTG Statistics." Now the FTG is trying to work through the suggestions and implement them.
- *SIMSTAT.* Another Eurostat funded project was completed titled "Redesign of Intrastat." This report put forth an alternative to reliance on SIMSTAT for import data in the FTG. It recommends a stratified sampling approach to the collection of import data that has important advantages over a single flow (i.e., the replacement of import data with SIMSTAT) solution.
- *Large Enterprises.* The FTG staff have also been communicating with the Large Enterprise Unit (LEU) to obtain information on large enterprises that either fail to respond on time or respond with faulty data.
- *Methodologists.* The FTG expanded their team of methodologists from two to four in order reduce the risk that methodological help is unavailable at critical times during the year.
- *Unit Nonresponse.* The process for imputation in the case of unit nonresponse has been simplified by reducing the number of models from 13 to 5. This simplification came with no apparent loss in estimation quality.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Estimation below the Cut-off.* As noted in our Round 5 review, the sampling cut-off for Intrastat was doubled to SEK 9 million for imports in 2015 to reduce respondent burden measured in the aggregate. The effects of this change on model/estimation bias have not been assessed. Further, the imputation process below the threshold is neither well documented nor well understood. We recommend that the FTG (a) evaluate the effect of raising the cut-off for Intrastat imports on estimator bias and (b) substantially improve the documentation of the estimation process for enterprises below the cut-off.
2. *Statistical Value Survey*. We encourage the FTG staff to continue to pursue redesign options for the Statistical Value Survey that will reduce respondent burden. For example, it may be

possible to use alternative sources of data to estimate statistical value rather than the survey. The FTG should pursue a research agenda in this area with possible funding from Eurostat and close consultation with the National Accounts.

3. *Respondent Burden*. There has been much effort devoted to reducing respondent burden; however, it is not clear how the FTG staff defines burden or what the current levels of burden are by whatever definition makes the most sense. For example, Statistics Sweden's standard definition of burden used in the Annual Report to the government is one definition. However, it may be appropriate to consider an alternative definition based upon the perception of the enterprise of the burden in responding to the FTG. Whatever definition is used, it is not possible to show effectiveness in reducing burden without being able to quantify the burden before and after some reduction measure is applied. The FTG should consider what definition of burden is appropriate for their work and how they might measure it.

4. *Respondent Burden Perception*. In that regard, has respondent burden really been reduced by the Intrastat Data Entry Package (IDEP.web)? The FTG should consider this question and try to provide evidence that respondents regard the system positively and they perceive that their burden has been reduced. In addition, we encourage the staff to mount an evaluation to see if there is any impact on the accuracy of responses.

Other Areas for Consideration

- *Separating Trade in Goods and Services.* The effects on FTG and FTS estimates due to the inability to accurately separate trade in goods from trade in services for some products.

- *Accuracy of Commodity Codes.* The accuracy of the FTG statistics is highly dependent on the accuracy of the commodity codes that enterprises assign to a good. There is a need for an evaluation of commodity coding error for the most problematic CN8 and CN6 codes.

Exhibit 8. Foreign Trade of Goods (FTG), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	Specification error	57	58	○	○	◒	◒	◒	◓	M
	Frame error	57	57	○	○	◒	◒	○	○	L
	Non-response /Missing data	62	62	◒	◒	◒	○	○	○	M
	Measurement /Content	63	65	◒	○	◒	◒	◒	○	H
	Data processing error	65	65	◒	◒	◒	◒	◒	◓	H
	Model/estimation error	73	73	◒	◒	◎	◎	◒	◓	M
	Revision error	67	67	◒	◒	◒	◒	◒	◓	H
	Total score	64,1	64,6							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.7 STRUCTURAL BUSINESS STATISTICS (SBS)

CONTEXT

The SBS gradually improved its capabilities of capturing data and processing it with higher accuracy. Significant efforts have been put into preparations for the new IT-system and planning for more coherent business statistics. Although there has only been nine months since the last round of APSIRE, important progress in some areas relating to quality have been made.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Effects of Editing*. A promising study on the effects of editing on the quality of the estimates has started. A first preliminary report has studied the effects of editing on a few core variables (output, intermediate consumption and value added). The experiment was carried out by first removing the edits for the 600 largest enterprises only and then removing the edits for all enterprises. The conclusion was that editing of the largest enterprises had most impact. At the aggregate level, there was, however, not much effect on value added, but on the industry level there were significant effects. There is, however, a lack of data on the subprocesses within editing and the amount of resources put into editing, so it is difficult to give explicit guidelines about what changes to editing should be made.
- *Profiling of Large Enterprises*. Regarding the profiling of large enterprises, there are now explicit suggestions on how to create Kind of Activity Units (KAUs) when all the required financial data are unavailable. At present less than 50 enterprises have been split into KAUs and previously there was an analysis suggesting that about 200 enterprises should be divided into several KAUs. One suggestion is to use models for some of the variables making it possible to handle some 800 enterprises instead, without significant increase in the respondent burden.

Other Accomplishments

- *Reduced Respondent Burden.* A study has reached the conclusion that one could reduce the number of enterprises who fill in the complete form from 500 to 300 without much loss of accuracy. This would save resources both from the 200 firms not having to answer an extensive survey and SBS staff ′s time for editing.
- *Electronic Reporting.* A further increase in the number of businesses providing their data electronically.
- *Improved Response Rates.* The response rate has increased in some industries, e.g. taxi drivers and restaurants. This could be explained by the work done by the Tax Agency, but Statistics Sweden has also modified the letter of introduction so that it now is clearer on the obligation to provide data.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Increase Profiling of Large Enterprises.* SBS should aggressively promote and actively support Statistics Sweden efforts to increase the number of large enterprises that are profiled to ensure the NACE classifications are accurate in SBS and National Accounts (NA) statistics. The work should be done in cooperation with the BR, Large Enterprise Unit (LEU) and the NA in order to increase the number of KAUs.

2. *Processing Data on Editing.* SBS should obtain more quantitative data on the subprocesses and costs of editing that would help to evaluate the effectiveness of editing. The aim is to save resources that could be used elsewhere to improve accuracy.
3. *Quarterly SBS.* Begin preparing for quarterly structural business statistics relying on an integrated quarterly survey of investments, inventories and intermediate consumption. There may also be a scope for integrating other current business surveys in order to increase coherence and reduce the respondent burden.

Other Areas for Consideration

- *Prepare for the New BR.* Statistical improvements in the BR have no clear timetable, but the SBS should anyway be thinking about the move to the new BR. There are likely to be discontinuities in the SBS data series and some thought should be given on how to manage these discontinuities and whether any additional information is required to help manage the discontinuities. For example, over-coverage because of inactive units may be significantly reduced with the new BR.

Exhibit 9. Structural Business Statistics (SBS), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	**Specification error**	57	57	○	○	◒	○	◒	◓	M
	Frame error	60	60	◒	◒	◒	○	○	◓	M
	Non-response /Missing data	70	70	◒	○	◒	◒	◒	○	M
	Measurement /Content	53	55	○	○	○	○	○	○	H
	Data processing error	55	58	◒	○	◒	○	◒	◓	H
	Sampling error	85	85	◎	◒	◒	◎	◎	◒	M
	Model/estimation error	48	48	○	○	○	◓	○	◓	H
	Revision error	53	55	○	○	◒	○	◓	○	H
	Total score	58,7	59,7							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.8 BUSINESS REGISTER (BR)

An accurate BR is essential to the quality of economic statistics. Nevertheless, despite the important improvements, we remain concerned about some aspects of the BR. This has been reinforced by a Total Survey Error study undertaken by SBS during 2015 and repeated this year. This study showed that the two most important sources of error for that survey were due to the BR: (1) the number of inactive units on the Register and (2) insufficient division of large businesses into KAUs.

There is enough information available on the various registers to indicate whether a business is active or not. This information has been used to reduce the impact of (1). With respect to (2), there now appear to be more definite plans to address using data available through the tax system as much as possible and utilising a modelling approach for those enterprises where the complete set of financial information is not available at the KAU level.

We should note that the error source "Frame Error: Duplication", which was designated as "N/A" (not applicable) in Round 5, was reactivated in Round 6. This change was also made retroactively to Round 5 thereby neutralizing the effect of this change when comparing the scores from Round 5 to Round 6.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Development of the new BR System*. Work has continued on schedule on the development of the new BR and it should be in operation soon. With a reduced budget, the focus has been on introducing the new IT system but deferring the statistical enhancements development work. Consequently, there are no definite plans for developing a version of the BR that would cover the areas affecting the accuracy of the BR for statistical requirements. A plan should be prepared for extending the BR system to incorporate the highest priority statistical developments (see recommendations below).
- *Profiling of Large Enterprises*. There are more definite plans to profile large enterprises and create KAUs later this year. It is recognised that modelling will need to be used in some cases. The opportunity will also be used to create enterprises in line with EU regulations.

Other Accomplishments

- *Collaboration with the Tax Agency.* The closer collaboration with the Swedish Tax Agency has continued with several meetings at the management level. The Tax Agency is reviewing how it will use NACE codes and this might be an opportunity to improve the quality of coding.
- *Register Maintenance Meetings.* All BR staff working on register maintenance are now meeting 4 to 5 times per year to discuss quality issues. These meetings are similar to the "quality circle" meetings popularized in the total quality management literature. These group efforts at addressing quality issues should lead to improvements.
- *Local Units.* An annual survey is sent to multi-location enterprises; however, a reactivated routine based on administrative sources from the Tax Agency (called the SKD) is used to complement the survey and has actually reduced the undercoverage by adding 600 new local units belonging to multi-location enterprises.
- *Accuracy of Estimates of Employment Change*. The reactivated SKD routine has also been used to check the accuracy of estimates of employment size (full time equivalents or FTEs)

especially for large enterprises listed on the BR. Updates to employment data where the analysis indicated there might be a problem.

- *Accuracy of Postal Addresses*. A one-off study was undertaken to see the error in the postal addresses – about 1.6 percent were found to be in error indicating this information is relatively accurate.
- *Quality Indicators in the BR.* Work has continued on the relatively new study on 'Quality Indicators in the Base Registers'. Its objective is to measure and quantify the uncertainties in the content of the Base Registers including the BR. The indicators are organized according to whether they represent Coverage, Linkage, Classifications and Contact errors.
- *Business Units.* The choice of sample unit seems to vary considerably across surveys within economic statistics. While this may make sense to the individual collection areas when looking at their surveys in isolation, it is counter to the Coherence dimension of quality and thus may not be a sensible approach from an organisational perspective. This has been discussed further with one consideration being to use different statistical units for sample selection, data collection and publication. Some of these discussions have taken place within an "Objects and Populations Group" and we support this initiative.
- *Internal Users*. The User Group for Internal Users continues to meet on a regular basis. It should be used to suggest priority areas for improvement of the BR.

Dependent Survey Feedback

Although it is contrary to Statistics Sweden policy, there have been revisions to the NACE codes and other enterprise data in light of more recent information obtained from enterprise surveys – a practice known as "dependent survey feedback". This is understandable given that otherwise enterprises could be allocated to the wrong industry stratum causing inaccuracies in the estimates as well as causing the enterprise to be confused by having received an inappropriate questionnaire for their industry. Some of this feedback is not to the BR itself but to local Registers maintained by the collection areas. The use of survey dependent feedback is fine for contact information and for other information in the 'take all' strata but it potentially creates biases for sampled strata. Studies in Statistics Sweden have shown that this practice does indeed bias survey estimates but it reduces variances. A study has shown that the preferred solution is to have two registers – one including corrected information and the other including the uncorrected information. The former would be used for data collection activities whereas the latter would be used for estimation. In practice, there would only be one physical register with the other being a virtual register based on indicators on the register. It may not be possible to implement this until the next stage of the BR systems development project but we would regard it as a high priority.

However, analysis by BR staff suggest the problem may not be as bad as first thought as many of the changes, identified through survey feedback, are actually activated on the BR through updated administrative sources. However, it is recognised that the changes would be implemented more quickly if survey feedback were allowed.

1. *Profiling of Large Enterprises*. There are plans to increase number of profiled units among the very largest and most complex enterprises. The work will commence this year. At least 200 are being considered recognising that a modelling approach based on partial data will be used by some enterprises. We believe this work should be treated as a high priority.
2. *Development of the new BR System*. A detailed plan for the statistical and other improvements for the revised BR System should be developed as soon as possible. The plan should emphasize the most important quality improvements such as eliminating inactive units (overcoverage), supporting improved NACE coding, and adjusting for dependent survey feedback. Although it may be several years before these changes can be implemented, the research work could start in the near future on the underlying algorithms to support the system changes.
3. *Development of the new BR System*. Furthermore, the new BR System should support the creation of a BR specifically for statistical purposes (i.e., the Statistical BR). This should not be a separate physical register. Rather, it should be a virtual register that can be created from the BR using the information contained on it. An obvious example is to eliminate businesses that, although registered, are highly likely to be inactive when creating the Statistical BR.
4. *Accuracy of NACE Coding*. The level of error in NACE coding should be monitored on an ongoing basis through an independent coding study, possibly using data from the SBS. The results of these studies should be made available to users, especially internal users. A strategy for addressing the most important inaccuracies in the NACE codes should be developed. The Construction Consultancy industry was suggested as a high priority.
5. *Quality Indicators on Base Registers*. The "Quality Indicators on Base Registers" project has reached the proof of concept stage. Work will be limited without additional funding. The results are important to the registers as they indicate whether their quality processes are having the desired result or not. They are also useful to the users of the register. Additional funding for this project would be a worthwhile investment (see Section 4.7.2 for a number of related recommendations).

Exhibit 10. Business Register (BR), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	Specification error	58	58	○	○	◒	◒	◒	◓	M
	Frame overcoverage	58	58	○	○	◒	○	○	○	H
	Frame undercoverage	47	52	○	◓	○	○	○	○	H
	Frame duplication	57	57	○	○	◒	◒	○	◓	L
	Non-response /Missing data	47	47	○	○	○	○	○	◓	L
	Measurement /Content	55	55	○	○	◒	○	○	◓	H
	Total score	53,8	55,0							

Scores					Levels of Risk			Changes from round 5	
●	◓	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.9 TOTAL POPULATION REGISTER (TPR)

CONTEXT

The TPR's ratings are unchanged from last year's ratings in most areas. However, this result belies a fairly robust program of research in the areas of Overcoverage and Missing Data. This work was adequate for maintaining Very Good ratings for the former and Good ratings for the latter error sources.

We should note that the criteria related to "planning towards risk mitigation" and "effectiveness of risk mitigation" for the error source "Frame Error: Duplication" that were designated as "N/A" (not applicable) in Round 5, was reactivated in Round 6. The change was made retroactively to Round 5 to reduce the effects of the change on the change in ratings between Rounds 5 and 6.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Quality Evaluations of Core Variables*. TPR staff have done quite a bit in developing quality indicators that reflect the extent of missing data for key register variables. They show for example that missing data rates of country of origin and citizenship is increasing over time. The causes of these increases are unknown and need to be investigated.
- *Overcoverage Research*. The TPR staff will work with the data collection department on a pilot survey using propensities of eligibility during the data collection. In addition, useful information, that can aid validating the modelled propensities, can be obtained from data collection. This and other work on overcoverage will be presented at the Nordic meeting among statisticians in August this year.
- *Communications with Tax Agency*. The Tax Agency has carried out an overcoverage investigation among registered persons without valid residence permits. Among 13 500 persons, 90 percent were deleted from the population register. The Agency has asked Statistics Sweden to provide information about the magnitude of the overcoverage problem in order for the Tax Agency to try to receive funding to repeat this study in the coming years.

Other Accomplishments

- *Internal User Group.* TPR staff has established an internal user group comprised of representatives from Population Unit, Microdata Unit, Labour Force Surveys, Forecast Institute, Regional Services and Planning, Public Finance and Microsimulations, Method Unit Individuals and Households and the Department for Data Collection from Individuals and Households. So far, they have met twice.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Overcoverage*. The TPR staff should continue to promote their modelling approach for identifying likely nonresidents on the TPR. It is important for the TPR to provide information to users regarding how to use the overcoverage propensities for uses such as nonresponse mitigation, field collection prioritisation and weighting adjustments.
2. *Evaluation of Overcoverage Model*. Although the validity of the overcoverage model was studied in few years ago, the study did not evaluate the model classification error (false

positive and false negative error) rates. It would be fruitful to do this in the near future because that would better inform the uses of the model. In addition, such an evaluation will be necessary if the method is to be published in a referred journal, as it should be.

3. *Impact of Overcoverage on Survey Estimates*. We encourage the TPR staff to work with survey areas that use the TPR to study the impact of overcoverage and missing data on the survey estimates. This would allow the TPR staff to be rated more highly on the sixth criteria – Effectiveness of Risk Mitigation Efforts. We understand that some discussions are being held with the Data Collection Department who may make use of the overcoverage indicator in the prioritization of nonresponse followup. It is important for the TPR staff to pursue these investigations and collaborate in demonstrating the effectiveness of the nonresponse follow up.

Other Areas for Consideration

- It has been about five years since the quality declaration was translated into English. Non-Swedish users, including EU members, as well as the ASPIRE team would benefit from an English version of the quality declaration. We suggest this be given a high priority in the coming year since much as happened on the TPR in recent years that impact data quality.
- Continue working with the Tax Agency to identify nonresidents on the TPR.
- Communicate with the LCS regarding the new variable on the TPR that identifies registered persons that reside in the same household. The LCS also creates a household variable and it would be useful to compare the two variables in order to assess the quality of both.
- Investigate the possibility of modelling the propensity of undercovering or overcovering geographic subpopulations (municipalities) as well as demographic subpopulations.

Exhibit 11. Total Population Register (TPR), Ratings for Round 6

	Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compli-ance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	Specification error	52	52	○	○	○	○	○	◔	M
	Frame overcoverage	65	65	◒	◒	◒	◒	○	○	H
	Frame undercoverage	57	57	○	○	◒	◒	◔	◔	L
	Frame duplication	57	57	○	○	◒	◒	◔	◔	L
	Non-response /Missing data	60	60	○	○	◒	○	○	○	M
	Measurement /Content	58	55	○	○	◒	○	◔	◔	L
	Total score	59,0	58,7							

Scores					Levels of Risk			Changes from round 5	
●	◔	○	◒	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

3.2.10 QUARTERLY GROSS DOMESTIC PRODUCT (GDP(Q))

The quarterly GDP estimates are produced from a very large and complex set of inputs from Statistics Sweden and other external sources. For our review, as with previous rounds, we could only look at a small number of the data sources that provided the greatest risk to the accuracy of GDP covering both the production and expenditure side. These are shown in Exhibit 12.

Last year was the first time we reviewed the input data sources for the expenditure-based estimates. We identified significant areas of potential improvement in collaboration with the National Accounts (NA) staff. The current situation with these data sources is:

- *Household consumption*. Although the quarter-to-quarter data on turnover may appear reliable, there are problems with outdated annual benchmarks. The Household Budget Survey is potentially a very important source but the 2016 survey was cancelled because of significant nonresponse problems and other data quality issues. It is not clear when the next survey will be conducted.
- *Investments*. The lack of data for the second quarter is an issue. Forecasts are used instead. This is a volatile part of the accounts and there are large annual revisions when SBS data is introduced. This problem should reduce when the quarterly SBS is introduced in 2017.
- *Research and Development*. On quarterly basis this is model based using the trend of value added in the relevant industry as well as foreign trade of services in R&D. A survey on R&D is conducted every 2 years. It is an item where large revisions are expected and improvement of quarterly data is important. There has been no real progress since the last review.
- *Foreign Trade*. Companies in certain industries (for example, IT) are having increasing difficulty distinguishing between goods and services. This distinction is not easy and by having different surveys for goods and services, problems with the double reporting of transactions or missing transactions can occur. The work of the Large Enterprise Unit has helped to reduce this problem.

SELECTED ACCOMPLISHMENTS

Progress toward Prior Recommendations

- *Replacement IT System.* Work has continued on a replacement of the Swedish NA IT-system scheduled for implementation in 2019. There are also plans for the considerable transition work that needs to take place.
- *Quarterly SBS.* There are now definite plans to conduct a quarterly SBS for the largest enterprises to obtain estimates of intermediate consumption and to be supplemented by VAT data for the smaller enterprises. It will be introduced in 2017. This will also allow investments data to be collected four times per year. It should also reduce the size of the revisions when quarterly data is benchmarked to data from the annual accounts.
- *Sensitivity Analysis.* Work on the sensitivity analysis has been continuing (see Section 4.6). Although the initial focus is on annual GDP, there are implications for quarterly GDP. Some of results have already been used. For example, the results of the study of the beneficial impacts of double deflation have reinforced the importance of the quarterly SBS mentioned above in order to support estimates of intermediate consumption. The study has also shown the importance of considering exchange rate movements when examining indexes of services

imports and exports. Initial research has also taken place on automatic balancing of the annual national accounts taking account of assessments of the relative uncertainty of the estimates for the different components. It shows sufficient promise for the work to be continued. A number of papers have been prepared on this work and will be presented in relevant conferences.

- *Inventories.* Quarterly data on inventories in service industries are to be introduced.

Other Accomplishments

- *Harmonisation.* The harmonization of the industrial and services production indexes has been completed with the harmonized survey introduced in 2015. This has enabled compilation of a Production Value Index from the second quarter in 2016. This enables the capture of services and trade margins in the manufacturing industries as well as trade margins rather than total turnover from trade in the service industries. Adjustments for changes in inventories have also been made.
- *Merchanting.* With the support of the large enterprise unit, the quality of the company reporting of merchanting on quarterly basis has improved.
- *Balance of Payments.* A working group has been established between national accounts and balance of payments to investigate data source, conceptual and modelling differences between the two collections of the current account. Reconciliations of these differences should lead to more reliable data for both collections.
- *Deflation.* Prices indexes for deflation are available for four new areas of services production.

RECOMMENDATIONS

Recommendations for Coming Year

1. *Training*. There needs to be more formality in the training while making greater use of new technologies to deliver that training. Self-paced training courses supplemented by coaching/tutoring by NA staff may be one possibility. There are existing NA training packages in other countries that could form a base for what is done in Statistics Sweden. These courses may also be of interest to users (e.g. the Riksbank) and those areas providing data to the national accounts (e.g. Economic Statistics Department).
2. *Sensitivity Studies*. We strongly recommend the continuation of the sensitivity studies. The work done to date has highlighted a number of important changes to the producer price indexes indirectly benefitting the NA. Preliminary work on the automatic (or objective) balancing of the annual NA shows great promise that this will indeed be feasible. It is recommended that this work be continued rather than risk losing the knowledge gained in recent months.
3. *Research and Development Estimates*. The models for quarterly R & D expenditure should be reviewed given the significance of this relatively new item in the NA. These models may require an increase in the frequency of the data collection possibly through the SBS. Software development is possibly the area most in need of requiring a new approach.
4. *Consistency with Balance of Payments*. In theory, there should be consistency between the relevant parts of the NA and the Balance of Payments, e.g. the current account. We support the establishment of a Working Group to gain further understanding of the reasons for existing differences with the objective of reconciling as many as possible. It may not be

possible to reconcile all the differences but they should be explained to users through a reconciliation table.

Other Areas for Consideration

- Consider speeding up the preparations for the implementation of a new IT-system for the NA. Decisions need to be taken on how much of the compilations should be in the new harmonised system or in supporting systems. If too many aspects are pulled into the new system, it will take more time and resources to develop. Crucial decisions about this need to be taken during coming years.
- We have supported the development of standardized or objective principles and methods for balancing the quarterly GDP estimates whilst recognizing there will always be an element of human judgment involved in the balancing process. The work associated with the sensitivity analysis has shown that objective and automatic balancing appears feasible for the annual accounts. This important research work should continue.

Exhibit 12. Quarterly Gross Domestic Product (GDP(Q)), Ratings for Round 6

		Error Source	Average Score Round 5	Average Score Round 6	Knowledge of risks	Communi-cation	Available expertise	Compliance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control over error sources)	Production Side	**Input data, ISP**	60	62	○	○	◡	◡	◡	◠	H
		Input data, IIP	60	62	○	○	◡	◡	◡	◠	H
		Input data, Merchanting (including royalites, licensing, R&D)	50	50	◠	○	○	○	○	◠	H
	Expenditures Side	**Input data, Turnover**	55	55	○	○	◡	◡	○	◠	H
		Input data, Government	57	57	○	○	◡	○	○	◠	M
		Input data, Investments	52	52	○	○	◡	○	○	◠	H
		Input data, Inventories	53	55	○	○	◡	◡	○	◠	H
		Input data, Net Exports in Goods and Services	52	53	○	○	◡	◡	○	◠	H
		Compilation (modelling)	47	47	○	○	○	○	◠	◠	H
		Compilation (data processing)	52	50	○	○	○	○	○	◠	H
		Deflation	55	60	◡	○	◡	◡	○	◠	H
		Balancing	50	52	○	○	◡	○	○	◠	H
		Revision	55	57	◡	○	◡	○	○	◠	M
		Total score	**52,9**	**54,0**							

Scores					Levels of Risk			Changes from round 5	
●	◠	○	◡	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

4 CROSSCUTTING ISSUES AND RECOMMENDATIONS

4.1 STATISTICAL COHERENCE WORK

One of the crosscutting recommendations in Round 4, reiterated in greater detail in Round 5, was to seek greater coherence across Statistics Sweden's statistical products. It was noted that Coherence could be regarded as a comprehensive indicator of Accuracy. Indeed, a discrepancy in the estimates of the same population quantities produced by different systems is an indicator of the magnitude of the systematic errors inherent in the systems. As an example, the "statistical discrepancy" in the estimates of GDP(P) and GDP(E) exists because of errors in the processes that generate the two GDP estimates. Reducing the statistical discrepancy is best achieved by reducing the errors in one or both of the GDP estimates. This is one of the primary goals of ASPIRE for the GDP.

In this round, the ASPIRE team met with several methodologists and management staff to continue discussions of the issues surrounding Coherence of statistical products at Statistics Sweden. Among the issues discussed in this meeting were:

- Definitions and concepts of Coherence including the concepts of absolute, trend and change coherence[1] that can be associated with a data series.
- Metrics for measuring absolute, trend and change Coherence for two or more data series.
- The distinctions between Coherence and Comparability and the applications of these dimensions to statistical series published by Statistics Sweden as well as comparisons with data series for other countries.
- The primary nonsampling error risks to Coherence, in particular, Specification Error: i.e., the difference in two data series due to inconsistencies in the concepts being measured; for example, the total employment figures produced by the LFS versus those obtained in the SBS.
- Methods for removing Specification Error from two data series in order to measure the incoherence due to measurement errors and other nonsampling errors.
- The responsibilities of a national statistical office to explain to users why two data series lack Coherence and how to interpret departures from Coherence. For example, reconciliation tables might be used for this purpose especially when variable specifications differ somewhat between the two data series.

RECOMMENDATIONS

1. *Sharing Knowledge.* Methodologists and analysts should continue to meet (at least quarterly) to identify, document and address the key issues concerning Coherence at Statistics Sweden. This group should reach out to its counterparts in other Nordic countries who are also deliberating these issues for their own countries.

[1] Absolute coherence is coherence in the level of estimates, trend coherence is coherence in the general trends suggested by the time series of estimates and change coherence is the coherence in the estimates of change between the same two points in time from two or more data systems.

2. *Best Practices, Metrics and Standards.* Once the issues have been sufficiently vetted, it would be useful to have a joint meeting of these groups with the objectives of (a) defining terms and best practices for addressing incoherent data series or disconcordant components of data series; (b) developing methods, *indicators* and other metrics to gauge degrees of Coherence, (c) formulating solutions to the issues of Coherence, and (d) agreeing on standards for communicating with users regarding Coherence.
3. *Reconciliation Table.* Work should commence on designing a reconciliation table for Employment which is one of the variables causing most concern to users.

4.2 INTEGRATION AND COORDINATION OF ECONOMIC STATISTICS

The coherence of economic statistics is of particular importance especially to the National Accounts. We have commented on integration of economic statistics in each of our reviews. In the last Report, we identified the following nine areas as being of critical importance for achieving economics statistics that are of high quality and coherent:

1. *Business Register.* A good quality Business Register (BR) is fundamental. When large businesses are complex and have significant activity in two or more industries, it is important to profile these businesses so that separate data can be obtained for these industry segments, perhaps modelled based on partial information.
2. *Common Business Framework (CBF).* A CBF should be derived from the BR to support (a) quarterly and (b) annual surveys. This requires the same units to be used across as many surveys as possible. Using a CBF helps coherence across the surveys. It ensures deficiencies in the framework (for example, inactive units) are dealt with in a consistent way. In addition, if desired, it makes it possible to rotate selected units out of the sample after an agreed period of time and minimise the chance of the same business being selected in multiple surveys. These two desirable attributes will not be possible for the very largest businesses.
3. *Standardized Classifications.* There should be common standards and classifications and the facilities available to support this. It is important that classifications of industry, geography, commodity and institutional sector be consistent or, if not, there is concordance among these classifications across collections. For example, National Accounts will want to utilise industry data from a range of Statistics Sweden collections and it is important that there is a concordance with the industry disaggregation that is used in supply-use tables for example. Other users may be interested in geographic data and a common geographic classification (especially if used outside Statistics Sweden as well) enables data from different collections to be brought together. Common classifications might be supported by coding frameworks that are shared by different areas and that improve consistency in the way the classifications are interpreted.
4. *Standardized Concepts.* The definitions of statistical concepts that are used in multiple collections should be centrally coordinated. In addition, there should be some agreement on the preferred approaches for collecting data on these concepts. For example, having easily accessible question ‘banks’ allows multiple collections to use the same questions. More generally, metadata should be able to be shared across data products.
5. *Standardized Methodology*. There needs to be a consistent approach to methodology such as the treatment of frame deficiencies (through CBF), treatment of nonresponse, changes in industry codes on the BR, survey feedback, etc. It is often surprising to observe the extent of differences due to different methodological approaches.
6. *Consolidation of Surveys*. The consolidation of surveys into fewer collections is one approach for achieving integration. Consolidation relieves respondent burden and saves costs.
7. *Input Data Warehouse (IDW).* An IDW is a central repository of data that facilitates sharing of data across the product areas and ensures consistent inputs for compilations.
8. *Revision Policy*. There should be a consistent revisions policy across collections and aligned with the National Accounts revision policy.

9. *Governance*. There will be issues of contention that need to be resolved from time to time. There needs to be governance arrangements that bring together the different stakeholders – it should operate at both the strategic and operational levels.

Statistics Sweden already has in place a large degree of integration of economic statistics and it was pleasing to see that further progress has been made since the last review in some areas (e.g. an agreement to increase the number of profiled businesses, a synchronized revision policy, quarterly SBS as a consolidated survey in 2017). There are further improvements that could be made but there is some uncertainty about next steps and some resistance to change. Some specific proposals are made below but we first outline our understanding of the current situation in Statistics Sweden using the same nine headings as above.

1. *Business Register*. Statistics Sweden has a good Business Register with excellent source data for maintaining the register. A SBS study has shown that the main deficiencies are inactive units on the BR (which are not known at the time of sample selection) and the lack of industry profiling for many large businesses. A new Register System is being developed and progress is well advanced but it does not incorporate the so-called statistical improvements in the first stage which would be need to be implemented in the later stages of the development. We also note the plans to significantly increase profiling in the coming year using a model-based approach for those large enterprises where the full set of accounts are not available. We also note that there has been some work on removing the number of inactive units on the BR.
2. *Common Business Framework*. Statistics Sweden uses a CBF for both its quarterly and annual collections; the main issue appears to be the different units that are used by different collections often for historical reasons. We note the suggestion to assess whether it is possible to use different units for sample selection on the one hand and publication purposes on the other and this is worth investigating.
3. *Standardized Classifications*. We know that Statistics Sweden has standard classifications but we were informed that they are not used consistently across products. We believe there is a need to improve consistency across collections. The main problem is with the use of different groupings across the statistical product areas. As an example, inconsistencies in commodity classifications between the CPI and HBS can cause difficulties for the NA when trying to use these data. There are also inconsistencies in the use of industry classifications between economic statistics and labour market statistics. It would be good to have one single corporate source for accessing standard classifications and their groupings at more aggregated levels.
4. *Standardized Concepts*. We are aware of the concerns about the different concepts of employment causing confusion to users. There may be other areas of concern about inconsistency of the standardized concepts. The last sentence on the previous point is also relevant for standard statistical concepts.
5. *Standardized Methodologies*. These methodological decisions are made on a collection-by-collection basis. There is scope for more collaboration among the methodologists to develop a consistent approach.
6. *Consolidated Surveys*. An annual consolidated survey of the business sector (SBS) has been in place for some time; a quarterly SBS, which would also be a consolidated survey, will be introduced in 2017.

7. *Input Data Warehouse*. This is now under consideration but would require some significant re-engineering of the design of surveys but can be implemented gradually.
8. *Revision Policy*. A common revisions policy has been adopted but it is not yet fully adopted for all products.
9. *Governance.* At the operational level, communication between the different stakeholders has improved considerably since we started the ASPIRE reviews. Things like the SLAs with the National Accounts, the BR User Group, etc. have aided this. There are now regular meetings with the department heads for Data Collection, Economic Statistics and National Accounts to discuss a range of issues related to the co-ordination of economic statistics, including strategic issues.

RECOMMENDATIONS

Our discussions with senior management suggested that there was some uncertainty about next steps as well as some resistance to change from the product areas. How can this be overcome? It is always more difficult if there are not strong pressures to make change. For example, if there are critical problems with economic statistics and the National Accounts in particular, it is easier to convince the product areas that there has to be change.

Without this driving force another approach has to be taken. One suggestion is to determine where Statistics Sweden would like to be in 5 years' time in terms of integrated economic statistics and then develop a strategic plan to make that happen. Perhaps an external expert could be engaged to help with this process. However, all the relevant products areas should also be involved. As well as providing greater ownership, such an approach would enable the discussions to benefit from their knowledge and experience. After agreement is reached on the 'future state', the next step would be to determine the most important projects to progress towards that ideal future state and to arrange appropriate project plans, resourcing, etc. The National Accounts should provide the overarching framework for determining what the future state should look like.

Our other recommendations are as follows. Some are also mentioned in the product reviews.

1. *Business Register*. When the current development of the Business Register (BR) is complete, work should commence on including the information within the BR to enable a Statistical Business Register to be extracted. For example, this would exclude businesses that are registered but inactive. More generally, the next version of the BR should be designed so that it can better manage the major quality concerns.
2. *Profiling of Large Enterprises.* There are plans in place to ensure that the largest and most complex enterprises are profiled into KAUs so that significant industry activities within the enterprise are identified. Where the full set of financial information is unavailable at the KAU level, it has been agreed that models utilising available information should be used to provide estimates at the KAU level. This work should be given a high priority. Furthermore, steps should be taken to ensure the KAUs are used uniformly across business surveys in Statistics Sweden through the application of common business frameworks.
3. *Harmonisation of Business Units.* As far as possible, harmonize the selection of business units across the business surveys especially for those surveys that contribute to the National Accounts. It has been suggested that different units might be used for selection than for publication and this would be worth considering. Before a change in unit definition is

adopted, the transition issues should be carefully considered because they could be disruptive for some collections.

4. *Standard Classifications.* There should be a study to assess where more should be done to implement standard classifications in a way that would support national accounts and other users.
5. *Methodological Decisions.* Ensure that key methodological decisions, such as adjustments for nonresponse, are performed in a consistent way. A current methodological decision of interest is whether to use dependent survey feedback (see Section 3.2.8) and an agreed position should be reached on this.
6. *Rationalisation of Collections.* Given the extensive uses of administrative data, Statistics Sweden still conducts a large number of data collections. There is also some duplication resulting in inconsistent or incoherent estimates. There is scope for rationalisation (that is, re-evaluating whether each collection is necessary), reducing respondent burden and possibly freeing resources for other activities. An example of inconsistent estimates is in the area of employment estimates and one early task might be to rationalise the multiple labour collections that are currently conducted, not always giving coherent results.

4.3 OUTSOURCING EXPERIMENT WITH THE LFS

Response rates for the LFS have continued their downward trend and are now about 58 percent. In addition, data collection costs have continued to increase. Statistics Sweden conducted a test of whether LFS data quality could be increased while data collection costs are reduced by outsourcing part of the data collection. Beginning in July 2015, Statistics Sweden contracted with EVRY (a privately owned call centre in Sweden) to interview approximately 20 percent of the LFS sample (about 5,350 interviews) each month. A number of comparisons of data quality and costs metrics between Statistics Sweden and EVRY were made to address the objectives of the study. Some of the key results are:

1. Although they are statistically significantly higher, EVRY response rates were not practically different from Statistics Sweden response rates since the differences were small.
2. EVRY invoiced costs were only a fraction of Statistics Sweden's real costs, according to a preliminary report, although differences in real costs could be substantially smaller.
3. As explained below, a number of quality issues and 'house' effects were identified for both EVRY and Statistics Sweden.

With regard to (1), although the EVRY response rates have not been dramatically higher than Statistics Sweden's, the results are nonetheless encouraging. We believe the prospects for continual improvement of both response rates and overall data quality are greater now that outsourcing has become part of the LFS operation.

With regard to (2), the ASPIRE team are incredulous of EVRY's costs given our experience with the cost of household surveys elsewhere. We believe their actual, steady state costs must be considerably higher than the invoiced costs for the study. It is possible, that, as Statistics Sweden becomes more and more reliant on EVRY to meet their production needs, data collection costs could sharply increase to reflect EVRY's true costs.

With regard to (3), in particular, we are quite concerned that, in relative terms, there are fewer unemployed (statistically significant), more employed and permanent employees (not statistically significant) and less temporary employees (statistically significant) in the EVRY sample while, at the same time, the proxy rate (i.e., the proportion of interviews that are conducted with a household informant rather than the sample member) is about three times higher than Statistics Sweden's rate. Prior studies have shown (see, for example, McGovern and Bushery, 1999) that partners and other proxy respondents are not always aware of the subject respondent's activities to look for work. This can result in respondents who should be classified as unemployed being misclassified as not in the labour force. Note also that obtaining more proxy response will elevate response rates while reducing callbacks and data collection costs.

RECOMMENDATIONS

The ASPIRE team has the following recommendations with regard to the EVRY results.

1. Statistics Sweden should continue to work with EVRY to
    - Reduce the proxy interview rate to the Statistics Sweden rate (which is currently about 2 percent)
    - Increase the EVRY response rate; a reasonable target is 63 percent which is about 5 points greater than the Statistics Sweden response rate.
    - Improve the balance indicators (Särndal and Lundqvist, 2014) and on the EVRY sample to no greater than ½ Statistics Sweden levels.

- Obtain costs estimates that accurately reflect EVRY's true costs given these quality enhancements.

2. Regarding the preliminary analysis of the EVRY results, some areas to further investigate include:
   - The cause of the greater number of unemployed (and higher number of employed) in the EVRY sample. What will be the effects on the LFS estimates when the EVRY sample is 20 percent and 50 percent of the full LFS sample?
   - Greater use of call monitoring for both Statistics Sweden and EVRY interviewers to identify quality differences between EVRY and Statistics Sweden interviewers and approaches for reducing interviewer effects for both call centres.

3. Finally, with regard to the idea of expanding the EVRY sample beyond 20 percent, we advocate a phased approach. First, Statistics Sweden should better understand the costs and data quality characteristics of the EVRY sample before expanding the sample. Then a phased expansion of about 10 percent additional sample households per year may be prudent to allow a smooth transition up to a maximum of 50 percent of the total LFS sample in three years. Each year, the expansion should be accompanied by a comprehensive review of data quality and costs and steps toward continual improvement in both facilities.

4.4 NONRESPONSE IN HOUSEHOLD SURVEYS

Nonresponse in household surveys continues to be a topic in which considerable staff time and importance is placed, especially for the Labour Force Survey (LFS). There is a good reason for this as declining nonresponse rates increases the risk of nonresponse bias in the estimates. Substantial resources have been allocated to the nonresponse problem over the last 5 years. Knowledge of the problem has increased through literature review and nonresponse studies and work was initiated on improved communication with respondents to help motivate response. During 2015, the use of mixed mode data collection, typically web and telephone, to improve response rates was implemented on several surveys with some success. Response rates and nonresponse bias continue to be important study domains for ongoing and periodic data collections.

To better focus effort on different aspects of the nonresponse problem, the crosscutting project of previous years was split into several projects – one focused on experimentation with mixed mode data collection, another on improving data collection processes, and others on responsive design and data base implementation for process data. The stronger delineation of responsibilities for individual projects with identified leadership, responsibilities and goals is intended to help ensure the success of the projects.

The direction this year focuses on planning for the offering of a web questionnaire to persons holding steady jobs in rotation groups 2-7 of the LFS, developing additional information on the sources of nonresponse across several surveys, developing more cost efficient treatment of the "non contacts", and identifying processes for better managing interviewer resources (using LFS as the target survey). With respect to the latter, work has begun on developing models to predict unsuccessful call outcomes using paradata. These applications to data collections are important to pursue. They focus primarily on data collection efficiency, a useful objective as call attempts, noncontacts, and nonresponse increase. The study in which the effort to use the data collection history of response/nonresponse patterns of rotation group 8 of the LFS suggests that some efficiencies in data collection are possible using such models. The ASPIRE team is pleased that this line of research follows previous ASPIRE recommendations. Nevertheless, much remains to be done as the nonresponse problem is not diminishing.

We noted previously that Statistics Sweden's administrative registers provide substantial auxiliary information to make effective use of calibration methods to reduce nonresponse bias in critical estimates. While some data products acknowledge this and have taken an active role in using both the data and the methods, others have been more passive. It is important that data products that successfully implement such methods and quantify their reduction in nonresponse bias in some estimates continue their research and share their results with other data products. All data products associated with household surveys should be actively conducting data product-specific research to reduce nonresponse bias through statistical modelling.

During Round 5 of ASPIRE, the ASPIRE team was told of a new management structure within the data collection department. The management structure was intended to provide greater supervision for interviewers and, consequently, more efficient data collection processes and better quality data. An assessment of the effectiveness of the organizational changes as they relate to the intended goals of the reorganization has not been undertaken, and indeed, may be difficult to achieve. However, understanding the impact of the changes will be useful to developing the next steps to be considered from the data collection point of view in the mitigation of nonresponse and nonresponse bias in household surveys. Furthermore, the EVRY experiment has shown, according to a preliminary

report, that the cost structure of Statistics Sweden is high so it is important to be able to demonstrate that the new management structure is cost effective especially in terms of its impact on quality.

RECOMMENDATIONS

The program on nonresponse in household surveys has progressed, but a number of observations made previously remain and ought to be addressed:

1. *Mixed Mode Data Collection.* Continue research to fully understand the effects of the mixed mode data collection on the quality of the respondent-reported data, especially for important subgroups. Quantifying potential mode effects in surveys that offer both web (or self-administered) questionnaires and interviewer-assisted questionnaires is important for understanding the quality of survey data.
2. *Better Understanding of Noncontacts.* Develop protocols to reduce the number of noncontacts in a survey and develop a deeper understanding of who the noncontacts are and how nonresponse bias can be mitigated.
3. *Call Management System.* Continue to work toward improving the call management system to reduce noncontact and nonresponse rates.
4. *Web Questionnaires in First Wave.* Investigate how best to offer a web questionnaire in the initial wave of a panel survey and its effect on response rates, nonresponse bias, and data quality in general.
5. *Optimising Calls.* Continue to research the modelling of the probability of an unsuccessful call outcome using paradata and register data. In the LFS context, more research on the use of seven waves of contact history in a panel to predict the unsuccessful outcome in the eighth interview should be carried out. Additional model development using six waves of a contact history should be considered and evaluated. More generally, household surveys need to take advantage of paradata and register data to bring additional efficiency to their data collection operations.
6. *Multiple Telephone Numbers.* Continue research on the problem of multiple telephone numbers per household. An approach to develop procedures to better identify a productive phone number is needed.
7. *Reducing Nonresponse Bias at the Estimation Stage.* More emphasis should be placed in all household surveys to quantify and characterize nonresponse bias for critical survey variables. In this regard, bias-mitigation modelling and weighting research is extremely important since it is the last opportunity to reduce nonresponse bias after all efforts to increase response rates have been tried.
8. *External Review of Data Collection Department.* As discussed in detail in Section 5.1, we believe an external expert who has experience at running an efficient and effective interviewing operation should review the processes of the Data Collection Department for Individuals and Households. This review should be coordinated with current plans to internally review data collection operations later this year.

4.5 ROLE OF METHODOLOGISTS AT STATISTICS SWEDEN

The ASPIRE process affords the external consultants with many opportunities to connect with the methodologists who work with the various products in the ASPIRE review. In addition, we have had a number of group sessions and discussions with methodologists in Stockholm and Örebro since 2011. We observe that when methodologists are involved in the planning and implementation of product improvements as well as their evaluations, the improvement activities tend to be more successful and there is more attention paid to assessing the effectiveness of the improvements. However, too often, we observe that methodologists are not involved in the improvement activities that could benefit from their attention, nor are their opinions always sought when new improvement projects are being plan. We often hear that when methodologists offer ideas for low cost but potentially high impact improvements, a frequent response is "we don't have money for that". This response can be quite discouraging and we sense that methodologists generally have some frustration with the current state of affairs.

RECOMMENDATIONS

1. Generally, there should be greater involvement of the methodologists in the planning, implementation and evaluation of improvement projects. We recommend that, during the 2016 planning activities, some attention be given to how methodologists can have greater involvement in these activities. In addition, we suggest that the effectiveness of efforts to utilise methodologists more extensively be objectively measured and evaluated.
2. One area where methodologists can be particularly useful is to suggest approaches for evaluating the effectiveness of quality risk mitigation activities. This is an area where methodologists can become the local experts, assisting not only in planning and evaluation activities, but also in the training of product staff in the many techniques for demonstrating risk mitigation effectiveness.
3. Finally, most methodologists are trained experts in total survey quality – a term that encompasses both user and producer dimensions of quality. As such, they can be expected to contribute to the producer-user interactions; for example, by suggesting topics for meetings of the User Councils and for soliciting input from users regarding the various trade-offs among quality dimensions. Some examples of such trade-offs, include timeliness versus accuracy; the costs of accessibility and clarity of data files; comparability of time series versus error reduction by incorporating improved methodologies, and so on.

4.6 SENSITIVITY ANALYSIS IN ECONOMIC STATISTICS

During recent years, many innovative and important projects have been carried out under the heading sensitivity analysis of the GDP. These projects have had the purpose of building knowledge about how sensitive estimates in the national account are to uncertainties in the input data.

The sensitivity studies on input data sources, especially producer prices, for the national accounts have already delivered substantial insights that should lead to improvements in the accuracy of both the national accounts and producer prices. It is important that the fresh knowledge is expanded to other areas and that the results are implemented in production mode.

Although the initial focus has been on annual GDP, there are implications for quarterly GDP and many other variables within the national accounts. In addition, the work has been extended to examine balancing and how to improve those processes.

RECOMMENDATIONS

1. *Resourcing.* Statistics Sweden should continue to provide resources for the sensitivity analysis in economic statistics. Although there is probably more to be done with the PPI, other input data sources should be investigated as well.
2. *International Collaboration.* The international exchange of the experiences from the sensitivity analysis project should be continued. Similar work may be going on in other countries and it should be very valuable to learn from others work in this area as well as getting their reactions to Statistics Sweden's work.
3. *External Visibility*. There have been a number of publications and presentations emanating from the sensitivity analysis work which increases the scientific stature of the individuals involved as well as Statistics Sweden. These professional activities should continue and be strongly encouraged and supported, as they will no doubt provide useful feedback on Statistics Sweden's work on sensitivity analysis. We also suggest that such accomplishments enjoy greater visibility within Statistics Sweden as exemplary of the kinds of professional stature activities in which all product areas should be engaged.

4.7 METHODS AND METRICS FOR EVALUATING THE EFFECIVENESS OF MITIGATION ACTIVITIES

4.7.1 ENHANCING EVALUATION

We first raised this issue in our 2015 Report. We noted that Statistics Sweden did not have a strong evaluation culture even though there are outstanding quantitative skills within Statistics Sweden. There is evaluation work undertaken but there is considerable scope for improvement particularly in documentation and follow-up action. In our 2015 Report, we made the following recommendations:

1. With each improvement activity, staff should consider how the effectiveness of the activity could be demonstrated. In particular:
    a. Measureable objectives should be clearly stated at the start of an activity and metrics aimed at verifying effectiveness of each activity should be identified.
    b. Following implementation, the metrics should be analysed to determine the degree to which each objective was met.
    c. The results of this analysis should be documented.
2. Statistics Sweden methodologists should develop training and offer assistance and guidance in the assessment of effectiveness of an intervention designed to improve data quality. To increase their capability to this, there would be merit in using external experts to conduct a workshop with Statistics Sweden's methodologists on methods for evaluating the effectiveness of quality improvements projects.

There have been some evaluation studies undertaken by some products since our last Review but these recommendations remain valid and are largely repeated below.

The Evaluation literature is immense. There is a vast range of evaluation methods that can be used. Some will be more relevant to Statistics Sweden than others will. Some are more expensive to implement than others are. The focus of an evaluation activity might be to identify the effectiveness of current mitigation strategies with a view of identifying what changes might be made. In the ASPIRE schema, this would improve the ratings for 'Knowledge' and potentially 'Communication'. The alternative focus might be on the evaluation of the effectiveness of new mitigation measures. In the ASPIRE schema, this would improve the ratings for 'Effectiveness'. This discussion is relevant to both applications of Evaluation methods but the main focus is on the effectiveness of quality improvement efforts.

How one should go about evaluation varies widely across products, error sources and data quality improvement objectives. Perhaps the easiest way to demonstrate effectiveness is "before and after" analysis. This approach compares a (proxy or direct) measure of the error prior to and after the intervention that is intended to improve the measure. However, this approach can be risky to the extent that uncontrolled factors could influence the post-intervention measures, thus confounding the before and after comparisons. One example we mentioned in our last Review was whether efforts to improve response rates are effective or not. This could be done by comparing response rates before or after. Analysis should examine at whether there is evidence that the mitigation strategy either has increased response rates or prevented them from declining further. However, it is often difficult to take account of other factors that might influence response rates irrespective of the mitigation actions.

A preferred approach is the controlled experiment. While this approach reduces confounding of comparisons of the experimental treatment with the control, experiments can be costly, potentially risky and difficult to conduct in a real-time production setting.

There is a need to identify the most relevant evaluation methods from Statistics Sweden together with their strengths and weaknesses. They could be considered part of an 'evaluation toolbox'. These might be documented in a Manual or Guidelines of Best Practice on Evaluation Methods. There should be emphasis on the development of metrics that support evaluation. Case studies should also be widely used in the Manual.

It is not sufficient to just have the Manual. There needs to be supplementary support and training arrangements. The methodologists are best placed to provide the technical support and the technical training should focus on them. One way of undertaking this training should be a two-day workshop facilitated by external experts familiar with evaluation in a national statistical office.

There is also a need to address the cultural issues. Once the Manual is available, there should be planned socialisation activities with the product areas. Key issues to address in this socialisation are the importance of evaluation (regard it as an investment rather than a cost), the need to plan and resource evaluation activities, the importance of engaging methodological support, and the need for documentation. Knowledge from evaluation studies should be shared with other product areas and perhaps the wider statistical world. There should be a structured repository of evaluation studies that supports meta-analysis of what can be learned by combining the knowledge of a range of evaluation studies.

RECOMMENDATIONS

1. *Evaluating Effectiveness.* We repeat our previous recommendation that staff should consider how the effectiveness of improvement activities could be demonstrated by:
    a. Clearly stating measureable objectives at the start of an activity as well as metrics aimed at verifying effectiveness of each activity should be identified.
    b. Following implementation, the metrics should be analysed to determine the degree to which each objective was met.
    c. Documenting the results of this analysis.
2. *Manual on Evaluation Methods.* A Manual on Evaluation Methods for Statistics Sweden should be developed. The Manual should include Case Studies to help illustrate the application of these methods.
3. *Developing Methodological Capability in Evaluation Studies.* As previously recommended, Statistics Sweden's methodologists should be trained in methods for assessing the effectiveness of interventions designed to improve data quality. To increase their capability to do this work, there would be merit in using external experts to conduct a workshop with Statistics Sweden's methodologists on methods for evaluating the effectiveness of quality improvements projects.
4. *Socialisation.* There should be socialization of these methods among the product areas with a strong emphasis on the importance of undertaking of evaluating studies and documenting and analysing the results.
5. *Documentation.* There should be a repository of the documented evaluation studies to assist with the sharing of gained knowledge and to support meta-analysis.

4.7.2 QUALITY INDICATORS FOR BASE REGISTERS

Complete and accurate administrative registers are critical to the production of official statistics in Sweden. Registers are used (a) to compute official statistics, (b) as controls in the final statistical adjustment stage and (b) as a sampling frame for the selection of sample units. Statistical applications using administrative registers rely on their accuracy and completeness. Previous rounds of ASPIRE encouraged the evaluation of the accuracy of critical stratification or auxiliary variables, such as, age, country of origin, gender, marital status, and region, that reside on three registers: Total Population Register, Business Register, and the Real Estate Register. During the past year, staff conducted a pre-study that identified a selection of quality indicators for registers, then implemented and reported on a number of such indicators. Staff focused on quality indicators that used variables related to coverage, linkage, classification and contact information. The pre-study also identified contact information variables for study.

The pre-study was conducted without a special resourcing allocation because of significant staff interest and the need to inform register users of the quality of variables on the register. Many data quality indicators can be defined and implemented, and a selection of such indicators was presented to the ASPIRE team. Further study of quality indicators is needed to determine the most informative indicators for the various classifications of variables on the register. This work holds promise of being useful and informative.

RECOMMENDATIONS

We recommend continuing support for this project and suggest the following activities:

1. Ensure continuing communication, cooperation and input from each register's staff,
2. Develop an understanding of quality indicators for registers used by other countries,
3. Study the information provided by the quality indicators and assessing their importance and usefulness (identifying high, medium, and low priority problems),
4. Identify boundaries/limits to determine when the indicator is not "in control", and
5. Select a subset of quality indicators for each type of variable on the register for monitoring purposes and for informing the register's staff and its user community.

5. SUMMARY AND RECOMMENDATIONS FOR FUTURE ROUNDS

5.1 SUMMARY

As stated in our previous reports, Statistics Sweden is a world-class organisation and in each ASPIRE round this fact is reinforced and verified. In most of the products we evaluated we saw improvements with very few deteriorations. Nevertheless, there have been a few areas where quality has deteriorated compared to Round 5 and these have been identified in this report. One product where overall quality has deteriorated according to the ASPIRE criteria is the LCS/SILC where a strong emphasis on operational matters has meant that quality issues have not received sufficient attention. We believe the balance should be adjusted.

Exhibits 2a and 2b shows the current ratings, prior year ratings, and the improvements by product. Exhibit 2c provides a summary of the ratings since Round 1 in the form of a bar chart. Justifications for the rating changes are summarized to some extent in the product reviews whereas details of each change are provide in rating change tables for each product that are available separately upon request from Heather Bergdahl.

With a maximum possible score of 100 percent (indicating perfect quality), the product scores in Exhibit 2a under the revised criteria ranged from 48.2 percent (for the LCS/SILC) to 64.6 percent (for the FTG) with an average rating of 57.9 percent. This does not include the GDP(Q) with a score of 54.0 overall. The average improvement in ratings for the products in Exhibit 2a was 0.8 percent this round compared to 0.6 percentage points in the last round. We note that the intervening period between Rounds 5 and 6 was only nine months rather than 12 months as in the first 4 rounds. If the period between the reviews was a full 12 months you would expect further quality improvements. However, in assigning ratings, we have attempted to take into account shorter period.

Following six rounds of ASPIRE, scores for Knowledge, Communication, Expertise, Compliance with Standards and Best Practices seem to have stabilised somewhat. Consequently, products are finding it increasingly difficult to increase their scores without implementing further evaluation studies to increase their knowledge of the risks as well as identifying risk mitigation strategies that result in real, demonstrative improvements. Such activities require a culture that supports the resourcing, conduct and analysis of evaluation studies and the resourcing of the key accepted recommendations that emanate from these studies. Notwithstanding the relatively small increase in average scores for this round, there has still been a substantial percentage point increase since ASPIRE started in 2011 (see Exhibit 2c) consequently of a substantial improvement in average quality for the products that have been continually reviewed.

The ASPIRE process has been modified and improved over the last six rounds and seemed to work quite well in the current round. Although it could be improved further, the revised criteria that were introduced last round seemed to capture the information Statistics Sweden seeks regarding risk mitigation effectiveness. We continue to be pleased that many products have taken up our recommendations from prior rounds to conduct which are highly innovative and informative studies.

For Round 7, we propose a similar approach to Round 6 although we will adjust the approach based on any feedback from Statistics Sweden. We will be able to take greater advantage of the web interface developed by Statistics Sweden (see Section 2.1) which facilitates product completions of checklists. This will provide even better checklists with less effort from the product areas. The system will also allow us to follow developments, ratings and comments for individual error

sources over time in a more efficient way. It is not always clear under which error source to which a particular risk should be allocated. As a result of our efforts on 'delineation of error sources' this will become clearer. Prior to Round 7, there should be consultations with the product areas to ensure that they agree with the proposed delineation.

In preparing for their Round 7 ASPIRE reviews, we hope staff will consider the product-specific recommendations we have made and make progress to the extent resources and time allow. In addition, as we proposed last year, we suggest greater consideration be given to demonstrating the effectiveness of the improvement efforts rather than simply relying on reasoning that an intervention that was designed to address some quality issue, actually achieved the desired effect.

In the discussion of the reviews for each of the products, we have identified the highest priority areas for improvement. In general, the highest priority should be given to error sources with high risk ratings combined with quality criteria with below average ratings. Some desired improvements are crosscutting or general in nature and we have discussed these in Section 4 of this report. These recommendations require consideration by top management rather than the individual product areas. Most will require some allocation of funding so there may need to be priority decisions made by top management to determine funding allocations. This year, as requested by management, we have identified what we consider the highest priority general recommendations.

Some of the highest priority improvements for the products might require additional funding although products should be encouraged to do as much as possible from existing funds. As previously suggested, it may be worth considering a pool of funding for quality improvements. Bids could be made against this pool and funds allocated to those proposals that are judged to be the highest priority based upon their impacts on quality, costs, and probabilities of succeeding.

Finally, we would like to thank Statistics Sweden for enabling us to work on this important and interesting project. In particular, we would like to thank Heather Bergdahl and Mikaela Järnbert for their tireless and professional support and the excellent co-operation from all the Statistics Sweden staff with whom we had contact.

5.2 HIGHEST PRIORITY RECOMMENDATIONS

We regard the following recommendations as highest priority among all the recommendations stated in the report. Priorities were assessed based on impact and viability with cost being an important aspect of viability.

1. *Interviewing Operations.* The EVRY experiment with the data collection for the LFS suggests, according to preliminary findings, that the cost structure for Statistics Sweden's interviewing operations is exceedingly high. This is not sustainable especially as there is no clear evidence that data quality is substantially better. For example, the LFS nonresponse rate for Statistics Sweden was slightly higher when compared with EVRY. In addition, violations of good interviewing practices appear to be higher among Statistics Sweden's interviewers. There have been past reviews of interviewing operations but they have not resolved the problems. There are plans to review of the Data Collection Department and we think it would benefit from the active participation of an external expert who has experience at running an efficient and effective interviewing operation. Aspects that might be considered are:
    a. Efficiency, effectiveness and quality of decentralized and centralized interviewing compared with "best in class" call centres.

b. Practical solutions that address the inability of current field structure to meet industry standards for costs, handling variable workloads, quality monitoring, call management and basic reports on process statistics.
c. Call center management structures that will address current issues with high costs and poor quality.
d. Physical design of the centralized facility in Örebro, particularly with respect to the use of open bays and cubicles where supervision is easier, as opposed to the current close office concept.

2. *EVRY Experiment*. Closely related to Recommendation 1 is the need for further evaluation of the EVRY experiment. The preliminary evaluation presented in Japec, et al (2016) is an excellent start, but left some questions unanswered. Although it suggested the quality of the EVRY interviewing at least matched that of Statistics Sweden interviewing, we are concerned that the EVRY estimates of unemployment is lower and employment is higher. The differences are significant and may be related to EVRY's much higher rate of interviews conducted with proxy respondents. These differences are concerning and warrant further analysis before expanding the EVRY sample.

3. *Mixed Mode*. Statistics Sweden has undertaken some important research into the use of mixed mode for interview surveys. This research should continue particularly on the use of the web mode in traditionally telephone surveys such as the LFS. However, Statistics Sweden should consider testing the use of PAPI (paper and pencil interviewing) as a further alternative to address the needs of households who do not have web access as well as those with access but who do not want to use the web. Studies in the United States have shown this may add as much as 10 percentage points to the response rate. Even more importantly, it may also lead to better sample representativity.

4. *Call Monitoring*. Monitoring of the telephone interviewing is an effective way of improving the quality of field operations and is used extensively by the best call centres. In particular, it identifies weaknesses in interviewing process which can allow training to focus on the highest priority problem areas. Statistics Sweden's call monitoring operations are not up to industry standards and need to be brought up to standards. The review of the Data Collection Department should also examine the telephone monitoring operations and suggest improvements and the external expert mentioned under Recommendation 1 should be able to give advice on this. We think there should be an increased use of call monitoring for centralized interviewers for all surveys with more frequent feedback. (For example, rather than once per year, as is currently given, feedback should be given about once per month on average with greater frequency to new or poor performing interviewers.) Monitoring of decentralized interviewers should also be initiated with feedback to them. In addition, new questions that will be introduced in ongoing surveys should be frequently monitored for quality issues.

5. *Alternative Sources of Household Budget Data*. The 2016 Household Budget Survey (HBS) was discontinued because of very high nonresponse rates. It is unlikely that a traditional HBS will be able to be conducted in the future. Alternative approaches should be investigated possibly in collaboration with other Nordic countries that are facing similar problems. The CPI and National Accounts are among the main users of HBS data but their prime interest is estimates of aggregate household consumption by expenditure category and do not require individual household data. Ideally, they would like this data annually as they revise their weights with this frequency. Perhaps it might be possible to get reasonable estimates using alternative data sources including scanner data from a range of sources. If the other main uses

of the HBS are to support micro-simulation analysis, it may be possible to rely on a quota sample to ensure representation rather than a probability based sample as at present.

6. *LCS/SILC*. Statistics Sweden should convene a technical advisory committee (TAC) to provide guidance on the redesign of the LCS/SILC/Children's Survey trilogy. This TAC should have representatives from subject matter areas, survey methodology, statistics and the user community. Some questions that the TAC might address include:
    a. What minimum data requirements on living conditions and quality of life will meet the needs of both national and the international users?
    b. How will the recommendations on quality of life measures prepared by external reviewer, Professor Emeritus Robert Erikson (http://www.regeringen.se/contentassets/dbb4c911287747b3943b4f61cf2b344f/far-vi-det-battre-om-matt-pa-livskvalitet-.pf) be addressed in the redesign?
    c. Can respondent burden and nonresponse be reduced by reducing the LCS and still maintain relevance?
    d. Should the Children's Survey continue as a supplement or should it be independent of the LCS?
    e. Noting the reductions made a few years ago, can the LCS/SILC be consolidated into one SILC survey with supplemental questions as needed?

7. *Business Register*. The BR is fundamental to high quality, integrated economic statistics. Progress on developing a new BR system has been advancing well. Whilst this work has been in progress, there have been a number of improvements made to the operations of the register. However, some important improvements still need to be made. The first improvement is to increase the number of large businesses that are profiled into KAUs to create purer industry statistics. This will improve the accuracy of a number of important economic statistics especially the National Accounts. To achieve this objective, the KAUs would have to be used consistently across the quarterly and annual collections. The second improvement is to determine and then incorporate statistical improvements in the next phase of the development of the BR system. Based on analysis undertaken by the SBS team, we believe the most important improvement is to create a Statistical BR that will enable BR users to eliminate inactive units from the scope of their collections.

8. *Co-ordination of Economic Statistics*. In addition to the BR, other important initiatives will improve co-ordination or integration of economic statistics. A high priority might be to start work on the rationalization of business surveys. Given its high use of register data, Statistics Sweden still seems to conduct a large number of collections. Rationalizing and possibly consolidating the number of collections will reduce the load on respondents and reduce the cost of collection activities, creating resource capacity that can be used for other purposes. The quarterly SBS integrates some surveys but perhaps others could be added. One initial area of investigation may be the various labour collections that can also give inconsistent results, confusing users. There are other forms of possible rationalization, including:
    a. Reducing the frequency of collections (e.g. monthly to quarterly) where there is not a strong need for monthly data and
    b. Ceasing collections where the uses do not justify the costs to Statistics Sweden and respondents.

9. *Evaluation*. Although Statistics Sweden has a strong methodological background, there is not a strong culture of evaluation. To some extent, this is because of a lack of awareness of

evaluation techniques many of which can be relatively low cost. Several steps can be taken to redress this, including:

a. Prepare a manual on evaluation techniques for Statistics Sweden.
b. Arrange for intensive training of methodology staff as they will be responsible for technical support on evaluation studies.
c. Conduct seminars and other activities that involve product area staffs and representatives from management to promote the importance of evaluation studies and to present case studies that demonstrate relevant evaluation methods. Create opportunities where the knowledge gained from evaluation studies can be shared among all Statistics Sweden staff. This might also include relevant studies from other statistical offices.

10. *Methodology*. We believe that some products underutilize staff with methodological skills partly because of the way methodologists charge their time to projects and partly because some products do not fully appreciate the benefits methodologists can bring to their work. Statistics Sweden should investigate whether current labour billing procedures are an issue and if so, how they might be modified to encourage greater use of methodologists.

11. *User Communication*. The effectiveness of user communication should be evaluated especially regarding user priorities. Do technology developments provide more opportunities for user input? Should there be more workshops of the most important internal and external users to consider issues of significance? Should there be a facilitator of these workshops independent of the product area to ensure new ideas are given proper consideration?

6. REFERENCES

Biemer, P. and Lyberg, L. (2003). *Introduction to Survey Quality*, John Wiley & Sons, NY

Biemer, P. P., & Caspar, R. A. (1994). Continuous quality improvement for survey operations: Some general principles and applications. *Journal of Official Statistics, 10*(3), 307–326.Biemer, P. and Trewin, D. (2012). "Development of Quality Indicators at Statistic Sweden," Internal Statistics Sweden report.

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

Biemer, P., Trewin, D., Bergdahl, H., and Japec, L. (2014). "A System for Managing the Quality of Official Statistics," *Journal of Official Statistics,* 30(3).

Biemer, P. and Trewin, D. (2014). "A Third Application of the ASPIRE Quality Evaluation System for Statistics Sweden," Internal Statistics Sweden report.

Biemer, P. and Trewin, D. (2015). "A Fourth Application of the ASPIRE Quality Evaluation System for Statistics Sweden," Internal Statistics Sweden report.

Biemer, P.,Trewin,D., Kasprzyk, D. and Hansson, J. (2015). "A Fifth Application of the ASPIRE Quality Evaluation System for Statistics Sweden," Internal Statistics Sweden report.

Bushery, J., McGovern, P. (1999) "Data Mining the CPS Reinterview: Digging Into Response Error," Internal U.S. Census Bureau Report, Washington, D.C.

Janssen,B. (2011). "Implementing Web Data Collection in the Labour Force Survey: An Experiment," downloaded from https://www.destatis.de/EN/AboutUs/Events/LFS/PapersP/C3_ImplementingWebDataCollection_Janssen.pdf?__blob=publicationFile

Japec, L., Lundqvist, P. and Westling, S. (2016). "An Evaluation of the LFS Data Collection, a Comparison of SCB/DIH and EVRY (External Provider)," Powerpoint Presentation presented to the ASPIRE Team on May 23, 2016.

Särndal, C. and Lundquist, P. (2014). "Accuracy in Estimating with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation," *Journal of Survey Statistics and Methodology*, 2, 361-387.

ANNEX 1 – CHECKLISTS FOR ACCURACY DIMENSION OF QUALITY

Accuracy Dimension Checklist. *For each applicable error source, indicate either compliance or noncompliance with an item in the checklist by marking "Yes" or "No," respectively. In order to achieve a higher rating for a criterion, all items for that higher rating must be checked. You may use the "Comments" field to provide comments you deem necessary to explain your response to an item.*

Knowledge of Risks	Check Box	Comments
1. Documentation exists that acknowledges this error source as a potential risk.	☐ Yes ☐ No **Fair**	
2. The documentation indicates that some work has been carried out to evaluate the effects of the error source on the key estimates from the survey.	☐ Yes ☐ No **Good**	
3. Reports exist that gauge the impact of the source of error on data quality using proxy measures (e.g., error rates, missing data rates, qualitative measures of error, etc.)	☐ Yes ☐ No **Good**	
4. 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 ☐ No **Very Good**	
5. Existing documentation on the error source is of high quality and explores the implications of errors on data analysis.	☐ Yes ☐ No **Excellent**	
6. There is an ongoing program of research to evaluate the components of the MSE that are relevant for this error source.	☐ Yes ☐ No **Excellent**	

Communication	Check Box	Comments
1. Data users have been informed of the risks from this error source to data quality through verbal communications, reports, websites and other formal and informal means.	☐ Yes ☐ No **Fair**	
2. Likewise, for data providers whose inputs pose some risk to data quality from this error source, there have been communications regarding these potential risks.	☐ Yes ☐ No **Fair**	
3. These communications to data users and providers have explained the risks in terms of the potential degradation to overall accuracy of the estimates.	☐ Yes ☐ No **Good**	
4. The potential impacts on users have been conveyed using sampling errors and/or proxy measures of bias and variance components. The measures have also been interpreted in a satisfactory way in order to facilitate the users' understanding of these risks.	☐ Yes ☐ No **Good**	
5. User documentation speaks clearly, comprehensively, and with appropriate detail on the size of the MSE components for the target audience.	☐ Yes ☐ No **Very Good**	
6. Provider communication is sufficiently detailed regarding the effects of errors including the quantification of impacts, and provides adequate information to enable the data providers to develop mitigation strategies that have real impacts on product quality.	☐ Yes ☐ No **Very Good**	
7. Based upon the communications they have received, users should be able to act appropriately regarding the risks from this error source when analysing the data.	☐ Yes ☐ No **Excellent**	
8. There is evidence that data providers have been intimately involved in the process of mitigating the risks of error from this error source resulting in a significant reduction in the risk from this error source. Communication has been ongoing, positive, productive, and produced important changes in the inputs resulting in a significant reduction in the risk from this error source.	☐ Yes ☐ No **Excellent**	

Available Expertise	Check Box	Comments
1. The product staff, or those areas servicing the product, include at least one person who is quite knowledgeable about methods for controlling or reducing the effects of the error source.	☐ Yes ☐ No **Fair**	
2. Expertise for this error source is adequate in most areas that are relevant for this collection (design, data collection, estimation, analysis, and data dissemination).	☐ Yes ☐ No **Good**	
3. At least some members of the product staff are adept at communicating risks for this error source to the both data users and providers clearly and concisely.	☐ Yes ☐ No **Good**	
4. The expertise could be made available if required and Communication is good across the internal groups that need to coordinate to reduce the risks from this error source.	☐ Yes ☐ No **Very Good**	
5. A good working relationship exists between the product staff and external groups who are key to reducing the error from this error source and their impact on SCB statistics.	☐ Yes ☐ No **Very Good**	
6. The key experts frequently participate in conferences, workshops, and other venues where approaches for minimizing the risks of error from this error source are pursued.	☐ Yes ☐ No **Excellent**	

Compliance with Standards and Best Practices	Check Box	Comments
1. Staff are aware of internal and external standards that apply as they pertain to this error source.	☐ Yes ☐ No **Fair**	
2. Key staff members are aware of best practices in the field that apply as they pertain to this error source.	☐ Yes ☐ No **Fair**	
3. Current activities for controlling or minimizing data quality risks from this error source comply with all appropriate standards.	☐ Yes ☐ No **Good**	
4. There are no serious violations of standards and best practices as they relate to this error source.	☐ Yes ☐ No **Very Good**	
5. The steps that have been taken to comply with standards and to minimize the risk from this error source may be regarded as state of the art and represent current best practices. Compliance with best practices is routinely monitored.	☐ Yes ☐ No **Excellent**	
6. Key staff actively read the literature as it pertains to this error source and some staff members are actively contributing to best practices in this area through conference presentations and publications.	☐ Yes ☐ No **Excellent**	

Planning Towards Error Mitigation	Check Box	Comments
1. Documented discussions are being held with appropriate staff with the objective to control or reduce the risks from this error source.	☐ Yes ☐ No **Fair**	
2. A written plan has been drafted that lays out a clear and effective strategy for mitigating the risks to data quality from this error source.	☐ Yes ☐ No **Fair**	
3. If applicable, a Service Level Agreement (or its equivalent) with the source data providers is being drafted that specifically targets this error source.	☐ Yes ☐ No **Fair**	
4. The written plan with measurable objectives has been approved by management. The plan adequately addresses the work required for mitigating the risks of poor data quality for this error source.	☐ Yes ☐ No **Good**	
5. If applicable, a Service Level Agreement (or its equivalent) with the source data providers has also been approved by management that specifically targets this error source.	☐ Yes ☐ No **Good**	
6. Appropriate resources have been allocated and Progress toward achieving the goals of the risk mitigation plan is regularly reviewed and compliance with the plan is appropriately monitored.	☐ Yes ☐ No **Very Good**	
7. Considerable progress has been made and the plan and SLA (if applicable) are updated appropriately as work progresses and new knowledge is gained regarding the error source.	☐ Yes ☐ No **Very Good**	
8. Mitigation plans have been fully implemented or well underway. Information has been provided to users/providers regarding progress toward risk mitigation.	☐ Yes ☐ No **Excellent**	
9. Accountability measures are in place to ensure compliance with the plans.	☐ Yes ☐ No **Excellent**	

Effectiveness of Mitigation Measures	Check Box	Comments
1. There have been some current efforts to mitigate the risk of error from this source.	☐ Yes ☐ No **Fair**	
2. As a result of these efforts, current proxy measures of the error from this source suggest that the error risks were mitigated to some extent. Further, these efforts have been well-documented.	☐ Yes ☐ No **Good**	
3. The work undertaken to reduce the error from this source has resulted in significant reductions in the error risks based upon both proxy error measures as well as some direct measures of the MSE components. These improvements efforts have been well-documented.	☐ Yes ☐ No **Very Good**	
4. Direct estimates of the MSE components associated with this error source indicate that substantial reductions of the error risks were the result of current mitigation efforts. These accuracy improvements have been documented, have been discussed with key users and are publically available.	☐ Yes ☐ No **Very Good**	
5. There is strong evidence based upon direct estimates of the MSE components that current mitigation efforts have substantially reduced the risks of error from this error source resulting in important improvements in accuracy. The evaluation has also considered the possibility that other errors sources may have been adversely affected by these mitigation efforts and no such unintended consequences were identified. These results have been thoroughly documented and are publically available.	☐ Yes ☐ No **Excellent**	
6. In addition, key users have confirmed that the mitigation measures have succeeded in providing them with statistics that are more accurate and fit for purpose.	☐ Yes ☐ No **Excellent**	