A THIRD APPLICATION OF ASPIRE FOR STATISTICS SWEDEN

Paul Biemer and Dennis Trewin

January 20, 2014

TABLE OF CONTENTS

1 Executive Summary 3

2 Background and Introduction 6

3 Product Quality Assessments and Monitoring 8
3.1 The ASPIRE Model 8
3.2 Scope of the review 10
3.3 Evaluation Criteria 13
3.4 Application to the Products 15
3.5 Limitations of ASPIRE 16

4 Findings for the Ten Statistical Products 18
4.1 General Observations 18
4.2 Product by Product Ratings 22
4.2.1 Annual Municipal Accounts 22
4.2.2 Consumer Price Index 25
4.2.3 Foreign Trade of Goods 27
4.2.4 Labour Force Survey 30
4.2.5 Structural Business Statistics 34
4.2.6 Living Conditions Survey 37
4.2.7 Business Register 40
4.2.8 Total Population Register 43
4.2.9 Quarterly Gross Domestic Product 45
4.2.10 Annual Gross Domestic Product 49

5 General Recommendations 51
5.1 Progress on Round 1 and Round 2 Recommendations 51
5.2 New Recommendations 56

6 Summary and Conclusions 58

7 References 60

Annex 1 – Checklists for the Accuracy Dimension of Quality 61
Annex 2 – Product Specific Rating Changes between Rounds 2 and 3 67

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, 2012). The initial application of ASPIRE (referred to as Round 1) was applied to eight products, viz.: Annual Municipal Accounts (RS), Consumer Price Index (CPI), Foreign Trade of Goods Survey (FTG), Labour Force Survey (LFS), Structural Business Statistics (SBS), Business Register (BR), Total Population Register (TPR) and the National Accounts (NA). In 2012, ASPIRE was improved and applied to the same eight products evaluated in Round 1. The most significant improvement was a substantially revised set of evaluation criteria for the NA that was tailored to the unique error structure of both the annual and quarterly gross domestic product (GDP) estimates. Moreover, ASPIRE was applied to one additional product – viz., the Survey of Living Conditions (ULF/SILC). For each of these ten products, Accuracy (or data quality) was assessed for all error sources that were applicable in each product. Also in Round 2, a number of so-called user dimensions of survey quality were assessed for two products. In particular, Relevance/Contents and Accessibility & Clarity were assessed for the CPI and Timeliness & Punctuality and Comparability & Coherence were assessed for the LFS.

This report summarizes the results from the Round 3 of ASPIRE which was conducted in November 2013. For this round, we slightly revised the evaluation criteria (primarily to incorporate criteria for communication with suppliers in the production process) and several additional aspects of ASPIRE were refined. The Accuracy dimension for the ten products in Round 2 was re-evaluated; however, evaluations of the user dimensions for the LFS and the CPI were not repeated. The current report does not directly consider the user dimensions of any product in this review.

As in the prior rounds, the evaluation for each product involved a self-assessment, extensive reviews of relevant documentation, a comprehensive interview of key staff, and a staff review of the preliminary evaluation results with feedback. As in Round 2, each product was scored (on a 10-point scale) using five criteria which 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.

With a maximum possible score of 100 percent (indicating perfect quality), the product scores ranged from 51.1 percent (for the ULF/SILC) to 67.6 percent (for the FTG) with an average rating of 59 percent. (Exhibits 4a and 4b in the report provide the scores for each product by error source.) Most of the products reviewed in Round 2 increased their scores in this round with a substantial increase (9 points) for ULF/SILC. The exceptions were SBS which showed a decrease and the BR which showed no change from Round 2. Detailed justifications for the ratings for each product are provided in the section of the report devoted to that product. The average improvement in ratings over all products and error sources was about 2.7 percentage points. When combined with the 3.2 percentage point increase in Round 2, there has been a 5.9 percentage point increase since ASPIRE started in 2011 which represents roughly an 11 percent average improvement in quality for these 10 products.

Some of additional findings from the reviews include the following:

- As in Rounds 1 and 2, measurement error appears to be the error source with the highest intrinsic risk; it was rated a "High" for six of the eight products where that error source is relevant.
- Model/estimation error replaces measurement error at the bottom of the ratings for two reasons: measurement error ratings improved and model/estimation ratings dropped for the LFS and the CPI.
- 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.
- Most of the improvement is under the Planning/Mitigation criteria. Most of this is for planning rather than mitigation.

In addition, the following general findings are notable:

- The nonresponse rates for household surveys continue to deteriorate despite the considerable effort put into addressing this problem.
- Last year, 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, as reported last year, most quality evaluations tend to focus on error rates and indirect measures rather than direct error measures such as bias, validity, and reliability.
- Furthermore, Statistics Sweden does not take full advantage of the evaluation studies that are undertaken. Knowledge transfer could be improved which would be aided by better archiving of evaluation studies. This is discussed further in Section 5.

The main report provides specific comments on each product, justifications for any changes in the ratings since Round 2, and some suggestions for improvement. Finally, in our previous reports, we laid out recommendations to improve quality that cut across all products in these 13 areas:

1. Greater Integration of Economic Statistics
2. Increasing Cooperation between the NA and Statistical Areas
3. Improving the Accuracy of NACE Coding
4. Need for Additional Evaluation Studies
5. Reducing Nonresponse in Household Surveys
6. Improving the Relationship with the Tax Agency
7. Improving the Policy on Continuity of Statistical Series
8. Improving the Relationship between IT and their Client Areas
9. Addressing the Lack of Telephone Interviewing Monitoring
10. Development of Improved Quality Profiles for Key Products
11. Increase the Focus on Coherence between Relatable Statistics
12. Initiate Succession Planning in Some Important Statistical Areas

Although progress has been made in many of these areas, with significant progress in some areas, more improvement is needed and the work should continue to progress in the highest priority areas. These decisions should be made at the corporate level. One approach might be for management to spend a day considering the recommendations in this report, further specifying and explicating them, ranking them in order of organizational priority, and making arrangements to develop the action plans to address the highest priority areas.

In addition, four new recommendations are added as a result of the current review. These are:

1. Develop well-specified criteria for deciding whether and how large enterprises should be profiled and implement other steps to reduce the number of large enterprises that have too few Kind of Activity Units (KAUs).
2. Develop a comprehensive, cross-unit plan for phasing out of Visual Basic 6 (VB6) that is widely-supported and well-communicated to all departments affected by the plan.
3. Develop a systematic approach for archival and retrieval of manuscripts and reports that document quality improvement projects that are authored or co-authored by Statistics Sweden staff.
4. Launch an annual process for planning and monitoring projects that specifically address the recommendations in the annual ASPIRE reports.

Our report also addresses the issue of household survey nonresponse specifically for the LFS and the ULF/SILC as well as more generally for all household surveys at Statistics Sweden. We are concerned by (a) the high costs and apparent ineffectiveness of current efforts, (b) how these collective efforts are organized and coordinated, and (c) the opportunity costs for other quality improvement efforts that may have greater impact on product quality. We provide recommendations regarded to these concerns in Section 5.

The revised ASPIRE approach for Accuracy worked very well for most products. The revised error structure developed for the NA in 2012 was an important innovation that greatly improved the evaluation of GDP estimates. Criteria and checklists that address the unique characteristics of the GDP error components worked quite well. However, additional improvements are planned to enhance the criteria and checklists for the NA and the other products. Specifically, the Planning and Mitigation criteria need refinement to better reflect whether or not mitigation efforts have been successful.

2 BACKGROUND AND INTRODUCTION

The government of Sweden stated in Statistics Sweden's appropriations directive for 2011 that the agency was required to complete ongoing work within the area of quality and that significant quality improvements were to be reported to the government at end of 2011 and every year following. In this context the government has requested a report in the form of specific indicators that signify any quality improvements that are occurring in pre-specified, key programs.

Up until 2008 Statistics Sweden monitored the quality of statistical programs by way of a self-assessment questionnaire to which survey managers responded annually. The results of these assessments were traditionally included in the agency's annual report to the government. However, because of the inherent bias in self-assessments, the process did not yield the informative and accurate measures of data quality needed for effective, continual quality improvement. The self-assessment process was thus discontinued and Statistics Sweden has not quantified progress on product quality for the annual report since then.

The Research and Development Department (R&D) was commissioned by the Director General of Statistics Sweden during the year to develop a model that will capture quality changes in the agency's statistical programs. This led to us to undertake a review of eight products in the period of November/December 2011 using an approach referred to in this report as ASPIRE (A System for Product Improvement, Review, and Evaluation). Our report was finalised in January 2012 (Biemer and Trewin, 2012) and provided a baseline for these products. That work will be referred to as Round 1.

The 2011 evaluation process worked very well for all products except for NA. To improve the process for the NA (NA), an alternative approach was devised for 2012 that was customized to the unique error sources associated with NA products – specifically quarterly and annual gross domestic product (GDP) estimates. This approach effectively created a new baseline evaluation for the NA for these two time series. The other seven products were evaluated and a new product, the Survey of living Conditions (ULF/SILC), was also evaluated for the first time in 2012, bringing the total number of products under the ASPIRE process to ten. The 2012 evaluation is referred to as Round 2.

Also in 2012, four so-called *user* dimensions of quality were also evaluated for two products – the Current Price Index (CPI) and the Labour Force Survey (LFS). Statistics Sweden has over the past two decades worked quite actively with quality concepts in official statistics providing definitions and recommendations for producers firstly to aid them in the actual development of statistics and secondly to help them in their communication with the users by way of quality declarations. In the ASPIRE system, five dimensions of total survey quality are defined – Accuracy, Relevance/Contents, Timeliness & Punctuality, Comparability & Coherence, and Accessibility & Clarity[1]. The latter four dimensions are referred to as user dimensions because they are primarily focussed on user needs and concerns. Thus, Accuracy, Timeliness & Punctuality and Comparability & Coherence were evaluated for the LFS while Accuracy and the remaining two user dimensions were evaluated for the CPI. The review was limited to only two

[1] These quality dimensions differ somewhat from the dimensions that are currently in use by SCB, viz., Contents, Accuracy, Timeliness, Comparability & Coherence, and Availability & Clarity. (See *Quality definition and recommendations for quality declarations of official statistics*, MIS 2001:1). In this report, we have replaced "Contents" by "Relevance/Contents" and "Availability" by "Accessibility" following the Code of Practice within the European Statistical System.

products because it was intended primarily to the criteria, guidelines, and checklists that were developed specifically for evaluating the user dimensions within ASPIRE.

For this round (Round 3), the focus of ASPIRE returned to Accuracy only, dropping the user dimensions in order to reduce the level of effort for the external reviewers. The criteria developed for user dimensions in Round 2 can feasibly be applied by internal reviewers, thus saving costs to the ASPIRE process. The same ten products reviewed in Round 2 and documented in Biemer and Trewin (2013) were reviewed again for this round.

The objective for Round 3 was to identify areas within each of the ten products where clear improvements had been made since the previous evaluation. As we did in Round 2, any the ratings that were assigned in Round 2 that were determined to be incorrect (usually because the information upon which the prior rating was based was either incomplete or flawed) were corrected. Such corrections are important because current process improvements, which are judged relative to the previous ratings, assume that the prior ratings are accurate. Discussions of quality improvements in this report will clearly distinguish between original, corrected and new current ratings. Our report also identifies the highest priority areas for improvement both at the product level and across products where cross-cutting issues can be identified.

The ASPIRE process that was applied in this review is described in the next section including a discussion of some of the improvements made to the original approach and suggestions for further improvements. Section 4 summarises the results of the quality evaluations for the ten products (treating quarterly and annual GDP as separate products). Section 5 summarises some cross-cutting methodological and other findings including non-response for household surveys. Finally, Section 6 provides our recommendations and conclusions.

3 PRODUCT QUALITY ASSESSMENT AND MONITORING

3.1 THE ASPIRE MODEL

In Biemer and Trewin (2012) (i.e., ASPIRE: Round 1), we developed an approach for evaluating the accuracy of official statistics produced by Statistics Sweden referred to in this document as ASPIRE. This approach is general in that it can be applied to a specific statistical estimate such as the monthly unemployment rate, a range of products produced by a data collection program such as the Municipal Accounts (RS), a frame or register such as the Total Population Register (TPR), or a compilation of a number of statistical inputs such as the system of NA. ASPIRE is also comprehensive in that it considers the errors in official statistics arising from all major error sources from the design of the data collection to final publication or data release.

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

The ASPIRE model assesses product quality by first decomposing the total error for a product into major error components. It then evaluates the potential for these error sources to affect data quality (referred to as "the risks of poor quality") according to five quality criteria which will be described in Section 3.3. Well-specified guidelines are used to evaluate these risks with a high degree of inter-rater reliability. To explain further, suppose $\hat{Y}$ denotes a survey estimate that is subject to errors from a number of sources. One can conceive of an "error-free" version of $\hat{Y}$ denoted by Y; i.e., if the processes producing $\hat{Y}$ were error free and ignoring possible sampling errors, the estimate ($\hat{Y}$) and the error-free parameter (Y) would be equal. The difference, i.e. $\hat{Y}-Y$, is then due to errors in the processes that produce $\hat{Y}$ (referred to as the *total survey error*). The total survey error (TSE) includes both the *nonsampling* errors and the sampling error, if applicable, of a product. In the ASPIRE system, the TSE is decomposed into seven components: frame error, nonresponse, measurement error, data processing error, sampling error, model/estimation error, and revision error. These errors will be now be defined.

Frame error arises in the process of constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. It includes the inclusion of non-population members (*overcoverage*), exclusions of population members (*undercoverage*), and duplication of population members, which is another type of overcoverage error. Frame error also includes errors in the auxiliary variables associated with the frame units (sometimes referred to as *content error*) as well as missing values for these variables[2]. *Nonresponse error* encompasses both unit and item nonresponse. *Unit nonresponse* occurs when a sampled unit does not respond to any part of a questionnaire. *Item nonresponse* occurs when the questionnaire is only partially completed

[2] In our approach, missing information for frame variables is distinct from missing information for variables collected during a survey. The latter is referred to as survey item nonresponse.

because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. *Measurement error* includes errors arising from respondents, interviewers, survey questions and factors which affect survey responses. *Data processing error* includes errors in editing, data entry, coding, computation of weights, and tabulation of the survey data. *Modelling/estimation error* combines the error arising from fitting models for various purposes such as imputation, derivation of new variables, adjusting data values or estimates to conform to benchmarks, and so on.

Finally, *revision error* is the error in a preliminary, published estimate from a survey that is later revised. It can be shown to be a component of the total error of the preliminary estimate. To see why, let $\hat{Y}_p$ denote the preliminary, published estimate whereas $\hat{Y}$ is the final estimate. Then the total error in $\hat{Y}_p$ given by $\hat{Y}_p - Y$ can be rewritten as $\hat{Y}_p - \hat{Y} + \hat{Y} - Y$ where $\hat{Y}_p - \hat{Y}$ is the revision error and $\hat{Y} - Y$ is the total error in the final published estimate as described above. Because Statistics Sweden is very interested in reducing the error in all published estimates, not just the revised one, we focus on both preliminary and revised estimates in our evaluation of Accuracy. Furthermore, considering revision error as a distinct error source reflects the view that large revisions, regardless of their reasons, are undesirable from the user's perspective and should be avoided. Thus, an important quality goal for Statistics Sweden is to reduce the size of the revisions which is facilitated by emphasizing revision error whenever it is applicable.

Note, however, that revision error is somewhat unusual because it reflects the combination of all other error sources on the preliminary estimate. For example, the preliminary estimate may differ from the final estimate as a result of late respondents (i.e., nonrespondents at the preliminary deadline) whose characteristics may be estimated in the preliminary estimate while their reported values are used in the final estimate. Likewise, revisions may correct for other nonsampling errors such as measurement, data processing, or modelling/estimation errors that are identified after the preliminary deadline. In this way, revision error may account for error sources that have already been considered in the assessment of data quality for the revised estimate. However, the revised estimates may also use updated post-stratification or other adjustment factors that are based upon data that were unavailable when the preliminary estimates were published. Such corrections cannot be readily attributed to other error sources and therefore are not considered in the assessment of other error sources.

For our review, we do not attempt to decompose revision error into its associated subcomponents (nonresponse error, data processing errors, etc.) because the errors that affect the preliminary estimates also affect the final estimates, although presumably to a somewhat smaller extent. The other error components are considered in detail in our evaluation of the revised estimates. Rather, our primary interest for the preliminary estimates is on the size of revision error, i.e., $\hat{Y}_p - \hat{Y}$ and what steps can be taken to reduce it and/or its impact on data users.

For most products, an eighth error source – referred to as *specification* error – is also applicable. Specification error arises when the observed variables, *y*, differs from the desired construct, *x* – i.e., the construct that data analysts and other users prefer. In survey literature (see, for example, Biemer 2011), *x* is often referred to as a *latent* variable representing the true, unobservable variable and *y* is often referred to as an indicator of *x*. As an example, in the FTG, the invoice value of goods is collected from enterprises (*y*) while the statistical value (*x*) (which excludes shipping costs within Swedish borders), is preferred for most statistical uses of the data. Thus, specification error may be defined as the difference between *y* and *x* (see, for example, Biemer and Lyberg, 2003).

Specification error biases the estimates of population parameters. Let X denote the true population parameter which is a function of x. Then the total survey error in $\hat{Y}$ can be written as

$$\hat{Y} - X = (Y - X) + (\hat{Y} - Y), \text{ or, in words,}$$

TSE = (specification error) + (other sampling and nonsampling errors)

Under this model, the TSE of an estimate includes specification error as well as the other aforementioned sampling and nonsampling errors. Thus, the specification error in the aggregate, $\hat{Y}$, is essentially the difference between the expected value of $\hat{Y}$ conditioned on the concept implied by the survey instrument (Y) and the population parameter under the preferred concept (X). As an example, Y may be the total invoice value while X is the total statistical value for imports for a given commodity. One way to identify and prevent specification error is have subject matter experts and other data users review the survey instrument to ensure that the concepts underlying each data item conforms to the concepts that are implied in the use of the data items.

Although the TSE components were defined for surveys, they can also be used for compilations and registers, with some modifications. For compilations, the TSE components pertain primarily to input data sources, many of which are derived from survey data. However, as described below, the GDP estimation process is quite complex and addition error sources are needed to fully represent its error structure. For registers, frame error, which can also be an important error source for the survey products, was expanded to include its major subcomponents, viz., overcoverage, undercoverage, duplications, content error, and missing data. The use of the term "content error" for registers rather than "measurement error" emphasizes that, when register data are in error, the cause of the error (albeit the measurement process, data processing, modelling, imputation, etc.) is often not known. Likewise, the cause of missing data in the register cannot always be attributed to nonresponse. Therefore, it will be referred to simply as "missing data" for purposes of register evaluation.

3.2 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 3). The error sources that are associated with these products are shown to the right of these products. Likewise, middle panel shows the two registers included in this review and their error sources. Finally, the bottom panel shows the 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 show on the right that panel. As we previously noted, all of these products were evaluated in Round 2 and those results are documented in Biemer and Trewin (2013).

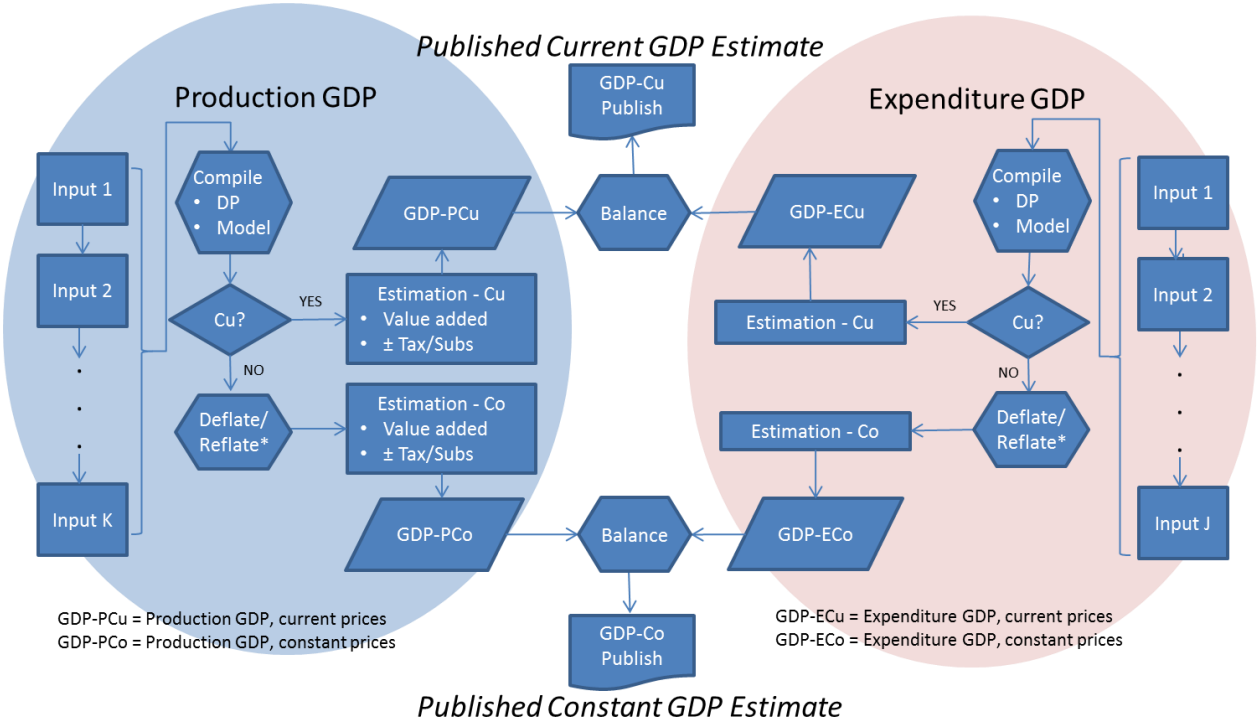
As in the Round 2 review, the focus of the current review is on product improvements and deteriorations in data quality over the last year and how those changes affect the prior round's rating for each error source. Also, with regard to the NA products, the current review, like Round 2, focused somewhat 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).

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

Exhibit 2 provides a flow diagram that attempts to capture the major activities associated with the estimation of GDP. As shown in this exhibit, the GDP estimation process incorporates two somewhat independent approaches for estimating GDP. These are referred to as the production (shown on the left of Exhibit 2) and the expenditure approaches (shown on the right). Both approaches begin with a number of inputs that must be assembled, processed, and compiled to prepare them for the next step in the process. The "Compile" stage includes data processing, which may be simply entering the inputs into an Excel spreadsheet but may also include some editing as well as modelling/estimation. This latter process may involve combining multiple inputs to create derived variables as well as modelling the data to reduce specification and other errors. For producing GDP in current prices, these compiled inputs proceed through an estimation stage which, for the production approach, involves adding taxes and deducting subsidies (subs) appropriately. (There are some situations where current price estimates are estimated by reflation of constant price estimates.) For constant prices, the current prices must be "deflated" using the appropriate prices indices before adjustments for taxes and subsidies. Both the production and expenditure approaches will produce interim estimates of GDP (both current and constant prices) which must then be "balanced" or forced into agreement as the economic theory dictates (see, for example, Lequiller and Blades, 2006). This balancing process produces the preliminary estimates of GDP for both current (denoted by Cu in the exhibit) and constant (denoted by Co) prices. The latter differs from the former primarily by a deflation/reflation process that adjusts prices to a common base-year. The preliminary estimates are subsequently revised when addition data become available. Thus, the error sources associated with the GDP estimation process are as shown in Exhibit 1, bottom panel.

Exhibit 2. Process for Estimating GDP by Current and Constant Price Approaches



•NOTE: Some items follow the deflation process in the opposite direction and are complied starting with information on volume change from the previous year. The volume estimate i s then reflated with the price index in order to come to the current price estimate. Items within the Energy sector is one such example.

As shown in bottom panel of Exhibit 1, the evaluation of the GDP estimation process is confined to the production side of Exhibit 2 including balancing and final publication of the estimates. We elected to focus on the production approach because several important inputs to this process were already included in the evaluation process – viz., the Structure Business Statistics (SBS), the Annual Municipal Accounts (RS) and the Consumer Price Index (CPI). In addition, the evaluation team also held meetings with the producers of the two most important additional inputs to the production approach – the Service Production Index and the Industry Production Index.

In the evaluation of production GDP, considerable attention is given to the error in the inputs and their effects on the error in the GDP estimates. Priority is given to inputs that posed the greatest risk to GDP error. To illustrate this approach, suppose the *K* inputs shown on the left of Figure 2 give rise to *P* input variables denoted by $y_1, y_2, \cdots, y_P$. The estimate of GDP, denoted by GDP′, is some function of these input variables; i.e.,

$$\text{GDP}' = f(y_1, y_2, \ldots, y_p) \qquad (0.1)$$

Depending upon the data source, each of these variables is subject to error from numerous sources (for e.g., the components of TSE that are applicable) which, for the *p*th variable, will be denoted collectively by ε_p. Let x_p denote the value of y_p that would be observed if these errors were negligible; i.e., if ε_p were essentially 0. Thus, we can write

$$y_p = x_p + \varepsilon_p, \text{ for } p = 1, \ldots, P$$

which means that the observed input variable is equal to the true value of the variable plus an error. Of course, x_p is a theoretical true value because it is always observed with some amount of error. Indeed the goal of many evaluation studies associated with the other products in ASPIRE is to evaluate ε_p .

Likewise, the theoretical true value of GDP can be expressed as some function of the true values of the input variables, say

$$\text{GDP} = g(x_1, x_2, \ldots, x_p) \tag{0.2}$$

and thus, we can write

$$\text{GDP}' = \text{GDP} + e \tag{0.3}$$

which means that the estimate of GDP is equal to the true value of GDP plus some unknown error, *e*. In our evaluation of the GDP input data sources, we are particularly interested in determining which ε_p's contribute most to the error, *e*, in the GDP estimation process. Note that the most influential errors for estimating GDP may not be associated with the variables that have very large errors. A large error in a variable that plays a small role in the calculation of GDP may also have small influence on *e*. In addition, an influential variable having a large error may have a small influence on the GDP error, *e*, if its error contribution is limited in the estimation process; i.e., through the function *g*. Thus, we are also interested the potential contributions of *g* on *e* where *g* includes compilation (both model/estimation and data processing error), inflation/reflation, balancing, and revision stages of the estimation process. In terms of the input data sources, we have done this subjectively in collaboration with the NA. There may be ways of doing this more objectively but it would not be a straight-forward exercise.

3.3 EVALUATION CRITERIA

In addition to decomposing total error for a product into its component sources, the risks associated with each source are further subdivided into five risk categories (represented by the five quality criteria) and explicit guidelines were developed to aid the assessment of current quality and quality improvements. As for Rounds 1 and 2, we have used five criteria; viz., Knowledge of Risks, Communication, Available Expertise, Compliance with Standards and Best Practices, and Achievement Towards Improvement Plans. In Round 3, the guidelines for these criteria have been further enhanced and improved. One significant change was the addition of "communication with suppliers" of data and information under the Communication criteria. Prior rounds only assessed "communication with users" regarding the error sources for a product. .

The application of these guidelines is facilitated by the use of checklists for each criterion (see Annex 1). The checklists are generic in that the same checklist could be applied to each relevant error source. Moreover, we believe the simple "yes/no" format used for the checklists eliminates much of the subjectivity and inter-rater variability associated with the quality assessments. In addition, the checklists incorporate an implied rating feature so that upon completing the checklist for a criterion, the rating for that criterion is largely pre-determined based upon the last "yes"-checked item in the list.

As was done in previous rounds, a two-step rating process is used to assign ratings on a 10-point scale for each error source by criterion combination. First, a given criterion is assigned a qualitative rating of Poor (1-2), Fair (3-4), Good (5-6), Very Good (7-8), and Excellent (9-10). In

the second step, these qualitative ratings are then refined by choosing between low or high numerical point ratings within each of the five categories. Note that for some checklists in Annex 1, a particular qualitative rating may be associated with two checklist items rather than one. Depending upon whether one or both items were answered "yes," a refined numerical rating can be determined. For example, for the Knowledge of Risks checklist, items 2 and 3 both map to a "Good" rating. If the answers to item 2 is "yes" and item 3 is "no," a numerical rating of 5 is implied. Otherwise, if item 3 is "yes" and item 4 is "no," then a numerical rating of 6 is implied.

An option that was introduced in Round 2, and repeated in this round, is allow a "not applicable (n/a)" rating in cases where the context of the error source is such that a criterion rating does not make sense. For example, if an error source poses a very small risk to quality for a product, it is often imprudent to invest resources in risk mitigation or improvement planning as this could divert resources from higher priority areas. In such cases, an "n/a" rating would be more appropriate for "Achievement Towards Improvement Plans" than a rating of "poor" which is viewed somewhat stigmatically.

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

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

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

Inherent risk is an important component of a product's overall score because it determines the weight attributed to an error source in computing a product's average rating. Residual risk does not play an active role in the evaluation and is generally not noted in the evaluation. Rather, its primary purpose is to clarify the meaning and facilitate the assessment of inherent risk. In at least one case (LFS), the residual risk will be discussed because its level has reached a critical or "crisis" level (see Section 4.2.4 for more discussion).

A product's *error-level score* is just the sum of its ratings (on a scale of 1 to 10) for an error source across the five criteria divided by the highest score attainable (which is 50 for most

products) and then expressed as a percentage. A product's overall score, also expressed as a percentage, is then computed by following formula:

$$\text{Overall Score} = \sum_{\text{all error sources}} \frac{(\text{error-level score}) \times (\text{error source weight})}{10 \times (\text{number of criteria}) \times (\text{weight sum})}$$

where the "weight" is either 1, 2, or 3 corresponding to an error source's risk; i.e., Low, Medium, or High, respectively, and "weight sum" is the sum of these weights over all the product's error sources. In most cases, the "number of criteria" that are applicable for an error source is 5; however, in a few cases, "Achievement Towards Improvement Plans" is not applicable (N/A) for reasons that will be described in the discussion of each product affected. For those cases, the value of "number of criteria" is 4.

3.4 APPLICATION TO THE PRODUCTS

Similar to the process in Biemer and Trewin (2012, 2013), the application of this model to the ten products in Exhibit 1 follows a three-step approach consisting of (a) pre-interview activities, (b) an interview of product staff to assess product quality, and (c) post-interview activities. These are described below.

PRE-INTERVIEW ACTIVITIES

Pre-interview activities include two primary activities. In Round 1, the evaluators (i.e., Biemer and Trewin) received an extensive list of materials (some in Swedish) for each of the products. These materials were reviewed in the weeks preceding the quality interview. In Round 2, the review process was considerable facilitated by the existence of QDs for all products which, in some cases, were substantially expanded and improved since Round 1. For the current round, the evaluators reviewed all the documentation for Rounds 1 and 2 and any other documentation that had been developed since Round 2.

In addition, each product was asked to complete a self-evaluation by completing the criteria checklist for each error source. The evaluators reviewed these checklists and developed a list of questions to discuss during the quality interview where these information contained on the checklists was reviewed.

THE QUALITY INTERVIEW

As in prior rounds, quality interviews were conducted in both Stockholm and Orebro. These interviews occurred during the period from November 4 – 15. Each interview took approximately four hours to conduct. The meetings were organized into essentially five parts:

a) discussion of any notable changes that have occurred during the preceding 12 months that may have some effect on data quality,
b) review of the QDs focusing on clarifications of the processes associated with product design, data collection, data processing, estimation, and reporting and emphasizing changes occurring within the past year,
c) progress that was made on the recommendations from Round 2
d) assignment of preliminary ratings for each criterion by error source using the quality checklists, and
e) review of all assigned ratings with a discussion of the results and recommendations for improvement.

Detailed minutes were kept of all interviews. These minutes provided a record of the proceedings and were used extensively in refining the ratings as well as in the writing of this report.

POST-INTERVIEW ACTIVITIES

Shortly after the interviews, the evaluators reviewed the minutes of the evaluation meetings and refined their ratings. Considerable care was taken to identify and address any apparent inconsistencies in the ratings within and across products. Some adjustments were necessary; however, we noted that the ratings appeared more consistent than they were in Round 2. We believe this is due primarily to the use of the checklists as well our greater familiarity with the products.

Following this rating reconciliation period, staff who attended the quality interviews were sent their semi-final ratings along with the narratives explaining the ratings, and were asked to correct any inaccurate or misleading information and identify ratings that they believed were not well-founded. Based upon this input, the ratings were further adjusted, the rating narratives were revised, and the contested ratings were further supported and adjudicated. This process produced the final ratings that appear in this report.

FUTURE REVIEWS

We anticipate that the ASPIRE process will be repeated in the next year for these products in order to monitor continuing quality improvements efforts and to provide feedback – both positive and negative – regarding were future improvement efforts should be directed. Additional products may be added to the process as they were for Round 2.

3.5 LIMITATIONS OF ASPIRE

Any method for evaluating the quality of processes as complex as those associated with these ten products will be subject to some limitations and imperfections. Measuring the true accuracy (for example, all components of the TSE) of a statistic such as the CPI or quarterly GDP is virtually impossible because the data necessary to estimate the total error are unavailable. Moreover, data that are available for bias and variance calculations are themselves subject to error. The ASPIRE approach does not purport to provide direct measures of the total error in a product. Rather, the goals of ASPIRE are to:

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

We believe that product Accuracy will improve to the extent that these three goals are met and as efforts to achieve these goals continue. The ASPIRE approach is capable of achieving these goals provided that the inputs to the process – in particular, the information needed to accurately assess each criterion – are accurate, complete, timely, and accessible by the evaluators. Continuing to update and improve the documentation of quality is an important determinant of the success of ASPIRE to achieve its goals. We further believe that the quality ratings assigned by ASPIRE are correlated with the level of quality risks in the sense that changes in the ratings for a product predict real changes in the risks of poor data quality.

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

One weakness of the model is that it is, at best, a proxy measure for product quality. As previously mentioned, ASPIRE cannot provide a direct measure of the total error of a variable, estimate, or product. It relies on the assumption that reducing the risks of poor data quality and improving process quality will lead to real improvements in data quality. Another weakness of the approach is that it is somewhat subjective in that it relies heavily on the knowledge, skill, and impartiality of the evaluators as well as the accuracy and completeness of the information available to the evaluators. Significant improvements were made in the documentation in ASPIRE Round 2 as the information contained in the QDs was "lifted" for a number of products. However, as we will discuss further in Section 5 more work is needed to enhance the completeness and clarity of these QDs. Progress between Rounds 2 and 3 was disappointing. We believe the QDs should be revised and updated dynamically as new information becomes available so that they reflect the current state of quality for the products. At a minimum, there should be an annual review.

Finally, comparisons of improvements in ratings across products may be difficult to interpret without taking into account the some measure of the resources required to achieve those improvements. For example, two products, say A and B, may show the same increase in overall ratings for an evaluation year. However, the resources consumed by Product A to achieve the ratings increase may have been much larger than Product B. In other words, the "ratings increase per krona" is much higher for Product A compared to Product B. It may be reasonable to interpret this ratio as a measure of how effectively quality improvement resources are being spent by product, or to determine which products still have so-called "low hanging fruit" to harvest and should be encouraged and supported in those directions. ASPIRE does not currently report the improvement costs measures but may add such measures in the future.

The next section provides the results of the reviews for the 10 products evaluated in this round. Section 4.1 summarizes our general observations from the evaluations and Section 4.2 provides the more detailed product by product reviews.

4 FINDINGS FOR THE TEN STATISTICAL PRODUCTS

Exhibits 4a and 4b provide the overall scores for eight products (excluding the NA products) by error source. A discussion of the NA is deferred to Section 4.2.1. 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.

Note that the interpretation of the error sources (see Section 3.1) 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 y and x in the notation of Section 1. 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.

4.1 GENERAL OBSERVATIONS

Before discussing each product's detailed ratings, some general observations regarding the results in Exhibits 4a and 4b and 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. This tendency should be resisted as the model was not developed to facilitate inter-product comparisons. For example, the total scores reflect 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 official statistics. Further, although we have attempted to achieve some degree of consistency in ratings among products, inconsistencies may 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. We believe subjectivity has been considerably reduced with the development of the check list as discussed above. Nevertheless, 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.

Close inspection of scores in Exhibits 4a and 4b yield the following observations:

- For the eight products in Exhibit 4a, the overall mean quality rating is 59 compared to 57 in Round 2 and 54 in Round 1. Thus, it appears that overall quality continues to increase for these products.

- The last row of each exhibit shows the Round 2 to Round 3 changes in the overall quality ratings by product. It is notable that SBS quality actually deteriorated slightly since Round 2. This is the result of deteriorations in three areas: frame error, model/estimation error, and revision error.
- Also notable is the substantial improvement in the ULF/SILC – up nine points which is the largest increase of any product in the evaluation.
- Focusing on the mean ratings by error source (last column in Exhibit 4a), model/estimation has the lowest mean rating at 54. This error source is medium to high risk for the eight products in Exhibit 4a and high risk for the NA products in Exhibit 4b.
- As in the prior rounds, the 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 last year.
- 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.
- Last year, 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, as reported last year, most quality evaluations tend to focus on error rates and indirect measures rather than direct error measures such as bias, validity, and reliability.

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 12 cells in Exhibit 4a that satisfy these criteria and they are:

- frame error undercoverage – BR
- nonresponse/missing data – LFS and ULF/SILC
- measurement/content error – SBS, ULF/SILC and BR
- data processing error – RS
- 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 4b be given high priority in the coming year.

Exhibit 4a. Product Error-Level, Overall Level, and Error Source-Level Ratings with Risk-Levels Highlighted and Comparisons to Round 1 Overall Ratings

Error Source/Product	RS	CPI	FTG	LFS	SBS	ULF/ SILC	BR	TPR	Mean rating
Specification error	N/A	**72**	**58**	70	**58**	**58**	66	**58**	63
Frame error	60	**64**	64	58	**60**	**42**	54	63	58
overcoverage							**58**	**58**	
undercoverage							**42**	60	
duplication							63	70	
Nonresponse error /Missing data	**60**	55	**68**	**52**	**70**	**46**	48	**66**	58
Measurement error/Content	**62**	**68**	**64**	**68**	**56**	**52**	**52**	62	61
Data processing error	**54**	**76**	**66**	**62**	**60**	50	N/A	N/A	61
Sampling error	N/A	**70**	N/A	**80**	**86**	**62**	N/A	N/A	75
Model/estimation error	**38**	**44**	**82**	**64**	**48**	**50**	N/A	N/A	54
Revision error	63	N/A	**72**	N/A	**54**	N/A	N/A	N/A	63
Round 3 mean rating	55,0	65,2	67,6	64,3	60,1	51,1	52,7	61,4	60
Round 2 mean rating (re-rated if relevant)	51,5	62,3	64,7	60,9	60,8	42,1	52,7	58,8	57
Change (improvement/deterioration)	3,5	2,9	2,9	3,4	-0,7	9,0	0	2,6	2,9

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

Error source	Quarterly GDP	Annual GDP
Input data source (Average)	**56**	**66**
Structural Business Survey (SBS)	N/A	**66**
Index of Service Production (ISP)	**62**	N/A
Index of Industrial Productions (IIP)	**62**	N/A
Merchanting Service of global enterprises	**44**	N/A
Compilation error - modelling	**48**	**50**
Compilation error - data processing	**52**	**52**
Deflation error (including specification error)	**48**	**48**
Balancing Error	**56**	**58**
Revisions Error	**58**	**56**
Round 3 mean rating	53,6	54,9
Round 2 mean rating (re-rated if relevant)	51,8	53,2
Change (improvement/deterioration)	1,8	1,7

Exhibit 4c. 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.)*

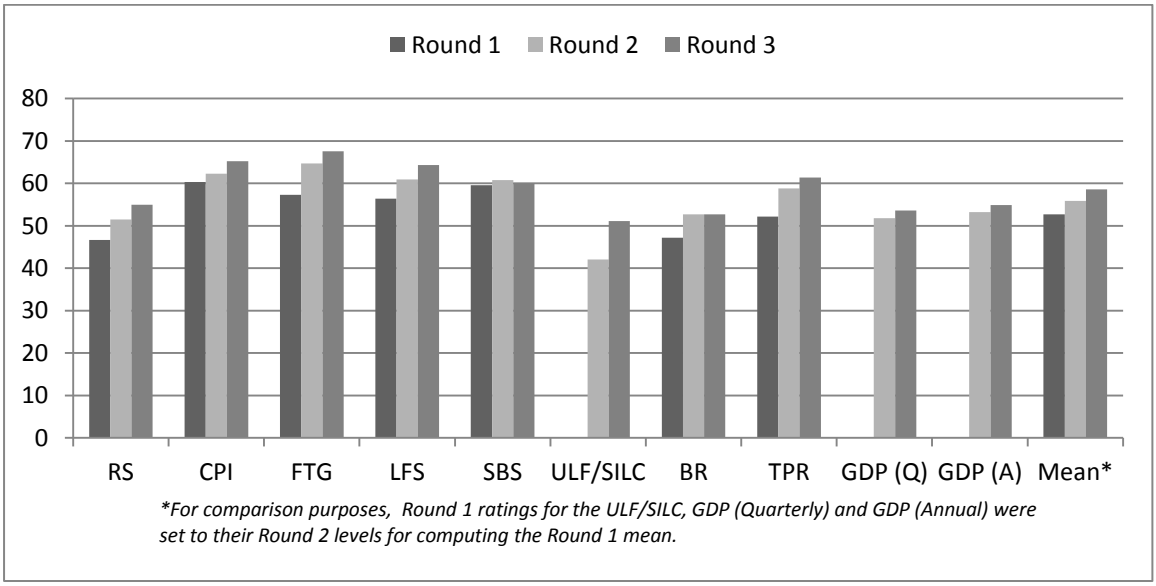


Exhibit 4c shows the overall ratings by product for the three evaluation rounds. With the exception of SBS and BR, all products have steadily improved during the last three years. The mean ratings (last set of bars) show a 3.2 points increase from Round 1 to Round 2 and a 2.7 points increase from Round 2 to Round 3. This constitutes a total 5.9 points increase from Round 1 to the present. However, we caution against interpreting these results 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 discussed in Section 3.5, we can 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.

In the next section, we discuss the detailed ratings for all ten products individually. These ratings, with accompanying comments, appear in Annex 2.

4.2 PRODUCT BY PRODUCT RATINGS

In this section, we review the progress over the past 12 months for the ten products shown in Exhibit 1 using the checklist that appear in Annex 1. 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.

4.2.1 ANNUAL MUNICIPAL ACCOUNTS (RS)

For the RS in 2013, some notable problems occurred in that survey responses from two county councils were corrupted during the data editing process. Unfortunately, these errors were not discovered until after the RS results were published. However, on the positive side, the RS has learned from this event and has taken steps to avoid similar problems in the future. In addition, there are several other fronts where efforts to mitigating the risks of error have been made. Some of these are noted below.

- Continuing the questionnaire redesign effort that was completed in 2011 for municipalities, the county council questionnaire has now also been redesigned. In addition, the new production system developed last year for municipalities was adapted for use with the county councils.

- As for municipalities, the Cognitive Laboratory was consulted in the redesign of county council questionnaire and will be used again in the coming year to debrief county councils on the RS process.

- Considerable effort was directed toward the editing process. A "quality circle" approach was implemented where data editors meet daily during the editing process to discuss current issues they are encountering with specific cases. This has provided greater consistency in the way such editing problems are handled, thus reducing editor variance. This so-called "agile" approach is less reliant on the judgment of a single editor for resolving difficult editing issues and is expected to result in greater accuracy in editing.

- The problem of potentially over-editing the data from municipalities has also received some attention in the previous year. A symptom of over-editing is generating a large number of edit failures which are unproductive in that post-editing changes tend to be minor. New selective editing rules were put in place to reduce the number of unproductive failed-edit alerts. As a result, the proportion of alerts that lead to meaningful changes and data quality improvements was also increased.

- Regulations have been established for making the survey mandatory for municipal associations and for collecting preliminary annual figures from municipalities and county councils. The latter mandate is primarily to inform the spring budget and Excessive Deficit Procedure figures for the Ministry of Finance's report to the EU. This should have the effect of providing more complete and timely information for these reports.

In our evaluation, we decided to downgrade the intrinsic risk level for Model/Estimation Error to medium (M) after it was better understood that the NA estimates were not so importantly affected by the costs disaggregation models as was thought in prior rounds. This downgrade in intrinsic risk should be regarded as a correction and not due to a change in processes.

We commend the RS staff for the good progress that has been made during the last year to improve data quality. The effects of these and other improvements on the ratings can be seen in Exhibit 5 where we show ratings from Round 2 compared to this round's ratings.

We have several recommendations to offer for future research.

1. The errors in the county council data noted above suggest the need for more stringent quality control during the editing process and publication process. The RS staff should mount a review of their editing and publishing processes, not only for county councils but more generally, in order to further mitigate the risks of publishing erroneous results. At least part of this review should identify errors in the post-editing results, back-tracking these to discover their origins and root causes so that these can be appropriately addressed with quality control measures.

2. As noted in our review from Round 2, more research should be devoted to understanding 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 the RS, there has not been much effort in quantifying the errors nor understanding how important users such as the NA are affected by them.

   For example, a relatively simple way to understand the effects of editing on the RS data is to consider the change in various key RS estimates before and after editing. If the difference is sizeable for some estimates, one can conclude that editing is having a sizeable effect on these estimates. These results can then be used to direct further study to examine the errors associated with editing for these estimates and their potential influence on the NA.

3. Also mentioned in our Round 2 recommendations is the need to understand the risk of error when municipalities allocate common costs to various sub-activities. For example, 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. The remaining municipalities allocate common costs according to their own model. However, there has been no study to quantify the error associated with these allocations even though there potential impact on the accuracy of the relevant items is high. The RS should mount such a study in the coming year.

   One way to begin to, at least partially, examine common costs allocation error is to apply the Statistics Sweden model to the 20 percent of municipalities that do not use it and then try to understand the differences in observed to the extent that they are sizeable.

4. With regard to the redesign, one goal 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 the Round 2 report, 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 5. Annual Municipal Accounts (RS), Ratings for 2013

	Error Source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (Control 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	56	60	○	○	◡	○	○	**M**
	Measurement error	58	62	○	○	◡	◡	◡	**M**
	Data processing error	48	54	○	○	◡	○	○	**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	58	63	○	○	◡	◡	N/A	**L**
	Total score	51,5	55,0						

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

4.2.2 CONSUMER PRICE INDEX (CPI)

There have been a number of improvements to the CPI over the last 12 months. Those that we noted and resulted in increased ratings, as shown in Exhibit 6 were:

- The release of the QD in February 2013. This included updated information on sampling errors. Statistics Sweden is one of the few official statistical agencies able to provide estimates of sampling errors on their CPI.

- The extension of the use of scanner data to cover other retail chains. As well as increasing the size of the sample in some important segments, the prices provide for discounts which are otherwise difficult to collect. "Web scraping" is also used to collect price data for some commodities. As a consequence, more of the price data collection is now being undertaken centrally where quality is easier to manage.

- 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. Further improvements in procedures are planned for next year.

- Although there is no specific budget for methodology, a number of studies were undertaken during the year (e.g. selection bias as encouraged by the CPI Board, sample design, changes in methodology to support the introduction of scanner data).

- There have been studies of the price collection methods for mortgage interest rates including subsidies on interest payments.

In Biemer and Trewin (2013), we thought the error risks that most need addressing were (a) the size of the sampling errors in the CPI, (b) potential bias in adjusting for quality change in new products, (c) potential bias in measuring price change in the conceptually difficult area of owner occupied housing, and (d) measurement errors in the data collection process. Good work has taken place in all these areas over the last two years.

With respect to (a), sampling errors have been reduced through a combination of doubling the number of products and outlets in the sample, and the use of scanner data and data scraping in certain commodity groups. The reduced sampling errors seem to be acceptable to users even though month to month changes are mostly not statistically significant. With respect to (b), there have been a number of initiatives to address this problem although the impact has not been quantified. New products are introduced cautiously. On (c) work has commenced over the last year to address this important problem. With respect to (d) there have been many steps taken to reduce measurement errors due to price collector error on assessing quality change including the introduction of centralised collection where it is easier to manage quality.

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 but, in making them, note that Swedish CPI is of a very high standard especially when compared with most other countries. The first four recommendations are modified versions of the recommendations from last year. The modifications are largely because work had commenced on addressing the previous recommendations. The last three recommendations are new.

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. New products and quality change are areas of particular interest.

2. Continue the introduction of scanner data and 'web scraping' to reduce sampling errors in the relevant components but, perhaps more importantly, reduce the measurement errors especially those associated with assessing discounts. In making this recommendation we note the leadership role Statistics Sweden has been taking globally on the introduction of scanner data.

3. Although we agree that quality assessment and selection bias may be more important issues, some consideration of the efficiency of the current sample design should also be undertaken especially with the introduction of the large scanner data sets. We are told that product varieties are the greatest contributor of sampling variability but the recent increases in sample size were in outlets and products.

4. Statistics Sweden has excellent expertise in methods for the CPI and has had for several years. Several of the most experienced staff have either retired or plan to do so over the next few years. This might considerably reduce the expertise unless steps are taken to build up this expertise in new staff. We strongly support the plan to introduce a training program for the staff including the methodologists working on the CPI.

5. Research into methods for measuring quality adjustment should continue as this may well be the most important influence on accuracy.

6. There is a lot of dependency on the work of the data collectors. Their work should be monitored from time to time. We support the planned use of the hand held computers for this purpose.

7. The Household Budget Survey (HBS) has a significant influence on the weights used in the CPI. There is an allowance for the high sampling variability by averaging data over three years which seems sensible. Data from other sources is used for items like tobacco and alcohol. An issue of potential concern is the increasing non-response rate in the HBS. There should be a sensitivity study to understand whether this is an issue of real concern or not.

Exhibit 6. Consumer Price Index (CPI), Ratings for 2013

	Accuracy - by error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	Specification error	68	72	◒	○	◎	◒	○	**H**
	Frame error	62	64	◒	◒	○	◒	○	**M**
	Non-response error	55	55	○	◓	◒	◒	N/A	**L**
	Measurement error	62	68	◒	○	◎	○	○	**H**
	Data processing error	74	76	◒	○	◎	◒	◒	**H**
	Sampling error	66	70	◒	◒	◎	◒	○	**H**
	Model/estimation error	44	44	◓	◓	○	◓	○	**H**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total Score	62,3	65,2						

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

4.2.3 FOREIGN TRADE OF GOODS (FTG)

For Round 3, the FTG has continued the high level of performance it established in the prior ASPIRE rounds. The following are some of the noteworthy quality improvement activities that have occurred since Round 2:

- Every five years, the FTG staff conducts a survey of enterprises that will help to update the models used to relate invoice values to statistical values. Data from this survey are used to recalibrate the models for converting invoice value to statistical value and vice versus. This survey was mounted in 2013 and was completed in November 2013. It is anticipated that new adjustment factors will be available in March 2014.

- An important study was completed that provides more information regarding measurement errors. These results were documented in the following report written by Frank Weideskog: "Record check - praktisk tillämpning". This paper reports on a record check using VAT figures to evaluate measurement errors in the Intrastat invoice values. The results show relatively large net measurement errors in the data.

- Some effort is underway to prepare for SIMSTAT, an electronic data base of imports and exports at the micro level shared by 17 EU states that will ultimately replace the current system for collecting import and export data within the EU from these states. The intent of SIMSTAT is to reduce burden on enterprises as well as to increase data quality although the latter may not occur if the data of the other EU is of lower quality than that for Statistics Sweden. SIMSTAT will be tested in 2015.

- Approximately, half of data are collected using the software IDEP, a program respondents download and used to enter their FTG data. FTG staff are preparing to replace this software with a web version of IDEP which will reduce respondent burden to some extent.

- A user forum for Foreign Trade statistics was established for power users of Foreign Trade Statistics. The forum attempts to capture users' suggestions for improvements and to inform FTG staff regarding how the data are being used. The forum also informs users regarding coming changes in the statistics for both the trade of goods and services (for e.g., on-going projects, changes in the manuals, and so on) that can influence users.

- To facilitate communication with a key user of the FTG estimates, the FTG staff now hold regular meetings with the NA staff in conjunction with the FTG quarterly reports. Among other things, 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.

This last point is particularly important given that foreign trade amount to about 30 per cent of GDP so even a 1 percent revision in foreign trade estimates can make a significant impact on GDP growth estimates, possibly even changing the direction of quarterly change. There have been recent quarters when most of the revisions in the estimates of GDP growth are due to the revisions in FTG. In our FTG review, some FTG staff indicated that they do not consider revision error a problem for the NA. Yet NA staff commented that FTG revision errors continue to be problematic despite efforts by FTG to communicate with them more regularly. This suggests a persistent lack of communication between the NA and the FTG staffs on the topic of revision error. Thus, one of our recommendations for the coming year is for the FTG and NA staffs to specifically discuss revision error and its importance to GDP.

The current and previous round's ratings are shown in Exhibit 7 as well as the current ratings in graphical form. We commend the FTG staff for their excellent progress during the past 12 months.

In planning for 2013 and beyond, we offer the following recommendations:

1. Work on moving from the current Visual Basic 6 (VB6) based IT system to a Windows-based system should be high priority in the coming year given that the Microsoft support for the current system will be phased out in 2015.
2. Reducing the size of the revisions should be a high priority for future research. It is important to understand what level of revision error is acceptable in terms of its effects on the GDP estimates which are currently not well-known. It is possible that this research is important to other EU countries as well. If so, some collaboration with other EU countries is encouraged.
3. With the launch of the new web version of the 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.
4. More research is needed to better estimate the trade below the cut-off limit for Intrastat for reassurance that it is insignificant.
5. We applaud the efforts of FTG staff to understand the effects of CN8 (Combined Nomenclature Goods Codes) coding error on the trade statistics through the asymmetry studies that have been conducted. Additional studies are needed especially because of the plans to move to the SIMSTAT system in 2015.
6. We believe the QD should be updated annually at least to include the findings of the many research studies that have been undertaken. Furthermore, it should speak more directly regarding size of revision error and its affects. One useful addition would be a comparison of the revision error for Statistics Sweden foreign trade statistics and those of other EU countries. In addition, errors in the industry coding and their potential effects on estimates of foreign trade by industry need more discussion in the QD.
7. FTG should plan to meet with larger enterprises whose late responses are an important cause of revision error. While one meeting was held with a tardy enterprise in spring 2013, more are needed. In addition, the results from these meetings should be well-documented.

The FTG staff noted that the conversion from VB6-based to web-based IT system presents a rare opportunity to considerably improve the current production processes. Greater efficiencies and data quality improvements could be made by implementing the many ideas the staff have collected ideas over the years. In addition, any impact of the new production system on the FTG statistics would need to be studied. Currently, no resources have been allocated to the staff to implement the necessary conversion and evaluation steps.

Exhibit 7. Foreign Trade of Goods (FTG), Ratings for 2013

	Error Source	Average Score round 2	Average Score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control for error sources)	**Specification error**	58	58	○	○	◡	◡	○	**M**
	Frame error	58	64	○	○	◡	◡	◡	**L**
	Non-response error	66	68	◡	◡	◡	○	◡	**M**
	Measurement error	62	64	◡	○	◡	◡	○	**H**
	Data processing error	60	66	◡	◡	◡	○	◡	**H**
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Model/estimation error	80	82	◡	◡	◎	◎	◡	**M**
	Revision error	70	72	◡	○	◡	◡	◡	**H**
	Total Score	64,7	67,6						

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

4.2.4 LABOUR FORCE SURVEY (LFS)

In our Round 2 report, we noted that response rates for the LFS have deteriorated over the years. As we expected, the response rate has continued its decline over the last 12 months, from 71 percent to 69 percent and even lower for some months. Generally, response rates have declined of 2 percentage points per year since 2006. The LFS staff that we spoke with in the Orebro call centre believe this is reflective of changes in society: people tend to avoid phone calls from unfamiliar numbers and when they do answer, they tend to be uncooperative. However, the LFS staff in Stockholm believe the downward trend in response rates have more to do with a combination of organizational and workplace cultural issues. Among these issues are: (a) inability of the current call scheduling system to schedule calls as necessary to achieve high contact rates; (b) insufficient staffing in the call centre; (c) workplace agreements that do not provide sufficient flexibility for the existing interviewing staff to work the hours required to optimize contact and cooperation rates, (d) a supervisory field staff lacking the knowledge, expertise and tools to manage the workload in a way that maximises response rates, and (e) a culture that believes that falling nonresponse rates are inevitable.

We believe the root causes of the declining response rates are related to both societal and organizational issues. Indeed, response rates to household surveys, particularly for telephone surveys, have dramatically decreased worldwide over the last 10 years. Both noncontacts and refusals have increased in telephone surveys even when optimal telephone call-back protocols are followed. However, our review of LFS data collection processes suggests that organizational issues may be the primary contributor to the problem. In order to maximise response rates, a fully functional call scheduling system that uses state of the art approaches for predicting contact probabilities for each case at each time slot is essential. Staffing in the call centre must be sufficient at times of peak calling activity to ensure that cases that are assigned to a given time slot are called in that time slot. Further, the supervisory staff should be motivated to achieve high response rates and knowledgeable about how to optimally manage the call centre to achieve its full capabilities. These areas appear to be lacking at Statistics Sweden and, unless they can be adequately addressed, response rates will continue to decline.

We believe it should be possible to obtain substantially higher response rates if the organizational issues were resolved. However, even then, response rates could still continue to the decline due to issues in the external environment. Using adaptive design and other innovative approaches could slow this decline, but that would require considerable effort and ingenuity.

We were informed that Statistics Sweden has let a call-for-tender to outsource 5 000 interviews, of 29 000 in total, for the LFS. Our understanding is that the primary objective of this project is to determine what response rate is achievable by an external organization that is not burdened by the aforementioned organizational issues that exist at Statistics Sweden. It could also relieve the workload pressure at the Orebro facility which could positively impact response rates for the main LFS.

Although much can be learned from this experiment, we believe the results, including costs, should be interpreted very carefully. For example, a confounding factor in the experiment is that the contractor's workload (5 000 cases) is less than 20 percent of the LFS main survey workload. Response rates for smaller scale operations tend to be higher due to the ability to find highly qualified interviewers in sufficient numbers to handle a small project. Thus, the analysis, before continuing or extending the use of the contractor, should consider the differences in interviewer capabilities between Statistics Sweden and the contractor's facility when interpreting the results.

We were also quite interested in learning about the concerted efforts to understand and mitigate the risks of nonresponse for household surveys that are proceeding under the Nonresponse Project. The goals of this project are to (a) reduce nonresponse rates for the LFS and other demographic surveys (b) achieve greater control over the telephone data collection process, and (c) reduce the costs of data collection. We have the following points to make with regard to these efforts. First, we believe goal (a) is somewhat misguided. Rather than reducing nonresponse rates, the goal should be to reduce nonresponse bias. These are very different goals when one realizes that reducing nonresponse rates could actually increase nonresponse bias. Further, the reduction of nonresponse bias is sometimes accompanied by a slight increase in nonresponse rates. In this regard, the focus of the nonresponse initiatives should be on the reduction of nonresponse bias in the published labour force estimates. For example, rather than increase the unweighted response rates, emphasis should focus on *weighted* response rates as the latter is better measure of the risk of bias in the estimator[3]. In addition, efforts to increase response rates should target groups known to substantial biases in the final estimates such as temporary workers and the youth.

With regard to (b), priority should be given to the strategy for contacting cases and the capability of the current data collection system to implement an optimal call strategy. Our discussions with staff suggest that the current call scheduler has important limitations for optimally allocating calls across time and that staffing levels in the call centre are not fully controllable. Likewise, there are no records of the time and day of the call attempts made by field interviewers; thus, it is impossible for supervisors to review call histories for field cases or to develop an optimal call strategy for field interviewing with accountability. Thus, some of our recommendations are aimed at correcting these critical deficiencies with the current system.

Finally, with regard to (c), the costs associated with even small reductions in the nonresponse rate can be considerable and, from a total survey error perspective, may not be optimal in the sense that investing the same resources in other areas of the LFS would produce much greater gains in data quality.

Some questions that the LFS staff should try to address are:

- What are the biases due to nonresponse in the adjusted estimates of labour force status? Is the current bias acceptable? If not, what level of bias in the estimates is acceptable?

- If response rates were increased by five percentage points, how much might the nonresponse bias be reduced? Unless, the effort is effectively targeted, there may be no actual reduction in nonresponse bias. Is achieving that milestone worth the costs which are likely to be considerable?

- What are the opportunity costs of substantial investments in nonresponse reduction? That is, what other important improvements in data quality, not just for the LFS but for other Statistics Sweden data products, are being suspended as more and more resources are being consumed by the nonresponse problem?

[3] The weighted response rate is defined as the sum of the selection weights of the eligible respondents divided by the sum of the selection weights of the eligible respondents and nonrespondents.

Notwithstanding these important organizational and system deficiencies relating to nonresponse, there have been a number of notable improvements in the LFS since our last review. These include the following:

- A previously mentioned, a project was launched aimed at reducing nonresponse and data collection costs in the LFS and to exert greater control over the field work although it is not clear whether the team members can devote sufficient time to the research activities.
- In addition, a call-for-tender went out to determine the extent to which Statistics Sweden's organizational issues are suppressing response rates.
- A reinterview study of 2000 responding households has been completed and an evaluation of the measurement error in labour force and other LFS statistics is now underway.
- A number of reports were published during the year including seasonal adjustment, linking of time series 1970-1986, youth unemployment, and a new indicator for "unemployed part-time workers seeking jobs".
- The sample was somewhat redesigned to reduce the number of strata which is expected to increase the stability of the estimates.
- A study of new auxiliary variables for nonresponse and coverage adjustments was completed that focused on auxiliary variables primarily from registers that heretofore have not been used in the adjustment process.
- A promising study of measurement error using Markov latent class analysis was conducted and paper documenting the preliminary findings was written.
- A new project is being launched to investigate the use of R-indicators during data collection to achieve greater sample representativity and a reduction in nonresponse bias.
- A cognitive evaluation of the LFS questionnaire focusing on specification error and measurement error was completed and a report summarizing the results was written.
- A new estimator of number of temporary workers was developed based upon a special sample of 8 000 unemployed persons. This estimator has much better statistical properties than the estimator it replaced. A report documenting the results is being written.

Exhibit 8 displays the changes in ratings between Rounds 1 and 2 resulting from these improvements as well as from the deteriorating state of the LFS nonresponse problem. We have the following recommendations for improvements:

1. The objective of the nonresponse work should be to reduce nonresponse bias, and not necessarily to reduce the nonresponse rate. Research should focus on quantifying the nonresponse bias, understanding its major determinants, and reducing the bias by increasing the weighted response rates especially for population subgroups most responsible for the bias.
2. In that regard, a better measure of the risk of nonresponse bias is the weighted response rate, rather than the unweighted rate. The LFS should begin monitoring weighted response rates (or nonresponse rates, if preferred) in addition to their unweighted counterparts.

3. The call monitoring system should be evaluated for its impact on cost, respondent burden and data quality. Statistics Sweden should investigate how call monitoring could be done less obtrusively and with much greater unpredictability by the interviewers being monitored.
4. Address the organizational issues noted above that are causing nonresponse rates to be much lower than could otherwise be attained.
5. Conduct a desktop study to evaluate the costs associated with using face to face interviewing for at least a portion of the LFS sample. For example, cases having low predicted response propensity at Wave 1 might be sent to the field to be interviewed by face to face. Subsequent waves could either continue with face to face or switch to the telephone or internet according to the respondent's preference.
6. Relatedly, conduct studies that seek to evaluate the bias in the fully weighted and adjusted LFS estimates. How effective are the nonresponse adjustments at compensating for nonresponse? Are better methods available that would lower the residual risk of nonresponse bias?
7. Conduct studies of rotation group bias to examine the extent to which it exists in the LFS and its causes. Our understanding is that a study was carried out in 1999 but it has not been well publicized, nor is there any mention of it in the QD and it may have changed since 1999 given the changes in interviewing mode over that time.

In the prior ASPIRE rounds, we noted that nonresponse and measurement error were high priorities for future research. These continue to be high priorities in the coming year and we are sure nonresponse will receive top priority. We commend the efforts of the LFS staff to address these issues, particularly the exemplary work on measurement error using both the reinterview methods and the latent class analysis of the LFS longitudinal data files.

Exhibit 8. Labour Force Survey (LFS), Ratings for 2013

	Error Source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	70	70	◡	◡	◡	◡	◡	**L**
	Frame error	58	58	◡	◡	◡	◠	○	**L**
	Non-response error	52	52	○	○	○	○	○	**H**
	Measurement error	56	68	◡	◡	◡	○	◡	**H**
	Data processing error	62	62	○	○	◡	◡	◡	**M**
	Sampling error	78	80	◡	◎	◡	◎	◡	**M**
	Model/estimation error	60	64	○	○	◡	◡	◡	**M**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total score	60,9	64,3						

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

4.2.5 STRUCTURAL BUSINESS STATISTICS (SBS)

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

The areas of improvement were as follows.

- The QD was published in 2013 resulting in users being much better informed on the quality of SBS statistics and so being able to use these statistics in a more informed way. There is scope to improve the QD, particularly in the material describing the survey structure which is a little difficult to understand, and we would encourage this during 2014. More quantitative material from research studies could be added.
- Electronic data transfer of SBS questionnaires from respondents has continued to increase. This should lead to higher quality statistics although it has not yet been proven.
- There have been experiments with the earlier despatch of the annual returns. This has led to quicker and higher response.
- There is some additional delivery of data from Tax Agency to enable the refinement of calendar/reference year estimates. This will support more accurate estimates of annual SBS especially in the event of an unstable economy.
- Work has commenced with use of the cognitive laboratory to better understand measurement error on the SBS questionnaire. Improvements to the web questionnaire have been made as a result.
- There is now some verification of the keying of paper forms.
- There is a project in place for developing better structured metadata, with a supporting IT-system, offering potential gains in data processing quality, more transparency and more possibilities to monitor processes.
- The Service Level Agreement with the NA department has been further developed although co-operation was already good.
- Although over-coverage in the BR, because of inactive units, is still a problem it is becoming less of a problem because of the reduced number of inactive units on the BR most of which have zero employees. Furthermore, there is a project in collaboration with the BR on how to reduce this problem in the new BR.

The areas of deterioration were as follows.

- The number of profiled businesses is continuing to decline resulting in some serious deficiencies in the industrial classification of SBS including breaks in series. For this and other reasons the number of Kind of Activity Units (KAUs) continues to decline even though one might expect it to increase.
- The threshold for the BR maintenance group's validation work has been raised from businesses with 10 employees to 15 employees. This can potentially affect the quality of statistics for small businesses in the service sector in particular.
- There is concern that the new BR will not support as many statistical improvements as previously anticipated with possible quality impacts. SBS were asking themselves whether they had not been strong enough in arguing for the importance of these improvements. Communication between the SBS and the BR does not seem as strong as it might. The BR staff are proposing to re-introduce the User Group and this is a good move.

Last year, we made six recommendations. Two of those recommendations have been implemented or are in progress. These are (i) the recommendation to do further questionnaire testing using the cognitive laboratory, and (ii) to increase the electronic submission by enterprises of their data hopefully based on their own chart of accounts.

The focus of further improvements should be on those areas of higher risk where the rating is relatively low. The following four recommendations from last year are still valid.

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.

2. Although the statistical improvements in the BR have been delayed, SBS should start thinking about the work required for moving to the new BR and what the implications are for survey continuity. There are likely to be discontinuities in the SBS data series and some thought should be given to how to manage these discontinuities and whether any additional information is required. For example, over-coverage because of inactive units may be significantly reduced with the new BR.

3. SBS should obtain more quantitative data that would help it evaluate errors from editing, imputation and the modelling of the more detailed items required by NA.

4. The EU standard on revisions is that "Revisions are regularly analysed in order to improve statistical processes". It appears that more analysis could be undertaken to understand the nature of revisions and how to possibly reduce them. The earlier involvement of NA may assist with the reduction of revisions. Their work enables them to have a good overview of the economy.

In addition, we propose the following recommendations.

5. There is a project to improve the structure and storage of metadata. This has been approved by SBS management but has not yet been supported financially. This is an important project for the quality of SBS including accessibility. A way should be found to support this initiative.

6. As noted above, the number of questionnaires collected electronically has increased. Studies show the data is different to when it was 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 9. Structural Business Statistics (SBS), Ratings for 2013

	Error Source	Average Score round 2	Average Score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	**Specification error**	54	58	○	○	◒	○	◒	**M**
	Frame error	64	60	◒	◒	◒	○	◓	**M**
	Non-response error	70	70	◒	○	◒	◒	◒	**M**
	Measurement error	52	56	○	○	○	○	◒	**H**
	Data processing error	60	60	○	○	◒	○	◒	**H**
	Sampling error	84	86	◎	◒	◒	◎	◎	**M**
	Model/estimation error	56	48	○	○	○	◓	◓	**H**
	Revision error	56	54	○	○	◒	◓	◓	**H**
	Total score	60,8	60,1						

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

4.2.6 LIVING CONDITIONS SURVEY (ULF/SILC)

The Survey of Living Conditions (ULF/SILC) is a long-standing survey dating from mid-1970.The survey has undergone a number of expansions, most notably the merging of the Eurostat SILC survey with the older ULF survey. As noted in our review last year, the cumulative effect of these changes produced a survey design that was quite complicated and unwieldy. We noted that the interview administration is quite complex. We also noted that an important deficiency of the ULF/SILC design was that selection probabilities are unknown requiring statisticians to assign equal selection probabilities where they are clearly unequal. We also noted a number of other important methodological issues facing the producers and users of the ULF/SILC. Some of these include:

1. The interview, which is conducted by telephone, averages 35 minutes but can be more than one hour for some situations. In such long telephone interviews, the reliability of the data, particularly for items placed at the end of the interview, is suspect.

2. Attempts have been made to adjust for nonresponse using calibration methods based upon demographic variables. However, unlike the LFS, the ability of such variables to adequately compensate for nonresponse bias in the key survey estimates has never been evaluated.

3. Children as young as 10 years old are interviewed for an average of 20 minutes by phone. Data collected from children are subject to reliability issues and this is exacerbated by the telephone mode.

4. Response rates, which average between 55 and 58 percent, have declined steadily over the years and tend to vary considerably by interview component.

5. Given the long history of the survey, the questionnaire is sorely in need of refreshing and updating. We noted that specification error posed a considerable risk to data quality primarily because an expert review of the survey questions had never been undertaken within the last 20 years.

6. Frame error is an important concern. Both undercoverage and overcoverage are important issues for the ULF/SILC yet the error sources have never been evaluated. Collaborative studies with the TPR staff are needed and should be given a high priority.

In the present review, we were pleased to learn that the ULF/SILC has responded very positively to the comments in our prior review and, as a result, the survey holds distinction as the product showing the greatest improvement in ratings of the ten products in Round 3. The staff of the ULF/SILC are to be commended for the impressive improvements that were made and are being planned for the future. Among these are the following.

- To address point 1, the length of the interview will be reduced by an average of 5-10 minutes in the coming year.

- The longitudinal component of the ULF will be dropped and thus the issue noted above with regard to unknown selection probabilities for part of the sample will be eliminated.

- The supplementary sample of persons 65 years and older will be dropped due to concerns about the quality of this portion of the survey.

- Beginning in 2015, the 10 to 11 year olds will no longer be interviewed, partly addressing the problems noted in point 3.

- Interviewers in six regions have received special training on how to increase contact and cooperation rates in the survey. These training sessions seem to have had a positive effect on the interviewing although there has been no formal evaluation.

- In response to point 5, the ULF questionnaire was reviewed by the Cognitive Laboratory who suggested a number of changes to the questionnaire. This review could not be extended to the SILC, however, since its content is regulated by the EU.

- As we recommended, the QD for the survey was revised and improved but has not yet been published.

There are a number of other improvements that are mentioned in the ratings table that appears in Exhibit 14.

In the next few years, the ULF/SILC faces some important changes that are mandated by the EU. 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 also would like the micro-data to 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. For example, due to the risks of seasonal effects on the data, deciding when to cut-off data collection so that a micro-data file can be prepared poses a problem. If data collection were concentrated in the 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.

1. Given that the nonresponse rate for the survey is relatively high and increasing, there is an important need for an analysis of 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.

2. As noted in the discussion of the LFS, adaptive design approaches during data collection could reduce the risks of nonresponse bias. There has been some work in this area for the ULF/SILC. This work should be continued to develop an implementation strategy.

3. There is much concern among the ULF/SILC staff regarding the risks of interviewer coding errors because interviewers must make quick judgments to code open-ended responses in the field. The potential for coding classification errors as well as recency and primacy effects (see, for example, Tourangeau, Rips, and Rasinski, 2000) is quite high yet these risks have never been evaluated. Evaluating and reducing the errors in the field coding process should receive a high priority in the coming year.

4. The effects of overcoverage of the TPR on the estimates have never been evaluated. It is a concern for the ULF/SILC and the effects of overcoverage on the nonresponse adjustment process should be investigated in the coming year.

5. The household composition information that is becoming available on the TPR could be exploited for use in the ULF/SILC given the dependence of the survey on household rostering approaches. For example, the TPR household data could be used to evaluate the accuracy of the ULF/SILC roster information. These areas should be investigated in the coming year.

6. 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.

7. As noted in Round 2, Appendix 16 from the so-called Appendix series on *The Swedish Survey of Living Conditions – Design and methods* should be updated. This document contains valuable information about the survey design but it is many years old. There is no document that provides this level of detailed information for the current ULF/SILC design.

8. Despite the increase of age-eligibility from 10 to 12, the data obtained in the children's survey is still a concern. A study of the reliability of these data is sorely needed.

9. An evaluation of the item nonresponse for the survey is needed given the extent of the problems in this area.

Exhibit 14. Living Conditions Survey (ULF/SILC), Ratings for 2013

	Error Source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	34	58	○	○	◡	○	◡	**M**
	Frame error	42	42	◠	◠	◡	○	◠	**M**
	Non-response error	40	46	○	◠	○	○	○	**H**
	Measurement error	46	52	◠	◠	◎	○	○	**H**
	Data processing error	42	50	○	◠	◡	○	○	**L**
	Sampling error	54	62	◡	◡	◡	○	○	**M**
	Model/estimation error	38	50	○	◠	◡	○	○	**H**
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
	Total Score	42,1	51,1						

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

4.2.7 BUSINESS REGISTER (BR)

There have been some important improvements over the last 12 months as noted below. However, there have been some deteriorations over the last year which we have also noted. There is no change on the overall ratings – the improvements and deteriorations cancel out. Some of the deteriorations are not direct responsibility of the BR Unit. However this review is of the BR as a product, not the Unit itself.

- There has been continuing work on the development of the new BR. Unfortunately, lower priority has been given to those developments which cover the areas impacting on the accuracy of the Register for statistical requirements and there is no definite plan for addressing these.

- The new BR system will have greater flexibility including the content of the Register.

- A closer relationship with the Swedish Tax Agency seems to have developed especially on reducing the problem of missing NACE codes. Since introduced in 2012, it is simply not possible to complete the registration form without a NACE code being inserted. They are also doing work to reduce the number of inactive units on their Register. As a consequence, the number of enterprises without a NACE code has continued to decline.

- Work has begun on a Study 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 at the time the statistical frames are provided.

- Steps are being taken to re-establish the User Group for internal users. This would be a positive step as communications with users are not always as good as they should be.

The number of missing NACE codes has halved over the last 12 months. It is now at a very acceptable level especially when you consider most of those without NACE codes have zero employees. This improvement is largely due to action taken in association with the Tax Agency.

Nevertheless, despite these improvements, we remain concerned about some aspects of the BR. It seems to have deteriorated in some aspects since our last review. Specifically, the number of inactive units on the Register seems to be increasing despite the efforts to reduce this number and there is some uncertainty about the extent of inaccurate NACE codes. Both these seem to be causing problems to the statistical areas who use the BR that we spoke to. These are the same two problems we referred to in the last two years and we are not convinced that sufficient action has been taken yet to address them especially the former.

However, the biggest concern seems to be the significant and continuing reduction in the number of kind of activity units which seems to be due to the reduction in profiling rather than any change in reality. This is causing a loss of accuracy of industry coding in important collections like SBS and consequently the NA. It is the responsibility of the Large Enterprise Group to do the profiling but they seem to be too ready to accept business arguments that it is difficult to provide information on the desired activity unit basis. Even if full accounts are not available on activity unit basis it may be possible to obtain partial information to support splits. This will provide more accurate statistics than the assumption that all of the business is part of a single industry. At present, only 40-50 units are being profiled. It is thought that the number should be much higher than this.

There was also concern expressed about the temporary change in the cut-off for maintenance of legal units from 10 employees to 15 employees. It was thought this might lead to further inaccuracies in the activity units. Although the change is only meant to be temporary, it is often difficult to reverse these changes in times of resource constraints.

As we noted last year, the preparation of the QD was an important development. In particular, it should help internal users understand the strengths and weaknesses of the Register especially if it contained more quantitative information. It has not been updated since last year. It should be treated as a dynamic document and updated as new information becomes available.

Some suggestions for future improvements are outlined below. These try to focus on the error sources of highest risk and where the rating is relatively low. Some are similar to last year. The only new recommendation is the first and possibly most important recommendation.

1. The procedures used by the Large Enterprise Unit for creating activity units need to be revised to ensure reasonable industry purity is obtained in business surveys and indexes. The number of profiled units needs to increase. If full data is not available for the desired active units, partial data should be obtained so that Statistics Sweden can impute for these industry dissections on an informed basis.
2. Planning needs to start shortly on the statistical improvements for the revised Business Register System as soon as possible so some definite milestones can be established. The emphasis should be on the most important quality improvements such as eliminating non-active units (overcoverage), supporting improved NACE coding, the introduction of new establishments for multi-establishment enterprises (undercoverage), and the introduction of a Common Business Framework. Unless the first three issues are addressed there will be continual deterioration in the quality of the BR.
3. 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 the statistical uses of the BR suffer as a consequence. This should be possible with the additional flexibility in the new Business Register System.
4. Although the relationship seems very sound at present, a Memorandum of Understanding should be developed with the Swedish Tax Agency to ensure both parties understand the modalities of the co-operation between Statistics Sweden and the Swedish Tax Agency.
5. The level of error in NACE coding should be monitored on an ongoing basis through an independent coding study. Can data from SBS be used to undertake some independent checking? The results of these studies should be made available to users, especially internal users. Methodologists at Statistics Sweden can assist with the design of the studies.
6. Descriptive information on industry should be obtained to support these evaluation studies and allow the NACE codes to be revised where necessary for the more significant enterprises. This would also enable the Tax Agency to audit the industry codes as there is some tax concessions are based on the industry classification.
7. The current arrangement of revising NACE codes when detected in the SBS potentially introduces biases. For example, if it is more likely that an enterprise coded to manufacturing will have its NACE code revised to a non-manufacturing enterprise than the reverse, this can create biases. (We are now advised that, although there is conjecture that this might happen, studies have suggested it is not a problem). These biases might be quite small but the significance of this potential bias should be evaluated to see whether it is important or not. If it

is important, then these NACE codes should only be changed for those enterprises in the completely enumerated strata or if the industry information is obtained from a source other than a sample survey. Regardless, some agreed operational rules should be established for when to revise NACE codes.

8. We encourage the work leading to the development of BR quality measures. Perhaps, the measures used in other statistical offices might be considered as part of this.

9. There should be some evaluation of the quality of employment data derived using models to assess whether the models are reliable or need to be revised in some way. There are four different sources for employment data. A statistic known as the Cohen Kappa might be useful for better understanding the level of consistency across the sources.

Exhibit 10. Business Register, Ratings for 2013

	Error Source	Average Score round 2	Average Score round 3	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	○	○	◡	◡	◡	**L**
	Frame error - overcoverage	56	58	○	○	◡	○	○	**H**
	Frame error - undercoverage	46	42	◠	◠	○	○	◠	**H**
	Frame error - duplication	63	63	○	○	◡	◡	N/A	**L**
	Missing data	48	48	○	○	○	○	◠	**L**
	Content error	50	52	○	◠	◡	○	○	**H**
	Total score	52,7	52,7						

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

4.2.8 TOTAL POPULATION REGISTER (TPR)

During the past year, TPR staff were quite involved with work on the 2011 Census and preparing those results for Eurostat who require them in March 2014. With the implementation of the 2011 Census, a major new initiative was also implemented – the development of a dwelling register based upon the assignment of four-digit dwelling number to each person on the TPR. This initiative represents an important improvement in the content of the TPR – the addition of household membership information.

The new dwelling unit information is not without its problems. The dwelling unit number is missing for about 320 000 persons as a result of nonresponse on the dwelling number. Evaluations that have been conducted so far suggest that the number of small households is underestimated while the number of large households is overestimated. Moreover, preliminary estimates suggest that the dwelling unit number is wrong for about 20 percent of population. These errors are expected to be reduced over time using daily updates from the Swedish Tax Agency.

As noted in previous rounds, overcoverage of the population is another important source of error for the TPR. A project that began in October 2013 and will be completed in April 2014 is attempting to quantify the extent of overcoverage. The hope is that a new variable can be created for each person on the TPR indicating the probability the person is not a Swedish resident. This information would be quite valuable to surveys using the TPR as a frame because overcoverage is confounded with survey nonresponse currently.

Approximately four times per year, two TPR staff members meet with a group of government authorities that includes the Swedish Tax Agency, Lantmäteriet (the Swedish mapping, cadastral and land registration authority) and the Swedish Association of Local Authorities and Regions (SALAR) to discuss issues related to the TPR and its quality. Such outreach efforts are a valuable source of information and communication for the TPR as well as for the agencies involved.
The Round 2 to Round 3 changes are shown in Exhibit 11 as well as the current ratings.

We include the following recommendations for the coming year:

1. A high priority should be given to improving the dwelling unit indicator that was added as a result of the 2011 Census. As noted above, there are a number of issues that need to be further explored and addressed in order to improve this information. Chief among these is to reduce the classification error resulting from wrong unit numbers being assigned as well as the amount of missing information in the variable.
2. We encourage the TPR staff to consider the use of logistic regression or equivalent approach for estimating the probability that an individual on the list is a resident. Such a model could be estimated based upon information on noncontacts, register activity or inactivity, personal characteristics and so on.
3. For studying overcoverage, it is not enough to simply report the overall rate of overcoverage in the TPR. The rate will vary considerably for important subgroups and these too should be estimated.
4. 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 effectiveness of the 'probability' indicator (see recommendation 2) in reducing overcoverage error should also be considered as part of this analysis.

5. 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.

With regard to (5), validity may be defined simply as the correlation between the register value of a characteristic and the true characteristic. Since the true characteristic will usually not be known, estimating validity can be quite difficult. However, some information on validity can be gleaned from the corrections that are continuously made to the TPR that flow from the Tax Agency, users, individuals, and other sources. The number of changes that occur per year and the magnitude of the changes could be tracked and reported. It may also be possible to form estimates of reliability of the data on the same variables that may be available from other registers.

Finally, as noted in our Round 2 report, TPR error evaluations should not proceed independently of the main users. It is important to understand how errors such as overcoverage affect the main uses of the TPR in order to assign an appropriate risk level and priority to the error source. In addition, working in collaboration with users can provide a better understanding of the issues that need to be addressed as well as their solutions. Therefore, we encourage the TPR staff to lead error evaluation projects in collaboration with main users of the TPR especially users within Statistics Sweden.

Exhibit 11. Total Population Register (TPR), Ratings for 2013

	Error Source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control for error sources)	**Specification error**	50	58	○	◔	○	○	◒	**M**
	Frame error: overcoverage	56	58	○	○	○	○	○	**H**
	Frame error: undercoverage	60	60	○	○	◒	◒	N/A	**L**
	Frame error: duplication	70	70	○	○	◒	◒	N/A	**L**
	Missing data error: item and variable	66	66	○	○	◒	○	◒	**M**
	Content error	58	62	○	○	◒	◒	○	**L**
	Total score	58,8	61,4						

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

4.2.9 QUARTERLY GROSS DOMESTIC PRODUCT (GDP)

The quarterly NA estimates are very complex products that rely 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 and GDP in particular. We also only looked at the production side of the quarterly NA. 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. The first two were chosen largely because of the significant contribution they make to the quarterly GDP whereas merchanting was chosen because it had been making a significant contribution to estimates of change in GDP and questions were being asked about the reliability of this data. All three were considered again in this round but the services and industrial production indexes were considered in much greater detail including a discussion with the product areas.

In addition to input data sources, we looked at errors from modelling, data processing, deflation, balancing and revisions.

The major improvements made over the last 12 months were as follows.

- Further work has taken place on the harmonization of the industrial and services production indexes.

- A number of macro edits are being developed and will be implemented next year in the estimation of quarterly GDP. When implemented they should reduce the extent of balancing that is required to synchronize the production and expenditure estimates.

- There has been a pre-study of the Finnish NA processing system and its potential for introduction to the processing of the Swedish NA. The IT staff were also involved in this evaluation. The better relationship between NA and the IT staff was noted.

- 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. The results of the study are encouraging.

- The Service Level Agreements have been further developed.

In our last report we stated that we believed the areas most in need of improvement, in priority order, were (1) a robust processing system for the NA that includes the time series dimensions, (2) evaluation of the models used for the important areas of intermediate consumption and construction, (3) review of the methodology for estimating merchanting services, (4) sensitivity studies on errors in the industrial production index, the services production index and the indexes used for deflation.

We are pleased to see that over the last year, work has taken place on (1) and (2) although further work needs to be done. However, no work has taken place on (3) and (4) and we still think these are important.

Last year, we strongly supported the short term economic statistics project which will integrate or harmonise those surveys supporting the industrial production index and the services production index. We are pleased to see this work has continued. It has become even more important with the reduction in profiling of large enterprises. This increases the risk of significant services activity being included in the industrial production index and significant industry activity being included in the services production index. There are often significant discontinuities in the two indexes at the time of the annual reselection. A major contributor may be the changes in profiling of large enterprises which are realized at the time of the reselection.

We also supported the development of standardized or objective methods for balancing the quarterly GDP estimates recognizing there will always be an element of human judgment involved in the balancing process. Statistics Sweden's practice of publishing the discrepancy prior to balancing, and the influence of different stages in the balancing process, is an excellent example of transparency in statistics. We were not aware of the latter at the time of our last report.

We have noted that there have been several research studies in NA that are not followed through to the implementation stage. This may be because of the limited capacity of NA staff to do research work. The many tasks involved in the compilation of the quarterly NA estimates have to take priority and it is difficult to dedicate much time to research activity. In the last report, we expressed concern about the proposal to discontinue the NA research group. It is important to have a group that can research NA although they don't organizationally need to be part of the NA staff. For example, in the ABS, NA research was undertaken by a Special Analysis Group that also researches price indexes, models, etc. that are used in economic statistics. It is easier to develop a critical mass this way although a close relationship with the NA and other users of their services is crucial. We understand this approach was tested in Statistics Sweden but did not work because the necessary strength of relationship between the two areas was not developed. There may be other possibilities for developing research capacity such as utilizing the research capability of the National Institute of Economic Research. In our view a NA research capability is essential for the long term welfare of the Swedish NA and the best way of providing this capability needs to be investigated as a matter of urgency.

Of concern is the level of experience, and possibly expertise, in NA staff with the large number of retirements in recent years as well as those in the future. We support any steps Statistics Sweden takes to build up this expertise in an area of statistics that is so crucial to the reputation of Statistics Sweden. Training programs exist but they are relatively short in duration and there is a lot of reliance on 'on the job' training. In our view, there would be benefits in a more formal approach and the development of an on-line training program that staff could undertake at their own pace but with tutorial type support from more experienced NA staff or those that have recently retired. Such training resources already exist (e.g. Eurostat) so it should not be a massive task for Statistics Sweden to adapt the existing resources.

With respect to improvement area (1) noted in the fourth paragraph, there is some urgency in developing an IT system because VB6 will not be supported in the future. The lack of a modern IT system is also an area of weakness in the Swedish NA although some improvements have been made to the current system in recent years. Without seeing the report that is due in January, the adaptation of the Finnish system seems like a good option. It may not be perfect for Sweden's use but it works for Finland. The use of the Finnish system is likely to be cheaper, be able to be implemented in a much shorter time frame, and with somewhat less risk. There are also benefits in being able to share knowledge with another official statistical agency at a similar level of

development and who work in a similar environment. The IT staff at Statistics Sweden should be involved in the adaptation.

As we mentioned last year, with respect to improvement area (2), questions marks have been raised about the validity of the model used to estimate intermediate consumption. For example, in times of declining economic activity it over-estimates intermediate consumption and therefore under-estimates GDP. The opposite occurs in periods of rapidly increasing economic activity. There has been research that looks at the possibility of using VAT data to estimate intermediate consumption rather than the somewhat simple models that are used at present (in Sweden and many other counties). This research is promising but effort needs to be put into this research so that a clear implementation strategy can be developed and activated.

For Construction, models have been used because it is difficult to get reliable estimates directly from surveys. It is an important sector of the economy and a strong indicator of general economic activity so it is important to have reliable estimates for this industry sector. We are pleased that there are definite plans to introduce SBS data to estimate some parts of the construction industry in the annual NA. This should provide a more reliable benchmark from which the quarterly estimates can be derived and this should be investigated.

With respect to improvement area (3), merchanting is a new area of statistics so it is not surprising there is some uncertainty. Statistics Sweden has now had several years of data collection experience so it would be timely to review the methodology perhaps in collaboration with another country with data collection experience with merchanting. We made this recommendation last year and it appears to be appropriate again.

With respect to improvement area (4), it is not always easy to understand the impacts on the NA of inaccuracies of the source data especially given the complexity of the processes used included the balancing processes. As we mentioned last year, one possibility is to use sensitivity studies where an error is introduced into a particular data source and the impact on GDP is assessed. This could be done for each of the key data sources in turn. The deflation indexes should be a priority as the producer price indexes are based on relatively small samples and may be somewhat volatile. This is likely to be an expensive operation so should be seen as a one-off exercise or one that is only undertaken every now and then. We do not know whether it is feasible or not and there may be other methods for approximating the impacts. However, it is worth investigating and the NA methodologist could be asked to investigate this. The objective is to assess the relative importance of the different input data sources to help focus data development effort. If necessary, we can make suggestions on the design of the sensitivity studies.

In conclusion, our first two recommendations are modifications of the equivalent recommendations last year; the next two recommendations are the same as last year but still valid in our view, and the next five recommendations are new.

1. Continue to investigate the adaptation of the Finnish NA system as a replacement system for the Swedish NA.

2. Take to the next stage the research on the use of VAT data for estimation of intermediate consumption. Develop an implementation strategy for the use of VAT data to estimate intermediate consumption. Examine revised approaches for the estimation of quarterly construction activity.

3. Review the methodology for estimating merchanting services.

4. Undertake sensitivity studies of the relative importance of the different source data on the accuracy of the GDP estimates. Put priority on the sensitivity to the deflation indexes.

5. Investigate the most appropriate mechanism for developing some dedicated research capacity in the NA staff.

6. Prepare a formal training strategy for new staff in the NA, based on training resources that are available both internally and externally.

7. The revised European System of National Accounts will be introduced next September. This will be the culmination of a lot of work and some unanticipated additional work will be required close to implementation. This will draw resources from the compilation of the quarterly NA putting their accuracy at risk. Some supplementation of NA resources should be considered during this crucial period.

8. We heard about biases in the indexes for the manufacturing industry. Alternative models are being examined. Given the importance of manufacturing to the GDP estimates, this work should be completed as a matter of priority.

9. We support the development of principles and guidelines for objective balancing as well as the further development of the quarterly supply use tables.

Exhibit 12. Quarterly GDP, Ratings for 2013

	Error Source	Average score round 2	Average score round 3	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	60	62	○	○	◡	◡	○	H
	Input data source - Index of Industrial Production, IIP	60	62	○	○	◡	◡	○	H
	Input data source - Merchanting Service of global enterprises (also covers royalties, licensing and R&D)	44	44	◠	○	○	◠	○	H
	Compilation error (modelling)	48	48	○	○	○	○	◠	H
	Compilation error (data processing)	44	52	◡	○	○	◠	○	H
	Deflation error (including specification error)	48	48	◠	◠	◡	◡	◠	H
	Balancing Error	56	56	○	○	○	○	○	H
	Revisions Error	56	58	◡	○	◡	○	○	M
	Total score	51,8	53,6						

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

4.2.10 ANNUAL GROSS DOMESTIC PRODUCT (GDP)

As is the case with the quarterly NA, the annual NA estimates are very complex products that 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 risk to the annual NA estimates and GDP in particular. As with the quarterly GDP, we only looked at the production side of the annual NA.

In addition to the input data source, we looked at errors from modeling, data processing, deflation, balancing and revisions.

The most important areas of improvement over the last 12 months were:

- Further development of the detail in the SLA with SBS.
- The work leading up to the planned introduction of SBS data into part of the construction industry estimate from the second quarter in 2014, replacing estimates previously used through modelling.
- A pre-study into the Finnish NA processing system with a view to adapting it for the Swedish NA system. The ICT Department was involved in this study.
- The completion of the standardized spreadsheets will reduce the risk of processing error.
- Macro edits are to be introduced in 2014 which will assist with the balancing as well as picking up errors earlier in the processing.

On the other hand, the reduction in activity units on the BR is causing deterioration in the accuracy of the industry data for SBS which will also impact on the accuracy of the annual GDP estimates even though changes in NACE codes of large enterprises are taken into account by the NA staff when compiling the annual accounts.

We also noted several evaluation studies undertaken by NA staff in their limited spare time. However, resource constraints seemed to limit their ability to implement findings.

We believe the areas most in need of improvement are (1) a robust processing system for the NA estimates that includes time series dimensions, (2) evaluation of the models used for estimating the trade margins which appears to be the area of greatest weakness in modeling, (3) sensitivity studies on errors in the indexes used for deflation especially the producer price indexes where the samples are relatively small, (4) given the loss of experienced staff, an upgrade in the extent of training provided to new staff and (5) an increase in the research capability for the NA staff. The first listed is the highest priority.

The first three recommendations were the same areas identified last year but only (1) has been addressed over the last year due to resource constraints.

We note the plan to introduce the new European System of NA during the September 2014. This is an important and significant task and will severely limit the capacity to make other developments with respect to the annual NA estimates. During this crucial period it might be

prudent to supplement NA resources. The size of the NA group is relatively modest compared with other developed countries.

With respect to improvement area (1), the suggestions are the same as for the quarterly NA estimates.

With respect to improvement area (2), the estimates derived from the SBS are unrealistic so other methods are used. It is maybe unrealistic to expect accurate estimates to be obtained direct from the SBS. However, it would be worthwhile investigating the SBS to see whether any design changes or additional content are required to obtain better estimates of the trade margins and we understand some work is taking place in this respect. We note that the ABS periodically conducts a detailed survey to estimate margins at the product (group) level to assist with the estimate of trade margins. A study of international practices may be worthwhile as part of this investigation. The trade industries are important, especially in measuring changes in GDP, so it is worth the effort of investigating improved practices.

With respect to improvement area (3), as mentioned for the quarterly NA it is not always easy to understand the impacts on the accuracy of GDP of inaccuracies of the source data especially given the complexity of the processes used included the balancing processes. However, it may be more straightforward when just looking at the deflation process by changing deflators by a certain amount and looking at what the difference in GDP estimates. The volatility of the deflators also has to be taken into account in deciding the size of the 'error' to introduce into the sensitivity study. Also, the focus should be on those deflators where there is most concern about accuracy. The NA methodologist could assist with the design of these studies.

With respect to improvement areas (4) and (5), the comments made in respect to the quarterly NA also apply to the annual NA.

Exhibit 13. Annual GDP, Ratings for 2013

	Error Source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality
Accuracy (control over error sources)	**Input data source - Structural Business Statistics, SBS**	66	66	◡	○	◡	◡	○	**H**
	Compilation error - modelling	48	50	○	◠	○	◡	◠	**H**
	Compilation error - data processing	44	52	◡	○	○	◠	○	**H**
	Deflation error (including specification error)	48	48	◠	◠	◡	◡	◠	**H**
	Balancing Error	58	58	○	○	◡	◡	◠	**H**
	Revisions Error	56	56	○	◡	○	○	○	**M**
	Total score	53,2	54,9						

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

5. GENERAL RECOMMENDATIONS

5.1 PROGRESS ON ROUND 1 AND 2 RECOMMENDATIONS

The general recommendations made in Rounds 1 and 2 are still relevant for this round. Although progress has been made on most recommendations, there is still much more to do. Efforts related to these recommendations should, therefore, be considered as a process of continual improvement. Below we list the recommendations and describe the current state of play as far as we can assess.

5.1.1 Need for Integration of Economic Statistics

A number of initiatives could be taken to improve the integration of economic statistics. There is still much work that needs to be done but we were pleased to see:

- Commencement of work on the establishment of a Common Business Framework (CBF) although we thought the objectives of the CBF might be stronger than proposed at present and the implementation of this aspect of the new BR has been delayed.
- The work on the integration of the surveys supporting the Services Production Index and the Industrial Production Index is well advanced.

This recommendation remains valid. In particular, it is important that a revised time line is set for the work on the design of the CBF.

5.1.2 Lack of Co-operation between the NA staff and Statistical Areas

There has been significant improvement in the relationships over the last two years in part due to the continuing work on the Memoranda of Understanding and more frequent meetings with source data areas. Furthermore, our judgment is that the relationship was good for each of the key input data sources that we considered.

It could be said that this recommendation is well on the way to being implemented but good co-operation requires on-going effort. As noted in the reviews, there are still areas where communication between the NA staff and the statistical areas could be better.

5.1.3 Evaluating the Accuracy of NACE Coding

This is a continuing concern of the statistical areas using the BR. They felt the problem was getting worse. As we noted in our last report, we were pleased to see that an evaluation study had been undertaken of the accuracy of NACE coding by registered enterprises. However, we had criticisms of the nature of the study especially the reliance on dependent coding. This approach has been shown to lead to an under-estimation of coding errors. We suggested that the methodology group be asked to assist in the design of a new coding study that uses independent coding. There appears to have been no work on assessing the accuracy of NACE coding since our last report.

As discussed below, the reduction in the extent of profiling of large enterprises is a further adverse influence on the accuracy of industry statistics. In particular, it has resulted in some discontinuities in important series when samples are refreshed.

This is a difficult area but it is recommended that Statistics Sweden develop a strategic plan for maintaining adequate accuracy of NACE coding. This should cover matters such as:

- Evaluation studies that provide insights into the sources of errors in NACE coding as a starting point to improve the procedures for obtaining more reliable codes.
- When to use survey feedback to revise NACE codes.
- The need for special studies from time to time to update NACE codes especially for the larger businesses.
- The reasons for the reduction in the number of KAUs especially for large enterprises. (See also recommendation 5.2.1 which further elaborates on this idea.)
- The procedures the Large Enterprise Unit should adopt for identifying KAUs when not all the required data are available.
- The need for an ongoing evaluation study of the accuracy of NACE coding.

5.1.4 Need for Additional Evaluation Studies

We are pleased that some new evaluation studies have taken place since we started ASPIRE, particularly for measurement error in household surveys. There is scope to improve the design of the studies and greater involvement of the methodology group is recommended. Better coordination of the evaluation studies is recommended. If the studies are well designed and the results accumulated, it may be possible to generalize the findings of the studies for wider application through Statistics Sweden. This also facilitates the use of meta-analysis which takes advantage of common findings across studies.

Statistics Sweden has a proud history of methodological research. It has been one of the leaders in the official statistics world. There have been many past studies but we have been surprised about how difficult it has been to find documentation on some of these past studies. There needs to be a system for archiving them and we would recommend the methodology group take this responsibility (see recommendation 5.2.4).

5.1.5 Increasing Nonresponse Rates in Household Surveys

Since our last report, response rates for household surveys have continued to deteriorate despite the very significant efforts devoted to ameliorating this problem. Although declining response rates increase the risk of nonresponse bias, the magnitude of the bias depends upon both (a) the nonresponse rate and (b) the differences between respondents and nonrespondents. Note that (b) can be made small even though (a) is large 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 on attempting to increase response rates, particularly for the LFS. Apparently, 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 result.

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. However, we believe the work would benefit by engaging an external consultant to advise the project team. We recommend that a strategic review of 2 to 3 months duration be established to address this problem and make recommendations of the future Statistics Sweden strategy. This should be a dedicated team that largely builds on existing knowledge. It should be facilitated by the external expert who would bring their specialist knowledge to the review.

We attended a presentation by the current nonresponse team and liked the fact that the team was focusing on only three key objectives. This was different from the previous year where we felt that the objectives of the nonresponse study were too diffused. The three objectives are:

1. Improve response rates in the LFS.
2. Obtain full control over the data collection process.
3. Reduce the cost of data collection.

In our view, the first objective should not simply focus on nonresponse rates. For example, during our review visit in November, the October LFS results were released. We were shown the effects of special efforts to increase the response rates among the younger age groups during the last days of data collection. This effort had virtually no effect on the bias ostensibly because these late respondents were more like the existing respondents than the remaining nonrespondents they were intended to represent. This incident demonstrates why simply aiming for higher response rates for under-represented demographic subgroups may not be an effective strategy.

We suggest that objective 1 be modified to "Stabilise nonresponse rates and reduce the bias in the household survey estimates, particularly for the LFS." This would involve altering the data collection procedures to achieve household survey samples that are less subject to nonresponse bias before post-survey adjustments have been applied. Such a strategy would rely less on post-survey adjustments to reduce the bias and could even make such adjustments more effective. Certainly, developing more effective nonresponse adjustment approaches is an important sub goal that is embedded in this restated objective.

We note that a number of positive initiatives have taken place and these need to be continued. As we have mentioned previously, the call scheduling is far from optimum especially in regard to evening and weekend calling. In part, this is due to limitations specified in the workplace agreements with the household interviewers. We support the steps to revise the agreement for centralized interviewers and the plans to do the same for the field interviewers. Given the significance of the problem, it is a worthwhile experiment to use a private firm where the interviewers have less restrictive working conditions. However, the experiment needs to be carefully evaluated to ensure that costs and quality implications of the tendered sample can be correctly interpreted and that changes in the error profile of the published estimates are well-understood.

The culture within the household interviewing team can be a big influence. In our last report we commented on the defeatist attitude among the field management team in that they were very pessimistic about being able to improve response rates and how this attitude is likely to be passed on to interviewers and can become a self-fulfilling prophecy. We note that there have been changes in the field management team and addressing the culture among the interviewers has to be one of their most important tasks.

The world has changed with respect to response rates in household surveys. The global trend has been for response rates to decline due to the increasing difficulties of contacting people by telephone and their propensity to refuse once contacted as a result of the increasing rate of unsolicited telephone contacts (for e.g., telephone marketers). This is why we believe more effort should be put into reducing nonresponse bias through improved data collection approaches such as adaptive design and by statistical adjustments rather than just rely on reducing nonresponse rates.

It may well be that the nonresponse bias problem is not as great as perceived by many users when the data are appropriately weighted. We recommend using the weighted nonresponse rate, rather than the unweighted nonresponse rate, as the indicator of the risk of the nonresponse bias. For

directing nonresponse follow up efforts, there are other (but difficult to compile) indicators such as the R indicator which might also be considered.

5.1.6 Improving the Relationship with the Tax Agency

Over the last two years, the relationship with the Tax Agency seems to have improved for both the BR and the TPR. There also seem to have been more regular meetings. However, we still think the development of a Memorandum of Understanding worthwhile. This means the arrangements are well understood as staff change over time as inevitably is the case. As staffing at the Tax Agency changes, the new staff may have a less positive attitude towards supporting Statistics Sweden.

5.1.7 Establishing a Policy on Continuity of Statistical Series

We understand there is no policy yet but we still think a policy would be worthwhile. As previously stated, we suggest the Statistics Sweden policy specify that every major redesign include some provision for bridging the series before and after the redesign unless an explicit exemption is granted by the Director General. Besides affecting the accuracy of trend lines and estimates of temporal changes, this policy is needed to address Comparability & Coherence, which are also critical dimensions of survey quality.

5.1.8 Improving the Relationship between IT and their Client Areas

We did not specifically consider this in the current review but we sensed there had been some improvement. In the past reviews, the NA staff have been the most critical of the relationship with IT but they were much more positive on this occasion and noted that some IT staff accompanied NA staff on their visit to Finland for a pre-study of the suitability of the Finnish NA system for Statistics Sweden. About 12 months ago the IT department, with the encouragement of top management, appointed contact persons with each of the other departments. This may have been an important influence. Given there has been some improvement in this area we make no specific recommendations in this round other than to continue this progress.

5.1.9 Lack of Telephone Interviewing Monitoring

We were very pleased to see the introduction of telephone monitoring for the LFS and ULF/SILC. As we mentioned in last year, we did have concerns that the interviewers were pre-warned that there was a 50% chance that designated interviews were to be monitored. We retain those concerns as it may result in their behavior being different for those interviews compared to all their other interviews and that the telephone monitoring would not pick up all the weaknesses in the interviewing system. We were given some anecdotal advice that this was the case.

The normal practice is to warn interviewers that some of their work would be monitored to understand weaknesses in the system, retraining, etc. but not to specify which interviews were liable to be monitored. We understand there may be some staff union issues to be negotiated and possibly some legal issues when it comes to informing respondents of the monitoring. Nevertheless, it is important to assess the effects of the monitoring alerts on monitoring effectiveness by comparing obtrusive and unobtrusive approaches and we retain our previous recommendation.

Moreover, although approximately 5 percent of the interviews are being monitored, the monitoring results are not being used effectively to improve interviewing and reduce

measurement errors. There should be substantial effort in the coming year aimed at optimizing the use of the monitoring results for improving the performance of the interviewer as well as for identifying problems in the questionnaire or interviews that should be addressed by a revision of procedures.

5.1.10 Development of Quality Profiles for Key Products

Quality declarations (QDs) exist for all the products we examined except the NA which utilizes the material they provide to Eurostat in the form of GNI Inventories in lieu of a specific quality declaration document. At the time of our last review, there had been improvements in the QDs for all the products we reviewed and many were published in the early part of 2013. We were disappointed that, except for LFS and ULF/SILC, there had been no updates to the QDs. These should be treated as dynamic (electronic) documents where new information is added as it becomes available. For most of the products, new information had been obtained during the year on quality. As previously mentioned, the most important improvement is to include more quantitative information on what is known about different aspects of quality particularly for those aspects where there is high risk.

5.1.11 Increase the Focus on Coherence between Relatable Statistics

This recommendation arose last year as a consequence of our review of Coherence with the LFS. As Coherence was excluded from the scope of this quality review, we did not examine factors that were directly relevant to this particular recommendation.

5.1.12 Initiate Succession Planning in Some Important Statistical Areas.

The two statistical product areas where this was of most concern were CPI and NA where a number of very experienced, capable statisticians had retired or were soon to retire. We suggested the process for identifying suitable replacements should begin now. In both areas the transition seems to have been managed reasonably well to date although it is important to provide for sufficient resources for the necessary training of new staff to take place. The use of recently retired staff can be an effective way of providing this training.

5.2 NEW RECOMMENDATIONS

As can be seen from section 5.1, there has been progress against many of our cross-cutting recommendations from the last two years. Nevertheless there is much work to be done with respect to most of these recommendations. This is where the focus should be and there are only a small number of new recommendations which are outlined below.

5.2.1 Provide clear instructions for the profiling of large enterprises

As noted in 5.1.3, the number of KAUs has declined over recent years apparently as a consequence of a reduction in the extent of profiling of large businesses. We did not talk to the Large Enterprise Unit but we were advised this was largely due a reluctance to create KAUs unless a reasonable amount of the required financial data are conveniently available for the KAU. The reduction in the number of KAUs will affect the accuracy of industry statistics and their continuity over time. For example, we were advised how it impacted the Services and Industrial Production Indexes when significant services activities were included in the Industrial Production Index because a separate KAU for these services activities was not created for a very large enterprise.

It may still be worthwhile creating a KAU even when financial data are not readily available. As an example, there may be ways to model the split among KAUs for financial data that are aggregated at higher levels. This needs to be negotiated with the enterprises themselves, presumably by the Large Enterprise Unit. It is recommended that a set of principles and rules be developed for when KAUs should be created. Furthermore, responsibilities need to be clarified as these are not clear at present. Although the Large Enterprise Unit should have the implementation responsibility, the business statistics areas and the BR staff should be involved in establishing the principles and rules.

5.2.2 Develop a top down plan for the phasing out of Visual Basic 6 that is widely-supported and well-communicated to all departments affected by the plan.

Support for VB6 is being phased out in about 18 months. VB6 is used extensively in Statistics Sweden so a lot of IT work is required to replace it. Our impression is that the IT department has developed a cross-cutting plan for replacing VB6 but it has not received wide support by the affected departments. In addition, the subject matter departments seem now to understand that responsibility for phasing out VB6 is their responsibility. The current situation appears to be confused and in a state of flux. We suggest that a well-publicized, widely-supported central plan for phasing out VB6 be developed for the following reasons.

- Re-invention of the solution to the same problem should be avoided. There are advantages in developing a common approach to the phase out.
- It misses the opportunity to develop more common approaches across the Statistics Swedish IT environment.
- In some cases, the best solution may be to replace the whole IT system for a product rather than just the VB6 component especially if the IT system is in need of a redesign. Given the higher cost of this “whole system replacement” approach it may only be feasible for a very limited number of products and this should be a corporate decision.

Resources are limited and the changes for some product areas may be higher priority than others. For example, it may not be possible to complete all the required changes before VB6 is phased out. In this case the recommended strategy is not to make any changes to the systems to minimize

the chance of VB6 support being required. This restriction will be more viable for some products than others.

5.2.3 Develop a systematic approach for archival and retrieval of manuscripts and reports that document quality improvement projects and that are authored or co-authored by Statistics Sweden staff.

As mentioned above, Statistics Sweden has a proud history of methodological research. It has been one of the leaders in the official statistics world. There have been many past studies but we have been surprised about how difficult it has been to find documentation on these past studies. There needs to be a system for archiving them and we would recommend the methodology group take the responsibility for developing a systematic approach.

5.2.4. Launch an annual process for planning and monitoring projects that specifically address the recommendations in the annual ASPIRE reports.

This report contains 16 recommendations (including this one) for improving the quality of the 10 products in the ASPIRE review. 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 an enormous 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. These work plans should specify clear, individual-level responsibility, actionable goals with timeliness, and realistic resource allocations. Progress should be monitored at the highest levels in the organization to ensure that work progresses in a timely, effective and efficient manner.

6. SUMMARY AND CONCLUSIONS

As we stated in our previous reports, we believe Stat Sweden remains a world class organisation. In most of the products we evaluated for the second or third time we saw improvements with very few deteriorations. Nevertheless there have been a number of areas requiring improvement and these have been identified in this report.

We have reviewed the Accuracy of seven products for the third time and three products for the second time. As a result of further information available this time we have corrected some of the ratings. In the report, we have distinguished the corrections from improvements and Exhibits 4a and 4b shows the current ratings, prior year ratings, and the improvements by product. Justifications for the rating changes are summarized by product in tables that appear in Annex 2.

With a maximum possible score of 100 percent (indicating perfect quality), the product scores ranged from 51.1 percent (for the ULF/SILC) to 67.6 percent (for the FTG) with an average rating of about 59 percent. Products generally increased their scores in this round; quite substantially in the case of the ULF/SILC (a full 9 points!). The exceptions were SBS which showed a decrease and the BR which showed no change from Round 2. The average improvement in ratings over all products and error sources was about 2.7 percentage points. When combined with the 3.2 percentage point increase in Round 2, there has been a 5.9 percentage point increase since ASPIRE started in 2011 (see Exhibit 4c) which represents roughly an 11 percent average improvement in quality for these 10 products.

We reviewed the Accuracy of the NA estimates using the same approach as last year. Our analysis is somewhat restricted in that we have only reviewed GDP compiled from the production point of view. However, we have analysed the quarterly and annual accounts separately although, not surprisingly, there is a strong correlation between the ratings for the quarterly and annual accounts.

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 5 of this report. There is considerable overlap with the cross-cutting recommendations in Biemer and Trewin (2013). 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.

The household survey nonresponse issue is probably the major quality concern for Statistics Sweden at present. It is impacting on the accuracy of statistics but the perception of the decline might be greater than the reality. However, the costs are increasing to such extent that work in other parts of Statistics Sweden is also impacted. We have made a number of suggestions for addressing the problem. Most importantly, we think there should be a dedicated strategic review, supported by an external expert, over something like a 2 month period with the objective of deciding what household surveys should look like in about 3 years or so and the steps needed to

get there. The support of the Swedish Government should be obtained for this review as some of the findings may have funding implications.

For this round, we did make some further improvements to the methodology for the Accuracy reviews based on our experience with the previous reviews. However, the methodology was largely the same as we used for Round 2 including the use of checklists against which the products could answer “yes” or “no.” This worked well and provided a number of important advantages.

- It enabled us to make more objective assessments.
- It enabled us to make more consistent assessments across products.
- It provided additional information which was useful to us in our quality reviews.

No doubt there will be opportunities for further improvement but we expect the changes will be at the margin rather than to the basic approach.

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.

7. REFERENCES

Biemer, P. (2011) *Latent Class Analysis of Survey Error*, John Wiley & Sons, Hoboken, NJ.

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

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.

Lequiller, F; Blades, D. (2006). *Understanding National Accounts*, Paris: OECD 2006, http://www.eastafritac.org/images/uploads/documents_storage/Understanding_National_Accounts_-_OECD.pdf

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*, Cambridge: Cambridge University Press.

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**	

Exhibit 2.1 RS Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	There is essentially zero risk of specification error so the Risk level has been changed to not applicable (N/A). Specification error is no longer an issue as a result of the new system recently implemented. All costs are now reported as accrued costs which is what is needed by the NA. No other areas of the survey were subject to specification error because data are reported directly from the governments accounts which follow the standarized chart of accounts definitions.
	Frame error	60	60	5	5	7	7	N/A	L	Frame error is only applicable to the municipal associations. It is quite small and the staff have a good knowledge about the prcess generating the frame. It is a low risk error source affecting only about 3% of the total. Given the low risk of frame error, further planning to mitigate risks is unnecessary and is not applicable.
	Non-response error	56	60	5	*5→6	7	*~~4~~ 6	*5→6	M	Nonresponse concerns primarily item nonresponse in the sections on educational activities and care for the disabled and elderly as well as social work in the summary accounts. No study has been done to quantify this risk. There is no imputation for item nonresponse at present. Communication has improved because the relationship with to data providers (ie the Data Collection Unit) is now included. Compliance to Standards re-rated because of consideration for both unit nonresponse and item nonresponse (previously only item nonresponse). Plans have made progress. Survey now mandatory for preliminary data (municipalities and county councils) as well as in general for municipal associations.
	Measurement error	58	62	5	5	7	7	*5→7	M	Cognitive laboratory evaluation enabled changes to the questionnaire which will reduce the risk of measurement error. The Cognitive Lab will be used in 2014 as well.
	Data processing error	48	54	*4→5	*3→5	7	5	5	H	Work on the reduction of alerts for editing to increase the cost-effectiveness of editing. Implementation of agile working methods in the editing work of the data collection unit. Good collaboration with data providers in data collection unit. Plans to review editing and publishing process next year approved by management.
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Model/estimation error	38	38	3	3	7	3	3	H →M	Regards common cost allocation as well as activity allocation. Change in risk level because the models do not affect NA who are the most important users of these statistics.
	Revision error	58	63	5	*3→5	8	7	N/A	L	Communication raised because data providers are now included in communication and there is good collaboration with data providers concerning revisions.
	Total Score	51,5	55,0							

Exhibit 2.2 CPI Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	68	72	8	6	9	*7→8	*4→5	**H**	Use of scanner data for every day goods will eliminate the problem of discounts prices for these goods. List prices still used for some goods although investigation has been done to use actual prices which is the required standard. Implementation only partially implemented for actual prices. No plans for new cars which would be good to develop. Have implemented a change when it comes to flights where the staff actually go through the process of booking flights up until till final step, to see what the actual price will be, which differs often from the listed price.
	Frame error	62	64	7	7	5	7	*5→6	**M**	Some progress has been made on coverage issues for the 60 centrally-collected surveys. The overview project of these was completed in 2012 but implementation work has spilled over to 2013 and will continue even next year regarding coverage issues as well as other quality issues. Progress has been made on including businesses with sales over the internet as well even though this work is also still in progress.
	Non-response error	55	55	~~3~~ 5	3	~~9~~ 7	7	N/A	**L**	Re-rating to show that some knowledge exists and some proxy measures are in place for non-response. Expertise re-rated because the internal working relationship between internal groups could be better. Nothing has changed from the previous year in respect of these two points. Nonresponse issues surrounding the Household Budget Survey are now covered under Modelling Error.
	Measurement error	62	68	*7→8	5	9	*5→6	*5→6	**H**	The main issues here are Selection Bias and the difficulties faced in making quality adjustments. A Study has been done on Selection Bias and presented to the Advisory Board. This has given insight into collectors' behavior and input to how to improve instructions to collectors. The increasing use of scanner data is considered as best practice as well as the fact that Stat Sweden hosted an international workshop on the subject. The data collection is not monitored.
	Data processing error	74	76	7	6	9	8	*~~8~~ 7→8	**H**	Efforts have mainly be made in maintaining the relatively new quality control system. There has been some improvement in the editing of the data from hand-held computers having to do with the updating of acceptance limits between the hand-held computers and the production system. As well, to date, approx. 95% of all Excel spreadsheets are standardized which is a higher figure than last year.
	Sampling error	66	70	7	7	9	*6→7	*4→5	**H**	Compliance to best practice in having a probability sample as well as computing confidence intervals for the CPI. Also progress has been made in decreasing the variance by 15% for 30% of the CPI by adjusting the definition of the elementary aggregate only by product group and not in combination with industry. There are plans to adjust the QD with information on interest payments for mortgages.
	Model/estimation error	44	44	~~5~~ 3	~~5~~ 3	6	4	6	**H**	Knowledge and Communication are re-rated as there are no real measures of error here. This has not changed since last year. Not much is known and communicated about the error in the Hedonic models or how nonresponse in the HBS affects the modeling of the product group weights.
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Total Score	62,3	65,2							

Exhibit 2.3 FTG Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	58	58	5	5	7	7	5	M	No change compared to last round.
	Frame error	58	64	5	5	7	*5→7	*7→8	L	Increase because of the use of alternative registers to strengthen the framework for the special movements survey.
	Non-response error	66	68	7	7	7	5	*7→8	M	Increase due to the introduction of process data on the delivery reports.
	Measurement error	62	64	*7→8	5	7	7	5	H	Knowledge of measurement error has increased due to the work on the record check for the central project on "Measure and Reduce Measurement Errors".
	Data processing error	60	66	7	7	7	*3→5	*6→7	H	keying procedures meets standards but not best practice plans to do away with paper forms plans for new IT-system do away with VB6. - The risk associated with Data processing error is already High but would become even higher with the phase out of Visual Basic. IT-system for volume index has been introduced to eliminate manual work previously undertaken which would have been more error prone.
	Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Model/estimation error	80	82	8	7	9	9	*7→8	M	Increase because Survey of Statistical Values has been undertaken. There have also been investigations into the use of VIES data to improve imputation of commodity level data when it is missing.
	Revision error	70	72	7	*~~7~~ 6	7	*~~9~~ 7→8	8	H	Communication was re-rated to reflect that better understanding is needed of the impacts of revisions for Foreign Trade of Goods on GDP.
	Total Score	64,7	67,6							

Exhibit 2.4 LFS Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	70	70	7	7	7	7	7	L	Classification of status in the labour market is the main focus - LFS has followed ILO recommendations. Alignment was made to the EU-recommendations since 2007. Study going on of questionnaire promoted by Eurostat to be delivered in april 2014. Risk for specification error is very low.
	Frame error	58	58	7	7	7	3	5	L	Overcoverage is an issue in TPR but stable over time 25-50 000 individuals - registered in Sweden but not living in Sweden. Undercoverage exists but is low. LFS goal population is individuals registered in Sweden. The ILO recommendation is the resident population.
	Non-response error	52	52	6	5	5	5	5	H	Non-response is rising in the LFS as with other surveys directed towards individuals. Planning seems to lacking focus. Plans to test to outsource data collection which purpose is to increase knowledge of reasons for the nonresponse problems.
	Measurement error	56	68	*5→8	*5→7	*5→7	5	*8→7	H	Reinterview survey has been conducted during 2013. Cognitive Lab has been very involved in mentioned study. Work by Pär Karlsson using MLCA to evaluation measurements in LFS is a major contribution to knowledge and achievement Planning/mitigation rating was reduced because, at the time of this review, there is still uncertainty that the work on reinterview and latent class analysis will continue. There should be some plans to ultimately use the results of these studies to mitigate the risk of measurement error.
	Data processing error	62	62	5	5	7	7	7	M	Coding study means meeting standards in a good way.
	Sampling error	78	80	7	9	7	9	*7→8	M	Work done in sampling design helps to maintain high scores in this area. The precision was improved by eliminating age strata which caused small stratum sample sizes.
	Model/estimation error	60	64	5	5	*6→7	7	*7→8	M	Further development going on with auxiliary variables. Access to internal seasonal adjustment expertise now. Plans are in place to develop the seasonal adjustment
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Total Score	60,9	64,3							

Exhibit 2.5 SBS Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	54	58	5	5	7	5	*5→7	M	This is mostly concerned with differences between business accounting and statistical concepts Production values vs value added. Increased rating on Planning/Mitigation because of advances in electronic reporting of businesses provides more certainity about the data actually provided.
	Frame error	64	60	7	7	7	6	*5→3	M	The main areas of risk are a) inaccuracies in NACE coding and b) overcoverage because register includes inactive businesses. Reduced rating because there is less certainity about how the new BR will address the issue of over-coverage. Also, the number of profiled businesses and activity units continues to decline.
	Non-response error	70	70	7	6	7	7	8	M	Nonresponse is reltively small in terms of contribution to estimates but requries considerable effort. The approved plan has made progress but has become of lesser priority compared to last year, according to management.
	Measurement error	52	56	6	5	5	5	*5→7	H	The main issue is with the detailed items required by NA that are not necessarily in the Chart of Accounts. EDT progress and cognitive lab interview is in progress leading to improved rating.
	Data processing error	60	60	5	5	7	6	7	H	Editing and data imputation are the main sources of risk. No change since last year.
	Sampling error	84	86	*8→9	8	*9→8	9	*8→9	M	Although sample design is very professional there is still a risk of sampling error but knowledge of sampling errors is excellent. Their use to adapt the sample on an annual basis is also excellent. Expertise has lower rating because of fears of further deterioration of sampling frame as a result of decisions on the priorities for the new BR.
	Model/estimation error	56	48	5	5	*8→6	4	*~~8~~ 6→4	H	The main issue is with modelling of data items when estimates cannot be provided by respondents. Declining of the business profiling affects expertise and planning and is reflected in reduced ratings. The problem has worsened since last year and plans are not yet approved by management for the mitigation of this risk. The cooperation is not as good with the Large co. Unit and the BR to profile enterprise and create more KAUs. Perhaps the SBS staff has failed to convince managment of the importance of the issue of profiling. There are plans to pick up the project to structure and store metadata in a better way. These plans are approved by management but the question of financing is not solved. Ratings for last round were revised downward as we were not aware that the previous metadata project was no longer planned.
	Revision error	56	54	6	5	8	4	*5→4	H	Revisions exist between preliminary and final estimates although steps are being taken to reduce these. Shift in priorities so plans for work on revisions is of lesser priority due to the lack of resources and work on IT-system.
	Total Score	60,8	60,1							

Exhibit 2.6 ULF/SILC Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	34	58	*3→5	*3→5	7	*3→5	*1→7	M	Questionnaire been reviewed by subject matter experts and Cognitive Lab together. Support exists from top management for the questionnaire redesign. There is a continuing dialog between the survey management and external experts regarding the content of the questionnaire and risks of specification error
	Frame error	42	42	3	3	7	5	3	*~~H~~→M	Change in risk level is due to a better understanding of the risk of over- and under-coverage. The impact of overcoverage on response rates has never been evaluated but it is expected not to pose a high risk to accuracy. There should be some planning to look at the impact of overcoverage on data quality. As it now stands, most of what is known and communicated is speculation.
	Non-response error	40	46	5	3	5	*4→5	*3→5	H	NR is relatively high and growing. More reporting in QD especially on partial non-response. Methodologist plans to do some computation of indicators for Responsive Design which would facilitate design changes to reduce non-response bias. The efforts to train interviewers on the theory of survey cooperation is commendable but what has been observed thus far may be Hawthorne effects. There should be more work on providing knowledge, information, tools, and motivation to address nonresponse in the field.
	Measurement error	46	52	3	3	9	*3→5	*5→6	H	Call monitoring system recently implemented but there are some problems with the way it is designed. Moreover, the results of the monitoring are not being used effectively. Questionnaire reviewed for child interviews and age limit raised from 10 to 12 to avoid measurement error. Studies of response reliability are needed that look at within person response variation and/or internal consistency among similar questions.
	Data processing error	42	50	5	3	7	5	*1→5	L	Initial Studies undetaken on field coding although further work needs to be done. Interviewer coding of some open ended questions subject to high risk of error 5% recoding. Thus, there are plans to investigate potential primacy and recency effects associated with long list of response categories that are coded on the fly by interviewers.
	Sampling error	54	62	7	7	7	*3→5	*3→5	M	Some components of the ULF were eliminated which remove the problem of unknown selection probabilities. The new sampling design can now compute selection weights.
	Model/estimation error	38	50	5	3	7	*3→5	*1→5	H	Calibration modeling for reducing bias and variance are sophisticated and but is providing some strange results. Some changes are needed and perhaps the number of calibration variables needs to be reduced. Methodologists will be working on calibration in 2014.
	Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Total score	42,1	51,1							

Exhibit 2.7 BR Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	66	66	5	5	7	8	8	L	The main cause of concern was that the Tax Agency provided details of legal units which Stat Sweden had to convert to enterprises Multiple sources exist for one variable, like employment which is defined in different ways. There did not seem to have been much change in the siuttion over the last year.
	Frame error- over coverage	56	58	6	*5→6	7	5	5	H	The risk is because some units on the BR are actually inactive. Data providers are included now in Communication which gives a different consideration for Communication.
	Frame error- under coverage	46	42	3	3	6	6	*5→3	~~M~~ H	The main risk is now with the reduction of business profiling and smaller number of KAUs. A secondary risk is with businesses with several localities. Current procedures do not update this information. Intrinsic risk rating is changed to high because the Tax Agency only provides information on legal units. Coordination with data providers but other internal coordination with statistical areas.
	Frame error - duplication	63	63	5	5	7	8	N/A	L	Duplication only seems possible at the local unit level. Not much change over the last year.
	Missing data error	48	48	5	5	5	5	4	L	The main issue is the missing NACE codes although this seems to be mainly for enterprises with zero employees. The number has continued to decline.
	Content error	50	52	*~~3~~ 5	*3→4	7	5	5	H	The main issue is the impact of inaccurate NACE codes which may be a growing problem. The earlier study of the accuracy of NACE coding is not optimal as it uses dependent coding. The increase in Content is based on re-assessment of previous rating. The increase in communication is because provider communication is now included.
	Total Score	52,7	52,7							

Exhibit 2.8 TPR Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Specification error**	50	58	*4→ 6	4	6	6	*~~3~~ 5→ 7	M	Correction to last year's rating for Planning is because the evaluation team did not understand how the Census 2011 work would effect Specification Error, particularly for the identification of households and families. Planning rating increased to reflect current plans along the same lines. Knowledge rating increased to reflect increased knowledge regarding specification error brought about by the Census work.
	Frame error: overcoverage	56	58	6	6	*5→6	6	5	H	The working group with representation from the Swedish Tax Agency, the Swedish land registration, the dwelling register, and the Swedish association for local government who will meet four times per year to discuss quality issues and plan for quality improvements. The STA report was shared with the TPR staff and the TPR staff were able to corrobarate those results with their own. This demonstrates an improvement in collaboration with the STA reflected by improved rating in Expertise.
	Frame error: undercoverage	60	60	5	5	7	7	N/A	L	No change but this is low risk and low priority.
	Frame error: duplication	70	70	6	6	8	8	N/A	L	The risk of duplication is very low except when a person has two different personal identification numbers in two different registers and the registers are merged. No change since last year.
	Missing data error: item and variable	66	66	6	6	7	6	8	M	Ratings would be higher with more documentation in the QD on improvements in missing data for e.g. the improvement in this area on the dwelling numbers
	Content error	58	62	*5→6	*5→6	7	7	5	L	The paper "Methodological Experiences from a Register-Based Census" by Claes Andersson, Anders Holmberg, Ingegerd Jansson, Karin Lindgren, and Peter Werner (2013) was published which demonstrates an improvement in Knowledge and Communication for Content Error.
	Total Score	58,8	61,4							

Exhibit 2.9 GDP Quarterly Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Description of the error source	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Input data source - Index of Service Production, ISP**	60	62	*4→5	~~5~~ 6	7	7	6	**H**	1) service industry kind of activity units in manufacturing enterprises are not included (may affect the extrapolation factors) 2) for some industries e.g. real estate because of measurement error 3) sampling effects are seen in some smaller industries	Knowledge improved through pre-study of intermediate consumption using VAT data. There was also study of size of changes when a new sample is introduced. Good communication with ISP increases communication. A SLA exists. Expertise rating declines because of loss of experienced staff and the limited training that is able to be given to new staff.
	Input data source - Index of Industrial Production, IIP	60	62	*4→5	~~5~~ 6	7	7	6	**H**	1) service activity in manufacturing kind of activity units is missing (e.g. merchanting) 2) sampling error is potentially high for industries with predominantly smaller enterprises 3) measuring "deliveries" (instead of turnover) which could be a specification error 4) estimation for below cut-off enterprises 5) could be measurement error - enterprises could include more or less than what is required.	Knowledge improved through pre-study of intermediate consumption using VAT data. There was a study of the size of changes when a new sample is introduced but the change showed no impact. Good communication with ISP increases communication. A SLA exists.Expertise rating declines because of loss of experienced staff and the limited training that is able to be given to new staff.
	Input data source - Merchanting Service of global enterprises (also covers royalties, licensing and R&D)	44	44	3	~~5~~ 6	5	3	5	**H**	The data source is Foreign Trade with Services (quarterly survey with the largest enterprises) which also covers licenses, royalties and R&D. The SBS is the annual source. The figures from the smaller enterprises are modelled from the SBS (year t-1). There are primarily measurement and coverage errors involved here.	Communication is revised because communication with data provider is now included. Expertise rating declines because of loss of experienced staff and the limited training that is able to be given to new staff.
	Compilation error (modelling)	48	48	5	5	5	6	3	**H**	Models - strong dependency on the work of the analysists. 1) intermediate consumption 2) construction 3) financial services 4) real estate 5) insurance 6) energy 7) water supply 8) hidden and illegal economy 9) seasonal adjustment	No change at this the use of VAT data, coupled with a survey of the largest enterprises (ie Quarterly SBS, may allow improved models for intermediate consumption in the future.
	Compilation error (data processing)	44	52	7	~~3~~ →5	~~3~~ 5	~~3~~ 4	*3→5	**H**	1) spreadsheets 2) IT-system (objective is to have a more automated process to exclude manual work with input data, also traceability) 3) more compilations in SAS 4) there have been several false starts at developing a new system. Accessibility to IT resources has been a big issue.	Improvement to Communication is because of greater involvement/ interest of IT staff including the investigations into the use of the Finnish system. Planning/mitigation improved because of the introduction of macro output edits and the pre-study into the Finnish IT system for national accounts.
	Deflation error (including specification error)	48	48	4	3	7	7	3	**H**	1) possible high sampling errors in some of the producer price indexes 2) wage indices have to be used in some cases 3) insufficient adjustment for quality in general 4) complex products pose difficulties in measuring change over time	No change although the study into the use of the Domestic Supply indexes was noted. There were no immediate plans to introduce this revised approach.
	Balancing Error	56	56	5	5	*7→6	6	*5→6	**H**	Dependency on experience of analysts and lack of standardized / formal methods but formal methods are now being developed.	Improvement in planning/mitigation because the introduction of macro output edits should reduce the balancing error. Also, it has been agreed to have principles and guidelines for objective balancing. These are currently under development. Expertise reduced because of the loss of a skilled resource in this part of the National Accounts compilation.
	Revisions Error	56	58	7	5	7	5	*4→5	**M**		Improvement in planning/mitigation as the pre-study into intermediate consumption should lead to a revised approach will reduce revisions. The proposed increase in transparency of future revisions was noted.
	Total Score	51,8	53,6								

Exhibit 2.10 GDP Annual Rating Changes between Round 2 and Round 3

	Error source	Average score round 2	Average score round 3	Knowledge of Risks	Communica-tion	Available Expertise	Compliance with standards & best practices	Plans or Achievement towards mitigation of risks	Risk to data quality	Description of error sources	Improvements compared to round 2 Deteriorations compared to round 2 Corrections to round 2 scores **Comments**
Accuracy	**Input data source - Structural Business Statistics, SBS**	66	66	7	*5→6	7	7	*7→6	**H**	The main issues were (1) estimates of margins from SBS for the trade industries seemed unreliable, (2) inaccurate estimates for some industries (eg Construction) requiring the use of models, and (3) potential problems from over-coverage and under-coverage. Inconsistency of NACE coding from one year to the next causes some problems. SBS is generally regarded as a reliable data source.	Communication with users at the level of good but SBS is Very Good with partial SLA. Average score is used. There is certain quantification in the process tables but not quantifying impacts on GDP. Slight decrease in Expertise. Decrease in number of KAUs will impact on the quality of industry data from SBS.
	Compilation error - modelling	48	50	5	4	5	7	*3→4	**H**	Modelling 1) trade margins 2) construction 3) financial services 4) real estate 5) insurance 6) energy 7) hidden and illegal economy	Improvement in Planning/Mitigation is because of plans to use SBS for part of construction industry which means a model is no longer required for this component.
	Compilation error - data processing	44	52	*~~5~~ 7	~~3~~ →5	~~3~~ 5	~~3~~ 4	*3→5	**H**	Data-processing 1) spreadsheets 2) IT-system (objective is to have a more automated process to exclude manual work with input data, also traceability) 3) more compilations in SAS	Revised score for Knowledge, Expertise and Compliance is to ensure consistency with quarterly national accounts. Improvement to Communication is because of greater involvement interest of IT staff including the investigations into the use of the Finnish system. Planning/mitigation improved because of the introduction of macro output edits and the pre-study into the Finnish IT system for national accounts.
	Deflation error (including specification error)	48	48	4	3	7	7	3	**H**	1) possible high sampling errors in some producer price indexes 2) wage indices are used for collective public consumption and some services 3) insufficient adjustment for quality in general 4) complex products pose difficulties in measuring change over time 5) the models used in constant price estimation for goverment	There has been a change to no longer have a quality adjustments for the volume measures for the constant price calculation. This is in accordance with revised Eurostat directive. This also pleases Swedish users.
	Balancing Error	58	58	*~~5~~ 6	*~~5~~ 6	*~~5~~ 7	7	3	**H**	1. Objective Editing 2. Subjective Editing 3. RAS method Supply and use tables for the 400 products dependency on experience of analysts. Inconsistency between national accounts and other economic statistics is an issue for Statistics Sweden.	Although Process Tables have been in place for several years, the reviewers were not aware of them. They provide excellent transparancy of balancing processing and provide data that enable the impact on particular variables to be studied. This is the reason for the re-ratings.
	Revisions Error	56	56	5	7	5	6	*~~4~~ 5	**M**	3 yearly estimates are made 1) sum of 4 quarters t+ 60 days 2) t+9 months (revisions covering largely Government sector) 3) t+21 months (revisions cover largely the non-financial business sector with the SBS). Revisions are generally regarded as acceptable but the revisions for the public accounts between the first and second estimates are of greatest interest.	There is a written plan describing how the revision process is to be followed in Statistics Sweden. This was in place in 2012 but was not previously noted.
	Total score	53,2	54,9								