A FOURTH APPLICATION OF ASPIRE FOR STATISTICS SWEDEN

Paul Biemer and Dennis Trewin

January 9, 2015

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

In 2011, the Ministry of Finance directed Statistics Sweden to develop a system of quality indicators for a number of key statistical products. This system was to include metrics that reflect current data quality as well as capture any changes in quality that occur over time. In response, Statistics Sweden collaborated with two consultants (Paul Biemer and Dennis Trewin) to develop a quality evaluation approach that is referred to as ASPIRE (see Biemer and Trewin, 2013 and Biemer, Trewin, Bergdahl and Japec, 2014).

This report summarizes the results from Round 4 of ASPIRE which was conducted in December 2014. It covered the Annual Municipal Accounts (RS), Consumer Price Index (CPI), Foreign Trade of Goods Survey (FTG), Labour Force Survey (LFS), Survey of Living Conditions (ULF/SILC), Structural Business Survey (SBS), Business Register (BR), Total Population Register (TPR) and the quarterly and annual National Accounts (NA).

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) using five criteria that were identical for all relevant error sources. 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.

Each of the ten products showed an improvement in ratings although relatively small. The biggest improvements were in RS (2.4 percentage points) and TPR (2.0 points), both of which conducted insightful evaluation studies. The average improvement in ratings over all products and error sources was about 1.3 percentage points. When combined with the 3.2 percentage point increase in Round 2, and the 2.7 percentage point increase in Round 3, there has been a 7.2 percentage point increase since ASPIRE started in 2011.

With a maximum possible score of 100 percent (indicating perfect quality), the product scores ranged from 52.0 percent (for the ULF/SILC) to 68.6 percent (for the FTG) with an average rating of 60 percent. (Exhibits 2a and 2b in the report provide the scores for each product by error source.) Although not in this Report, we prepared a 'Change Matrix' for each product that provides explanations for any changes in ratings. They are available from Heather Bergdahl on request.

Some additional findings from the reviews include the following:

- Ratings for all ten products increased in this round, albeit some only marginally.
- For the eight products in Exhibit 2a, the overall mean quality rating increased by about 1 point as compared to almost 3 points in Round 3. For the NA products, the average increase was less than 1 point compared to almost 2 points in Round 3.
- The last row of Exhibits 2a and 2b shows the Round 3 to Round 4 changes in the overall quality ratings by product. Note that TPR and the RS improved the most. This was the result of new and innovative quality improvement initiatives that they completed in 2014.
- 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 2a and high risk for the GDP products in Exhibit 2b.

- 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 two rounds.

In addition, the following general findings are notable:

- The nonresponse rates for household surveys continue to deteriorate, and at a faster rate, despite the considerable effort and resources put into addressing this problem.
- 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.
- 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.
- The sensitivity studies on input data sources in the national accounts show real promise of providing insights that may lead to improvements in the accuracy of the national accounts.
- The 'before and after' study of editing for RS provided information which will enable the staff to improve the cost-effectiveness of their data editing processes.
- The innovative study undertaken by TPR to provide information on the overcoverage areas also provides extremely useful information for the users of the TPR.

We should also note the successful introduction of the new European System of National Accounts (ESA 2010) in the September 2014 quarter. This is a major achievement.

As in our previous reports, we laid out some general recommendations to improve quality that cut across all products. This year we have focussed on what we believe are the four most important recommendations rather than having a long list of recommendations (16 in Round 3). These general recommendations are discussed under the following headings.

1. Opportunities to Improve the Quality and Cohesion of Economic Statistics.
2. Managing Increasing Nonresponse Rates in Household Surveys.
3. Funding for Research and Development.
4. Responding to ASPIRE Recommendations.

We also believe that organizational structure at Statistics Sweden, particularly the very "flat" reporting hierarchy, makes it more difficult for the organization to introduce and successfully manage cross-cutting projects. For each of the first two recommendations, we have suggested some supplementation of existing governance arrangements to address these specific issues.

2 BACKGROUND AND INTRODUCTION

This is the fourth round of ASPIRE. The background to ASPIRE has been provided in previous Reviews (see, for example, Biemer and Trewin, 2014). As with the previous round, for this round (Round 4), the focus of ASPIRE was on the Accuracy quality dimension. The same ten products reviewed in Rounds 2 and 3, and documented in Biemer and Trewin (2013) and Biemer and Trewin (2014), were reviewed again for this round.

One of the main objectives of Round 4 was to identify areas within each of the ten products where clear improvements had been made since the previous evaluation. Our report also identifies the highest priority areas for improvement at the product level. Furthermore, some general recommendations are made for high priority cross-cutting issues.

The ASPIRE process, error sources and evaluation criteria that was applied in this review were identical to the previous round and described in Biemer and Trewin (2014). The main difference in this round was that the interviews focussed on the main changes since the previous review. As a consequence the interviews were shorter as was the overall time the external evaluators needed to spend in Sweden.

Section 3 summarises the results of the quality evaluations for the ten products (treating quarterly and annual NA as separate 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.

2.1 SCOPE OF THE REVIEW

On the top panel of Exhibit 1 are the six survey products that are included in the ASPIRE review in this review round (Round 4). The error sources that are associated with these products are shown to the right of these products. Likewise, the middle panel shows the two registers included in this review and their error sources. Finally, the bottom panel shows the National Accounts (NA) products which are compilations of various other product inputs and data sources. The errors sources associated with these NA products (which are discussed below) are shown on the right that panel. As we previously noted, all of these products were evaluated in Rounds 2 and 3 and those results are documented in Biemer and Trewin (2013 and 2014, respectively).

With regard to the NA products, the current review, like Rounds 2 and 3, focused somewhat narrowly on the estimation of quarterly and annual GDP and solely from the production perspective (i.e., the expenditure perspective was not within the scope of the review). Biemer and Trewin (2014) provides a discussion of the error structure we used for NA.

Exhibit 1. Sources of Error Considered by Product

Product	**Error Sources**
Survey Products Foreign Trade of Goods (FTG) Labour Force Survey (LFS) Annual Municipal Accounts (RS) Structural Business Statistics (SBS) Consumer Price Index (CPI) Living Conditions Survey (ULF/SILC)	Specification error Frame error Nonresponse error Measurement error Data processing error Sampling error Model/estimation error Revision error
Registers Business Register (BR) Total Population Register (TPR)	Specification error Frame: Overcoverage Undercoverage Duplication Missing Data Content Error
Compilations National Accounts (NA) GDP by Production Approach, Annual GDP by Production Approach, Quarterly	Input data error (up to four sources) Compilation error Data Processing Error Model/Estimation Error Deflation/Reflation Error Balancing Error Revision Error

3 FINDINGS FOR THE TEN STATISTICAL PRODUCTS

Exhibit 2a provides the overall scores for the six 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 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 its corresponding true value. 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 2b provides the scores for the two NA products. As discussed in Biemer and Trewin (2014), the error structure used in the evaluation of these products has been customized to reflect the unique operations associated with compiling the data and generating both quarterly and annual estimates of GDP. For that reason, the Accuracy of the NA products is treated separately from the other eight products.

Finally, Exhibit 2c summarizes the total scores for all ten products over all four 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 2a, 2b and 2c as well as a few cautions should be stated. First, there is a natural tendency to compare the overall scores across the products or to rank the products by their total score. However, the ASPIRE model was not developed to facilitate such inter-product comparisons and there are some risks associated with ranking products in this manner. For one, the total score for a product reflects a weighting of the error sources by the risk levels which can vary considerably across products. Products with many high risk error sources, such as the NA, 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.

In addition, 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. 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 Section 1, 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 2a and 2b yield the following observations:

- Ratings for all ten products increased in this round, albeit some only marginally.
- For the eight products in Exhibit 2a, the overall mean quality rating increased by about 1 point as compared to almost 3 points in Round 3. For the NA products, the average increase was less than 1 point compared to almost 2 points in Round 3.
- The last row of Exhibits 2a and 2b shows the Round 3 to Round 4 changes in the overall quality ratings by product. Note that TPR and the RS improved the most. This was the result of new and innovative quality improvement initiatives that they completed in 2014. (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 2a and high risk for the NA products in Exhibit 2b.
- 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 two 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.

Cells with ratings that are high risk (i.e. shown in red) and below average for the error source (last column) could be regarded as the quality concerns. There are 13 cells in Exhibit 2a that satisfy these criteria and they are:

- frame error overcoverage and undercoverage – BR
- nonresponse/missing data – LFS and ULF/SILC
- measurement/content error – SBS, ULF/SILC and BR
- data processing error – RS and SBS
- sampling error – CPI
- model/estimation error – CPI, SBS, and ULF/SILC
- revision error – SBS.

Depending upon the available resources and the priorities of the organization, a subset of these cells should be considered for quality improvements in the coming year. Likewise, for the NA products, we recommend that high risk error sources having a score of, say, 55 or less in Exhibit 2b should be considered as high priority in the coming year.

Exhibit 2a. Product Error-Level, Overall Level, and Error Source-Level Ratings with Risk-Levels Highlighted and Comparisons to Round 3 Overall Ratings

Error Source/Product	RS	CPI	FTG	LFS	SBS	ULF/ SILC	BR	TPR	Mean rating
Specification error	N/A	**70**	**62**	70	**60**	**56**	**66**	**58**	63
Frame error	60	**66**	58	58	**60**	**42**	54	65	58
overcoverage							**58**	**66**	
undercoverage							**42**	60	
duplication							63	70	
Nonresponse error /Missing data	**60**	56	**68**	**58**	**72**	**48**	50	**64**	60
Measurement error/Content	**62**	**66**	**66**	**70**	**56**	**54**	**56**	62	62
Data processing error	**62**	**76**	**72**	**62**	**60**	50	N/A	N/A	64
Sampling error	N/A	**68**	N/A	**80**	**86**	**62**	N/A	N/A	74
Model/estimation error	**38**	**52**	**80**	**64**	**48**	**52**	N/A	N/A	56
Revision error	62	N/A	**72**	N/A	**54**	N/A	N/A	N/A	63
Round 4 mean rating	57,1	65,8	68,6	66,0	60,5	52,0	54,8	63,4	61
Round 3 mean rating (re-rated if relevant)	54,7	65,2	67,6	64,3	60,1	51,1	53,7	61,4	60
Change (improvement/deterioration)	2,4	0,6	1,0	1,7	0,4	0,9	1,1	2,0	1,3

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

Exhibit 2b. Product Error-Level, Overall Level, and Error Source-Level Rating with Risk-Levels Highlighted and Comparisons to Round 3 for the National Accounts

Error source	Quarterly GDP	Annual GDP
Input data source (Average)	**57**	**66**
Structural Business Survey (SBS)	N/A	**66**
Index of Service Production (ISP)	**64**	N/A
Index of Industrial Productions (IIP)	**64**	N/A
Merchanting Service of global enterprises	**44**	N/A
Compilation error - modelling	**50**	**50**
Compilation error - data processing	**54**	**52**
Deflation error (including specification error)	**50**	**50**
Balancing Error	**52**	**58**
Revisions Error	**58**	**56**
Round 4 mean rating	54,3	55,3
Round 3 mean rating	53,6	54,9
Change (improvement/deterioration)	0,7	0,4

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

Exhibit 2c shows the overall ratings by product for the four evaluation rounds. Note that all products have improved during the last four rounds. The mean ratings (last set of bars) show increases from Round 1 to the present round of 3.2, 2.7 and 1.3, respectively. This equates to a total

7.2 points increase from Round 1 to Round 4. However, it is somewhat disappointing to observe that the magnitude of the average increase for the current round is less than half of what it was in prior rounds, as can be seen from the last set of bars in Exhibit 2c.

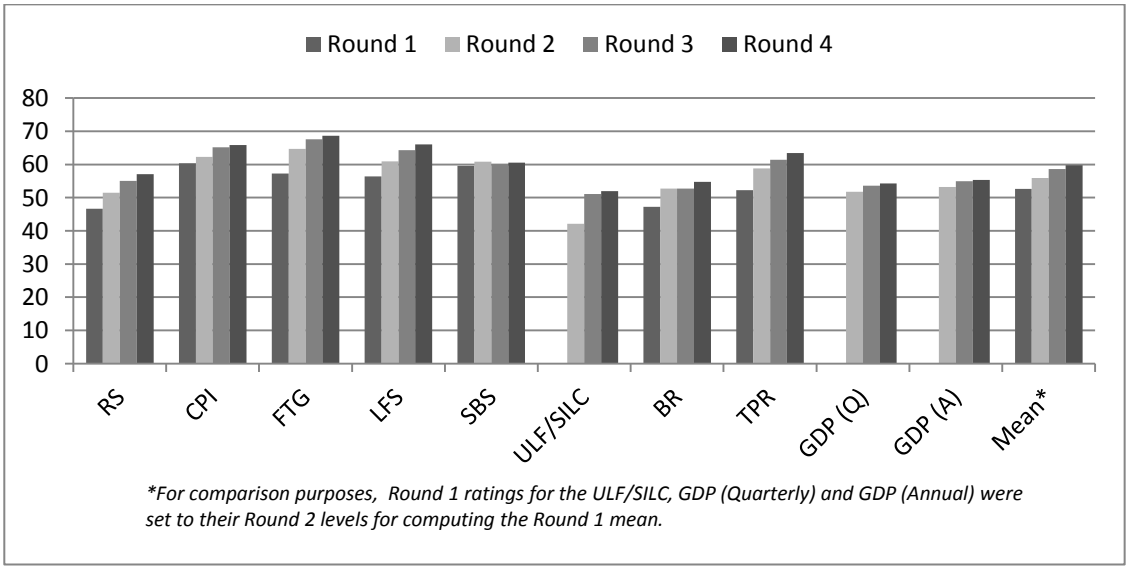
There are several possible explanations for this. One is that the so-called "low hanging fruit" of quality improvement (i.e., improvements that can be more readily accomplished with low budgets and minimal activity) was picked up in early rounds. Now, the achievement of further improvements will require 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 3 where implemented, the vast majority of them were not. In addition, there appears to be a lack of accountability regarding the treatment of the ASPIRE recommendations. Recommendation 4 in Section 4 addresses this concern.

We caution against interpreting the results in Exhibit 2c 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 five ASPIRE criteria. As previously noted, we can only say that data quality has been improved to the extent the five 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.

Exhibit 2c. Overall Quality Ratings for All Products by Round *(Note: ULF/SILC was not evaluated in Round 1. Also, the criteria for GDP (Quarterly) and GDP (Annual) were substantially changed after Round 1 so those ratings are also omitted from this chart.)*



**For comparison purposes, Round 1 ratings for the ULF/SILC, GDP (Quarterly) and GDP (Annual) were set to their Round 2 levels for computing the Round 1 mean.*

Some key observations to note when reading the product reviews are as follows:

- Work on updating the Quality Declaration (QD) documents continues to disappoint, with only a few QDs updated in the last two rounds.
- As was true in prior rounds, quality evaluation studies tend to focus only on error rates and indirect measures rather than direct error measures such as bias, validity, and reliability.
- The decline in response rates for the LFS, ULF/SILC and other household surveys continues to accelerate despite the considerable effort and resources devoted to reducing the decline.
- Much of the improvement observed in this round may be attributed to Planning/Mitigation and most of this is for planning rather than mitigation.

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. In addition, more emphasis is being given to the reduction of nonresponse biases rather than unweighted nonresponse rates.
- The sensitivity studies on input data sources in the national accounts show real promise of providing insights that may lead to improvements in the accuracy of the national accounts.
- The 'before and after' study of editing for RS provided information which will enable the staff to improve the cost-effectiveness of their data editing processes.
- The innovative study undertaken by TPR to provide information on the overcoverage areas. This also provides extremely useful information for the users of the TPR.

We should also note the successful introduction of the new European System of National Accounts (ESA 2010) in September 2014. This is a major achievement.

In the next section, we discuss the detailed ratings for all ten products individually. Detailed comments that support each rating change may be found in the rating change tables that were developed from each product. Due to their length, these tables are not provided in this report but may be obtained upon request from Heather Bergdahl.

In making recommendations for future improvements, our focus is on the areas of higher risk where the ratings are relatively low. In some cases, a recommendation from the last round is carried over to the current round due to the lack of efforts to address it and because we still consider it a priority. Other recommendations are either modified versions of the recommendations from last year to reflect the work that has been accomplished on them or new recommendations.

3.2 PRODUCT BY PRODUCT RATINGS

In this section, we review the progress over the past 12 months for eight of the ten products shown in Exhibit 1 using the checklist that appears in Annex 1. A slightly modified version was used for the two national accounts products. The ratings for each of the five 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 ANNUAL MUNICIPAL ACCOUNTS (RS)

SELECTED ACCOMPLISHMENTS

- Analysis of Editing Changes. A study was conducted that compared key RS variables before and after editing. This study quite successfully identified a number of strengths and weaknesses of the current editing approach. For example, some edited variables did not change much at all, while other variables (such as costs and revenues) experienced large changes, possibly due to erroneous inclusions. The study identified several areas where changes in the editing process are needed and these will be implemented in the coming year.
- New Transmission Interface. A new transmission interface for respondents was developed whereby respondents can upload their completed Excel forms to Stat Sweden's website rather than submitting them via email. This 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. Currently, 150 of the 290 municipalities have used the new interface; however, it will be mandatory for all 290 next year. Respondent feedback has been quite positive. Plans include expanding the scope of the feedback system.
- Improved Revision Policy. Respondents were advised that a summary of their responses will be published 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. It is also believed that the inputs have greater accuracy, although no formal evaluation of this has been done.
- Preprinted Information. Some changes were made to the preprinted information supplied to respondents to address inaccuracies in these data. As an example, rather than preprinting the estimate of the number of pupils enrolled by grade level, RS now provides the number of children in each age group which is much more accurate.
- Timeliness of Reporting. Only 30% of the municipalities send their forms before the deadline. An experiment involving 90 municipalities was conducted to increase this rate. This included contacting the tardy municipalities, reminding them of the deadline and offering assistance in completing their forms. This did not increase the timeliness of reporting but it rules out one hypothesis as to the cause of late responding. In 2011 RS studied how long it took to fill in the form – 3 weeks on average. There are ideas to reduce the burden but there are difficulties because many of the figures are not available in the accounting systems for the municipalities and county councils. The most likely cause for the delays is response burden given that municipalities say they require three weeks, on average, to complete their forms. Additional experiments are being planned to reduce response burden.

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.
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 error in 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 Round 3, we recommended a study of the allocation keys used to disaggregate common costs to various sub-activities. More than 80 percent of the municipalities allocate common costs to various activities using Statistics Sweden's automatic allocation key for common costs that is included in the form for municipal summary accounts while the remaining municipalities allocate common costs according to their own model. We repeat this recommendation, but also acknowledge that the biggest problem with cost allocation is the common costs figures that form the starting point for the allocation process. We strongly 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.
4. As noted in Round 3, one goal of the redesign was to simplify the questionnaire and to reduce some of the confusion among respondents with the old form. How well this was achieved should be evaluated. A simple indicator of the performance of the new instrument is the extent to which queries from respondents about how to complete the form have decreased after the new form was implemented. These data are currently available and it would not require much effort to tabulate and analyse them.
5. Finally, as noted in both Rounds 2 and 3, 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.

Exhibit 3. Annual Municipal Accounts (RS), Ratings for 2014

	Error Source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (Control of error sources)	**Specification error**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Frame error	60	60	○	○	◒	◒	N/A	**L**
	Non-response error	60	60	○	○	◒	○	○	**M**
	Measurement error	62	62	○	○	◒	◒	◒	**M**
	Data processing error	54	62	◒	○	◒	○	○	**H**
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Model/estimation error	38	38	◓	◓	◒	◓	◓	**M**
	Revision error	60	62	○	○	◒	◒	○	**L**
	Total score	54,7	57,1						

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

3.2.2 CONSUMER PRICE INDEX (CPI)

SELECTED ACCOMPLISHMENTS

- Revised QD. The release of an amended QD during the year which contained quite a bit of information to assess quality including sampling errors.
- Extended Use of Scanner and Internet data. As well as increasing the size of the sample in some important segments, scanner data provide prices that include discounts which are otherwise difficult to collect. "Web scraping" is also used to collect price data for some commodities although there were no significant developments in this area during the year. However, many more prices are now collected by the internet. As a consequence, more of the price data collection is now being undertaken centrally where quality is easier to manage.
- Quality Adjustments. Improved procedures for adjusting quality change are continuing to be introduced to provide better control over this important aspect of the accuracy of the CPI. Interestingly, the quality adjustments were dropped for some commodities as analysis showed the adjustments are not effective. Further improvements in quality adjustment procedures are planned for next year.
- Methodological Studies. Although there is no specific budget for methodology, a number of studies were undertaken during the year (e.g. shortfalls in coverage in products and services such as aged care services and furniture) and presented to the CPI Board. There were fewer studies than previous years because some development resources were being devoted to operations (see below).
- Hedonic Modelling. A study was done on the suitability of hedonic models to provide estimates of price increases for electronics and appliances. The work looks promising and was presented to the CPI Board.
- Constant Tax Index. At the request of power users, a Constant Tax Index has been compiled which will eventually replace the Net Price Index. This is a more accurate indicator of inflation as it does not include taxes and subsidies which are not technically part of inflation.
- Hand-held Computers. Updated hand held computers, similar to iPads, will be introduced by November 2015 following training in the previous months. It opens up a range of opportunities to improve the CPI which are discussed below.
- Selective Editing. Selective editing is going to be implemented next year as the development work has been done. The approach is to 1) focus on suspected errors, and 2) concentrate on the suspected error if only if it has considerable impact given the index. These two dimensions are combined to identify those data that most need attention.

Although this is a very respectable year's work on quality improvement, it has been constrained by reduced resource capacity. Because of the higher costs of the field work (20% increase in two years), development resources have been diverted to more pressing operational matters. Over the medium term this may have a serious impact on the quality of the CPI. Because of this, there has been a lowering of the ratings for 'Available Expertise' for a few error sources.

We believe the use of upgraded technology in the field has a number of potential benefits. The new technology will have a longer life than the previous version as well as some efficiency gains. Furthermore, it provides a number of possibilities to monitor the work better. This includes checks of whether field staff actually visit sites or not.

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

RECOMMENDATIONS

1. Redo the 1999 study on potential CPI biases as much has changed since then and CPI methods and revised procedures may mean that these biases are now different. Furthermore, if a Total Survey Error approach is taken to improve the accuracy of the CPI, this would provide the evidence base for deciding where to best place the research effort.

2. As part of this study, there should be some analysis of the sample design. The cost and error structures would have changed considerably since the last major examination of the sample structure and it may now be sub-optimal. For example, field costs have increased by 20% over the last two years. One possibility is to reduce the number of cities in the CPI given the objective is to obtain a representative sample of price relatives rather than prices. This would reduce the number of price collectors that need to be trained and equipped. Also, there is a possibility of having some price collectors focus on particular types of products.

3. 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. Furthermore, further use of the internet could be used to obtain price data. In making this recommendation we note the leadership role Statistics Sweden has been taking globally on the introduction of scanner data.

4. Research into methods for measuring quality adjustment should continue as this may well have the greatest influence on accuracy.

5. There is a lot of dependency on the work of the price collectors and their work should be routinely monitored. More up to date technology is being introduced to support data collection. The technology has the capability to collect 'operational' data as well as price data. This capability should be used to better understand the quality and effectiveness of the work of the price collectors.

6. The Household Budget Survey (HBS) has a significant influence on the weights used in the CPI. A remedy for the high sampling variability in the HBS is to average the data over three years which seems sensible. Data from other, more accurate sources, are used to derive the weights for items like tobacco and alcohol. An issue of potential concern is the increasing nonresponse rate in the HBS. A sensitivity analysis should be conducted to understand whether nonresponse bias is having a large impact on the weights and thus the quality of the CPI. The study should focus on those parts of the price regimen where price movements might be quite different to the rest of the CPI because errors in the weights may not matter for items where the relative price movements are much the same.

Exhibit 4. Consumer Price Index (CPI), Ratings for 2014

	Error source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	Specification error	72	70	◡	○	◡	◡	○	**H**
	Frame error	64	66	◡	◡	○	◡	◡	**M**
	Non-response error	56	56	○	◓	◡	◡	○	**L**
	Measurement error	68	66	◡	○	◡	○	○	**H**
	Data processing error	76	76	◡	○	◎	◡	◡	**H**
	Sampling error	70	68	◡	◡	◡	◡	○	**H**
	Model/estimation error	44	52	○	○	○	◓	○	**H**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total Score	65,2	65,8						

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

3.2.3 FOREIGN TRADE OF GOODS (FTG)

SELECTED ACCOMPLISHMENTS

- Statistical Value Survey. Every five years, the FTG staff conduct a survey of enterprises that will help to recalibrate the models used to relate invoice values to statistical values. This survey was completed in November 2013 and the new adjustment factors were implemented for both Intrastat and Extrastat in March 2014.
- SIMSTAT. The report titled "Use of administrative auxiliary information to improve the quality with respect to a future SIMSTAT system in SE" was published. This project was primarily focused on the future use of SIMSTAT data for estimating Intrastat trade. It also aimed to advance the work on asymmetries between Statistics Sweden and EU countries.
- Estimation below the Cut-off. The report titled "Improvement of the SE estimated Intrastat data by adding estimations based on VIES data" was published. This project considered using both the VAT and VIES data for estimating trade data for companies below the threshold for inclusion in the survey. This work is important because, beginning in 2015, the new threshold for Intrastat imports will be doubled to SEK 9.0 million and there has been no other evaluation of the impact.
- Sea Products. The report titled "Improve and develop new routines for data collection regarding the area Sea products within Specific movements of goods" was published. The work described approaches for obtaining better coverage of these shipments.
- Collaborative Meetings with the NA Staff. The FTG staff continued to hold regular meetings with the NA staff in conjunction with the FTG quarterly reports. These meetings have led to better understanding of the issues in the FTG that have an important impact on the NA, and effective means for addressing them. A key topic of these meetings is revision error in the FTG statistics and their impact on the national accounts. FTG staff will be involved in the balancing of NA estimates.

RECOMMENDATIONS

1. Raising the cut-off threshold will reduce respondent burden and allow the FTG staff to concentrate on obtaining responses from larger enterprises. However, there is also an increased residual risk of Model/Estimation error. The latter should be investigated in the coming year including how this error might be minimised.
2. Along these same lines, investigations of the use of VAT and VIES data in the estimation of trade below the cut-off threshold should continue. In particular, large revisions resulting from this estimation process should be investigated to better understand the causes of the revision and how to avoid them in the future.
3. Continue the close cooperation between FTG and NA. FTG staff should become more familiar with the process generating the GDP estimates and how their data are being used.
4. 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.

5. The planned measurement error studies, comparing VAT and Intrastat data, are strongly encouraged.
6. As we noted last year, the QD should be updated to include the findings of the many research studies that have been undertaken. It should also speak more directly regarding the size of revision error and its effects.

Exhibit 5. Foreign Trade of Goods (FTG), Ratings for 2014

	Error Source	Average Score round 3	Average Score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control for error sources)	**Specification error**	58	62	○	○	◡	◡	◡	**M**
	Frame error	64	58	○	○	◡	◡	○	**L**
	Non-response error	68	68	◡	◡	◡	○	◡	**M**
	Measurement error	64	66	◡	○	◡	◡	○	**H**
	Data processing error	66	72	◡	◡	◡	◡	◡	**H**
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Model/estimation error	82	80	◡	◡	◎	◎	◡	**M**
	Revision error	72	72	◡	○	◡	◡	◡	**H**
	Total Score	67,6	68,6						

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

3.2.4 LABOUR FORCE SURVEY (LFS)

Response rates for the LFS have continued their downward trend and are now about 65 percent. It accelerated somewhat over the last 12 months even though considerably more resources were devoted to interviewing. There are important cohort effects in the rate of decline. We believe this decline is due partly to societal and environmental factors and partly to organizational and workplace issues. As an example of the latter, an analysis conducted as part of the Nonresponse Project suggested that at least two-thirds of the contact attempts should be made during evenings and weekends to maximize the probability of a successful contact. However, this is not current practice because organizational issues make this very difficult. To circumvent some of the internal issues, the LFS is conducting an experiment to determine whether outsourcing data collection could increase response rates.

However, other research conducted as part of the Nonresponse Project suggests that increasing response rates may not have much impact on nonresponse bias. Consistent with our recommendations in Round 3, the Nonresponse Project is beginning to address questions regarding the bias due to nonresponse given the current response rates and how to minimize the effects of nonresponse on the estimates. We applaud these efforts and encourage more research along these lines in the coming year. We also applaud the efforts to consider a web questionnaire option as part of the LFS and other household surveys. While this may not increase response rates dramatically, it could result in a more robust sample of interviews that is less subject to nonresponse bias or that could be more effectively calibrated to mitigate bias. There should also be some savings in data collection costs.

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

SELECTED ACCOMPLISHMENTS

- Nonresponse Project. In 2013, a project was launched primarily focused on the LFS to increase response rates, reduce data collection costs and achieve greater control over the field work. The project has led to a better understanding of the factors associated with the declining response rates in the LFS. However, so far, the response rate decline has been unabated.
- Outsourcing Data Collection. The LFS staff has been quite involved preparing for the outsourcing of a total of 5300 LFS cases which will take place over six months beginning in January. This is a test to determine whether an external company with more flexible personnel management practices can achieve higher response rates than Statistics Sweden. Plans are in place to evaluate the impact on LFS on a reasonably continuous basis.
- Evaluation of Nonresponse Bias. As part of the Nonresponse Project, some innovative work looking at nonresponse bias in the post-adjusted LFS estimates has been conducted. The results suggest that for one variable (viz., income from work), nonresponse follow-up beyond 20 days does not reduce the nonresponse bias but increases costs substantially.
- Mixed Mode Research. Work has begun to explore the possibility of mixing web data collection with telephone interviewing in the Party Preference Survey with possible applications to the LFS.
- Reinterview Report. A report documenting the results of the 2013 reinterview study was completed. The results are expected to lead to future changes in the LFS questionnaire.

- Industry and Occupation Coding. A number of improvements have come about as the LFS works to comply with the ISO 20252 certification standards. One ongoing project is the verification of industry and occupation codes which has led to continuous improvements in the Industry and Occupation coding process.

RECOMMENDATIONS

1. Work should continue on the evaluation of residual nonresponse bias in the adjusted LFS estimators, particularly labour force estimates. This work should focus on quantifying the residual (i.e., nonignorable) nonresponse bias, understanding its major determinants, and reducing the bias. Bias reductions should be approached by more effective weighting adjustments as well as by more targeted nonresponse follow-up that increases the weighted response rates for population subgroups that contribute most to the bias.
2. As noted in the prior round, the call monitoring system should be evaluated for its impact on cost, respondent burden and data quality. Among other things, this evaluation should investigate whether and how call monitoring could be done less obtrusively and without informing the interviewers that they have a high chance of monitored.
3. We are particularly pleased that mixed mode (i.e., telephone/web) data collection is being considered for household surveys, including the LFS. While there are many issues to be resolved in moving the LFS to a mixed mode, this research is essential to meet future challenges to using the telephone exclusively in household surveys and this should be given a high priority.
4. While the results from the reinterview study are useful on their own, much more insight into the levels and causes of measurement errors could be gained if these results were combined with those of Karlsson's Markov latent class analysis. This combined analysis would not only add to the knowledge of labour force misclassification, but also of the methodology of combining reinterview and MLCA-based estimates of classification errors. We support the planned cognitive studies into the misclassification of labour force status, especially of the 'Not in the Labour Force' classification.
5. Although there have been some positive developments in the data collection area (both centralized and decentralized operations), there are still many problems needing resolution. For example, new metrics are being developed to gauge the efficiency and effectiveness of current staffing and call scheduling practices in the centralized facility. However, these have not yet been implemented. In addition, similar metrics do not exist for decentralized interviewing. Our current thinking is that the entire approach to telephone data collection should be considered for re-engineering, possibly in consultation with an external expert on large-scale telephone operations.

Exhibit 6. Labour Force Survey (LFS), Ratings for 2014

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

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

3.2.5 STRUCTURAL BUSINESS STATISTICS (SBS)

There have been some improvements in the Structural Business Statistics (SBS) over the last 12 months but, as noted below, some areas of deterioration due to BR deficiencies.

SELECTED ACCOMPLISHMENTS

- Electronic Data Transfer. Electronic data transfer of SBS questionnaires from respondents has continued to increase and is now about 95%. This should lead to higher quality statistics (e.g. inbuilt errors) although it has not yet been proven.
- Nonresponse Penalties. The threat of fines has successfully increased response rates. The focus has been on larger enterprises who had not responded for five years. Response rates are now at their highest ever level.
- Questionnaire Appraisal. The cognitive laboratory has been used to better understand measurement error on the SBS questionnaire. Changes to the instructions have been made.
- New Production System. Work has commenced on the development of a new meta data driven production system that will also be used for PRODCOM and two minor surveys that feed into the national accounts.

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 BR and Large Enterprise Unit 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 3.2.7), a modelling approach may be needed in order to achieve this goal.
2. Although the statistical improvements in the BR have been delayed indefinitely, 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 on 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 evaluate editing. 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 and gave them useful insights into the effectiveness of their editing.
4. As noted above, the number of questionnaires collected electronically has increased. Studies usually show the data are different to those collected through traditional mail questionnaires but there is no proof that the accuracy has improved. It would be expected that accuracy would be improved but there should be a research study to demonstrate that this is the case.

Exhibit 7. Structural Business Statistics (SBS), Ratings for 2014

	Error Source	Average Score round 3	Average Score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	**Specification error**	58	60	○	○	◡	○	◡	**M**
	Frame error	60	60	◡	◡	◡	○	◠	**M**
	Non-response error	70	72	◡	○	◡	◡	◡	**M**
	Measurement error	56	56	○	○	○	○	◡	**H**
	Data processing error	60	60	○	○	◡	○	◡	**H**
	Sampling error	86	86	◎	◡	◡	◎	◎	**M**
	Model/estimation error	48	48	○	○	○	◠	◠	**H**
	Revision error	54	54	○	○	◡	◠	◠	**H**
	Total score	60,1	60,5						

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

3.2.6 LIVING CONDITIONS SURVEY (ULF/SILC)

SELECTED ACCOMPLISHMENTS

- New System for Quality Assurance (QA). In compliance with the ISO 20252 standards, a new QA system was put into place in the Data Collection Department to guide the myriad of processes associated with the ULF/SILC. This system promises to increase standardization, reduce process variation and increase the quality of the ULF/SILC operations.
- Shorter Questionnaire. A new, shorter instrument that was revised based upon a thorough cognitive evaluation, was implemented in the ULF.
- LCA. A latent class analysis of the SILC labour force question was conducted. The results agreed closely with the results of a Markov latent class analysis of a similar question on the LFS.
- Estimation Improvements. Some improvements were made to the calibration model (viz., adding age by sex interactions and a new, improved income measure).
- Sample Size. The sample was increased from 19,000 to 19,400. However, there are important concerns that the size of the children cohort may be inadequate because 10-11 year olds were dropped from the survey.

As noted in Round 3, the ULF/SILC has been undergoing some important changes over recent years in order to simplify a very complex survey. Furthermore, some changes have been mandated by the EU and further changes are planned. The EU is requiring that the number of interview waves be increased from four to six. Containing the attrition bias as the number of interview waves is increased will be a challenge as attrition at waves 3 and 4 is already an important concern. The EU is also requiring that the micro-data be delivered in December of each year which is some months earlier than it is currently delivered. But because data collection continues throughout all months of the year, to deliver in December would require a considerable change to the interview calendar. Current plans are to conduct the survey over the first 6 months of the calendar year but this is not without serious quality concerns. With a data collection concentrated in the winter and spring, seasonal effects could bias estimates of health issues, leisure activities, and other behaviours and conditions that change by season.

In light of these concerns and issues, we have the following recommendations.

RECOMMENDATIONS

1. As discussed above, a key concern for the ULF/SILC is the change in the data delivery schedule being imposed by the EU. In addition to concerns regarding seasonal effects on the estimates, the increased workload in the call centre could affect response rates and response quality. The product staff admit that they have not yet looked at the redesign issues and that these should receive a high priority in the coming year.
2. The nonresponse rate for the survey continues to increase. While the Nonresponse Project searches for a solution, the ULF/SILC staff should analyse the nonresponse bias in the final, adjusted estimates. The evaluation should focus in part on the efficacy of the nonresponse adjustment procedures, the choice of auxiliary variables in the adjustment process, the GREG modelling approach, and the potential for new calibration methods that adjust for nonignorable nonresponse to reduce the bias. Changes to the CATI system to facilitate adaptive design approaches should also be developed and implemented.

3. There is a need for much better documentation of the sample design and the weighting methodology for both the ULF and the SILC. This task should receive high priority in the coming year.
4. The ULF/SILC staff expressed concern over the children's survey – the sample size, the questionnaire and the very low response rate. The need for a redesign of the children's survey should be evaluated in the coming year.
5. Although telephone monitoring has been implemented, it has yet to be used as a tool for improving data quality. The potential for telephone monitoring to improve interviewing technique, reduce interviewer variance, identify problem questions, and understand respondent concerns regarding key questions has not been exploited. More effort should be devoted on how to make the best use of monitoring results for improving data quality.

Exhibit 8. Living Conditions Survey (ULF/SILC), Ratings for 2014

Error Source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Specification error	58	56	○	○	◡	○	○	**M**
Frame error	42	42	◠	◠	◡	○	◠	**M**
Non-response error	46	48	○	◠	○	○	○	**H**
Measurement error	52	54	◠	◠	◎	○	○	**H**
Data processing error	50	50	○	◠	◡	○	○	**L**
Sampling error	62	62	◡	◡	◡	○	○	**M**
Model/estimation error	50	52	○	◠	◡	○	○	**H**
Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
Total Score	51,1	52,0						

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

3.2.7 BUSINESS REGISTER (BR)

SELECTED ACCOMPLISHMENTS

- New BR System. 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 "statistical" BR which would cover the areas impacting the accuracy of the BR for statistical requirements. There will be a transition to a new master frame commencing in 2015. The new BR system will have greater flexibility especially with respect to modifying its contents. Also, the new BR system will help to further reduce the number of enterprises with missing NACE codes.
- Missing NACE Codes. As a result of work done with the Swedish Tax Board, work has been done on reducing the number of enterprises with missing NACE codes. As a consequence, the number of enterprises without a NACE code has continued to decline.
- New Institutional Sector Classification. A more reliable institutional sector classification, based on new Eurostat standards (ESA 2010), has been introduced.
- Quality Metrics. Work has begun on the development of quality measures which will help users better understand the changes in the BR and how this might impact on their statistics. Initially statistics on changes in the BR will be provided quarterly two weeks prior to statistical frames being prepared.
- Internal User Group. The User Group for internal users has been reactivated. This should be a positive step as communications with users have not been as good as they should be during recent years.
- Methodologist Availability. There has been an increase in the methodologist hours for working on the Business Register.

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 which seem to have continued to deteriorate since our last review. Specifically, the number of inactive units on the Register seems to be increasing despite the efforts to reduce them. There are also concerns about the extent of inaccurate NACE codes which has also likely increased. Both of these seem to be causing problems to the statistical areas in our review that use the BR. Overcoverage and NACE classification errors are the same two problems we referred to in the last three years and we are not convinced that sufficient action has been taken yet to address them, especially the former.

As mentioned above, there will be a transition to a new master frame, utilising the new IT system, starting early 2015. There will be a parallel run with the existing system to reduce the risk of errors during the transition period which we agree is a good approach. However, it was not clear that any transition impacts on surveys have been thought through nor has any thought been given as to how any important effects should be handled.

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, dependent survey feedback has been done on a survey by survey basis rather than in a coordinated way. The use of survey dependent feedback may be fine for the 'take all' strata but is potentially biased for sampled strata.

A recent Stat Sweden study has shown that this practice does indeed bias survey estimates and has 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.

However, the biggest concern seems to be the significant and continuing reduction in the number of kind of activity units (KAUs). At present, slightly more than 40 enterprises are being profiled. The Large Enterprise Unit advised us that there were about 100 enterprises they would ideally like to profile. This is causing a loss of accuracy of industry coding in important collections like SBS and consequently the NA. We heard about one very large (unprofiled) enterprise that substantially changed the quarterly GDP estimates due to a change in classification from the manufacturing industry to the service industry. The enterprise was a major contributor to value added in both sectors. This is a consequence of structural changes within the enterprise that have accumulated over time which suddenly became evident. If this industry change had been made, it would cause a major disruption to a number of economic series including the national accounts. This would not have been the case if the enterprise had been profiled as both industries would have been recognised as separate KAUs.

One difficulty seems to be the very stringent Eurostat data availability standards for determining when new KAUs should be formed. These standards are being reviewed but we suggest, in the interim, that a modelling approach 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 all of the business is part of a single industry and is likely to be more consistent with a revised Eurostat standard. It was also suggested that the system used by The Netherlands for identifying which enterprises should be profiled could be adopted.

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. A report has been prepared on a Common Business Framework but no decisions have been made on the basis of that report.

RECOMMENDATIONS

1. The procedures used by the Large Enterprise Unit (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 especially the very largest and complex enterprises. One possibility is the modelling approach described above but the responsibilities for decisions on modelling need to be determined. In

addition, the LEU should also consider alternate strategies for gaining cooperation of large enterprises to be profiled. As an example, for some critical enterprises, a letter from the Statistics Sweden Director General to the enterprise CEO requesting cooperation with the profiling task could be highly effective.

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, identifying locations in new multi-establishment enterprises (undercoverage), and developing a Common Business Framework (a documented proposal exists).
3. Furthermore, the new Business Register System should support the creation of a BR specifically for statistical purposes.
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 or an approach similar to what was used in the LFS. 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.
5. A Stat 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.
6. The reactivation of the internal User Group is a positive step; however, it should meet on a regular basis. Currently, its work is at the operational level. It should be supported by a higher level strategic group that would meet relatively less frequently during the year.
7. There should be some analysis of whether there are any statistical impacts from the transition to the new BR. If so, there needs to be further analysis of how they are best managed.

Exhibit 9. Business Register, Ratings for 2014

	Error Source	Average Score round 3	Average Score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plan or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	**Specification error**	66	66	○	○	◡	◡	◡	**M**
	Frame error - overcoverage	58	58	○	○	◡	○	○	**H**
	Frame error - undercoverage	42	42	◠	◠	○	○	◠	**H**
	Frame error - duplication	63	63	○	○	◡	◡	N/A	**L**
	Missing data	48	50	○	○	○	○	○	**L**
	Content error	52	56	○	○	◡	○	○	**H**
	Total score	53,7	54,8						

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

Note that the "risk to data quality" for Specification Error was increased from "L" to "M," hence the yellow shading. This change is due to a correction rather than a change in the actual intrinsic risk.

3.2.8 TOTAL POPULATION REGISTER (TPR)

SELECTED ACCOMPLISHMENTS

- Overcoverage. We were very impressed with the innovative work that has been completed to estimate overcoverage at the micro level, which has been steadily increasing over the years. This study combined multiple indicators of nonresidency to form a theoretically more accurate indicator of nonresidency for each person who is a "suspected" nonresident subset of the TPR. The research is important for several reasons. First, it provides a propensity of being a nonresident at the individual level. The sum of these propensities is the overcoverage estimate. Second, it can be used to characterize the overcovered population. As an example, the method produced an estimate of TPR overcoverage of about 85,000 in 2014, the bulk of which appears to be relatively new immigrants from outside the EU region. The propensity to be a nonresident might also be used in current surveys to adjust for nonresponse. For example, 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. The TPR group noted that the LFS and the Tax Board are quite interested in these results.
- Dwelling Unit Classification. The number of persons having a missing dwelling unit number was reduced from 304,000 last year to 274,000 this year. This is owed to the Tax Board who now requires movers to report of the dwelling number. Approximately, 1.3 million TPR entries are classified as movers each year.
- Error Evaluation. The rate at which a given field on the TPR is corrected based upon information provided by the Tax Board can provide an indicator of the error rate for that field. The TPR has been accumulating this information and will publish these results in the next QD.

RECOMMENDATIONS

1. We encourage the TPR staff to continue their research on overcoverage. Two methods that they may consider for estimating overcoverage propensity are logistic regression and latent class analysis. The former approach, which is currently being explored by the staff, could address some of the problems they have with the current methodology that provides inaccurate estimates for the current year. The latter method is ideal for combining multiple indicators of nonresidency into a single indicator that has been corrected for measurement error.
2. As we noted last year, it is also important to understand what level of overcoverage is tolerable for most users of the TPR. This requires working with subject matter staff that represent the main user groups to understand the effects of overcoverage on key population estimates such as the unemployment rate. The ability of the "resident propensity" indicator to mitigate overcoverage error in the key population estimates should be part of this analysis. The establishment of an internal user group may assist with these deliberations.
3. Studies should be mounted that evaluate the validity 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 10. Total Population Register (TPR), Ratings for 2014

	Error Source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control for error sources)	**Specification error**	58	58	○	◔	○	○	◒	**M**
	Frame error: overcoverage	58	66	◒	◒	○	◒	○	**H**
	Frame error: undercoverage	60	60	○	○	◒	◒	N/A	**L**
	Frame error: duplication	70	70	○	○	◒	◒	N/A	**L**
	Missing data error: item and variable	66	64	○	○	◒	○	◒	**M**
	Content error	62	62	○	○	◒	◒	○	**L**
	Total score	61,4	63,4						

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

3.2.9 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 the previous round, we could only look at a small number of the data sources that provided the greatest risk to the accuracy of the NA products and GDP in particular. We also only looked at the production side of the quarterly GDP. Last year, 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) the survey of foreign trade in services which provides estimates of merchanting services as well as some other data that are used in the quarterly GDP estimation process.

SELECTED ACCOMPLISHMENTS

- ESA 2010. The ESA 2010 was successfully introduced for Q2 and the series backcast to the early 1990s. This was a major achievement. The main change in the Swedish context was Research & Development Expenditure which added considerably to the size of GDP (the highest in Europe in percentage terms).
- Harmonisation. Further work has taken place on the harmonisation of the industrial and services production indexes and the harmonised survey is planned to be introduced in 2015.
- NA Processing System. It has been decided to look more closely at the Finnish NA approach to process the national accounts but not to actually take over their system. It is an area of high priority because support for the existing system is becoming more difficult and VB6 needs to be phased out. Work has commenced but there does not seem to be a fully agreed work plan at this time so it is an area of risk.
- Estimates of Intermediate Consumption. 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 led to a funding proposal to conduct a quarterly SBS for the largest enterprises to be supplemented by VAT data for the smaller enterprises.
- Constant Price Estimates. Measures were developed for improving the constant price estimates for FISIM and inventories.
- Sensitivity analysis. Work on the recommended sensitivity analysis has commenced and 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.
- Flash Estimates. Some measures have been taken to reduce revisions of the flash estimates produced for Q2 to assist the government's budget process. This involves obtaining earlier responses from enterprises. The subsequent revisions were a little smaller as a consequence.
- Productivity Measures. Inconsistencies appear in the NACE coding in the NA and the LFS because the LFS is employee-based and the NA production data is enterprise-based (and may be impacted by inaccuracies in NACE codes on the BR). An exploratory study is being undertaken to see if the coherence can be improved because the LFS is the only source for hours worked. The objective of the study is to improve quarterly industry productivity measures. This stage of the work will be completed at the beginning of 2015.

We have supported the development of standardised or objective principles and methods for balancing the quarterly GDP estimates recognising there will always be an element of human judgment involved in the balancing process. The principles and methods used to date have not worked as well as hoped and are being reviewed which we support.

We were shown graphs that indicated that there seemed to be a systematic difference between the quarterly GDP series and the equivalent series after it was benchmarked to annual GDP a few years later. The graph suggested that, in recent years, there was a positive bias in the quarterly GDP estimates prior to the annual benchmarking. It had occurred since the last peak in growth at the end of 2010. Possibilities are:

- Single deflation rather than double deflation is used for much of the industry data in the quarterly accounts and the normal assumptions may not hold when GDP growth rates are declining.
- Different statistical units may be used in the main annual and quarterly data sources, especially given the decline in the number of KAUs in recent years.
- The treatment of frame deficiencies (for example, inactive units) may vary between collections.
- The treatment of nonresponse may differ between collections.

The data for the computer manufacturing and computer services industries has been very volatile and the seasonal pattern was unrealistic. The quarterly pattern has been reviewed and revised to provide a more plausible series. However, the pre-study for the sensitivity analysis (described in the review summary for GDP(A)) shows there may be significant problems with deflation which is dampening (already high) growth.

RECOMMENDATIONS

1. There is a need for a robust processing system for the NA estimates that includes time series dimensions and this should have a very high priority. It is actually not clear at present regarding the degree to which the Statistics Finland System can be adapted to serve the Swedish National Accounts even though the objective is to use it as a prototype. The risks associated with this systems redevelopment are therefore higher than what we first understood. There are distinct advantages in physically locating the IT staff with NA during this period to minimize communication issues.
2. If funding is received, focus on the successful implementation of the quarterly SBS supplemented by VAT data to obtain better estimates of intermediate consumption.
3. Review the methodology for estimating merchanting services.
4. There is relatively high turnover of staff in the National Accounts. Given this, there needs to be more formality in the training 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.
5. More sophisticated models should be developed for quarterly Research & Development Expenditure given its significance. These may require some additional data collection.
6. Develop an understanding of the reasons for the systemic differences in the quarterly GDP series and the equivalent series when it is benchmarked to the annual series a few years later.

7. We strongly recommend the continuing funding of the sensitivity studies (see review of annual national accounts).
8. Review the producer price indexes used for deflating the computer industries based on information gained from the sensitivity studies.

Exhibit 11. Quarterly GDP, Ratings for 2014

	Error Source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans and Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	**Input data source - Index of Service Production, ISP**	62	64	○	○	◒	◒	◒	**H**
	Input data source - Index of Industrial Production, IIP	62	64	○	○	◒	◒	◒	**H**
	Input data source - Merchanting Service of global enterprises (also covers royalties, licensing and R&D)	44	44	◓	○	○	◓	○	**H**
	Compilation error (modelling)	48	50	○	○	○	○	◓	**H**
	Compilation error (data processing)	52	54	◒	○	○	○	○	**H**
	Deflation error (including specification error)	48	50	◓	◓	◒	◒	◓	**H**
	Balancing Error	56	52	○	○	○	○	○	**H**
	Revisions Error	58	58	◒	○	◒	○	○	**M**
	Total score	53,6	54,3						

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

3.2.10 ANNUAL GROSS DOMESTIC PRODUCT (GDP(A))

As is the case with the quarterly GDP, the annual GDP estimates are very complex and rely on many input data sources from both within Statistics Sweden and from external sources. For our review, we only looked at the SBS as an input data source which was deemed to provide the greatest 'data source' risk to the annual NA estimates and GDP in particular. However, as noted below, producer prices may be another area of high risk.

SELECTED ACCOMPLISHMENTS

- ESA 2010. The successful implementation of the ESA 2010 was a major achievement although these efforts may have constrained efforts in other high priority areas.
- NA Processing System. It has been decided to look more closely at the Finnish NA approach to process the national accounts but not to actually take over their system. It is an area of high priority because support for the existing system is becoming more difficult and VB6 needs to be phased out. Work has commenced but there does not seem to be a fully agreed work plan at this time so it is an area of risk.
- Standardised Spreadsheets. The completion of the standardised spreadsheets in 2013, together with the locking of some ratios embedded in spreadsheets this year, will further reduce the risk of processing error.
- Sensitivity Analysis. Work on the sensitivity analysis has started with a pre-study and will continue during 2015. There will be a particular focus on deflation error.
- Relationship with Data Providers. Seminars were conducted by NA with a number of data providers to enable them to better understand how the National Accounts use their data.

Last year, it was intended to use the Construction Industry part of the SBS for construction industry estimates in the annual national accounts. This was not possible due to a number of problems including uncertainty around the treatment of joint ventures as well as the non-coverage of co-operative housing associations. These are out of the scope of the SBS due to the fact that they are not-for-profit.

There is some pressure to improve the timeliness of the annual accounts so that the benchmarking of the quarterly accounts can start earlier and prior to the quarterly accounts being used in the government's annual budget process.

It is not always easy to understand the impacts on the NA of inaccuracies in the source data especially given the complexity of the processes used included the balancing processes. We have been recommending the use of sensitivity studies, where an error is introduced into a particular data source and the impact on GDP is assessed, as a way of providing insights. We are pleased to see that a pre-study has commenced which also includes some work on understanding the relative importance of errors in the different variables used in the compilation of the national accounts. For example, the pre-study has shown the importance of the deflation of the rapidly growing computer manufacturing and computer services industries.

In making suggestions on areas for future improvements, the focus should be on the areas of higher risk where the ratings are relatively low. We offer the following suggestions. Most of the recommendations are modified versions of the recommendations from last year. The modifications are largely because work has commenced on addressing the previous recommendations.

1. There is a need for a robust processing system for the NA estimates that includes time series dimensions and this should have a very high priority. It is actually not clear at present regarding the degree to which the Statistics Finland System can be adapted to serve the Swedish National Accounts even though the objective is to use it as a prototype. The risks associated with this systems redevelopment are therefore higher than what we first understood. There are distinct advantages in physically locating the IT staff with NA during this period to minimize communication issues.
2. There should be some evaluation of the models used for estimating the trade margins which appears to be the area of greatest weakness in modeling.
3. A pre-study of sensitivity studies on errors in the indexes used for deflation (especially the producer price indexes where the samples are relatively small) is underway. The pre-study appears to have identified some very worthwhile areas for further research and practical implementation, covering both deflation and balancing. This additional work should be identified and a funding proposal prepared. The focus should be on identifying those areas of the national accounts where quality improvement effort is best concentrated.
4. There is relatively high turnover of staff in the National Accounts. Given this, there needs to be more formality in the training making greater use of new technologies. 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.

Exhibit 12. Annual GDP, Ratings for 2013

	Error Source	Average score round 3	Average score round 4	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	**Input data source - Structural Business Statistics, SBS**	66	66	◡	○	◡	◡	○	**H**
	Compilation error - modelling	50	50	○	◠	○	◡	◠	**H**
	Compilation error - data processing	52	52	◡	○	○	◠	○	**H**
	Deflation error (including specification error)	48	50	◠	◠	◡	◡	◠	**H**
	Balancing Error	58	58	○	○	◡	◡	◠	**H**
	Revisions Error	56	56	○	◡	○	○	○	**M**
	Total score	54,9	55,3						

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

4. GENERAL RECOMMENDATIONS

In previous reports we have had a relatively long list of general recommendations. Although most of the recommendations would still be valid, we have taken a different approach for this report and focused on a smaller number of what we believe are the most important recommendations.

1. *Opportunities to Improve the Quality and Cohesion of Economic Statistics*

A number of initiatives could be taken to improve the integration of economic statistics. These are listed below and discussed in more detail in the following paragraphs.

- Undertaking the systems development work required to enable the new BR system to separately recognize a Register designed to best meet statistical purposes.
- Undertaking the systems development work to enable the BR system to address the major quality concerns such as eliminating inactive units causing an overcoverage problem.
- Evaluating the accuracy of the NACE coding and taking the necessary steps to address the most important deficiencies in NACE coding.
- Reducing the variation in the (business) units used in the business surveys especially those surveys that contribute to the national accounts.
- Establishing a Common Business Framework (CBF) for both the quarterly and annual surveys especially those that contribute to the national accounts.
- Ensure that all the largest and most complex enterprises are profiled so that significant industry activities within the enterprise are identified.
- Ensure that methodological decisions such as the treatment of nonresponse are performed on a consistent basis.
- Put in place the governance arrangements so that strategic decisions that cut across Departments can be made.
- After the current work on establishing a revision policy is finalised, put in place arrangements to ensure that it is applied uniformly across Statistics Sweden.

The lack of coherence in economic statistics, and whether it is changing, can be assessed to a large extent by looking at trends in the balancing item of the quarterly and annual national accounts prior to the balancing taking place.

The new Business Register System should support the creation of a BR specifically for statistical purposes. At present the main objective is to maintain a register of all currently registered enterprises and this is what is to be implemented when the new BR system is launched in early 2015. The statistical uses of the BR suffer as a consequence as there will be many registered but inactive units on the Register. It should be possible to create a Statistical Register (as a subset) with the additional flexibility in the new Business Register System.

Overcoverage, because of inactive units, is a growing problem. The efficiency of the sample designs is affected. Furthermore, the survey area will not know whether a non-active unit is a nonrespondent (where imputed values will be used) or inactive (where zero will be imputed). There is enough information available from the various registers to determine whether an enterprise is inactive or not but this information cannot be used at present because the functionality to do so does not yet exist in the BR system.

There is likely to have been continued decline in the level of error in NACE coding and certainly this is the feeling of the areas using the BR. A strategy for addressing the most important inaccuracies in the NACE codes should be developed (refer section 3.2.7 for more detail).

There are many differences in the sample units used by the various collections. No doubt, the collections have chosen the units which they believe are most appropriate for their collections. However, a set of optimal local decisions does not necessarily result in the best solution for Statistics Sweden especially with respect to the national accounts and is likely to be a major source of incoherence. The units used by each of the collections should be reviewed and desired changes determined.

A Report has been prepared on the establishment of a Common Business Framework (CBF). However, no decisions have been made on the recommendations yet.

As discussed under the BR review, the procedures used by the Large Enterprise Unit (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 especially among the very largest and complex enterprises. A different approach is required as the existing Eurostat rules have shortcomings. If full data is not available for the desired activity units, partial data should be obtained so that Statistics Sweden can use a modelling approach to impute industry dissections. More details are provided in section 3.2.7.

There are a range of methodological decisions involved in sample surveys. For example, key decisions are the treatment of nonresponse and outliers. If these decisions vary somewhat from collection to collection, this will impact the coherence of economic statistics and the national accounts in particular. In the 1990s, there was an extensive study in the Australian Bureau of Statistics of the reasons for incoherence in the national accounts and methodological differences were a major explanation. The main issues were the treatment of weaknesses in the Business Register (for example, inactive units) and the treatment of nonresponse.

There are several departments involved in the development and production of economic statistics. Whilst this has many advantages, it can cause problems when decision making on strategic issues cuts across these departments. The responsibility for the final decision is not always clear. As a consequence, many important matters are left without a resolution. There is a need to establish governance arrangements that overcome this problem. The BR User Group has been reactivated and this is a positive step but this is a discussion forum and the participants are not senior enough to decide on the most strategic matters (but they can certainly inform the discussion). The strategic decision making group needs to be at the departmental head level but someone needs to be assigned the decision making authority, perhaps the new Deputy Director-General.

It is pleasing to see that work on a revisions policy is well-advanced. Again it is important that a consistent approach is undertaken across statistical collections and they are synchronized with the revisions policies of the national accounts. A particular issue is the treatment of discontinuities whether due to changes in standards, methodological changes or a major redesign. As stated in previous reports, we suggest the Statistics Sweden policy specify that for every major change of this type there be some provision for bridging the series before and after the change unless an explicit exemption is granted by the Director General. In some cases it will be necessary to splice (for example, CPI) or backcast the series (for example, the NA estimates).

2. Managing Increasing Nonresponse Rates in Household Surveys

Since our last report, response rates for household surveys have continued to deteriorate (at a slightly accelerated rate) despite the very significant efforts devoted to ameliorate 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 which can result in small nonresponse bias despite high levels of nonresponse.

In its efforts to address the nonresponse problem, Statistics Sweden has devoted considerable resources to increase response rates, particularly for the LFS. This has been at the expense of the budgets of other product areas and we have concerns that, from a total error perspective, the quality of other products may be adversely affected as a consequence. In the recent review, we heard about some examples where this could be the case.

Increasing nonresponse is a global problem, not just particular to Statistics Sweden and, to a large extent, reflects changes in society. A new approach is needed that focuses as much on reducing nonresponse bias as it does on reducing the declining response rates. Striving to achieve the highest response rates possible within available resources is certainly a reasonable goal; however, attempting to elevate response rates to their 2008-2010 levels may be futile and may not even result in significant bias reduction. Many countries are focusing on a mixed-mode approach – that is, combining telephone interviewing with web-based questionnaires. This has been shown to increase the diversity of the responding sample which can result in smaller differences between respondents and nonrespondents.

Statistics Sweden has the comparative advantage over most countries in that much information on nonrespondents is available from its administrative registers. This enables nonresponse bias to be more 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.

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. It needs to be more subtle than simply increasing the response rates. Statistics Sweden has devoted considerable resources to researching the nonresponse and much valuable information has been obtained. As we did in Round 3, we again attended a presentation by the current Nonresponse Project team to be debriefed on the Project. In contrast to our reaction in Round 3, we now believe the focus of the work is appropriate. The focus has changed somewhat from Round 3 and much additional knowledge has been gained about the nonresponse problem.

However, we are not clear whether the governance arrangements are in place to allow decisions to be made based on the findings of the Nonresponse Project. If not, they should be put in place. 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.

Some of the strategic areas where decisions need to be made were listed in Section 3.2.4 for the LFS but could just as appropriately be applied to demographic surveys in general at Statistics Sweden. Without repeating these recommendations, we want to emphasis three areas that are particularly important. First, we believe the research on using mixed modes holds much promise for the long-term future of household surveys at Statistics Sweden, but there are many issues to consider in moving to this approach. Particular attention needs to be given to mode effects which have been shown to be quite severe in surveys that mix interviewer assisted and self-administered modes.

Second, considerable resources have been directed towards mitigating nonresponse in household surveys, particularly the LFS. However, in terms of the "quality improvement per monetary unit," the return on investment (ROI) may be quite low relative to the ROI for reducing the error from other sources for the same expenditure. Redirecting a portion of the nonresponse reduction resources towards understanding the causes and reducing the effects of measurement error, modeling and estimation error and overcoverage might result in a much greater ROI. Unfortunately, the data necessary to compare these two ROIs are not available but could be obtained through appropriately designed evaluation studies. We disagree with the view that response rates must remain high to ensure confidence and credibility in surveys. However, it appears that the latter view is driving the decision to expend more and more resources to incrementally increase response rates, without assessing how this may improve total survey error.

Finally, a whole host of issues could be listed under the rubric "data collection management and operations." Statistics Sweden's data collection approach shows signs of continued deterioration: accelerated decline in response rates over the last 10 years, increased levels of dissatisfaction among internal users with call center management and performance, inability to affect changes in the operations that could lead to real quality improvements and so on. Outsourcing data collection, at least for some key surveys such as the LFS and HBS, is one approach but it comes with some clear risks to other quality dimensions such as Comparability, Coherence, Timeliness and possibly Accuracy. These issues will not be resolved in the coming year; however, it is important for Statistics Sweden to continue the progress made in the past year to:

- Understand the root causes of nonresponse and nonresponse bias and then address these with using targeted approaches that are tailored to reduce both. For example, we have noted that calibration is a powerful technique for reducing nonresponse bias that has yet to be fully exploited. In particular, the techniques described by Kott (2010) should be considered for the LFS and other demographic surveys.
- Adopt a management structure for telephone operations (both centralized and decentralized) that allows for rapid changes to the systems that address quality and cost issues. The current management structure seems convoluted and it is difficult to imagine how the current structure could sustain high quality interviewing. As an example, although call monitoring has been implemented for several years now, these data have not been utilized effectively.
- Along the same lines, some adaptive design functionality has been developed in the WinDati system for targeting cases for special contact and interview approaches. This work should continue. In particular, the ability to produce daily reports indicating how the interviewing effort and outcomes are distributed by time of day, day of the week and interviewer for both centralized and decentralized components would aid in creating more desirable outcomes in terms of efficiency, timeliness and data quality.
- Finally, a coordinated research agenda that considers the total survey error for all demographic surveys is needed so that results, both positive and negative, can be shared across surveys. This is already in place to some extent with the Nonresponse Project. However, this work could be expanded to other error sources such as measurement, frame and model/estimation errors. In addition, it may be possible to collaborate with other NSOs to both broaden the research agenda and create greater synergies that will energize and multiply Statistics Sweden's efforts.

3. 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 devoted to IT projects especially those associated with the phase out of Visual Basic 6. This is understandable but it would be appropriate if more funding could be devoted to statistical development projects as soon as possible.

The Innovation Laboratory is a good initiative and perhaps some of these resources could be devoted to some of the issues mentioned in General Recommendation 2 above. For example, the Australian Bureau of Statistics was able to use operations research techniques to significantly improve the efficiency of their household survey data collection operations.

4. Responding to ASPIRE Recommendations

This report contains a number of product level recommendations as well as the three general recommendations in this part of the Report. Some of these can be addressed with relatively little effort while others may require considerable investments in financial resources and human capital. Some may require an ongoing, multi-year project while others may only involve short-term efforts. Likewise, some are best addressed by cross-cutting, multi-unit coordination and collaboration while others may involve only the product staff and have only minor implications to other products. Nevertheless, taken as a whole, the recommendations represent a large amount of work – perhaps too much to consider for a single annual cycle. Deciding on how to best prioritize these recommendations can be a complex process that trades-off costs, risks and resource availability while considering Statistics Sweden's current strategic objectives, long-range plans, and the potential effects of anticipated or probable changes in the external environment.

For these reasons, we have not attempted to assign priorities to the recommendations although we believe that prioritization is an essential next step. Rather, we believe Statistics Sweden's top management should identify the highest priority recommendations and ensure that well-integrated, agency-level work plans for addressing them are developed as soon as possible.

Recognising that some of the projects can be done within existing resources, we recognise that there should be a formal response to each of the product level recommendations. There are a variety of responses that are possible such as:

(i) Rejecting the recommendation,
(ii) Accepting the recommendation but noting that it may not be possible to do work in the coming the year,
(iii) Modifying the recommendation in some way and outlining the work that is planned, and
(iv) Accepting the recommendation and outlining the work planned.

These responses should be signed off at the departmental head level as an appropriate response.

5. SUMMARY AND CONCLUSIONS

As we stated in our previous reports, we believe Statistics Sweden is a world class organisation and believe that even more strongly with each completed round of ASPIRE. 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 3 and these have been identified in this report.

This is the fourth ASPIRE review for seven products and the third review for three products. As a result of further information available this time we have corrected a small number of the ratings. In the report, we have distinguished the corrections from improvements and Exhibits 2a and 2b shows the current ratings, prior year ratings, and the improvements by product. 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 ranged from 52 percent (for the ULF/SILC) to 68.6 percent (for the FTG) with an average rating of about 61 percent. Products generally increased their scores in this round but the average improvement in ratings over all products and error sources was only about 1.3 percent this round compared to 2.7 percentage points in the last round. In Section 3.1 we provided some possible reasons for the reduced average improvement which may be a combination of (a) greater difficulty to improve after four years of ASPIRE, (b) reduced resources to address meaningful quality improvements and (c) lack of attention to the ASPIRE recommendations.

Clearly, (a) was evident in some reviews. Following four 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 (and their knowledge and possibly communication ratings) and real risk mitigation strategies. These require resources (b), which as previously noted in this report, have been more constrained this year for a number of products. In addition, the relatively small improvement this year may be partly explained by staff motivation and accountability (c). To address this, we have added Cross-cutting Recommendation 4 in Section 4 which assigns responsibility for disposing of the ASPIRE recommendations to upper management. Notwithstanding the small increase in average scores for this round, there has still been a 7.2 percentage point increase since ASPIRE started in 2011 (see Exhibit 3c) which represents a substantial improvement in average quality for these 10 products.

The ASPIRE process has been modified and improved over the last four rounds and seemed to work quite well in the current round. We were quite 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. These staffs should be commended for their inspiration and initiative.

As we think ahead to Round 5 of ASPIRE, some changes the scope of the process should be considered to increase the impact of ASPIRE on overall data quality for Statistics Sweden products. Some suggested changes include the following:

- Replace the Foreign Trade in Goods (FTG) with Foreign Trade in Services.
- Replace the CPI with the Producer Price Indices

- Delete the SBS, but retain meeting with them as part of the reviews for the BR and the NA. The Industrial and Services Production Indices could then be added.
- Delete the RS and add the Household Budget Survey which poses a considerably greater risk to Accuracy.

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 relatively low ratings (i.e. Fair, Poor or Good). 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. We note with some pride that a paper documenting the ASPIRE approach appeared in the international literature in 2014 (viz., Biemer, Trewin, Bergdahl and Japec, 2014). This paper has generated considerable international attention and "buzz" around Statistics Sweden's quality improvement initiatives. With the publication of this paper, Statistics Sweden has been established as a world leader in the area of official statistics quality management. The staff of Statistics Sweden should also take pride in these accomplishments because ASPIRE would not be possible without their inspiration, motivation to improve and willingness to share results both positive and negative on Swedish statistics quality.

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.

Kott, P. & Chang, T. (2010). "Using Calibration to Adjust for Nonignorable Unit Nonresponse," Journal of the American Statistical Association Vol. 105, 491.

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. 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 providers whose inputs pose some risk to data quality from this error source, there have been discussions regarding these potential risks.	☐ Yes ☐ No **Fair**	
3. These communications 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 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. Likewise, the level of detail that has been shared with providers regarding how their inputs affect data quality is sufficient for them to formulate and plan mitigation strategies (if applicable).	☐ Yes ☐ No **Good**	

6. User documentation speaks clearly, comprehensively, and with appropriate detail on the size of the MSE components for the target audience.	☐ Yes ☐ No **Very Good**	
7. Provider communication is sufficiently detailed regarding the effects of errors including the quantification of impacts, and provides adequate information to enable the providers to develop mitigation strategies that have real impacts on product quality.	☐ Yes ☐ No **Very Good**	
8. 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**	
9. There is evidence (in the form of emails and other forms of communication) that providers have been intimately involved in the process of mitigating the risks of error 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**	

Achievement towards Improvement Plans	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 has been approved by management.	☐ 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. 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. 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. Quality improvement strategies that have been implemented have been successful at minimizing the risk to data quality from this error source.	☐ Yes ☐ No **Excellent**	