A FIFTH APPLICATION OF ASPIRE FOR STATISTICS SWEDEN

Paul Biemer, Dennis Trewin, Dan Kasprzyk and Jesper Hansson

October 15, 2015

TABLE OF CONTENTS

1 Executive Summary 3

2 Background and Introduction 6

2.1 Changes to ASPIRE in Round 5 7
2.2 Scope of the Review 9

3 Product Reviews 10

3.1 General Observations 10
3.2 ASPIRE Reviews 17

3.2.1 Labour Force Survey 17
3.2.2 Living Conditions Survey 20
3.2.3 Consumer Price Index 22
3.2.4 Producer Price Index and Services Producer Price Index 24
3.2.5 Annual Municipal Accounts 26
3.2.6 Foreign Trade of Goods 28
3.2.7 Structural Business Statistics 30
3.2.8 Business Register 32
3.2.9 Total Population Register 35
3.2.10 Quarterly Gross Domestic Product 36

3.3 The Household Budget Survey Review 39
3.4 ASPIRE Checklists and Quality Declarations 41

4 Cross-cutting Issues and Recommendations 42

4.1 Integration of Economic Statistics 42
4.2 Nonresponse in Household Surveys 46
4.3 Telephone Call Monitoring 49
4.4 Funding for Research and Development 51
4.5 Evaluation of Effectiveness of Mitigation Measures 52

5 Summary and Recommendations for Future Rounds 54

6 References 56

Annex 1 Checklists for Accuracy Dimension of Quality 57

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. In response, Statistics Sweden collaborated with two consultants (Paul Biemer and Dennis Trewin) to develop a quality evaluation approach that is referred to as A System for Product Improvement Review and Evaluation or ASPIRE (see Biemer and Trewin, 2013 and Biemer, Trewin, Japec and Bergdahl, 2014).

This report summarizes the results from Round 5 of ASPIRE which was conducted in August/September 2015. It covers products that have been previously reviewed, namely the 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) and the quarterly National Accounts, production approach (GDP(Q)). In this round, we did not review the annual GDP, as we have in previous rounds, but we did look at the quarterly GDP from the expenditure perspective. There were two new products that were reviewed together as one product: the Producer Price Index and the Services Producer Price Indexes (referred to as PPI_SPPI).

Although not part of the formal review process, we did discuss the methodology for the Household Budget Survey (HBS) as well as plans for the 2016 survey at length. A summary of our findings with recommendations is included in Section 3.3.

For this round there were two additional external reviewers – Dan Kasprzyk and Jesper Hansson. As in the prior rounds, the evaluation for each product involved a self-assessment, reviews of relevant documentation, 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. However, in this round, we used six rather than five criteria, four of which were the same as in previous rounds and two criteria which essentially split one prior criterion into two. This will be discussed in more detail in Section 2.1. The use of quality criteria checklists greatly facilitated the application of the criteria and, we believe, provided 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.

All but one of the eight previously reviewed survey and register products showed an improvement in ratings, although the increases were relatively small. FTG was the exception with a slight decrease of 0.4 percentage points; however, it still maintained its position as the highest rated product in our review. The biggest improvement was in BR (1.7 percentage points) which had undertaken some studies in association with the Tax Office which resulted in accuracy improvements as well as improved knowledge of the accuracy risks. Still, there is much to do to improve the quality of the BR and this is discussed elsewhere in this Report. The average improvement in ratings over these eight products was about 0.6 percentage points. The National Account product GDP(Q) also noted an improvement in rating by 2.0 percentage point this year. In 2011 when ASPIRE started, the mean rating across all products was 54. Now, at approximately 61, the mean rating has increased about 7 points since ASPIRE's inception.

In this round, one prior round criterion – Plans or Achievements towards Mitigation of Risks – was replaced by two criteria: Risk Mitigation Planning and Effectiveness of Mitigation Measures. This change was needed because relatively too much emphasis was being placed on planning and not enough on the implementation of these plans. Exhibit 2 provides a check list for how we evaluated against the effectiveness criterion. To reach a "Good" rating, the products have to provide some data (actual or proxy) on the effectiveness of the mitigation measures. This happened rather infrequently and consequently the scores against criteria are lower than for the other criteria. Mostly it was the product area's judgement that the mitigation measures had been effective but otherwise there was not strong evidence. This is a cross-cutting issue for Statistics Sweden which is discussed in Section 4.5. To provide some continuity, we provided a statistical bridge between the old and new criteria by calculating scores for Round 5 using both the old criteria and the new criteria.

With a maximum possible score of 100 percent (indicating perfect quality), the product scores including under the revised criteria ranged from 49.4 percent (for the LCS/SILC) to 64.1 percent (for the FTG) with an average rating of 56 percent. (Exhibits 3a and 3b in the report provide the scores for each product by error source.) As PPI_SPPI (with an average rating of 51.6 under these criteria) is a new product, we expect significant improvements next year as happened with other products after they were first introduced. Although not in this report, we prepared a 'Change Matrix' for each product that provides explanations for any changes in ratings. They are available from Heather Bergdahl on request.

Some of additional findings from the reviews include the following:

- Under the previous criteria (with risk mitigation and planning combined into one criterion), ratings increased for eight out of nine products participating in both this and the previous round, albeit some only marginally. The average increase for these products is about 0.8 of a percentage point.
- Under the revised criteria (i.e. treating risk mitigation effectiveness separately), ratings for all nine products participating in both Rounds 4 and 5 decreased substantially; the average decrease is more than 3 percentage points.
- The last row of Exhibits 3a and 3b shows the Round 4 to Round 5 changes in the overall quality ratings by product. Using the old criteria, the BR and the RS improved the most this year. This was the result of new and innovative quality improvement initiatives that they completed in 2015. (Details are provided in the discussions for those products in Section 3.)
- As in prior rounds, model/estimation has the lowest mean rating. This error source is medium to high risk for all survey products in Exhibit 3a and high risk for the GDP(Q) in Exhibit 3b.
- Also, as in the prior rounds, measurement error poses the highest risk to products; however, its mean quality rating continues to improve as a result of the increasing risk mitigation planning and implementation activities that have taken place over the past several rounds.
- Not surprisingly, the error source with the highest quality score, and by a wide margin, is sampling error. This was also true in the prior rounds.

In addition, the following general findings are notable:

- The nonresponse rates for household surveys continue to deteriorate despite the considerable effort and resources put into addressing this problem. Also, it has been a strong focus of research.
- In Round 2, we noted that the documentation of quality was greatly improved owing primarily to enhancement in the Quality Declaration (QD) documents. Progress since then has been disappointing with only a few QDs updated. Most of the updates are not in English and therefore not accessible to most users outside Sweden.
- Unfortunately, most quality evaluation studies continue to focus on error rates and indirect measures rather than direct error measures such as bias, validity, and reliability.

We were particularly pleased with the results of several studies that were undertaken in respect of our recommendations.

- The nonresponse project continues to make findings that provide real insights into the problem and how the cost-effectiveness of the household survey data collection can be improved. However, a mechanism needs to be established for making some hard decisions on changes in methods. At present, responsibility is spread across several work units and it is not clear who should make these decisions. In most national statistical organisations, the product areas have this responsibility consulting with experts and other stakeholders as necessary.
- The sensitivity studies on input data sources, especially producer prices, in the national accounts show real promise of providing insights that may lead to improvements in the accuracy of the national accounts. We are also impressed by the willingness of the PPI_SPPI staff to be involved in these studies.
- The innovative study undertaken by TPR to provide information on the overcoverage areas provides extremely useful information for the users of the TPR. However, work needs to be done to establish how to best utilise this information.
- On the other hand, we are concerned that Statistics Sweden is not in a position to conduct a good quality HBS in 2016. We suggest Statistics Sweden defer the collection by one year to allow time for the necessary research and development work to take place.

As in our previous reports, we laid out some general recommendations to improve quality that cut across all products. These topics are discussed in Section 4.

1. Integration of Economic Statistics
2. Nonresponse in Household Surveys
3. Telephone Call Monitoring
4. Funding for Research and Development
5. Evaluation of the Effectiveness of Mitigation Measures

2 BACKGROUND AND INTRODUCTION

This is the fifth round of ASPIRE. In Round 4, we shortened the reporting process because the background and technical details of ASPIRE have been well-documented in prior reports including 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. Nine of the ten products reviewed in Rounds 2, 3 and 4 were reviewed again for this round. Only the Annual GDP estimation process was not reviewed again because it was felt that the reviews of the quarterly GDP cover many of the same issues as the annual GDP.

In addition, to these nine products, three new products were reviewed – the expenditure approach to estimating quarterly GDP (GDP(E)), the Producer Price Index (PPI) and the Services Producer Price Index (SPPI). GDP(E) was combined with the quarterly GDP using the production approach (GDP (P)) and evaluated as a single product under the heading GDP(Q). These two products share many methodological issues and are produced within the same unit at Statistics Sweden. Likewise, the PPI and SPPI were also reviewed as one product (denoted PPI_SPPI) for the same reasons. The 10 products or product pairs that comprise the scope of our review are listed in Exhibit 1.

As in prior follow up rounds, one objective of Round 5 for previously reviewed products was to identify areas where clear improvements (or deteriorations) had been made since the previous evaluation. For new products, the review focused on establishing a base-line evaluation from which future progress can be measured. 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 cross-cutting issues.

For the most part, the ASPIRE process, error sources and evaluation criteria that was applied in this review were identical to the previous round and described in Biemer, et al (2014) as well as the reports from rounds 2 and 3 (Biemer and Trewin, 2013, 2014, respectively). However, some important changes were implemented in this round that are described in the next section.

Exhibit 1. Sources of Error Considered by Product or Product Pairs

Product	**Error Sources**
Survey Products 1. Labour Force Survey (LFS) 2. Living Conditions Survey (LCS/SILC) 3. Consumer Price Index (CPI) 4. Producer Price Index (PPI) and Services Producer Price Index (SPPI) 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. National Accounts (NA) Quarterly Gross Domestic Product (GDP (Q)) – both GDP(P) and GDP(E)	Input data error Compilation error Data Processing Error Model/Estimation Error Deflation/Reflation Error Balancing Error Revision Error

2.1 CHANGES TO ASPIRE IN ROUND 5

In addition to the three new products that were added for this round and the suspension of Annual GDP review, two additional changes were made in this round. First, two new reviewers were recruited to participate as ASPIRE reviewers; viz. Dan Kasprzyk, Senior Fellow at NORC at the University of Chicago, and Jesper Hansson, Director of Forecasting at the National Institute of Economic Research (NIER). The new reviewers have many years of experience in the production or use of official statistics: Dan Kasprzyk with the production of population statistics and Jesper Hansson with the use of economic statistics. Adding reviewers trained in ASPIRE is essential to Statistics Sweden's plans to continue and expand the scope of ASPIRE going forward. After the new reviewers were trained in the ASPIRE approach, each conducted a product review and co-authored sections of this report.

In addition, a major change was implemented this round for the criterion "Plans or Achievements towards the Mitigation of Risks." In prior rounds, this criterion has been challenging to apply and the ratings have been difficult to interpret. This is because the criterion combined two very different activities – (a) planning toward risk mitigation and (b) the achievement or effectiveness of the risk mitigation activities. A product could receive a high rating for (a) even though the plans either were never implemented or they were implemented but not particularly effective at reducing error risks. Thus, round to round ratings could be quite variable and interpreting the rating changes was challenging.

To address this, the criterion was split into two criteria – one addressing (a) and another addressing (b). Exhibit 2 shows the guidelines and implied ratings for the two criteria.

Exhibit 2. Guidelines and Implied Ratings for Criterion 5 – Planning towards Error Mitigation and Criterion 6 – Effectiveness of Risk Mitigation Measures

Implied Rating	5. Planning towards Error Mitigation	6. Effectiveness of Risk Mitigation Measures
Fair	• Documented discussions are being held with appropriate staff with the objective to control or reduce the risks from this error source. • 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. • If applicable, a Service Level Agreement (or its equivalent) with the source data providers is being drafted that specifically targets this error source.	• There have been some current efforts to mitigate the risk of error from this source.
Good	• 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. • 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.	• As a result of these efforts, current proxy measures of the error from this source suggest that the error risks have been mitigated to some extent.
Very Good	• 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. • 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.	• 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. • Direct estimates of the MSE components associated with this error source indicate that substantial reductions of the error have been the result of current mitigation efforts. These accuracy improvements have been documented and these documents are publically available.
Excellent	• Mitigation plans have been fully implemented or well underway. Information has been provided to users/providers regarding progress toward risk mitigation. • Accountability measures are in place to ensure compliance with the plans.	• 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.

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 (Round 5). The error sources associated with these products are shown to the right of these products. Only the PPI and the SPPI, treated as one product under the heading PPI_SPPI, are new for this round. Likewise, the middle panel shows the two registers included in this review and their error sources which were reviewed in all prior rounds.

The bottom panel shows the National Account (NA) products which are compilations of various other product inputs and data sources. The error sources associated with these NA products (which are discussed below) are shown on the right panel. As we previously noted, GDP (P) was evaluated in Rounds 2, 3 and 4 and those results are documented in Biemer and Trewin (2013, 2014 and 2015, respectively). GDP(E) is being reviewed for the first time in this round and is assumed to have the same error structure as GDP(P). Biemer and Trewin (2013 and 2014) and Biemer, et al (2014) provide a discussion of the error structure for GDP. GDP (E) and GDP (P) are evaluated together as one product under the heading GDP(Q).

In addition to reviewing these 10 products, the ASPIRE reviewers were also asked to comment on four special areas of Statistics Sweden's work: (a) the Nonresponse Project, (b) current plans for the next Household Budget Survey (HBS), (c) progress on our Round 4 recommendations on telephone call monitoring and (d) progress on our Round 4 recommendations regarding economic statistics. Thus, in this report, we include our thoughts and recommendations on these four areas as well as the 10 ASPIRE reviewed products.

The next section summarises the results of the quality evaluations for the 10 products. Section 4 summarises some general recommendations on cross-cutting methodological and other issues. Finally, Section 5 provides our recommendations on the future directions of ASPIRE and conclusions.

3 PRODUCT REVIEWS

Exhibit 3a provides the overall scores for the seven survey products and two registers by error source. To facilitate the exposition of the results, the error sources were consolidated into a single list which appears in first column of the table. The other columns of the table refer to the particular product being evaluated. For each product, the red bold figures correspond to "High Risk" error sources, black bold corresponds to "Medium Risk," and non-bold corresponds to "Low Risk" error sources for a product.

As discussed in Biemer and Trewin (2014), the interpretation of the error sources and criteria may vary between surveys and registers. For example, for a survey, it may be appropriate to consider measures such as bias and variance because the products of surveys are estimates. This is not the case for registers which do not, themselves, produce official estimates. The quality of register data is concerned with the quality of the data or variables maintained on the register. Thus, it may be more appropriate to consider the validity and reliability of the register data because these quality concepts are appropriate for variables. Here, validity refers to the correlation between a variable on the register and a hypothetic error-free version of that variable – i.e., the correlation between the observed value (y) and its corresponding true value (x). Reliability is a measure of the "signal to noise" ratio of a variable – i.e. the ratio of the variance of x to the variance of y – which is the inherent population variation of the variable, compared with the variation among the variable's observed values.

Likewise, Exhibit 3b provides the scores for GDP(Q). 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 the NA products is treated separately from the other nine products.

Finally, Exhibit 3c summarizes the total scores for all ten products over all five ASPIRE rounds in the form of a histogram. All three exhibits will be discussed in some detail in the next section.

3.1 GENERAL OBSERVATIONS

Before discussing each product's detailed ratings, some general observations regarding the results in Exhibits 3a, 3b and 3c as well as a few caveats can be stated.

First, the introduction of the new sixth criterion had a profound effect on the ratings for all products. Some reasons for this as well as some remedies are discuss later in this section. However, as a result, ratings for the current round using the new criterion should not be compared with ratings from the prior round using the old criteria. Rather, we report ratings under both the new (w/Effectiveness) and old (w/o Effectiveness) criteria to bridge the Round 5 and Round 4 ratings. In this regard, most of the comparisons between Round 5 and Round 4 or earlier rounds will focus on the ratings using the old criteria. However, the revised criteria will become the standard in future rounds.

Second, 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.

Third, 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.

Close inspection of scores in Exhibits 3a and 3b yield the following observations:

- Under the old criteria (with risk mitigation and planning combined), ratings increased for eight out of nine products participating in both this and the previous round, albeit some only marginally. The average increase is 0.6 points for surveys and registers and 2.0 points for the National Account products.
- Under the revised criteria (i.e. treating risk mitigation effectiveness separately), ratings for all eight products participating in both Rounds 4 and 5 decreased substantially; the average decrease is more than 3 percentage points.
- The last row of Exhibits 3a and 3b shows the Round 4 to Round 5 changes in the overall quality ratings by product. Using the previous years' criteria, the BR and the GDP(Q) improved the most this year. This was the result of new and innovative quality improvement initiatives that they completed in 2015. (Details are provided in the discussions for those products below.)
- As in prior rounds, model/estimation has the lowest mean rating. This error source is medium to high risk for all survey products in Exhibit 3a and high risk for GDP(Q) in Exhibit 3b.
- Also, as in the prior rounds, measurement error poses the highest risk to products; however, its mean quality rating continues to improve as a result of the increasing risk mitigation planning and mitigation activities that have taken place over the past several rounds.
- Not surprisingly, the error source with the highest quality score, and by a wide margin, is sampling error. This was also true in the prior rounds.

To increase their ratings, products should concentrate on areas that are classified as high risk (shown in red in the exhibits) and have low ratings, say ratings less than some threshold value. For example, ratings below the average rating for the revised criteria - i.e., 58.0 for products in Exhibit 3a and 52.9 for products in Exhibit 3b - could be targeted for improvement. However, these are not necessarily the highest priority areas for quality improvement. Other factors need to be taken into account such as the feasibility of the improvements, the relative priorities of products involved and the need to improve Accuracy compared with other dimensions of quality.

There are 13 cells in Exhibit 3a have high intrinsic risk and ratings below 58.0. They are:

- specification error - PPI_SPPI
- frame error undercoverage - BR
- nonresponse/missing data - LFS and LCS/SILC
- measurement/content error -LCS/SILC, SBS and BR
- data processing error - SBS
- model/estimation error - LCS/SILC, CPI, PPI_SPPI and SBS
- revision error - SBS

Likewise, 6 cells in Exhibit 3b have high risk as well as ratings below 52.9. They are:

- Production Input Data - Merchanting Service of Global Enterprises
- Expenditure Input Data - Net Exports in Goods
- Expenditure Input Data - Investment Survey
- Compilation Error - both Modelling and Data Processing
- Balancing Error

Exhibit 3c shows the overall ratings by product for the five evaluation rounds. Under the old criteria, all but one previously reviewed products have shown some improvement during the five rounds of ASPIRE. The exception is FTG but the decrease was only 0.4 points. Restricting the average to only the 9 products that have been reviewed every round since Round 2 (i.e., excluding PPI_SPPI), the cumulative improvement is approximately 5 points. The product with the largest improvement over the last 4 rounds is the LCS/SILC with more than 10 points improvement.

It is somewhat disappointing to observe that the magnitude of the average increase for the current round is quite small in Rounds 4 and 5 compared to earlier rounds, as can be seen from the last set of bars in Exhibit 3c. It is important to note that the previous rounds of ASPIRE were conducted at 12 month intervals whereas, for the current round, only 9 months elapsed since the prior review. Thus, the time available to implement improvements was somewhat less.

Some additional possible explanations for the small average increase in ratings were also noted in our Round 4 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 much greater commitment of resources, personnel and innovative thinking.

In addition, we note that the ratings for the criterion, Available Expertise, have declined in general because of the reduced availability of research and development staff. It appears that these staff have been diverted to operational work as a result of the realignment of priorities and reductions in product budgets to raise support for these other priorities. This may have the effect of stifling progress on other criteria as well; particularly for Knowledge and Risk Planning/Mitigation.

A third possibility is that product staff, for reasons other than budgetary constraints, 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. For example, while a few recommendations from Round 4 were implemented, the vast majority of them were not.

In addition, maintaining a high rating (such as Very Good) for a particular error source may be quite difficult for products having a number of high risk error sources with lower ratings. While attending to improvements for these lower rated error sources, ratings for the higher rated sources could deteriorate if resources are not adequate to maintain the level of activity needed to maintain the high ratings. This should be expected because there are seldom enough resources to do all that is needed to achieve high quality in all high risk areas.

Finally, we caution against interpreting the results in Exhibit 3c as suggesting that *data quality* has been improved for all these products. Although that is the ultimate goal of ASPIRE, an improvement in ASPIRE ratings means that products have improved relative to the six ASPIRE criteria. As previously noted, 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. But actual error reduction may not be realized until these plans have been implemented. This is one of the reasons the sixth criterion for Risk Mitigation was introduced in this round.

Exhibit 3a. Product Error-Level, Overall Level, and Error Source-Level Ratings with Risk-Levels Highlighted and Comparisons to Round 4 Overall Ratings

Error Source/Product	LFS		LCS/ SILC		CPI		PPI_SPPI		RS		FTG		SBS		BR		TPR		Mean rating	Mean rating
	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness	w/o Effectiveness	with Effectiveness
Specification error	70	63	**56**	**52**	**72**	**68**	**42**	**40**	N/A	N/A	**60**	**57**	**60**	**57**	**64**	**58**	**56**	**53**	60	56
Frame error	56	52	**42**	**40**	**66**	**60**	**58**	**53**	60	60	58	57	**66**	**60**	57	57	68	64	59	56
overcoverage												62			**60**	**58**	**68**	**65**		
undercoverage												63			**46**	**47**	60	57		
duplication												65			65	65	70	70		
Nonresponse error /Missing data	**58**	**55**	**48**	**45**	56	52	**56**	**52**	**60**	**57**	**64**	**62**	**72**	**70**	50	47	**62**	**60**	58	56
Measurement error/Content	**70**	**63**	**54**	**50**	**68**	**65**	**50**	**47**	**62**	**57**	**68**	**63**	**54**	**53**	**58**	**55**	62	58	61	57
Data processing error	**62**	**57**	50	47	**74**	**67**	**64**	**58**	**64**	**60**	**72**	**65**	**60**	**55**	N/A	N/A	N/A	N/A	64	58
Sampling error	**80**	**78**	**62**	**60**	**70**	**65**	**66**	**63**	N/A	N/A	N/A	N/A	**86**	**85**	N/A	N/A	N/A	N/A	73	70
Model/estimation error	**66**	**63**	**54**	**52**	**52**	**48**	**50**	**47**	**42**	**40**	**80**	**73**	**52**	**48**	N/A	N/A	N/A	N/A	57	53
Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	62	58	**72**	**67**	**54**	**53**	N/A	N/A	N/A	N/A	63	59
Round 5 mean rating	66,1	61,9	52,4	49,4	66,4	61,8	55,2	51,6	58,4	55,0	68,9	64,1	61,4	58,7	56,5	54,5	63,2	60,7	61	58
Round 4 mean rating (re-rated if needed)	66,0	--	52,0	--	65,8	--	--	--	57,1	--	69,3	--	60,5	--	54,8	--	63,0	--	61	--
Change (improvement/deterioration)	0,1	--	0,4	--	0,6	--	--	--	1,3	--	-0,4	--	0,9	--	1,7	--	0,2	--	0,6	--

RED BOLD = HIGH RISK
BLACK BOLD = MEDIUM RISK
REGULAR FONT =LOW RISK
N/A= Not Applicable

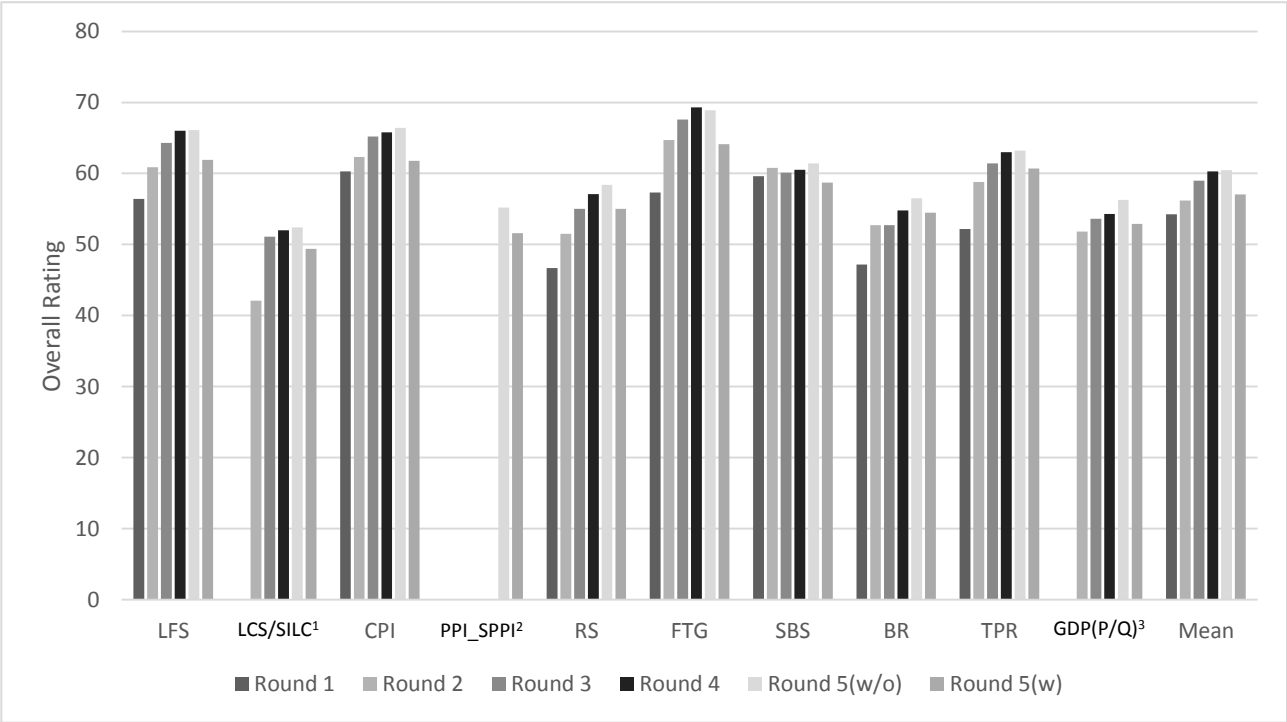
NOTE: Mean rating for Round 5 includes the PPI_SPPI which has below average ratings. The change in mean ratings (0.6 in the last row) pertains to the eight products evaluated in both Rounds 4 and 5.

Exhibit 3b. Product Error-Level, Overall Level, and Error Source-Level Rating with Risk-Levels Highlighted and Comparisons to Round 4 for the National Accounts

Error source /Product	GDP (Q)	
	w/o Effective-ness	with Effective-ness
Input data sources Production Side (Average)	**59**	**57**
Index of Service Production (ISP)	**64**	**60**
Index of Industrial Productions (IIP)	**64**	**60**
Merchanting Service of global enterprises	**50**	**50**
Input data sources Expenditure Side (Average)	**58**	**50**
Household Consumption -Turnover Statistics	**60**	**55**
Public Consumption -Municipalities	**60**	**57**
Investments -Investment Survey	**56**	**52**
Change in Inventories -Inventory survey	**58**	**53**
Net Exports -Exports in Goods -Exports in Services	**56**	**52**
Compilation error - modelling	**50**	**47**
Compilation error - data processing	**56**	**52**
Deflation error (including specification error)	**58**	**55**
Balancing Error	**52**	**50**
Revisions Error	**58**	**55**
Round 5 mean rating	56,3	52,9
Round 4 mean rating	54,3	--
Change (improvement/deterioration)	2,0	--

RED BOLD = HIGH RISK
BLACK BOLD = MEDIUM RISK
REGULAR FONT =LOW RISK
N/A= Not Applicable

Exhibit 3c. 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_SPPI was evaluated for the first time in Round 5.*
[3]*The GDP(P) component of GDP(Q) was evaluated in Rounds 1-4. 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.*

3.2 ASPIRE REVIEWS

In this section, we review the progress over the past 9 months for the eight products shown in Exhibit 1 that were also reviewed in Round 4 using the checklists that appears in Annex 1. The new product PPI_SPPI is also reviewed with the focus on establishing a reference point for future comparisons. Customized versions of the checklists were used for the two national accounts products to take into account their unique error structure (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)

Response rates for the LFS have continued their downward trend and are now about 62 percent. Addressing nonresponse continues to be a top priority for the LFS as well as for other demographic surveys at Statistics Sweden. There are a number of on-going projects aimed at either slowing the response rate decline or better understanding its causes. Since these are discussed in some detail in Section 4.2, they will not be repeated here. One that appears to be showing promise is the split-ballot experiment to compare LFS random samples collected by EVRY (a private company in Sweden) and by Statistics Sweden's own call centre. Another is the use of mixed-mode (web and telephone) which appears to increase response rates at later waves while not affecting the first wave response.

One area that we neglected to explore in prior rounds of ASPIRE is the use of dependent interviewing for collecting labour force status after the first interview. Evaluations of this approach are not mentioned in the Quality Declaration and we have some concerns that it could be biasing estimates of the levels and changes in labour status rates. Dependent interviewing is generally not used by other countries in their equivalent surveys. This is an area that should be explored in future research.

Next is a summary of other notable accomplishments for the LFS in this round.

SELECTED ACCOMPLISHMENTS

- Preparations are underway to conduct a cognitive study in the fall to follow-up issues identified in the reinterview study. In particular, the youth have problems correctly reporting their employment status and whether their job was permanent or temporary and this will be assessed in the cognitive study.
- Work was completed on an improved estimator of the correlation between two consecutive, annual, labour force estimates which is an important term in the variance of year-to-year change estimates.
- Work has progressed to investigate models used for nonresponse adjustments. Specifically, sensitivity analyses are being conducted to evaluate the effects of including or excluding certain auxiliary variables from the nonresponse adjustments.
- A feature has been added to the quality control systems used to automatically check the input data for the programs that seasonally adjust the labour force estimates.
- Plans are being developed to introduce mixed mode data collection in the LFS in 2016. It is anticipated that the data collection protocol will be similar to that used for the Party Preference Survey experiment.

RECOMMENDATIONS

1. The experiment with EVRY should provide valuable information regarding the causes of the high LFS nonresponse rates and could potentially have an enormous impact on the LFS and other demographic surveys at Statistics Sweden. It is important that the data from this experiment be analysed objectively using high quality statistical approaches. Thus, Statistics Sweden should develop a detailed analysis plan and arrange for staff who can quickly and independently analyse the data as they become available.
2. The sensitivity of the nonresponse adjustments to exclusion/inclusion of auxiliary variables should be very valuable in increasing knowledge of the residual bias. This work should be continued and documented.
3. 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 (i.e., the one tested in the Party Preference Survey) 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.
4. Further analysis of the paradata for contacts, noncontacts, interviews and refusals is needed to better understand the causes of nonresponse, particularly noncontacts. For example, it appears that no analysis has been conducted to evaluate nonresponse by time-in-sample (TIS) and to address questions such as:
    a. What are the nonresponse rates for TIS-*k*, for $k = 1,\ldots,8$ decomposed by component of nonresponse?
    b. What proportion of TIS-*j* noncontacts are contacted TIS-*k*, for $k>j$ and $j=1, 2,\ldots, 7$? What is known about why they were not contacted at TIS-*j*? The analysis should be repeated for refusals.
    c. How do these statistics compare for centralized vs. decentralized interviewers? Statistics Sweden vs. EVRY interviewers?
5. A project should be instituted to study the potential effects of dependent interviewing on estimates of the level and change in labour force status rates. As a starting point, LFS staff should conduct a literature review that summarizes the benefits and potential biases arising from this practice. On the basis of this review, recommendations could be formulated for future research.

Exhibit 4. Labour Force Survey (LFS), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	Available expertise	Compliance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy(control for error sources)	**Specification error**	70	70	63	◒	◒	◒	◒	◒	◓	**L**
	Frame error	58	56	52	◒	◒	◒	◓	◓	◓	**L**
	Non-response error	58	58	55	○	○	◒	◓	◒	◓	**H**
	Measurement error	70	70	63	◒	◒	◒	○	◒	◓	**H**
	Data processing error	62	62	57	○	○	◒	◒	◒	◓	**M**
	Sampling error	80	80	78	◒	◎	◒	◎	◒	◒	**M**
	Model/estimation error	64	66	63	○	○	◒	◒	◒	○	**M**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total score	66,0	66,1	61,9							

Scores						**Levels of Risk**			**Changes from round 4**		
N/A	●	◓	○	◒	◎	**H**	**M**	**L**			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.2 LIVING CONDITIONS SURVEY (LCS/SILC)

SELECTED ACCOMPLISHMENTS

- Research on a proposed redesign to satisfy Eurostat data availability requirements for the harmonised Survey of Income and Living Conditions (SILC) is under development and will be completed this year.
- Estimates were produced using an improved estimation scheme and a comparison showed that there was no change in the estimates when compared to the results using the old procedure but the effective sample size was smaller.
- Longitudinal weights were produced for Eurostat.
- The LCS/SILC staff appear willing to enter into an agreement with the Data Collection staff to develop statistics on the errors of measurement associated with field coding some questionnaire items in the LCS/SILC.

The LCS/SILC staff were engaged this past year in a number of operational concerns and products. Staff express the point of view that the statistical program is very complex. This is consistent with previous conversations concerning the surveys. Despite attempts to simplify the surveys, they retain their complexity, as there are three surveys conducted simultaneously – the LCS, the SILC, and the children's survey – with each survey having its own requirements in terms of data products and statistical estimation and statistical procedures. The European Union defines new requirements that are at times difficult for Statistics Sweden to comply with. For example, there is a requirement to provide Eurostat microdata by December of the data collection year. As reported last year, this requirement may affect the data collection period which now takes place throughout the calendar year. One alternative is to collect all the data in the first six months of the year; in this scenario Statistics Sweden must terminate data collection in the spring, resulting in seasonal effects in estimates of health, leisure activities, and other behaviours. Another example is the possibility that the European Union will require 6 waves of interviews for the longitudinal survey rather than the current four interviews. Changing requirements can bring opportunities for fostering a better understanding of the sources of error in the survey and the trade-offs among them.

RECOMMENDATIONS

1. Research plans to study the implications of the new data delivery schedule proposed by Eurostat must be formulated and implemented. Staff acknowledges the need to study possible seasonal effects in the estimates produced with a six month data collection period; they also acknowledge other data collection and estimation strategies ought to be reviewed to satisfy Eurostat's requirements. It is important this year to not only identify the research but also to identify and produce metrics that permit an understanding of the effects of changes to the current data collection and data delivery system. Documentation of these results, metrics and decisions is important for the LCS/SILC and should be incorporated into Quality Declaration.
2. The LCS/SILC has been simplified over time, but it still retains significant complexity. Additional research on appropriate methods and processes to integrate the three surveys into one survey is important. Results from this research have the potential to save resources and staff time as well as reduce complexity.
3. The sample size for the children's survey seems to be determined on an ad hoc basis. Precision requirements for the children's survey ought to be established based on the primary uses of the data.

4. 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

5. Two estimates of the Gini Coefficient exist for Sweden based on Statistics Sweden data, one published by Eurostat based on LCS/SILC and the other on Register data. This state of affairs is confusing for the data users. Understanding, resolving, and documenting the differences of these estimates is an important research task.

6. The LCS/SILC has high nonresponse rates as do other demographic surveys. We understand that discussions with the Process Department on ways to increase response rates and measure nonresponse bias have taken place and this is encouraging. It is important for the LCS/SILC staff to continue to engage in conversations and collaborate with both the Data Collection Department and the Process Department to identify strategies – in the spirit of adaptive design – for targeting priority sample cases for followup. In the same spirit, identifying and developing metrics that measure the results of these efforts is critical to understanding whether data quality is being improved. Work should also continue on the evaluation of residual nonresponse bias in the adjusted estimates, quantifying it and researching effective weighting approaches to reduce the bias in key LCS estimates.

Exhibit 5. Living Conditions (LCS/SILC), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	Available expertise	Compliance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy(control for error sources)	**Specification error**	56	56	52	○	○	◒	○	○	◓	**M**
	Frame error	42	42	40	◓	◓	◒	○	◓	◓	**M**
	Non-response error	48	48	45	○	◓	○	○	○	◓	**H**
	Measurement error	54	54	50	◓	◓	◎	○	○	◓	**H**
	Data processing error	50	50	47	○	◓	◒	○	○	◓	**L**
	Sampling error	62	62	60	◒	◒	◒	○	○	○	**M**
	Model/estimation error	52	54	52	○	◓	◒	○	○	◓	**H**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total Score	52,0	52,4	49,4							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◓	○	◒	◎	H	M	L			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.3 CONSUMER PRICE INDEX (CPI)

SELECTED ACCOMPLISHMENTS

- Sample Redesign. A sample redesign was undertaken leading to a more efficient two stage design to be introduced in 2016. Because of the greater clustering, the cost of data collection will be reduced. Further, because the intra-cluster correlation is quite small, this increase in clustering is expected to have little impact on sampling errors.
- Extended Use of Scanner and Internet data. As well as increasing the size of the sample in some important segments, scanner data can provide prices that include discounts which are otherwise difficult to collect. Scanner data use has been extended to include prices for alcohol and rail transport. As a consequence, 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. This enables the impact of quality adjustments to be assessed and should be an important macro editing tool. This tool might be useful for other price indexes, including the producer prices.
- Hand-held Computers. Updated tablet computers, similar to iPads, will be introduced in the coming months. Training on these computers is about to commence. Once there is more familiarity with the tablets, a range of opportunities to improve the CPI will be opened and some of these are discussed below.
- Interest Rates. Transaction prices, instead of list prices, are now used to get 'price' data for mortgage interest rates.

We believe the use of the upgraded tablet technology in the field has a number of potential benefits. The new technology will have a longer life than the previous technology as well as greater efficiency. Furthermore, it provides a number of possibilities to better monitor the work of the price collectors. We understand there are some restrictions on monitoring but, within these constraints, the possibilities need to be explored. For example, there are no quality checks in the field at present and it may be possible to imbed some capability for this in the new technology. Also, there is a camera in the tablet which could be used to photograph goods (with permission) or to capture bar codes.

More items are being purchased over the internet, some from outside Sweden. As a consequence more of an effort has gone into collecting internet prices. There are some difficulties in getting a frame of internet stores. There is also some discussion of whether items purchased from outside Sweden should be included although they are currently excluded.

One of the advantages of the new sample design is that it may create cost savings that can be devoted to quality work which has been restricted because of resource limitations.

In previous reviews we have commented on the Household Budget Survey (HBS) and the limitations on its use for the CPI. We made no CPI recommendation here as we have undertaken a separate review of the HBS (see Section 3.3). Among other things, we suggested that the HBS hold a workshop with CPI staff and other main users. In the case of the CPI, the objective would be to see how CPI requirements can be best satisfied.

We offer the following recommendations but note that the Swedish CPI continues to be of a very high standard especially when compared to those of other countries.

RECOMMENDATIONS

1. Redo the 1999 CPI error study. CPI methods and procedures have changed considerably since 1999 and therefore biases, variances, error sources and so on could be quite different now. Furthermore, a Total Survey Error approach should be taken for 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.

2. 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. Web scraping is in its infancy as a price collection tool. The methods are immature and price collection for some commodities (for example, air travel) can be quite complicated. It is recommended that some resources be devoted to the methodological issues associated with web scraping.

3. Research into methods for estimating quality adjustments should continue as this may be one of the most influential sources of error in the CPI. The Implicit Quality Index diagnostic tool should be invaluable in assessing the merits of the different methods.

4. There is a lot of dependency on the work of the price collectors and their work should be routinely monitored. More up to date tablet technology is being introduced to support data collection. The technology has the capability 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. We support the study to see what might be done using Google Analytics as well as other online tools for analysing price changes. In particular, we recommend that there be research into how technology might be used to improve quality. The technology has been introduced by other countries and there may be some benefits in sharing knowledge on this use of technology.

Exhibit 6. Consumer Price Index (CPI), Ratings for 2015

	Error source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	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)	Specification error	70	72	68	◡	○	◡	◡	○	○	**H**
	Frame error	66	66	60	◡	◡	○	◡	◡	◠	**M**
	Non-response error	56	56	52	○	◠	◡	◡	○	◠	**L**
	Measurement error	66	68	65	◡	○	◡	○	○	○	**H**
	Data processing error	76	74	67	◡	○	◡	◡	◡	◠	**H**
	Sampling error	68	70	65	◡	◡	◡	◡	○	◠	**H**
	Model/estimation error	52	52	48	○	○	○	◠	○	◠	**H**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total Score	65,8	66,4	61,8							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◠	○	◡	◎	**H**	**M**	**L**			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.4 PRODUCER PRICE INDEX AND SERVICES PRODUCER PRICE INDEX (PPI_SPPI)

These two products are new to the ASPIRE process this year. The major reason for including them, are that they are very important input to the National Accounts (NA) when calculating GDP in constant prices. We think that although they are two separate products, mostly for historical reasons, they are best reviewed together. The two products share many methodological issues and are produced within the same unit at Statistics Sweden. There is much less research and public debate about the measurement of producer prices (PPI) than of consumer prices (CPI), which we believe is unwarranted given the importance of the PPI for economic analysis using the NA.

SELECTED ACCOMPLISHMENTS

- The data collection system has been upgraded and now includes features producing automatic reports that are used in the macro-editing process.
- Two new product groups, SPIN 74 Professional, scientific and technical activities and SPIN 82 Office administrative, office support and other business support activities, is on a trial year in the SPPI this year. They will be included in the regular publications next year.
- Development of new SPPI and experimental collection of prices for four new product groups (licences, TV-broadcasting, etc.) are planned for 2016, for future inclusion in the SPPI.
- Participation in the project Sensitivity Analysis for GDP has generated increased knowledge, both at the Price Statistics Unit and at other units/departments at Statistics Sweden, about the importance of PPI and SPPI in the calculation of GDP growth in constant prices
- Cooperation with the Data Collection department on the contacts with enterprises started this autumn. It will create opportunities for improving the response rate, for instance by reminding enterprises earlier in the month. It will also free up resources at the Price Statistics Unit which could be used more productively.

We believe that the Swedish PPI and SPPI have a good, or very good, compliance with standards and best practices. Nevertheless, there is still room for improvements, and some ideas are suggested in the following recommendations. Compared to other products in the ASPIRE review, the average rating of the PPI_SPPI may seem a bit low (see Exhibit 7). We think that this is quite natural for new products that not have had the opportunity to respond to previous years' recommendations. It may, however, also reflect the fact that our knowledge of these new products is less than for other products that have been in the ASPIRE process previously. We fully expect the ratings for this product to dramatically improve in the next round.

RECOMMENDATIONS

1. Continue the gradual improvements of the data collection process. Develop quantitative measures that can be used to illustrate the results on accuracy of these improvements.
2. Conduct a benchmark study against other countries which are supposed to be at the frontier regarding best practice (e.g. Germany). Focus primarily on the non-sampling error sources (measurement, quality adjustments, coverage of products) where we think the scope for improvements are most likely.
3. Increase the knowledge of the latest 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.

4. Increase communication with core users of PPI and SPPI, the most important being the NA. Users need to be more involved in the production process, giving feedback on the deficiencies of the PPI_SPPI they are experiencing.

5. Develop a measure comparable to the Implicit Quality Index of the CPI for the producer price index. 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).

Exhibit 7. Producer Price Index and Services Producer Price Index (PPI_SPPI), Ratings for 2015

	Error source	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	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)	Specification error	42	40	◓	◓	○	○	○	◓	**H**
	Frame error	58	53	○	○	○	◒	○	◓	**M**
	Non-response error	56	52	○	○	◒	○	○	◓	**M**
	Measurement error	50	47	◓	◓	○	○	○	◓	**M**
	Data processing error	64	58	◒	○	◒	◒	○	◓	**H**
	Sampling error	66	63	◒	○	◒	◒	◒	○	**H**
	Model/estimation error	50	47	◓	◓	○	◒	○	◓	**H**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total Score	55,2	51,6							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◓	○	◒	◎	**H**	**M**	**L**			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.5 ANNUAL MUNICIPAL ACCOUNTS (RS)

SELECTED ACCOMPLISHMENTS

In the nine months since the Round 4 review, the Annual Municipal Accounts data collection has made a number of improvements to its data collection and editing process that should enhance data quality. These improvements have resulted in two higher ratings – data processing best practices (elevated to Good) and plans toward mitigation of risks for model/estimation error (elevated to Very Good). Some accomplishments of the RS in this round include the following.

- New Transmission Interface. The new transmission interface for respondents put into service last year is now used by all 290 municipalities. This system allows respondents to upload their completed Excel forms to Statistics Sweden's website rather than submitting them via email. The interface then provides respondents with immediate feedback regarding edit failures and warnings. Although such edit feedback was available with the prior method of transmission, it entailed a considerable response delay. Respondent feedback continues to be quite positive.
- Analysis of Editing Changes. The study conducted last year that compared key RS variables before and after editing has been further analysed. Based upon this study, the strengths and weaknesses of the current editing approach were identified. In the coming year, a two-year strategy for improving editing and other data processing activities will be drafted. In addition, the method of identifying the changes to the data as a result of editing will continue to be used and its results evaluated.
- Improved Revision Policy. The improved revision policy developed last year has been implemented. This policy publishes a summary of each respondent's responses on the RS website five times between April and August. During that period, respondents can correct their inputs before the final publication in August, after which there can be no further revisions. Besides improving the timeliness of revisions, this new policy can reduce burden for both respondents and RS staff. In the coming year, some indicators will be developed to gauge the degree to which data quality has improved as a result of the new policy.
- Timeliness of Reporting. As noted last year, only about 30 percent of the municipalities send their forms before the deadline. RS staff have been experimenting with ways to increase the proportion responding prior to the deadline by, for example, contacting the tardy municipalities, reminding them of the deadline and offering assistance in completing their forms. These interventions were not successful, however. Currently, respondent burden is believed to be the primary cause of respondent tardiness since it was estimated in 2011 that municipalities took an average of 3 weeks to complete the survey. In the coming year, some RS staff members plan to visit a few municipalities in order to study the primary sources of tardiness, including the response burden hypothesis. Hopefully this will lead to some improvements in the form that will increase respondent timeliness.

RECOMMENDATIONS

1. The review of data before and after editing appeared to be quite useful in identifying areas of improvement for the editing process. We encourage the RS staff to continue this work and follow through on changes to the editing system suggested by the before and after analysis. As a visual aid for identifying potential inefficiencies, redundancies and areas for improvement, the RS staff should consider developing a detailed flowchart of the editing process. An illustration of how process flowcharts (or maps) can be used can be found, for example, in Breyfogle (2003).

2. As noted in our reviews from prior rounds, more research is needed to understand the errors associated with the RS data and how these errors propagate through the NA to cause biases in the NA estimates. Although there has been considerable progress during the last year toward understanding the errors associated with data processing in the RS, there has not been much effort in quantifying the errors nor understanding how important users such as the NA are affected by them. For example, to address the problem of high, year-to-year volatility of the investment account data, the instructions for some of the items were clarified to reduce the risk of double-counting and erroneous inclusions. Thus, it seems that a fruitful area to explore is the sensitivity of the GDP estimates to these types of errors.

3. In Rounds 3 and 4, we recommended that the allocation keys used to disaggregate common costs to various sub-activities be evaluated. We noted that more than 80 percent of the municipalities allocate common costs to various activities using Statistics Sweden's automatic allocation key. The remaining municipalities allocate common costs according to their own model. To our knowledge, there has never been an evaluation of the Statistics Sweden allocation model. We again recommend that the RS staff mount an investigation of the accuracy of the common costs data, the inaccuracies in the cost allocation keys, and how these two sources of error may interact to generate important errors in the RS data. One approach for studying this allocation model is to apply the model for a sample of municipalities that use their own allocation method for common costs and then compare the two allocations results.

4. Also, as noted in prior rounds, there is the potential for important errors in RS for the disability care estimates. We noted that what a municipality reports on for these costs can directly influence the size of subsidy or fee municipalities receive. The RS should continue to monitor these estimates in the coming year and consider whether their current procedures for mitigating this risk can be improved.

Exhibit 8. Annual Municipal Accounts (RS), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	Available expertise	Compliance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (Control of error sources)	**Specification error**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Frame error	60	60	60	○	○	◡	◡	N/A	N/A	**L**
	Non-response error	60	60	57	○	○	◡	○	○	◠	**M**
	Measurement error	62	62	57	○	○	◡	◡	◡	◠	**M**
	Data processing error	62	64	60	◡	○	◡	◡	○	◠	**H**
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Model/estimation error	38	42	40	◠	◠	◡	◠	○	◠	**M**
	Revision error	62	62	58	○	○	◡	◡	○	◠	**L**
	Total score	57,1	58,4	55,0							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◠	○	◡	◎	H	M	L			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.6 FOREIGN TRADE OF GOODS (FTG)

SELECTED ACCOMPLISHMENTS

The FTG continues to be among the highest rated products in our review despite a slight dip in its ratings in this round (even when the new Mitigation Effectiveness criterion is not counted). For the past nine months, much effort has been devoted to IT-related issues which have absorbed both resources and attention to other possible quality improvements. The ratings show two improvements: Plans toward Mitigation Risks for Measurement Error increased to Very Good, primarily as a result of a Eurostat project that is being conducted that will yield information on the error risk. A report is due in December, 2015. In addition, Communication for Revision Error improved due to more elaboration on this error source in the Quality Declaration and continued improvements in communication with NA staff. Other improvements which were sufficient to sustain ratings in other areas include:

- The conversion to IDEP.web which replaced IDEP.net was completed this year. This system greatly simplifies the process respondents use to send their data to Statistics Sweden.
- The staff continues to study the possibility of using SIMSTAT data for estimating Intrastat trade. A Eurostat decision regarding how SIMSTAT data should be used is due next year and could have important implications for the FTG.
- We are particularly pleased that the Quality Declaration was revised and translated into English. This greatly facilitates communication between the FTG staff and the ASPIRE reviewers.

RECOMMENDATIONS

1. Estimation below the cut-off. Beginning in 2015, the sampling cut-off for Intrastat was doubled to SEK 9M for imports, while the cut-off was kept at SEK 4.5M for exports. Although this lessens the average burden on enterprises, it also increases the risk of model/estimation bias since enterprises below these cut-offs are imputed using VAT data. To date, there has been no formal study to evaluate this increased bias risk. Further, the imputation process is neither well-documented nor well-understood. We recommend that the FTG (a) evaluate the effect of raising the cut-off for Intrastat imports and exports on estimator bias and (b) substantially improve the documentation of the estimation process for enterprises below the cut-off.
2. Statistical Value Survey. At five-year intervals, the FTG staff conducts a survey of enterprises in order to recalibrate their models for converting invoice values to statistical values. The last survey was completed in November 2013 and the new adjustment model was implemented in March 2014. However, to our knowledge, there has never been a study to determine whether the recalibrated model is really much of an improvement over the old model. In the coming year, the FTG should consider how they might evaluate the statistical value models in order to show how much is gained by conducting these quinquennial efforts. For example, the staff should consider whether relying less on data collection and more on modeling could be just as effective but at a substantially reduced cost. If this investigation suggests a cost-effective approach, FTG staff should plan to launch such a study at some time in the near future.
3. Continue the close cooperation between FTG and NA. FTG staff should continue to become more familiar with the process generating the GDP estimates and how their data are being used.

4. Other areas that should be considered for future research are:

   - The effects on FTG and FTS estimates due to the inability to accurately separate trade in goods from trade in services for some products.
   - 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.
   - With the launch of the new web version of the Intrastat Data Entry Package (IDEP) data entry system, the FTG staff should evaluate its effects on respondents to determine respondents' reactions to the system and the extent to which respondent burden has been reduced. There should also be an evaluation to see if there is any impact on the accuracy of responses particularly as edits have not been introduced into IDEP yet.

Exhibit 9. Foreign Trade of Goods (FTG), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	Available expertise	Compliance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control for error sources)	**Specification error**	62	60	57	○	○	◒	◒	○	◓	**M**
	Frame error	58	58	57	○	○	◒	◒	○	○	**L**
	Non-response error	68	64	62	◒	◒	◒	○	○	○	**M**
	Measurement error	66	68	63	◒	○	◒	◒	◒	◓	**H**
	Data processing error	72	72	65	◒	◒	◒	◒	◒	◓	**H**
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Model/estimation error	80	80	73	◒	◒	◎	◎	◒	◓	**M**
	Revision error	72	72	67	◒	◒	◒	◒	◒	◓	**H**
	Total Score	69,3	68,9	64,1							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◓	○	◒	◎	**H**	**M**	**L**			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.7 STRUCTURAL BUSINESS STATISTICS (SBS)

SELECTED ACCOMPLISHMENTS

There have been several improvements in Structural Business Statistics (SBS) over the last 9 months.

- A TSE-focused study was completed that estimated the relative contribution of various error sources. It found that the most important error sources were (a) over-coverage because of the relatively large number of inactive units on the BR and, (b) the lack of kind of activity units (KAUs) for some enterprises in some industries.
- There has been greater involvement in the setting of accounting standards which should result in more accurate data for statistical purposes.
- The nonresponse rate continued to decline. This is owed in part to more aggressive enforcement of the mandatory participation requirements.
- The comparison of reported data with tax data obtained better insights into measurement errors.

Despite the above improvements, the number of profiled businesses is continuing to decline resulting in some serious deficiencies in the industrial classifications of large SBS enterprises.

RECOMMENDATIONS

1. SBS should collaborate with the Business Register (BR), Large Enterprise Unit (LEU) and National Accounts in order to increase the number of large enterprises that are profiled to ensure the NACE classifications are accurate in SBS and NA statistics. As discussed under the BR review (Section 3.2.8), a modelling approach may be needed in order to achieve this goal. The focus should be on those businesses that are most likely to distort industry statistics unless dissected into KAUs.
2. Although the statistical improvements in the BR have been delayed again, and with no clear timetable, SBS should start thinking about the work required for moving to the new BR and what the implications are for survey continuity. There are likely to be discontinuities in the SBS data series and some thought should be given to how to manage these discontinuities and whether any additional information is required. For example, over-coverage because of inactive units may be significantly reduced with the new BR.
3. SBS should obtain more quantitative data that would help it to evaluate editing decisions. One useful study may be to look at data before and after editing to study the net impact. This is similar to the study that was undertaken by RS in 2014 and gave them useful insights into the effectiveness of their editing.

Exhibit 10. Structural Business Statistics (SBS), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	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)	**Specification error**	60	60	57	○	○	◒	○	◒	◓	**M**
	Frame error	60	66	60	◒	◒	◒	○	○	◓	**M**
	Non-response error	72	72	70	◒	○	◒	◒	◒	○	**M**
	Measurement error	56	54	53	○	○	○	○	○	○	**H**
	Data processing error	60	60	55	○	○	◒	○	◒	◓	**H**
	Sampling error	86	86	85	◎	◒	◒	◎	◎	◒	**M**
	Model/estimation error	48	52	48	○	○	○	◓	○	◓	**H**
	Revision error	54	54	53	○	○	◒	◓	◓	○	**H**
	Total score	60,5	61,4	58,7							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◓	○	◒	◎	**H**	**M**	**L**			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.8 BUSINESS REGISTER (BR)

SELECTED ACCOMPLISHMENTS

- Work has continued on the development of the new BR. With a reduced budget, the focus has been on introducing the new IT system but without the statistical enhancements development work. As a consequence there are no definite plans for developing a version of the BR which would cover the areas impacting the accuracy of the BR for statistical requirements. We were previously informed that the new BR system will have greater flexibility especially with respect to modifying its contents. This may be useful for dealing with inactive businesses for example (see recommendations).
- As a result of work done in collaboration with the Swedish Tax Agency, there has been some work to assess the quality of the register. As a result of this study, it was concluded: (a) there is no serious overcoverage in the BR due to deliveries from the Tax Agency (although this primarily addresses duplications, not inactive businesses), (b) there is no serious undercoverage in the BR due to deliveries from the Tax Agency, and (c) there is good consistency in the variable content between the Tax Register and the BR.
- A new, computer-assisted editing (maintenance) system has been implemented involving about 10 full time staff and this has replaced the manual editing done on the basis of paper lists to a large degree.
- Work continues on deriving accurate estimates of employment size (full time equivalents or FTEs) for enterprises listed on the BR.
- Work was due to commence on an important new study referred to as 'Indicators in the Base Registers' which will measure and quantify the uncertainties in the content of the Base Registers including the BR.
- Also, "0 employees" was assigned to more than 7000 single establishment employers who had not shown any employment activity for several years.
- The User Group for Internal Users has been reactivated and has been meeting on a regular basis. It is starting to suggest priority areas for improvement of the BR.

An accurate Business Register is essential to the quality of economic statistics. Nevertheless, despite the improvements mentioned above, we remain concerned about some aspects of the BR. This has been reinforced by a Total Survey Error study undertaken by SBS. This study showed that the two most important sources of error for that survey were due to the Business Register – (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. It is a matter of finding a way of utilising this information. This is addressed in the recommendations below.

The other big concern seems to be the significant and continuing reduction in the number of KAUs. At present, slightly more than 40 enterprises are being profiled. This is causing a loss of accuracy of industry coding in important collections like SBS and consequently the NA. Each year we hear about another very large (unprofiled) enterprise that substantially influenced the National Accounts and other statistics because there is significant activity in industries other than the NACE to which the business has been allocated. These problems would not have been severe if the enterprise had been profiled as both industries would have been recognised as separate KAUs.

A past difficulty has been the very stringent Eurostat data availability standards for determining when new KAUs should be formed. These standards have been relaxed and now allow a modelling approach to be considered for cases when the fully detailed accounts are not available at the activity unit level. It should be possible, especially on the income side, to obtain partial information to support modelling of the data needed to partition an enterprise into two or more KAUs. This approach could provide more accurate statistics than assuming the entire business belongs to a single industry.

The Large Enterprise Unit (LEU) informed us that there were about 100 enterprises they would ideally like to profile. We are not sure that is necessary to profile (and maintain) that many businesses. A better approach might be to focus on a relatively small number of large complex businesses. National Accounts may be able to identify those industries causing most concern.

The definition of a business unit seems to vary considerably across surveys. 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 (see Section 4.1 for further discussion of this point). A stringent top down approach is necessary recognising that there will be transition issues that need to be managed if there is a change of unit in a particular collection.

Although it is contrary to Statistics Sweden policy, there have been corrections to the NACE codes and other enterprise data based on new information obtained through 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. Unfortunately, decisions on dependent survey feedback have been made on a survey by survey basis rather than in a coordinated way. Because its conditioning effects can bias inference based upon samples, the use of survey dependent feedback may be fine for the 'take all' strata but is potentially biased for sampled strata. A preliminary study showed that this practice does indeed bias survey estimates and recommended that the practice cease. However, this decision requires a careful analysis considering both the pros and cons at both national and industry levels. In particular, the survey estimation process may be compromised without survey dependent feedback due to more complex and variable weights being used when businesses are reclassified to their correct industry class after being sampled.

Perhaps there is a compromise solution. Certainly, enterprises should be sent the correct questionnaire or Statistics Sweden's reputation could suffer. A decision needs to be made at a relatively senior level after considering various approaches for dealing with errors on the BR identified through surveys. We understand the study is still progressing but it would be good if this issue were resolved in the coming year.

RECOMMENDATIONS

1. The procedures used by the LEU for creating activity units need to be revised to ensure reasonable industry purity is obtained in business surveys, business indexes and the national accounts. The number of profiled units needs to increase mostly for the very largest and complex enterprises. One possible alternative to profiling is the modelling approach but the responsibilities for decisions on modelling need to be determined.
2. A detailed plan for the statistical improvements for the revised Business Register 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 developing a Common Business Framework (a written proposal exists). Although it may be several years before these changes can be implemented, the research work could start in the near future. In particular, research could start on the algorithms for identifying businesses that are inactive or likely to be inactive.

3. Furthermore, the new Business Register System should support the creation of a BR specifically for statistical purposes. This should not be a separate physical Register. Rather, it should be a virtual Register that can be created from the Business Register 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 Register.
4. 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 Petroleum and Construction Consultancy industries were suggested.
5. A Statistics Sweden study showed that the use of dependent survey feedback potentially creates important biases in survey estimates. A more detailed analysis of the arguments for and against using dependent survey feedback is needed. In particular, conditions should be established regarding when dependent survey feedback can be used to correct erroneous NACE classifications on the BR.

Exhibit 11. Business Register (BR), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	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)	**Specification error**	66	64	58	○	○	◡	◡	◡	◠	**M**
	Frame error - overcoverage	58	60	58	○	○	◡	○	○	○	**H**
	Frame error - undercoverage	42	46	47	◠	◠	○	○	◠	○	**H**
	Frame error - duplication	63	65	65	○	○	◡	◡	N/A	N/A	**L**
	Missing data	50	50	47	○	○	○	○	○	◠	**L**
	Content error	56	58	55	○	○	◡	○	○	◠	**H**
	Total score	54,8	56,5	54,5							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◠	○	◡	◎	H	M	L			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.9 TOTAL POPULATION REGISTER (TPR)

SELECTED ACCOMPLISHMENTS

- The staff of the TPR continue to develop their approach for assessing overcoverage in the TPR. The study has been published on the Statistics Sweden website, distributed to key users in printed form and was orally presented to four of the nine user councils.
- A study of the residential population in Sweden revealed that undercoverage of the non-registered but still residents is significant. While overcoverage is approximately 64,000, undercoverage of residents was approximately 110,000, including students, asylum seekers and illegals.

RECOMMENDATIONS

1. We encourage the TPR staff to continue their research on overcoverage. One area that needs to be addressed is the accuracy of the overcoverage model. One approach would use historical figures to evaluate the most recent estimates.
2. As we noted in previous years, it is important for the TPR to provide information to users regarding how to use the overcoverage propensities for uses such as nonresponse mitigation and weighting adjustments. It may be useful to engage the internal users group in the application of this research. For example, the sum of the overcoverage propensities is an estimate of the true TPR overcoverage proportion. In addition, these propensities can be used to characterize the overcovered population. The propensity could also be used in current surveys to either prioritize cases for nonresponse followup or to adjust for nonresponse. For the latter, note that the probability of being a survey nonrespondent is equal to the probability of being a resident times the probability of being a nonrespondent given that the individual is a resident. This is one way of removing nonresidents from the nonresponse calculus. Some illustrations of these uses in actual applications would be very helpful to users.
3. Studies should be mounted that evaluate the accuracy of the "core" variables – i.e., important stratification and auxiliary variables used frequently in survey design and estimation such as age, country of origin, gender, marital status and region.

Exhibit 12. Total Population Register (TPR), Ratings for 2015

	Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	Available expertise	Compliance with standards & best practices	Plans towards risk mitigation	Effective-ness of mitigation measures	Risk to data quality
Accuracy (control for error sources)	**Specification error**	58	56	53	○	○	○	○	○	◓	**M**
	Frame error: overcoverage	66	68	65	◒	◒	◒	◒	○	○	**H**
	Frame error: undercoverage	56	60	57	○	○	◒	◒	◓	◓	**L**
	Frame error: duplication	70	70	70	○	○	◒	◒	N/A	N/A	**L**
	Missing data error: item and variable	62	62	60	○	○	◒	○	○	○	**M**
	Content error	62	62	58	○	○	◒	◒	○	◓	**L**
	Total score	63,0	63,2	60,7							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◓	○	◒	◎	H	M	L			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

3.2.10 QUARTERLY GROSS DOMESTIC PRODUCT (GDP(Q))

The quarterly GDP estimates are a very complex product that relies on many input data sources from both within Statistics Sweden and from 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 the National Accounts (NA) products and GDP in particular. In previous reviews we have only looked at the production side of quarterly GDP. Using the advice of the NA staff, we selected three input data sources: (1) the services production index, (2) the industrial production index and (3) those parts of the survey of foreign trade in services which provides estimates of merchanting services.

This year we also looked the expenditure side of quarterly GDP. Even more data sources are involved. We first disaggregated to the major expenditure categories and focused on the most important data source under each category. They are:

1. Household consumption: turnover statistics, including retail trade statistics
2. Public consumption: quarterly municipal accounts
3. Investment: investment survey
4. Change in inventories: inventory survey
5. Net exports: statistics on foreign trade in goods and services, respectively

SELECTED ACCOMPLISHMENTS

- The harmonization of the industrial and services production indexes and the harmonized survey has now been completed and the harmonized survey was introduced in 2015.
- Work has started on a replacement Swedish NA system. It is an area of high priority because support for the existing system is becoming more difficult. The initial work will replace the core data base system which is currently reliant on Visual Basic.
- There have been studies of the potential of using VAT data to estimate intermediate consumption to overcome the current modelling weakness of assuming a constant proportion of intermediate consumption to output. This has led to a funding proposal in the current planning round to conduct a quarterly SBS for the largest enterprises to be supplemented by VAT data for the smaller enterprises.
- Work on the recommended and innovative sensitivity analysis has been continuing. Although the initial focus is on annual GDP, there are implications for quarterly GDP. For example, the results of the studies of the impacts of double deflation and the reliability of producer price indexes will also have implications for quarterly GDP. The work has been extended to examine balancing and how to improve those processes.
- Through the work of the large enterprise unit, quality of the company reporting of merchanting on quarterly basis has improved.

This is the first time we have reviewed the input data sources for the expenditure based estimates. Significant areas of potential improvement that were brought to our attention are:

1. 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 when it is conducted but it has not been used to a very great extent because of significant accuracy concerns.

2. Investments: The lack of data for the second quarter appears to be 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.
3. 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.
4. 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.

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 principles and methods used to date are being reviewed which we support. A proposal to develop quarterly supply use tables is likely to eventuate. Also, the work on the sensitivity analysis might be useful for improving the objective balancing.

RECOMMENDATIONS

1. There is a need for a robust processing system for the NA estimates that includes a time series dimensions. The current system has a time series capabilities, but there is scope to make it easier and more effective. Work has commenced and this should continue to have a high priority. The first phase is limited in that it is replacing the core system only in order to overcome the problem of Visual Basic being no longer supported. We support the intention to look at re-engineering possibilities in the second phase. There are distinct advantages in co-locating the IT staff with NA during this period to minimize communication issues.
2. 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 will be existing NA training packages which could form a base for what is done in Statistics Sweden. These courses may also be of interest to users and those areas providing data to the national accounts.
3. We strongly recommend the continuing funding of the sensitivity studies. It should lead to more reliable national accounts and more effective use of resources in quality improvement effort.
4. Concerning GDP(P), a quarterly SBS, supplemented by VAT data, should be introduced to obtain better estimates of intermediate consumption. This should reduce the size of revisions when the quarterly accounts are benchmarked to the annual accounts.
5. Concerning GDP(E), the missing second quarter for the Investment Survey should be introduced to help reduce the volatility of these estimates. There should be investigations of the reasons for the large revisions following the annual introduction of the SBS estimates. The possibility of incorporating these data as part of the quarterly SBS should be investigated.
6. In addition, the models for quarterly Research & Development Expenditure should be reviewed given the significance of this new item. These models may require an increase in the frequency of the data collection.

7. Given its growing importance to both the National Accounts and the balance of payments, and the concerns about the reliability of the data, there should be a review of the survey of Foreign Trade in Services to see if it should be redesigned to better satisfy the most important needs. This should include issues such as coverage and the product classifications that are used.

8. Even if improved input data for both the GDP(P) and GDP(E) is likely to reduce the discrepancy between GDP estimated from the production side and the expenditure side, there will probably still be a need for balancing. The balancing process will always include some judgments, but they should as far as possible be based on objective criteria. The previous work on this topic should be updated.

Exhibit 13. Quarterly Gross Domestic Product (GDP (Q)), Ratings for 2015[1]

		Error Source	Average score round 4	Average score round 5 without Effective-ness	Average score round 5 with Effective-ness	Knowledge of risks	Communica-tion	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 source - Index of Service Production, ISP	64	64	60	○	○	◡	◡	◡	◠	H
		Input data source - Index of Industrial Production, IIP	64	64	60	○	○	◡	◡	◡	◠	H
		Input data source - Merchanting Service of global enterprises (also royalties, licensing and R&D)	44	50	50	◠	○	○	○	○	○	H
	Expenditure Side	Input data source Household Consumption -Turnover Statistics	--	60	55	○	○	◡	◡	○	◠	H
		Input data source Public Consumption -Municipalities	--	60	57	○	○	◡	○	○	◠	M
		Input data source Investments -Investment Survey	--	56	52	○	○	◡	◡	◠	◠	H
		Input data source Change in Inventories -Inventory survey	--	58	53	○	○	◡	◡	○	◠	H
		Input data source Net Exports -Exports in Goods -Exports in Services	--	56	52	○	○	◡	◡	◠	◠	H
		Compilation error (Modelling)	50	50	47	○	○	○	○	◠	◠	H
		Compilation error (Data processing)	54	56	52	◡	○	○	○	○	◠	H
		Deflation error	50	58	55	○	◠	◡	◡	○	◠	H
		Balancing Error	52	52	50	○	○	○	○	○	◠	H
		Revisions Error	58	58	55	◡	○	◡	○	○	◠	M
		Total score	54,3	56,3	52,9							

Scores						Levels of Risk			Changes from round 4		
N/A	●	◠	○	◡	◎	H	M	L			
Not Applic-able	Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations	Modifications

[1] In the ratings matrix for this product pair, the rows labeled "Production Side" correspond to the input sources for GDP(P) which are the same as those considered in prior ASPIRE rounds. The rows labeled "Expenditure Side" correspond to the input sources we considered for the evaluation of GDP(E). Together GDP(P) and GDP(E) comprise eight input data sources. Further, considering that seven of these eight input sources are rated as "H" and one is "M", the total weight attributed to input data sources is 7x3 + 1x2 = 24 points, or about 63 percent of the total product weight (i.e., 24+14 = 38). We believe that attributing 63 percent of the evaluation score to input data sources is not warranted because the other components of the evaluation should carry much more than 37 percent of the weight. Thus, although somewhat arbitrary, we have decided to halve the weights given to input data sources; i.e., reducing their combined weight from 24 points to 12 points. This gives about 46 percent (i.e., 12/(12+14) = 0.46) of the weight in the overall rating being to input data sources and 54 percent to the other components of total error which we believe is more appropriate.

3.3 THE HOUSEHOLD BUDGET SURVEY REVIEW

Budget surveys provide critical pieces of information to official statistics. The aim of these surveys is to obtain data on household expenditures on goods and services. These sample surveys are, however, difficult to conduct. The Statistics Sweden Household Budget Staff invited the ASPIRE Team to hear an update on their plans for the 2016 Household Budget Survey (HBS). Since 2012, the survey is conducted every four years; the next survey is scheduled to take place in 2016. The current design has 52 approximately equal samples – one for each week of the year.

An informant for a selected housekeeping unit is asked to report expenditures for the entire unit using a diary for 2 weeks. If preferred, informants may provide receipts for all qualified purchases instead of using the diary. About three fourths of respondents elect to provide receipts rather than use the diary. At the beginning of the two-week period, a telephone interview is conducted to obtain information on large expenditure items that were purchased in the past 12 months such as furniture and appliances costing at least SEK 500.

HBS experiences low response rates with both the diary and the interview survey. Measurement errors are also problematic, particularly for such items as alcohol, tobacco and minor purchases. Improving response rates and reducing respondent burden are key staff concerns. Plans are underway to attempt to address these issues in the 2016 HBS by reducing the amount of data requested in the diary, shortening the diary collection period to one week, and initiating the collection of data through a web diary. However, there does not seem to be any evidence that these measures will have much effect on bias; indeed, risks of exacerbating the HBS error appear to be high in our view. In the long run, the use of "big data" is seen as a potential solution to reducing response burden.

The 2012 HBS was not viewed as successful by several key users – namely, the Consumer Price Index (CPI) staff and the National Accounts (NA) staff. Both users noted that expenditure levels appear to be biased downward. The Total Survey Error paradigm is especially important in this survey context as a framework to understand and reduce sources of error. The staff would benefit from developing significantly more knowledge about the relative importance of the main error sources after attempts to reduce those error sources. For example, what is the relative importance of nonresponse bias after calibration adjustments are implemented? Is the larger source of bias observed with measurement error (that is, poor reporting resulting in low aggregate levels)? Additional analysis of the 2012 HBS would help the HBS better understand the sources of error in the survey.

A regular, ongoing dialog with the key users of the data is essential to clarifying the critical outputs for the survey. There are macro users, such as the CPI staff and the NA staff who are particularly concerned about aggregate data from the survey. There are also micro data analysts who use the income and expenditure data of different types of households for microsimulation and other policy research analysis. Staff acknowledge that they have not had significant contacts with users. Better communications will help to ensure the next HBS has survey content deemed essential by these users.

Staff have definite ideas about the short run design modifications and they are aimed at primarily reducing the nonresponse rate. These proposals seem to be in the early stages of planning and the benefits have not been thoroughly researched. For example, there is no evidence that the availability of a web diary will improve response rates. Other measures may be more effective and should be tested; for example, the use of incentive levels to reduce nonresponse bias. To achieve any

meaningful gains in response rates, the staff should consider the possibility of face-to-face interviews at least for large families. It is possible that a reduced sample size with better but more costly measurement could be a cost neutral alternative to the current approach while providing data of higher quality.

Further, although the effect of moving to a one week diary will have on survey estimates has been explored, the results did not confirm that accuracy would be improved when taking all sources of error into account. This is an important proposed design change done largely to reduce burden. However, while it may increase the diary response rate, the potential effects on estimation error may offset this benefit. These changes require testing in an experimental framework and the effects of the changes and the impact understood. For example, a report from a study undertaken by the product staff suggested that the next HBS contain an embedded experiment where some households may report on a 2 week diary while others report on a one week diary.

Many design options exist for the HBS; significant survey research is needed, backed by experimental evidence, to ensure good decisions are made to ensure a successful HBS.

RECOMMENDATIONS

1. Statistics Sweden should consider delaying the HBS at least one year to allow for the additional analysis, information gathering and experimentation that would better inform the design of the next HBS.
2. Additional analysis of 2012 HBS data may be warranted to better understand the magnitude of the various sources of error in the survey and use those analyses to inform design options that lead to a reduction in the mean square error of the estimates. For example, comparisons of the HBS data with external data sources could help identify areas of undercoverage or underreporting. Likewise, it is not clear whether the quality of the "receipt bag" data has been investigated. Although a simulation of the one week diary data has been completed and provides indications of potential reporting error, testing of a one week diary instrument in an experimental design should be undertaken.
3. A workshop with users – both internal and external to Statistics Sweden – should be held to determine substantive and analytic priorities for the survey. The content of the HBS should be adapted to these priorities.
4. The Bureau of Labor Statistics (BLS) in the United States is in the midst of a multi-year research project aimed at redesigning the U.S. Consumer Expenditure Survey. Their project has been supplemented with the work of an expert panel from the Committee on National Statistics (CNSTAT) in the United States, commissioned by the BLS. The next HBS will benefit by a staff review of the research conducted by the BLS and its CNSTAT panel, followed by in-person or telephone conversations with the BLS staff.
5. An external expert review panel ought to be formed as advisors to the HBS staff to consult, advise, and follow up on a research agenda that focuses on the reduction of total survey error through an understanding of the component sources of error.

3.4 ASPIRE CHECKLISTS AND QUALITY DECLARATIONS

The ASPIRE process is based on a Total Survey Error framework and reviews a comprehensive list of error sources as each pertains to a specific Statistics Sweden data product. Prior to the "Quality Interview" with the data product staff, the ASPIRE team relies on two approaches to obtain information about the statistical product: 1) the self-assessment checklist; and 2) the Quality Declaration document. Each contributes to the ASPIRE team's understanding of activities related to the statistical data product being evaluated as each provides important information in the conduct of an efficient "quality interview" and a more accurate assignment of evaluation scores for the component sources of error. Each approach is very important to the ASPIRE process and relies on the data product staff to provide accurate assessments and complete information.

We have observed some deterioration in the quality of the products reporting in the checklists. Some products have not filled in their assessment of all error sources and the box for comments could be used more intensively. This may be due to some fatigue with the ASPIRE process, but especially when a product suggests a higher rating, it is important that they provide some rationale for the change including references to the work that is done in the box for comments. We in the ASPIRE team will also try to develop a better interface for collecting this information. Maybe a web questionnaire could be developed in cooperation with Statistics Sweden.

The Quality Declarations have improved significantly since Round 1 of ASPIRE and many parts are also available in English. But there are multiple documents having similar contents; e.g. the description of the production of the statistics ("Statistikens framtagning" (SCBDOK), only available in Swedish) and the Quality Declarations (QDs) included in "Description of the Statistics". In addition to these two documents, which are updated annually for every product, there are other documents sent to Eurostat describing the products.

To reduce the work load of keeping the information in several documents up to date, Statistics Sweden should continue the ongoing work to create an integrated documentation system. The product documentation could include a non-technical summary and some other user friendly information that should be available in both Swedish and English. The more detailed description of the statistics, including the QDs, could possibly be in English only. Those who are using this documentation need fairly deep knowledge in statistics and they are likely to be able to read in English. This documentation is also essential for the ASPIRE work, but could also be very useful for other purposes including benchmarking activities against other countries. If Statistics Sweden decides to continue producing the documentation first in Swedish and then translate to English, this should be done at least every 3 years or upon completion of major design changes, significant research studies, or significant quality changes in the product.

4 CROSS-CUTTING ISSUES AND RECOMMENDATIONS

4.1 INTEGRATION OF ECONOMIC STATISTICS

Although the focus of ASPIRE is on the Accuracy dimension of quality, it is appropriate and essential to consider the Coherence of economic statistics since Coherence can be regarded as a comprehensive indicator of Accuracy. For example, for computing GDP, discrepancies between GDP(P) and GDP(E) – referred to as the *statistical discrepancy* – exist because the components of GDP both sides of the estimation process are subject to error. The balancing process is intended to reduce the difference although it may not necessarily improve GDP accuracy but it is likely to in most cases. The extent to which different products produce discrepant estimates of the same population quantities is an indication of the magnitude of the errors in the processes that produce these estimates.

The integration of economic statistics is key to the goal of providing a system of coherent official statistics. In pursuing this goal, we have identified the following nine areas as being of critical importance for achieving economics statistics that are of high quality and coherent, although those two attributes should be considered, in many ways, synonymous for the aforementioned reasons.

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 on the basis of partial information.
2. Common Business Framework (CBF). A CBF should be derived from the Business Register 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. Also, 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 either be consistent or 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. The common classifications might be supported by coding frameworks that can be used by the different areas to 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 non-response, changes in

industry codes on the Business Register, 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 but there are further improvements that could be made. These are discussed below but we first recognise our understanding of the current situation in Statistics Sweden including planned work. The same nine headings are used.

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 Business Register (which is not known at the time of sample selection) and the lack of industry profiling for many large businesses. We note the intention to increase profiling in the coming year. We also note the intention to study inaccuracies in the content of the Register.
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.
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. As an example, inconsistencies in classifications between the CPI and HBS can cause difficulties for the NA to use these data. There are also inconsistencies in the use of industry classifications between economic statistics and labour market statistics.
4. Standardized Concepts. Again, we are not aware of the Statistics Sweden practices.
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, is being seriously considered.
7. Input Data Warehouse. We are not aware of the details of what is happening in Statistics Sweden. Certainly it is adroit at combining administrative and survey data so we assume the data warehousing arrangements are in place to support this.
8. Revision Policy. Work is taking place now on a common revisions policy.
9. Governance. At the operational level, communication between the different stakeholders has improved considerably since we started the ASPIRE reviews. This has been aided by things like the SLAs with the National Accounts, the Business Register User Group, etc. However,

we feel that it is necessary to set a group that considers issues at a more strategic level including work. We were informed that there are now regular meetings with the department heads for Data Collection and Economic Statistics. This is a good start. However, it needs to operate in a way that strategic decisions could be made. Also, we would suggest that the department head for National Accounts be included in these discussions.

Consistent with the new criterion for measuring mitigation effectiveness (criterion 6), Statistics Sweden should periodically review its progress toward greater coherence and statistics integration to determine the effectiveness of its efforts. How does one measure Coherence and other objectives of statistics integration? There are a few measures that can be derived from the national accounts, including the following:

- Statistical discrepancy between the expenditure and production accounts before balancing.
- Discrepancy between the annual accounts and the sum of the quarterly accounts before they are benchmarked to the annual accounts.
- The volatility of some key ratios within the national accounts (e.g. savings ratio and capital investment to GDP ratios).
- The plausibility of implicit price deflators.
- Measures like labour productivity and multi-factor productivity which bring together national accounts and other data.

It is worth considering developing an agreed set of measures and monitoring them on a continuous basis. The measures are subject to error and volatility so interpreting movements of the indicators should take this into account and the primary focus should be on trends.

RECOMMENDATIONS

Here we mention a few additional recommendations. Some of these are also mentioned for the product areas as well as repeats from previous rounds.

1. 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 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. Ensure that the largest and most complex enterprises are profiled into Kind-of-Activity-Units (KAUs) so that significant industry activities within the enterprise are identified. The focus should be on the most important businesses rather than large numbers of enterprises which may difficult to maintain. Where the full set of financial information is unavailable at the KAU level, models utilising available information should be used to provide estimates at the KAU level.
3. Harmonize the selection of business units across the business surveys especially for those surveys that contribute to the National Accounts. It is always difficult to change historical decisions but a useful starting point might be to consider which units would be appropriate if you were designing the collection from the start. EU regulations would need to be taken into account. If there is a change in unit, the transition issues would need to be considered. It is appreciated that they might be considerable for some collections.
4. There should be a study to assess whether more should be done to implement standard classifications in a way that would support national accounts and other users.
5. Ensure that key methodological decisions, such as adjustments for non-response, are performed in a consistent way.

6. Introduce a quarterly SBS. This would greatly assist the estimation on intermediate consumption in the quarterly National Accounts which is believed to be an important source of bias in the quarterly National Accounts as well as one that seems to vary with the economic cycle. It would also improve the coherence between other variables collected in the survey and may mitigate other deficiencies such as the lack of investment data for the second quarter.
7. Build on the existing arrangements so that there are governance arrangements which decisions to be made on strategic issues that cut across Departments.

4.2 NONRESPONSE IN HOUSEHOLD SURVEYS

Response rates for household surveys have continued to deteriorate despite considerable efforts aimed at ameliorating this problem. Although declining response rates increase the risk of nonresponse bias, the magnitude of the bias depends upon both (a) the nonresponse rate and (b) the differences between respondents and nonrespondents. Note that (b) can be made small even though (a) is large. The result can be a small nonresponse bias even with high nonresponse rates or, conversely, a large bias even with low nonresponse rates. Nevertheless, reducing the nonresponse rate also reduces the risk of nonresponse bias and Statistics Sweden has devoted considerable resources toward this goal, particularly for the LFS.

As in Rounds 3 and 4, the ASPIRE reviewers again met with a few department heads, members of the Process Department and the Nonresponse Project team to discuss the current status of the nonresponse mitigation work at Statistics Sweden. It appears that the focus has changed somewhat from prior rounds as more knowledge has been gained about the nonresponse problem. Some of the current efforts underway include the follow:

- A new management structure within the data collection department that will provide greater supervision for interviewers
- Research on mixed mode data collection
- Greater focus on determining methods for reducing the noncontact rate
- Use of an external vendor to collect approximate 5,400 LFS cases to determine whether response rates can be increased by altering the mix of times at which interviewers attempt to contact respondents.
- Development of an improved system for call management (Copernicus) that will be implemented this Fall
- Further analysis to better understand the sources and causes of nonresponse and how to mitigate them

Because of its visibility both internally and externally, its increasing risks to data quality, and the considerable resources being spent to mitigate it, the nonresponse problem needs to be addressed with some urgency. Striving to achieve the highest response rates possible within available resources is certainly a reasonable goal; however, as we have noted in our previous reviews, attempting to elevate response rates to their 2008-2010 levels is futile and may not even result in significant bias reduction. Like a number of other countries (for example, The Netherlands), Statistics Sweden has begun to consider a mixed-mode approach – that is, combining telephone interviewing with web-based and postal questionnaires. This research suggests that response rates can be significantly increased. Based on the research to date, unfortunately, it does not appear that the composition of the responding sample is much affected by the mixed-mode approach based upon demographic diversity measures. Since sample representativity is a necessary condition for nonresponse bias reduction, these recent results suggest that Statistics Sweden's approach to mixed mode may not reduce nonresponse bias even if response rates can be elevated.

On the other hand, Statistics Sweden has a considerable amount of data for nonrespondents on its administrative registers. This enables nonresponse bias to be effectively mitigated through the use of calibration techniques that compensate for nonresponse. Studies undertaken as part of the nonresponse project have shown the calibration techniques are effective at reducing the nonresponse bias for some estimates. More research along these lines is needed.

Although there have been organizational improvements in the data collection area, it is not evident whether the governance arrangements are in place to allow decisions to be made based on the

findings of the Nonresponse Project. The decisions will impact several departments, and several products, so it is important that the governance arrangements allow for this. The statistical products are important stakeholders and they should be more involved in the Nonresponse Project than at present.

Moreover, in our interviews, we noted that both the LFS and the LCS/SILC consider the mitigation of nonresponse bias to be the sole purview of the Nonresponse Project. However, we believe that, while this may be true for increasing response rates, it should not be true when it comes to reducing nonresponse bias through post-survey weight adjustments. Consequently, these surveys, and perhaps others, are not actively pursuing methods for mitigating nonresponse bias via statistical modeling.

In this review, we want to emphasize six general areas that we believe are particularly important for the Nonresponse Project.

1. Mixed-Mode Data Collection Research. The research on using mixed modes holds much promise for the long-term future of household surveys at Statistics Sweden, but there are a number of critical issues to consider as this research moves forward. Three of these are:
    a. The current approach to mixed mode does nothing to increase response rates in the first wave of a panel survey. In most surveys, the initial wave has the highest nonresponse rate and, thus, provides the greatest risk to nonresponse bias. The research should investigate how mixed-modes can be introduced at the initial wave of a panel survey such as the LFS. Statistics Denmark has had some success in this respect.
    b. Particular attention needs to be given to possible mode effects which have been shown to be quite severe in some surveys that mix interviewer assisted and self-administered modes. The experience of Statistics Netherlands may be particularly relevant here. (See Janssen, 2011).
    c. Allowing persons to respond by mail (i.e., postal response) should increase response rates by several more percentage points since not everyone, particularly older adults, either has internet access or wants to respond by web. Besides increasing response rates, this could also help to recruit a more diverse and representative respondent sample and the research should confirm the extent to which this happens.
2. Case/Call Management. This important area has not received enough attention in our opinion. This includes analysis to inform call scheduling, optimal levels of effort (LOE) per case, and characterizations of contact/noncontact histories. In particular, modeling the propensity-to-be-contacted using past contact/noncontact histories as well as individual characteristics could prove to be a fruitful area of research. The research should also consider the appropriate LOE to expend for a case based upon these same characteristics so that precious data collection resources can be reserved for more productive cases as well as cases that contribute most to the nonresponse bias.
3. Two-Phase Sampling. In two-phase sample, only a random sample of nonrespondents are followed up rather than all nonrespondents. Two-phase sampling is not only economical but it can also increase the weighted response rate and reduce nonresponse bias. So far, Statistics Sweden has not investigated the use of two-phase sampling yet it could be an important tool for some surveys, including the LFS.
4. Data Collection Management. As noted above, the data collection department has increased the supervision of interviewers by reducing the span of control of its field managers. However, it is not clear how this will contribute to the goal of mitigating nonresponse. There

needs to a clear articulation of the goals of these organizational changes. Data should be collected that will provide an objective determination as to whether the changes have achieved those goals. In addition, activities that we believe may have more direct impacts on nonresponse include the following:

a. Develop management structures and systems for telephone operations (both centralized and decentralized) that allow for real-time, responsive deployment of resources to increase response rates and reduce nonresponse bias. This includes a flexible and adaptive call scheduling capability for assigning interviewers to time-slots based upon real-time reports of expected work-loads. In addition, a survey control system is needed that can provide real-time reports on the status of all cases in the system and can accurately predict interviewer staffing requirements at least one day ahead.

b. Develop a more strategic approach to dealing with the problem of multiple telephone numbers per household. For example, a sample person may have as many as five telephone numbers available and working all these numbers can be wasteful and unproductive. In addition, it can adversely affect the modeling of contact propensities so that the next best time to call is poorly predicted. A better strategy is needed to deal with this problem yet it appears that no research is currently being conducted on the issue.

5. Nonresponse Bias Analysis. All survey products should be concerned with nonresponse bias and how to mitigate it through bias-mitigation modeling and weighting. For example, in the last round we heard that the residual risks of nonresponse bias for LFS income estimates were substantially reduced by weighting. However, to our knowledge, this analysis has not been repeated for other variables and surveys, at least in recent times as the nonresponse rates have substantially increased. We believe bias-mitigation modeling and weighting research is extremely important since it is the last defense for reducing nonresponse bias after efforts (and resources) to increase response rates have been exhausted.

4.3 TELEPHONE CALL MONITORING

Call monitoring in centralized Computer Assisted Telephone Interviewing (CATI) facilities provides opportunities for observing and improving interview performance, detecting and deterring falsification, noting respondent concerns or questions and investigating questionnaire wording or formatting problems and so on. Centralized telephone monitoring with feedback can also help to standardize interviewer behaviour thus reducing interviewer variance which can substantially reduce the precision of the estimates. We suspect that interviewer variance may be quite high in Statistics Sweden's telephone facility although it has never been evaluated.

Recommendations from Round 4 ASPIRE made the point that "Data from call monitoring have not been utilized effectively in order to evaluate, for example, interviewer effects and measurement errors." Statistics Sweden Data Collection staff invited the ASPIRE Team to discuss call monitoring procedures and the staff's desire to identify a project that demonstrates the capabilities and potential of call monitoring. Statistics Sweden staff are interested in working on the Round 4 recommendation but funds to accomplish the work have not been available. Call monitoring systems used by companies in the United States, such as RTI, Westat, and NORC are sophisticated, flexible systems that permit a variety of statistics to be generated by day, by time slot, by question, by interviewer etc., if desired.

Statistics Sweden appears to have a number of difficulties to address with respect to more efficient use of call monitoring. Some of these include the following:

- For about 90 percent of all calls, Statistics Sweden interviewers know with certainty that they are not being monitored. For the remaining 10 percent, then are fairly certain (50 percent probability) that they are being monitored. Consequently, which calls are being monitored is quite predictable and interviewers can alter their "usual" behaviour for calls that are monitored. This means it is impossible to infer much regarding interviewer behaviour on the basis of the monitoring results. For example, interviewers that are motivated to falsify all or a portion of the interview, can do so with impunity for 90 percent of their calls.
- For calls that are likely to be monitored, interviewers are required to read a lengthy (about ½ minute) disclosure statement to the respondent. Respondents are required to provide active consent to be monitored before monitoring can proceed. Approximately 15 percent of respondents refused to be recorded which precludes their calls from being monitored. Possibly this rate could be reduced if disclosure statement were substantially shorter.
- The software Statistics Sweden uses for call monitoring does not appear to be as flexible as many other systems in use, particularly in the U.S. For example, the capability of capturing monitoring results electronically and producing reports in real time are not available.
- Very little use has been made of the call monitoring to code interviewer behaviour and code the respondent and interviewer reaction to sections of the questionnaire or questionnaire items.

Staff identified several potential projects for demonstrating the utility and value of call monitoring; for example, a project that evaluates measurement error in field-coded items in LCS/SILC.

RECOMMENDATIONS

1. In the coming year, at least one demonstration project aimed at showing value of call monitoring for improving questionnaire design and/or other aspects of the interviewing process should be implemented. In that regard, the LCS/SILC field coding project may be a good candidate. Another may be the LFS where questions about temporary versus permanent employment might be explored.
2. Statistics Sweden should consider estimating the interviewer variance for telephone interviewers. This should be possible without altering the current procedures for allocating cases to interviewers using the methods described in Stokes (1988). These results should help motivate the need for call monitoring to reduce the level of interviewer variance in the estimates.
3. The staff should study further capabilities of its call monitoring system by engaging in conversations with other organizations such as RTI, Westat, and NORC and others to determine the capabilities of their call monitoring systems with a view to proposing enhancements to Statistics Sweden capabilities.

4.4 FUNDING FOR RESEARCH AND DEVELOPMENT

Statistics Sweden does have funding for development projects. Proposals for funding can be developed and submitted for consideration. This is good practice and probably does not exist in many NSOs. However, at present, much of this discretionary funding is still devoted to IT projects especially those associated with the phase out of Virtual Basic 6. This is understandable but it would be appropriate if more funding could be devoted to statistical development projects as soon as possible.

There is some research capacity within the product areas, aided by their allocated methodology staff. We mention below that the effectiveness of mitigation measures is rarely quantified. This work should be given a higher priority. It should be led by the product areas utilizing their own resources as far as possible but we recognize that additional funding may be needed for more significant studies. Funding for these studies should be a consideration in the allocation of R&D resources.

Additional guidance to product areas undertaking these types of evaluation studies might help stimulate higher quality studies. We suggest the development of a manual or tutorial describing the various approaches of undertaking evaluation studies. This should be the responsibility of the Research and Development Department and we are prepared to assist with the development of such a document.

The Innovation Laboratory was a good initiative but does not seem to have advanced much since the previous round. We suggest the *modus operandi* for the Laboratory be reviewed to increase its effectiveness at providing inputs that are important to the future of Statistics Sweden.

RECOMMENDATIONS

1. Additional funding for research and development activities should be provided as soon as possible.
2. A document describing best practices for evaluation studies of the effectiveness of mitigation measures should be developed.
3. The *modus operandi* for the Innovation Laboratory should be reviewed to increase its effectiveness at providing inputs that are important to the future of Statistics Sweden.

4.5 EVALUATION OF THE EFFECTIVENESS OF MITIGATION MEASURES

As discussed in Section 2.1, Criterion 5 that was used in prior rounds tended to confound risk mitigation planning and achievement and as a result, a product could score high on the combined criterion by having very good plans that were neither (fully) implemented nor effective once implemented. Our solution was to separate "planning for risk mitigation" from "achievement or effectiveness of the risk mitigation efforts." Thus, in this round, former Criterion 5 was replaced by two new criteria referred to as Criteria 5 and 6 representing planning and achievement, respectively. New guidelines and a checklist were developed to guide the evaluation of Criterion 6. In applying these guidelines, we discovered that, although there has been considerable planning and some efforts to improve data quality, most ASPIRE products seldom evaluate whether these plans, once implemented, actually improve data quality. There are some exceptions which we noted in our reviews.

For example, it is not difficult to determine whether efforts to reduce sampling error are effective because one can simply observe where the published standard errors have been reduced. Likewise, efforts to improve response rates can be readily evaluated by comparing response rates before and after the nonresponse mitigation efforts have been applied. In both these examples, a product would be given at least a "Fair" rating for having attempted to reduce the error. If the measures were effective, then the rating would increase to at least "Good." However, if the efforts were unsuccessful, the rating for Risk Mitigation would remain at "Fair," although the rating for Knowledge may increase depending upon what was learned from these efforts. In the case of response rates, evidence should be provided that shows the current mitigation strategy either increased response rates or prevented response rates from declining further. The latter case would require that an experimental design be conducted so that the current method (including new methods) for reducing nonresponse could be compared with the counterfactual that does not apply the current (or status quo) approaches. Unfortunately, we see few cases where such experimental approaches have been used for the ASPIRE products.

How one should go about demonstrating the effectiveness of data quality improvement activities varies widely across products, error sources and data quality improvement objectives. Perhaps the easiest way to demonstrate effectiveness is the "before and after" analysis described above. 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 extent that uncontrolled factors could influence the post-intervention measures, thus confounding the before and after comparisons. A preferred approach is the controlled experiment. While this approach reduces confounding of comparisons of experimental treatment with the control, experiments can be costly, potentially risky and difficult to conduct in a real-time production setting. Other approaches for evaluating effectiveness are available and a good reference for this is Breyfogle (2003).

The new Mitigation Effectiveness Criterion was discussed at some length in the evaluation interview for each product and many ideas for improving the rating for this criterion in the next round were shared. We emphasized the need for the product staff to adopt a disciplined approach to quality improvement that combines good planning for risk mitigation and controlled implementation of those plans with a rigorous evaluation strategy to determine if the intervention was effective. At a minimum, this approach will provide valuable information to help better understand the underlying causes of the error.

RECOMMENDATIONS

1. With each improvement activity, staff should consider how the effectiveness of the activity can 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.

5. SUMMARY AND RECOMMENDATIONS FOR FUTURE ROUNDS

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 4 and these have been identified in this report. This round emphasized the effectiveness of risk mitigation efforts by evaluating it separately from risk mitigation planning. This had the effect of further attenuating ratings compared to Round 4.

At each round, new information about each product is revealed, some of which may suggest a rating given in an earlier round was in error. As a result, a small number of the ratings have been corrected. We have identified these ratings in the report to distinguish them from improvements or deteriorations. Exhibits 3a and 3b shows the current ratings, prior year ratings, and the improvements by product. Exhibit 3c 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.

With a maximum possible score of 100 percent (indicating perfect quality), the product scores in Exhibit 3a under the revised criteria ranged from 49.4 percent (for the LCS/SILC) to 64.1 percent (for the FTG) with an average rating of 57.5 percent. This does not include the GDP(Q) which scores 52.9 overall. When the old criteria are considered, products generally increased their scores in this round but the average improvement in ratings for the products in Exhibit 3a was only 0.6 percent this round compared to 1.3 percentage points in the last round. In Section 3.1, we provided some possible reasons for the reduced average improvement under the old criteria which may be a combination of (a) greater difficulty to improve after five years of ASPIRE, (b) reduced resources to address meaningful quality improvements and (c) lack of attention to the ASPIRE recommendations possibly due to other higher priority activities. In addition, the intervening time period between Rounds 4 and 5 was only 9 months rather than 12 months as in prior rounds. When the revised criteria are considered, all products show considerable deterioration as a result of much greater attention focused on risk mitigation effectiveness in this round.

Clearly, (a) was evident in some reviews. Following five 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 risk mitigation strategies that result in real, demonstrative improvements. Such activities require resources (b), which as previously noted in this report, have been more constrained in the last two years for a number of products. In addition, the relatively small improvement this year may be partly explained by staff motivation and accountability (c). Notwithstanding the small increase in average scores for this round, there has still been about a 7 percentage point increase since ASPIRE started in 2011 (see Exhibit 3c) which represents 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 five rounds and seemed to work quite well in the current round. Although it could be improve further, the revised criteria that was introduced this round seemed to capture the information Statistics Sweden seeks regarding risk mitigation effectiveness. We continue to be pleased that products such as TPR and RS took up our recommendations from prior rounds to conduct highly innovative and informative studies of overcoverage and editing error, respectively. We also pleased with the improvements reflected in

the ratings for the BR and GDP(Q). These product staffs should be commended for their inspiration, initiative and continued attention to the goals of ASPIRE.

In preparing for their Round 6 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, 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. Generally speaking highest priority should be given to error sources with high risk ratings (H) combined with quality criteria with below average ratings. Some desired improvements are cross-cutting in nature and we have discussed these in Section 4 of this report. The 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.

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. 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 for her tireless and professional support and the excellent co-operation from all the Statistics Sweden staff we had contact with.

6. REFERENCES

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.

Breyfogle, F. (2003). Implementing Six Sigma, Second Edition: Smarter Solutions Using Statistical Methods, John Wiley & Sons, NY

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

Stokes, L.S. (1988). "Estimation of Interviewer Effects for Categorical Items in a Random Digit Dial Telephone Survey," *Journal of the American Statistical Association*, 83, 623-630

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 analyzing 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	
2. As a result of these efforts, current proxy measures of the error from this source suggest that the error risks have been mitigated to some extent.	☐ Yes ☐ No	
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	
4. Direct estimates of the MSE components associated with this error source indicate that substantial reductions of the error have been the result of current mitigation efforts. These accuracy improvements have been documented and these documents are publically available.	☐ Yes ☐ No	
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	