# A SECOND APPLICATION OF THE "ASPIRE" QUALITY EVALUATION SYSTEM FOR STATISTICS SWEDEN

Paul Biemer and Dennis Trewin

January 25, 2013

# TABLE OF CONTENTS

# 1 EXECUTIVE SUMMARY

The Ministry of Finance directed Statistics Sweden (SCB) to develop a system of quality indicators that signify quality improvements in key statistical products. This system was to include metrics that reflect current data quality as well as changes in quality that occur over time. In 2011, Statistics Sweden in collaboration with two consultants (Paul Biemer and Dennis Trewin), developed a quality evaluation approach (referred to in this report as ASPIRE) for this purpose and pilot tested it on eight products (see Biemer and Trewin, 2012). This report re-evaluates seven of those products – viz., Annual Municipal Accounts (RS), Consumer Price Index (CPI), Foreign Trade of Goods Survey (FTG), Labour Force Survey (LFS), Structural Business Survey (SBS), Business Register (BR), and Total Population Register (TPR). It also proposes a new approach for one of them (the National Accounts) and applies ASPIRE to one additional product (the Survey of Living Conditions or ULF/SILC). For each of these products, Accuracy (or data quality) was assessed for the sources of error that were applicable for each program. In addition, the so-called user dimensions of Relevance/Contents and Accessibility & Clarity were assessed for the CPI and Timeliness & Punctuality and Comparability & Coherence were assessed for the LFS. The primary goal of this review was to test aspects of ASPIRE that were developed specifically for these dimensions.

For each product, the evaluation involved a self-assessment, extensive reviews of relevant documentation, a comprehensive interview of key staff, and a review with feedback of the final assessments. The Accuracy evaluation process described in Biemer and Trewin (2012) was improved and applied for this review. As in that review (referred to as Round 1), each product was scored (using a 10-point scale) according to five criteria which were the same for all applicable error sources. One new innovation that greatly facilitated the application of these criteria was the use of a checklist for each criterion. 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 the error source. With a maximum possible score of 100 percent (indicating perfect quality), the product scores ranged from 42.1 percent (for the ULF/SILC)) to 65.8 percent (for the FTG) with an average rating of 57 percent. (Exhibits 4a and 4b in the report provide the scores for each product by error source.) All the products reviewed in Round 1 increased their scores in this round. The average increase was about 4.6 percentage points.

Some of additional findings from the reviews include the following:

- As in Round 1, measurement error appears to be the error source with the highest risk; it was rated a high risk for six out of eight products.
- Measurement error still ranks among the bottom of the ratings; although, its rating has considerably improved from Round 1 primarily as the result of significant planning for risk mitigation for the coming year.
- The highest ranking error source by a wide margin is sampling error. Revision error is also high ranking although it only applies to three products.

In addition, the following general findings are notable:

- The documentation of quality was greatly improved owing primarily to enhancement in the Quality Declaration (QD) documents.

- 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.

The main report provides specific comments on each product, some justification of the low ratings for high risk error sources, and some suggestions for improvement. Finally, in our 2012 report, we laid out recommendations to improve quality that cut across all products in these 10 areas:

1. Greater Integration of Economic Statistics
2. Increasing Cooperation between the National Accounts 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 Office
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

Although some progress has been made in all areas, with significant progress in some areas, more improvement is needed and the work should continue to progress in these areas. In addition, three new recommendations are added as a result of the current review:

11. Increase the Focus on Coherence between Relatable Statistics
12. Improve Communications between Statistical Product Areas and Some Service Areas (e.g., Communications)
13. Initiate Succession Planning in Some Important Statistical Areas

The revised ASPIRE approach for Accuracy worked very well for most products. However, additional improvements are planned to enhance the criteria and checklists. The new error structure developed for the National Accounts was an important innovation that greatly improved the evaluation of GDP estimates. However, additional work is needed to develop criteria and checklists that address the unique characteristics of the GDP error components. The extension of ASPIRE to the evaluation of user dimensions was quite successful; although, like Accuracy, enhancing the criteria and the checklists is a priority for further development.

# 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 by the end of the year. 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 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 and the current work as Round 2.

The 2011 evaluation process worked very well for all products except for National Accounts for reasons described in Section 3.2. To improve the process for the National Accounts, an alternative approach was devised in the fall of 2012 that was customized to the unique error sources associated with National Accounts products – specifically gross domestic product (GDP) estimates. This approach, described Section 3.3, effectively created a new baseline evaluation for the National Accounts.

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. For our study we have used five dimensions of total survey quality – Accuracy, Relevance/Contents, Timeliness & Punctuality, Comparability & Coherence, and Accessibility & Clarity[1].

For this (second) round, the focus was on the following five activities:

1. An assessment of improvements in Accuracy relative to the baseline review for seven of the eight products reviewed in Biemer and Trewin (2012).
2. An application of an improved ASPIRE approach for the National Accounts,
3. An initial assessment of a product that was not reviewed in 2011 – the Survey of Living Conditions (ULF/SILC),

---

[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.

4. An extension of ASPIRE to the quality dimensions of Relevance/Contents and Accessibility & Clarity for the Consumer Price Index (CPI).
5. Likewise, an extension of ASPIRE to Timeliness & Punctuality and Comparability& Coherence for the Labour Force Survey (LFS).

These latter two activities was the first, formal attempt to incorporate the so-called user quality dimensions into ASPIRE.

Regarding (1), the objective was to identify areas where clear improvements had been made relative to the baseline evaluation. However, in the process of making those assessments, we found areas where the baseline ratings assigned in Biemer and Trewin (2012) were not accurate due to incomplete information or erroneous understanding of the processes. Because improvements relative to the baseline assume that baseline ratings are accurate, we provide corrected baseline ratings in a few instances. The discussions of quality improvements in this report will clearly distinguish between original baseline, corrected baseline 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 revised ASPIRE approach used in this report is described in the next section including a discussion of some of the improvements made to the original approach. Section 4 summarises the results of the quality evaluations for the ten products (treating quarterly and annual National Accounts as separate products). Section 5 summarises some cross-cutting methodological and other findings. Section 6 discusses our work on the User Quality Dimensions. Section 7 discusses further improvements in the quality evaluation model. Finally, Section 8 provides our recommendations and conclusions.

# 3 PRODUCT QUALITY ASSESSMENT AND MONITORING

## 3.1 THE ASPIRE MODEL

In Biemer and Trewin (2012), 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 National Accounts. 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 variation 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 data receive and its potential 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. 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

[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.

of a questionnaire. *Item nonresponse* occurs when the questionnaire is only partially completed 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

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

$$\text{TSE} = \text{(specification error)} + \text{(other sampling and nonsampling errors)}$$

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 ($X$). 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 with 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

The top panel of Exhibit 1 shows the six survey products that are included in the ASPIRE review in this review round (referred to as Round 2). As noted previously, all but one of survey (viz., the ULF/SILC) were reviewed in Biemer and Trewin, 2012 (referred to as Round 1). The focus of the review for the five returning surveys is on improvements and new developments since the last review. Because it is an initial review, the review of the ULF/SILC focuses on understanding the survey process, error risks, and obtaining a baseline quality level for determining current accuracy and future quality improvements. Also shown on the right side of Exhibit 1 are the error sources associated with these survey products.

***Exhibit 1. Sources of Error Considered by Product***

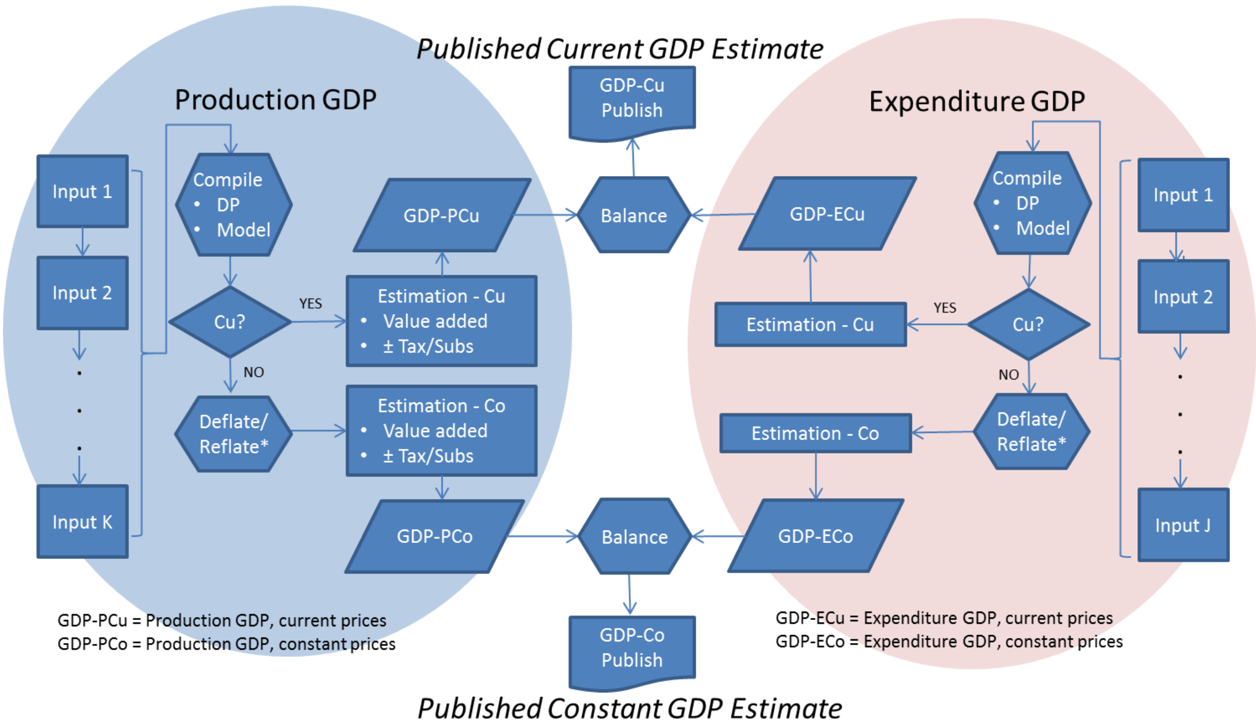
| **Product** | **Error Sources** |
|---|---|
| *Survey Products*<br>Foreign Trade of Goods (FTG)<br>Labour Force Survey (LFS)<br>Annual Municipal Accounts (RS)<br>Structural Business Statistics (SBS)<br>Consumer Price Index (CPI)<br>Living Conditions Survey (ULF/SILC) | Specification error<br>Frame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

The TPR and Business Register (BR), reviewed in Biemer and Trewin (2012), are again reviewed for this round. The middle panel of Exhibit 1 provides the error sources associated with these two registers. As noted above, the error sources for the two registers reflect their primary uses which are for sampling frames for surveys or censuses.

As previously noted, we believe our evaluation of the National Accounts in Round 1 did not satisfactorily consider the unique and complex error structure of the National Accounts products, particularly the GDP estimates. In addition, the numerous products that were included under the rubric “National Accounts” proved to be too extensive and complex to do them justice in the limited time available for our review. To address these difficulties, the focus of the Round 2 review was considerably narrower, focusing solely on the estimation of quarterly and annual GDP from the production perspective. In addition, the error structure of the GDP estimation process has been respecified to more precisely capture its major error sources.

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 Municipal Summary 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 is some function of these input variables; i.e.,

$$\widehat{GDP} = f\left(y_1, y_2, \dots, y_p\right) \tag{0.1}$$

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 $\varepsilon_p$. Let $x_p$ denote the value of $y_p$ that would be observed if these errors were negligible; i.e., if $\varepsilon_p$ were essentially 0. Thus, we can write

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

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 $\varepsilon_p$.

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

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

and thus, we can write

$$\widehat{GDP} = GDP + e \tag{0.3}$$

which means that the estimate of GDP is equal to the hypothetical 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 $\varepsilon_p$'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 *f*. Thus, we are also interested the potential contributions of *f* on *e* where *f* 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 National Accounts. 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 Round 1, Round 2 uses five criteria; viz., Knowledge of Risks, Communication with Users, Available Expertise, Compliance with Standards and Best Practices, and Achievement Towards Risk Mitigation or Improvement Plans. In Round 2, the guidelines for these criteria have been enhanced and improved. Exhibits 1.1a-1.1e in Annex 1 show the improved quality guidelines that were applied in Round 2 to each error source shown in Exhibit 1.

The application of these guidelines is facilitated in Round 2 by the use of checklists for each criterion (see also 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 pre-determined based upon the last "yes"-checked item in the list.

As was done last year, 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.

A new feature of ASPIRE in Round 2 is to 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 Risk Mitigation or 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 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 seldom 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 4.1.4 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 in Exhibits 1.1a – 1.1e (in Annex 1) 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 Risk Mitigation or 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), the application of this model to the eight products in Exhibit 1 follows a three-step approach described in the following.

### PRE-INTERVIEW ACTIVITIES

Pre-interview activities include two primary activities. First, each evaluator (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, review process was considerable aided by the existence of QDs for all products which, in some cases, were substantially expanded and improved since Round 1. Also during this period, the product's responsible staff were invited to a meeting that explained the evaluation model and its uses and any changes to the process that were made since Round 1. At this meeting, or subsequently, the staff used the model to perform a self-assessment of data quality using the newly developed quality checklists. This review of relevant materials and the self-assessments are essential steps leading to the main data gathering activity – i.e., the quality interview.

### THE QUALITY INTERVIEW

As for Round 1, the quality interviews were conducted in both Stockholm and Orebro. These interviews occurred during the period from November 28 to December 4, 2012. Each interview took approximately four hours to conduct. The meetings were organized into essentially five parts:

a) discussion of any notable improvements 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) review of the classifications of each error source into High, Medium, and Low categories of the inherent risk with corrections or other adjustments, if necessary,
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 1. 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 most of 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 in Round 2.

## 3.5 ASSESSING USER DIMENSIONS OF QUALITY

As noted previously, the ASPIRE system was expanded in Round 2 to incorporate a process for evaluating the four user dimensions of quality; viz., Accessibility & Clarity, Comparability & Coherence, Relevance/Contents, and Timeliness & Punctuality. The primary goal of this application was to develop a process for assessing other quality dimensions and to test how well the process works for two products. The framework for rating a product for these dimensions was modelled after the Accuracy framework; i.e., each dimension was decomposed into mutually exclusive components (analogous to the error sources defined for Accuracy) and quality was assessed according to five criteria similar to the five Accuracy criteria. These criteria are Knowledge of User Needs, Communication with Users, Available Expertise (to address user needs), Compliance with Standards and Best Practices, and Plans toward Addressing User Needs and were applied to each of the components under a dimension. The components associated with each user dimension appear in Exhibit 3. The generic criteria that were used for each component appear in Exhibit 1.2a – 1.2e in Annex 1. As for Accuracy, checklists were developed for each criterion and were generic across dimensions and components within dimensions. These checklists are also included in Annex 1.

***Exhibit 3. User Dimensions and their Components***

| |
|---|
| **Accessibility and Clarity**<br>• Level and timeliness of user support<br>• Ease of data access (including microdata where relevant)<br>• Documentation (including metadata)<br>• Availability of quality reports |
| **Relevance/Contents**<br>• Outputs (including microdata and other products)<br>• Inputs (content, scope, classifications, etc.) |
| **Timeliness and Punctuality**<br>• Timeliness of release of main aggregates<br>• Timeliness of release of detailed outputs (including microdata)<br>• Punctuality of data releases |
| **Comparability and Coherence**<br>• Comparability across geography, populations, and other relevant domains<br>• Comparability across time (including impacts of redesign)<br>• Coherence with other relevant statistics (including use of standard classifications, frameworks, etc.) |

The two products selected for this test were the LFS and CPI. The former product was evaluated for Timeliness & Punctuality and Accessibility & Clarity while the latter was evaluated for Relevance/Contents and Comparability & Coherence. The assessment process proceeded much like the process for Accuracy. The LFS and CPI staffs were asked to complete self-evaluation checklists prior to the quality interview. This information as well as information contained in the QDs regarding user-related quality improvements was provided prior to the quality interview. A separate quality interview was conducted that focused solely on the user dimensions. In this interview, the checklists, information contained in the QDs, and other information related to the user dimensions were reviewed and discussed. The interview concluded with the assignment of preliminary ratings for each component and criterion. As was done for Accuracy, the LFS and CPI staff were given an opportunity to review and comment on these ratings either by email or in an in-person meeting with the evaluators.

We believe these new features of ASPIRE worked very well for their initial application. However, the process can and will be improved for the next round. In particular, some refinements of the checklists and criteria for under each dimension are needed to better capture the risks to poor quality associated with each dimension.  Overall, the process yielded valuable information regarding what activities a product is undertaking for understanding and addressing the needs of user communities, what user needs are currently not being met, and where future activities should be focused.

## 3.6 LIMITATIONS OF THE 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 this ASPIRE round 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.

The next section provides the results of the reviews for the products evaluated in the round. Section 4.1 discusses the seven products whose accuracy was re-evaluated, focusing on any improvements that have occurred since Round 1. Section 4.2 considers the results for the newly reviewed products Quarterly and Annual GDP (using the production approach) and the ULF/SILC). Finally, some results from the assessment of quality related to the user dimensions will be provided in Section 4.3

# 4 FINDINGS FOR THE TEN STATISTICAL PRODUCTS

Exhibits 4a and 4b provide the overall scores for eight products (excluding National Accounts) by error source.  A discussion of the National Accounts 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 the variation among the variable's observed values.

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 National Accounts, 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, we are not confident that our efforts were successful and 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:

- As in Round 1, measurement error appears to be the error source with the highest risk; it was rated a high risk for six out of eight products.

- Measurement error still ranks among the bottom of the ratings; although, it's rating has considerably improved from Round 1 primarily as the result of significant planning for risk mitigation for the coming year.
- The highest ranking error source by a wide margin is sampling error. Revision error is also high ranking although it only applies to three products.
- The overall mean in this round is 57 compared to 54 in Round 1.
- The ratings for all seven products that were reviewed in Round 1 improved in the current round.
    - Average improvement for products reviewed in Round 1 is about 4.6 percentage points.
    - FTG showed the largest improvement (from 57.3 to 65.8).
    - TPR also showed substantial improvement (from 52.2 to 58.0).
    - SBS showed the smallest improvement (from 59.6 to 61.4).
- The ULF/SILC is the lowest ranking product.  It scored below average in all applicable error sources, including four deemed to be high risks.

In addition, the following general findings are notable:

- The documentation of quality was greatly improved owing primarily to enhancement in the QD documents.
- 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.
- In a few cases, the QDs contained too few (direct or indirect) measures of data quality relying instead on elaborate descriptions of processes to justify claims that the processes should have small residual risks of nonsampling errors.

***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 | RS | CPI | FTG | LFS | SBS | LCS | BR | TPR | Error Source Mean Rating |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | N/A | **68** | **58** | 70 | **54** | **34** | 66 | **46** | 57 |
| Frame error | 60 | **62** | 58 | 58 | **64** | **42** | 55 | 62 | 58 |
| Overcoverage | | | | | | | **56** | **56** | |
| Undercoverage | | | | | | | **46** | 60 | |
| Duplication | | | | | | | 63 | 70 | |
| Nonresponse error/Missing | **52** | 55 | **66** | **52** | **70** | **40** | 48 | **66** | 56 |
| Measurement error/Content | **58** | **62** | **62** | **56** | **52** | **46** | **46** | 58 | 55 |
| Data processing error | **48** | **76** | **60** | **62** | **60** | 42 | N/A | N/A | 58 |
| Sampling error | N/A | **66** | N/A | **78** | **84** | **54** | N/A | N/A | 71 |
| Model/estimation error | **38** | **52** | **80** | **60** | **60** | **38** | N/A | N/A | 55 |
| Revision error | 58 | N/A | **76** | N/A | **56** | N/A | N/A | N/A | 63 |
| **Round 2 Mean Rating** | 49,6 | 63,9 | 65,8 | 60,9 | 61,4 | 42,1 | 52,2 | 58,0 | 57 |
| **Round 1 Mean Rating** | 46,7 | 60,3 | 57,3 | 56,4 | 59,6 | N/A | 47,2 | 52,2 | 54 |
| Improvement | 2,9 | 3,6 | 8,5 | 4,5 | 1,8 | N/A | 5,0 | 5,8 | 2,5 |

**RED BOLD** = HIGH RISK
**BLACK BOLD** = MEDIUM RISK
REGULAR FONT = LOW RISK
N/A = NOT APPLICABLE

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

| Error Source | GDP Quarterly | GDP Annual |
|---|---|---|
| Input data source (Average) | **53** | **66** |
| Structural Business Survey (SBS) | N/A | **66** |
| Index of Service Production (ISP) | **58** | N/A |
| Index of Industrial Production (IIP) | **58** | N/A |
| Merchanting Service of global enterprises | **42** | n.e. |
| Compilation error (modelling) | **48** | **48** |
| Compilation error (data processing) | **40** | **35** |
| Deflation error (including specification error) | **48** | **48** |
| Balancing error | **56** | **50** |
| Revision error | **56** | **54** |
| **Round 2 Mean Rating** | 50,5 | 49,9 |

N/A = not applicable
n.e. = not evaluated

Those ratings that are high risk (i.e. shown in red) and having scores that are below average could be regarded as the quality concerns most in need of attention from the SCB Executive. The ULF/ SILC and the Quarterly and Annual National Accounts are the products with most number of ratings in this category.

In the next section, we discuss the detailed ratings for all ten products individually. These ratings, with accompanying comments, appear in the annexes at the end of the report.

## 4.1 CURRENT DATA QUALITY FOR THE SEVEN RE-EVALUATED PRODUCTS

In this section, we review the progress over the past 12 months for the seven products that we include in our 2012 report, excluding the National Accounts. Because the National Accounts evaluation process was completely revamped for this review, we shall treat it as a new product set. Our findings for the National Accounts, in particularly quarterly and annual GDP, are reported in Section 4.2.1.

### 4.1.1 ANNUAL MUNICIPAL ACCOUNTS (RS)

In Biemer and Trewin (2012), we stated that measurement error should have high intrinsic risk primarily because municipalities often do not keep accounting data at the level required for the RS. For example, home health care data may be combined with other home care making it difficult to separate the costs. In situations where the requested detailed data are not available, municipalities may use models to allocate costs – i.e., they disaggregate higher level data to estimate the more detailed accounting figures. For example, municipalities routinely allocate costs in the educational activities section and the social work and care activities section of the summary accounts. In such situations, modelling and estimation error is a more appropriate error source for capturing these error risks. It appears that this is the area of greatest risk for the RS. Thus, for the current round, the intrinsic risk for measurement error was downgraded to medium (M) while model/estimation error was upgraded from medium to high risk (H). In addition, the intrinsic risk for data processing error was upgraded from M to H to reflect the critical importance that data processing (most notably, editing) plays in the data collection process. These changes do not reflect a change in the RS. Rather they are corrections to the risk levels that were assigned in Round 1 that reflect a better understanding of the RS processes and the intrinsic risks they present.

One change to intrinsic risk was a direct result of the redesign of the RS instrument that occurred in 2011. Specification error, which was previously assigned a medium risk level, was downgraded to negligible intrinsic risk which effectively eliminates it from the list of RS error sources. In the new system, all costs are reported as accruals which are the form that National Accounts staff requires, thus eliminating this important cause of specification error. In addition, much of the information that municipalities provide for the RS is taken directly from standardized income statements and balance sheets. Therefore, it is highly unlikely for the concepts implied by the source data to deviate from the concepts underlying the survey questions and, thus, for a specification error to occur.

The following are noteworthy quality improvement activities that occurred in 2012:

- The RS instrument was evaluated by methodologists in Statistic Sweden's cognitive laboratory and a number of improvements were suggested. The RS instrument was revised on the basis of this evaluation.
- Various data collection procedures were redesigned and improved to reduce respondent burden and increase data accuracy.
- Important changes were made to the editing process to reduce editing error and increase editing effectiveness. For example, a "value stream" approach to editing is being implemented that includes selective editing, an editor team approach to editing, and more linking of related items to achieve greater internal consistency.
- The QD document was substantially revised and improved and plans are to release it in June 2013. However, see the recommendation below regarding needed further improvements.

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 5a where we show ratings from Biemer and Trewin (2012) compared to the current ratings.  Exhibit 5b repeats the current ratings in Exhibit 5a, but without the comparison to the prior year.

We have several recommendations to offer for future research. First, in Biemer and Trewin (2012), we noted that more research should be devoted to understanding the errors associated with the RS data and how these errors propagate through the National Accounts to cause biases in the National Accounts 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 National Accounts are affected by them.  Moreover, we noted above that the QD was substantially revised and improved. However, this document revealed that very little has been done to study, quantify, and document the key error risks for the RS.

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.

Likewise, it is important to understand the errors associated with the modelling of data. For example, RS staff reported that 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 of data quality is very high.  The RS should mount such a study in the next 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.

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 them.

Finally, in our 2012 report, we also mentioned the potential for catastrophic error in RS as a result of errors in the disability care estimates because what a municipality reports on this line as well as RS changes during the editing process can directly influence the size of subsidy or fee municipalities receive. The RS has been monitoring this problem and should continue to do so in the coming year.

*Exhibit 5. RS Accuracy Ratings for 2012*[3]

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification Error | 74 | N/A | N/A | N/A | N/A | N/A | N/A | **N/A** |
| Frame error | 43 | 60 | ○ | ○ | ◒ | ◒ | N/A | **L** |
| Non-response error | 52 | 52 | ○ | ○ | ◒ | ◓ | ○ | **M** |
| Measurement error | 52 | 58 | ○ | ○ | ◒ | ◒ | ○ | **M** |
| Data processing | 46 | 48 | ◓ | ◓ | ◒ | ○ | ○ | **H** |
| Sampling error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **N/A** |
| Model/estimation error | 38 | 38 | ◓ | ◓ | ◒ | ◓ | ◓ | **H** |
| Revision error | 58 | 58 | ○ | ◓ | ◒ | ◒ | N/A | **L** |
| **Total score** | **46,7** | **49,6** | | | | | | |

[3] Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. See Annex 2 for more details.

#### 4.1.2 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 updating of the QD to include material on sampling errors. This is to be released in February together with the January 2013 CPI.
- The commencement of the use of scanner data for one major retail chain. There are plans to extend the use to other retail chains. As well as increasing the size of the sample in some important segments, the prices will include discounts which are otherwise difficult to collect.
- Improved procedures for adjusting quality change are being introduced to provide better control over this important component. These will mean that more quality change is being assessed centrally rather than by price collectors.
- Revised estimates of sampling errors have been compiled.
- The new processing system has reduced the risk of error from this source.

In Biemer and Trewin (2012), 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.

With respect to (a), the problem still largely exists although more recent calculations suggest the sampling error is lower than previously thought – a sampling error of plus or minus 0.3% in absolute terms at the 95% confidence interval. Moreover, the use of scanner data will reduce the size of sampling errors in those commodity groups using scanner data and have some impact on the overall sampling error. With respect to (b), there have been a number of initiatives to address this problem although the impact has not been quantified. There has been no action with respect to (c). With respect to (d) there have been steps taken to reduce measurement errors due to price collector error on assessing quality change.

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.

1. Redo the 1999 study on potential CPI biases as much has changed since then with CPI methods and revised procedures may mean that these biases are now different.
2. Continue the introduction of scanner data to reduce sampling errors in the relevant components but, perhaps more importantly, reduce the measurement errors especially those associated with assessing discounts.
3. Review the efficiency of the current sample design especially with the introduction of the large scanner data sets. The emphasis should be on ensuring the most efficient design is obtained given the restricted budget for the CPI.
4. Review procedures for ensuring the providers of mainly internet based sales are included on the framework with appropriate probabilities.
5. Statistics Sweden has excellent expertise in methods for the CPI and has had for several years. Several of the most experienced staff may retire over the next few years. This might considerably reduce the expertise unless steps are taken to build up this expertise in new staff. This is strongly encouraged.

*Exhibit 6. CPI Accuracy Ratings for 2012*[4]

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification Error | 68 | 68 | ◡ | ○ | ◉ | ◡ | ◠ | **H** |
| Frame error | 62 | 62 | ◡ | ◡ | ○ | ◡ | ○ | **M** |
| Non-response error | 55 | 55 | ◠ | ◠ | ◉ | ◡ | N/A | **L** |
| Measurement error | 58 | 62 | ◡ | ○ | ◉ | ○ | ○ | **H** |
| Data processing error | 70 | 76 | ◡ | ○ | ◉ | ◡ | ◡ | **H** |
| Sampling error | 54 | 66 | ◡ | ◡ | ◉ | ○ | ◠ | **H** |
| Model/estimation error | 52 | 52 | ○ | ○ | ○ | ◠ | ○ | **H** |
| Revision error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **N/A** |
| **Total score** | **60,3** | **63,9** | | | | | | |

4 Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. See annex 2 for more details.

#### 4.1.3 FOREIGN TRADE OF GOODS (FTG2)

In our 2012 report, FTG's evaluation score was among the highest. We believe that the FTG has continued this high level of performance in the current ASPIRE round. The following are noteworthy quality improvement activities that occurred in 2012:

- Communication with users regarding survey error generally improved as a result of improvements to the QD (to be published in 2013) – particularly with regard to specification error, nonresponse error, data processing error, modelling/estimation error, and revision error.
- Three important studies were completed that provide more information regarding survey error. These results were documented in the following reports: "Improvement of the Work on Revisions in the Swedish External Trade on Goods," "Improving Macro-Editing in Intrastat," and "Improvement of the Distribution Keys for the Estimated Trade in the Swedish Intrastat."
- Swedish Customs adopted Statistics Sweden's editing system which demonstrates that it is a state of the art system.
- Plans are in place to better understand the causes of revision error, its impact on important users such as the National Accounts, and some effective means for reducing it over time.
- An asymmetry study with Finland was completed which focused on the effects of coding error on the trade statistics.
- Work is underway to replace the current Excel-based macro-editing software with much improved and flexible software written by IT professionals.
- Use of "The Standardized Toolbox" has increased leading to a number of improved practices.
- A new survey of statistical value is scheduled for 2013.

The current and previous round's ratings are shown in Exhibit 7 as well as the current ratings in graphical form. Note that the risk level for three error sources – frame error, data processing error, and revision error – were revised based upon new information we received in this round. Frame error was revised downward to Low based upon additional information obtained in this round that suggested that intrinsic risks of bias due to overcoverage error in the business register are lower than originally believed. Data processing error was raised to High intrinsic risk after realising the risks of editing to data quality. Likewise, revision error was raised to High based on conversations with the National Accounts staff regarding the impacts of revisions of FTG statistics on the estimates of GDP.

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. 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. This research is important to other EU countries as well so some collaboration with other EU countries would seem appropriate.
2. In addition to the National Accounts, FTG staff should reach out to other users to understand the impact revisions have on their users of the foreign trade statistics. Key among these are the Ministry of Finance, The Ministry of Enterprise, Energy and Communications, and the Riksbank.
3. The QD 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.

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 NACE coding error on the trade statistics through the asymmetry studies that have been conducted. Additional studies are needed, particularly at the CN8 level of coding as that is the level required for National Accounts estimation.
6. Improvements are needed in the modelling of statistical value. For example, currently models develop from Intrastat data are applied to Extrastat invoice values to derive Extrastat statistical values. However, since shipping costs are greater for Extrastat, these adjustments are likely inappropriate. This is but one area that needs further study.

Since revision error, is to a large extent, caused by a few large enterprises whose preliminary and final reports differ substantially, one activity related to (1) is to understand why this occurs and how Statistics Sweden can help these enterprises report more accurately. A related factor is late responders whose data are not available in time for the preliminary release. We understand that plans are underway to meet with a number of these enterprises to better understand their issues with reporting. We are supportive of this approach.

In Biemer and Trewin (2012), we recommended that the following error risks need immediate attention: (a) the misclassification of commodities (particularly in the paper reports), (b) the information on net weight (and other quantity measures) of shipments especially for textiles and chemicals, (c) errors in the editing process, (d) errors resulting from the methods used to convert invoice value to statistical value and (e) potentially missing data from the Extrastat component. We believe that the FTG has made good progress on (a) with the recently completed asymmetry study with Finland. Likewise, the macro-editing research addresses some of the concerns in (b) and (c) and there are plans to study ways to improve the models used for converting invoice value to statistical value – i.e., (d) and (e) and note the next survey of statistical value to be conducted in 2013. We encourage further research along these lines.

*Exhibit 7. FTG Accuracy Ratings for 2012*[5]

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification error | 58 | 58 | ○ | ○ | ◒ | ◒ | ○ | **M** |
| Frame error | 58 | 58 | ○ | ○ | ◒ | ○ | ◒ | **L** |
| Non-response error | 62 | 66 | ◒ | ◒ | ◒ | ○ | ◒ | **M** |
| Measurement error | 54 | 62 | ◒ | ○ | ◒ | ◒ | ○ | **H** |
| Data processing | 46 | 60 | ◒ | ◒ | ◒ | ◓ | ○ | **H** |
| Sampling error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **N/A** |
| Model/estimation | 66 | 80 | ◒ | ◒ | ◎ | ◎ | ◒ | **M** |
| Revision error | 62 | 76 | ◒ | ◒ | ◒ | ◎ | ◒ | **H** |
| **Total score** | **57,3** | **65,8** | | | | | | |

[5] Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. See Annex 2 for more details.

#### 4.1.4 *LABOUR FORCE SURVEY (LFS)*

As noted in the 2012 report, low response rates, which have further deteriorated in the ensuing 12 months, is a continuing concern for the LFS. While the *intrinsic* risk for this survey remained at High, the *residual* risk (i.e., the risk of error after accounting for all attempts to reduce the risk) is now at a *critical* or *crisis* level. In other words, we believe that, so far, attempts at reducing the risks of nonresponse bias have been insufficient to stem the rising tide of nonresponse. There is little cause to be optimistic that the situation will improve in the next few years.

There are several factors in the gradual erosion of response rates experienced over recent years. LFS staff reported that they believe a key factor is the telephone interviewer (TI). They note that a rapid increase in workload in 2009 from 21,500 to 29,500 interviews per month put heavier demands on the TIs, requiring that many new TIs be recruited. Prior to 2008, the TIs were largely hired on a temporary basis giving the agency quite a bit of flexibility in steering interviewer time to evenings and weekends when household contact rates can be optimized. However, new legislation, passed in 2008, required employers to transform temporary employment contracts, for those who had worked for more than two years with the agency, to permanent-term employment contracts. These new contracts allowed workers a minimum number of hours per week, a certain portion of which became regular office hours. In general, these new arrangements gave Statistics Sweden less flexibility for staffing interview time-slots compared with the previous arrangements, with the consequence that the agency increasingly faces difficulties in scheduling TIs at optimal contact times. Thus, the contact rate, which comprises about 50% of the LFS nonresponse rate, has decreased. In addition, the motivation to strive for higher response rates seems also to have eroded, most notably among the supervisory field staff who believe that these staffing issues may only be a small part of the problem. 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.

An important root cause to consider is the current approach of conducting the initial interview by telephone rather than by face to face. Although using in-person contacts for the first wave interviews would add costs, first wave response rates would be substantially higher – potentially as much as 10-15 percentage points. However, research has shown that response rates to later waves, which are still conducted by telephone, are also higher – perhaps by 5-10 percentage points because a relationship has been established. Further, the costs of re-contacting Wave 1 respondents in subsequent waves may be considerably lower because contact details can be obtained. If so, the cost of the initial face to face interview would be partially off-set by the subsequent gains in efficiency. Many countries still use face to face interviewing for the first contact for their LFS.

We do not purport to understand all causal factors of nonresponse for the LFS. There are many to consider and the situation appears to be quite complex, even to methodologists who work on the LFS routinely. To address the issue, a nonresponse project was commissioned in 2010 by the DG and led by the Deputy DG. This project has spawned an number of activities and plans to conduct experiments and studies aimed at understanding the causes of nonresponse in household surveys, especially the LFS, and what steps should be taken to reverse the downward trends.

Some of the notable improvements in the LFS since our last review include the following:

- Plans have progress to conduct a reinterview study of 2000 responding households aimed at evaluating the measurement error in labour force and other LFS statistics.

- Plans have also progressed to conduct a cognitive evaluation of the LFS questionnaire focusing on specification error and measurement error.
- There has been much planning for how to reduce/adjust for nonresponse bias and a number of error-reduction studies are in process. For example, a plan was developed to implement adaptive design to reduce nonresponse bias without necessarily increase response rates.
- Call monitoring has been implemented for both centralized and decentralized interviewing. This addresses one of the key areas of non-compliance with standards identified in the 2012 report.
- Innovative research on the use of GREG estimation has been completed and plans are now underway to implement this new estimation methodology.

Exhibit 8 displays the changes in ratings between Rounds 1 and 2 as well as the current ratings in graphical form. We have the following recommendations for improvements:

1. Conduct evaluation studies aimed at isolating the causes of nonresponse in the LFS. Possible causes are (a) the societal changes which have resulted in great difficulties in contacting respondents and persuading them to participate in surveys, (b) TI work hours preferences that tend to be 8 am to 5 pm on weekdays, (c) poor management strategies in the telephone centre, and (d) other causes related to current methodologies that are used in the LFS.
2. Conduct experiments to determine whether face to face interviewing at the initial LFS interview would increase response rates and the likely cost and benefits of such an approach.
3. Rather than emphasizing response rates in the evaluation of nonresponse, emphasize sample representativity. Conduct studies to improve sample representativity even in the face of declining response rates.
4. Relatedly, conduct studies that seek to evaluate the bias in the fully weighted and adjusted LFS estimates. How effective is the nonresponse adjustments at compensating for nonresponse? Are better methods available that would lower the residual risk of nonresponse bias?
5. Evaluate the effectiveness of the current approach to call monitoring. This research would determine how the fact that TIs know with some certainty which calls are monitored affects the ability of the monitoring approach to achieve its data quality goals.
6. 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 Round 1 review, we noted that nonresponse and measurement error were high priorities for future research. These continue to be high priorities in the coming year with even greater emphasis on nonresponse. We commend the LFS staff for its attention to measurement error and implementation of a call monitoring capability, albeit somewhat limited. We also commend the LFS staff for the degree to which the risks of nonresponse bias have been dealt with in the previous year. Unfortunately, even greater effort was apparently needed because the situation has deteriorated somewhat since our last review.

As noted in Section 3.4, two user dimensions of quality were also evaluated for the LFS. A discussion of these findings is provided in Section 4.3.2.

### *Exhibit 8. LFS Accuracy Ratings for 2012*[6]

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification error | 66 | 70 | ◒ | ◒ | ◒ | ◒ | ◒ | **L** |
| Frame error | 58 | 58 | ◒ | ◒ | ◒ | ◓ | ○ | **L** |
| Non-response error | 56 | 52 | ○ | ○ | ○ | ○ | ○ | **H** |
| Measurement error | 50 | 56 | ○ | ○ | ○ | ○ | ◒ | **H** |
| Data processing error | 54 | 62 | ○ | ○ | ◒ | ◒ | ◒ | **M** |
| Sampling error | 70 | 78 | ◒ | ◎ | ◒ | ◒ | ◎ | **M** |
| Model/estimation error | 50 | 60 | ○ | ○ | ○ | ◒ | ◒ | **M** |
| Revision error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **N/A** |
| **Total score** | **56,4** | **60,9** | | | | | | |

[6] Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. The blue shaded areas show where deteriorations have been noted. See Annex 2 for more details.

#### *4.1.5 STRUCTURAL BUSINESS STATISTICS (SBS)*

There have been some improvements in Structural Business Statistics (SBS) over the last 12 months.

- The complete overview of their QD (to be published in 2013), and the inclusion of new information, will result 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. We had difficulty understanding the material describing the survey structure until it was explained to us. Perhaps this suggests some modifications to this material in the QD.
- There have been developments in the IT systems used for coding and imputation in the SBS.
- There is improved documentation on potential errors in data processing.
- There has been investigatory work on the development and maintenance of statistical units in preparation for the Common Business Framework initiative.
- A service Level Agreement has been developed with the National Accounts although co-operation was already good.

In Biemer and Trewin (2012) we noted two areas of risk which were (a) data processing because it does not follow ISO standards in some respects and (b) revisions between preliminary and final estimates where there appear to be some systemic differences.

The risk from (a) was mainly concerned with the lack of checking on keying but the risk would have continued to reduce because 95% of data is now received electronically. The Statistics Sweden standard does not require checks on data keying accuracy in these circumstances.

There has been no further work on (b) but we may have under-estimated the changes that were made following a study "Use of administrative data for enterprise statistics" which was carried out in 2010 on behalf of Eurostat to study the differences between the preliminary and final results with the aim of improving the quality of the preliminary delivery. The study identified a number of problem areas that caused much of the differences, and led to some changes in the survey schedule, model assumptions and calculations to reduce these differences. Nevertheless, there may be scope for further improvement as noted below.

The focus of improvements should be on those areas of higher risk where the rating is relatively low. Based on the recent review, we suggest the following.

1. We strongly support the proposed investigatory work on compliance with questionnaire concepts using the cognitive laboratory.
2. We also support the continuation of experiments to upload enterprise accounts (compiled using a standard chart of accounts) to the SBS data collection system thereby further reducing the amount of data keying that is required.
3. SBS should collaborate with the 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 National Accounts statistics.
4. SBS should start thinking about the work required for moving to the new Business Register in 2014 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 may be significantly reduced with the new Business Register.

5. 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 National Accounts.
6. 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 National Accounts may assist with the reduction of revisions. Their work enables them to have a good overview of the economy.

***Exhibit 9. SBS Accuracy Ratings for 2012[7]***

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification error | 50 | 54 | ○ | ○ | ◒ | ○ | ○ | **M** |
| Frame error | 64 | 64 | ◒ | ◒ | ◒ | ○ | ○ | **M** |
| Non-response error | 70 | 70 | ◒ | ○ | ◒ | ◒ | ◒ | **M** |
| Measurement error | 50 | 52 | ○ | ○ | ○ | ○ | ○ | **H** |
| Data processing error | 54 | 60 | ○ | ○ | ◒ | ○ | ◒ | **H** |
| Sampling error | 82 | 84 | ◒ | ◒ | ◎ | ◎ | ◒ | **M** |
| Model/estimation error | 60 | 60 | ○ | ○ | ◒ | ◓ | ◒ | **H** |
| Revision error | 56 | 56 | ○ | ○ | ◒ | ◓ | ○ | **H** |
| **Total Score** | **59,6** | **61,4** | | | | | | |

[7] Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. The blue shaded areas show where deteriorations have been noted. See Annex 2 for more details.

### 4.1.6 BUSINESS REGISTER

There have been some important improvements over the last 12 months as noted below.

- There has been continuing work on planning for the development of the new Business Register which includes some of the areas impacting on the accuracy of the Register for statistical requirements.
- Studies of user requirements have been undertaken.
- The preparation of the QD is an important development. In particular, it should help internal users understand the strengths and weaknesses of the Register. However, it could benefit from the inclusion of more quantitative information.
- A closer relationship with the Swedish Tax Agency seems to have developed. The Tax Agency decided not to go ahead with some proposed changes as a result of representations by Statistics Sweden.
- There has been a study of the accuracy of the coding undertaken by enterprises although the dependent nature of the study raises question marks about its reliability. As discussed below, we would propose a different study design.
- The number of enterprises without a NACE code has continued to decline.

Nevertheless, despite these improvements, we remain concerned about some aspects of the Business Register. 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 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 Business Register that we spoke to.

These are the same two problems we referred to last year and we are not convinced that sufficient action has been taken yet to address them. There is some reliance on a new Business Register System to be constructed in 2014 but it will be three or more years before the new system can be used.

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.

1. The specifications for the revised Business Register System should focus 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 Business Register.
2. The new Business Register System should support the creation of a Register 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 Business Register suffer as a consequence.
3. 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. It may be necessary to meet more than twice a year, as at present, during the Business Register redevelopment project. It would be good to formalise the arrangements especially if part of the Tax Agency is to move out of Stockholm.
4. 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.

5. 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 dependency on industry for tax concessions.
6. The current arrangement of revising NACE codes when detected in the SBS introduces biases. For example, it is more likely that an enterprise coded to manufacturing will have its NACE code revised to a non-manufacturing enterprise than vice versa. 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 is obtained from a source other than a sample survey.
7. 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.

***Exhibit 10. BR Accuracy Ratings for 2012[8]***

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification error | 62 | 66 | ○ | ○ | ◒ | ◒ | ◒ | L |
| Frame error - overcoverage | 48 | 56 | ○ | ○ | ◒ | ○ | ○ | H |
| Frame error - undercoverage | 42 | 46 | ◓ | ◓ | ○ | ○ | ○ | M |
| Frame error - duplication | 55 | 63 | ○ | ○ | ◒ | ◒ | N/A | L |
| Missing data | 48 | 48 | ○ | ○ | ○ | ○ | ◓ | L |
| Content error | 42 | 46 | ◓ | ◓ | ◒ | ○ | ○ | H |
| **Total score** | **47,2** | **52,2** | | | | | | |

[8] Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. The blue shaded areas show where deteriorations have been noted. See Annex 2 for more details.

#### 4.1.7 TOTAL POPULATION REGISTER (TPR)

As we reported in Biemer and Trewin (2012), the highest risk area for the TPR is overcoverage of the population – essentially the inclusion of nonresidents. There has not been much progress to address this problem; however, plans have been laid to make more progress in this area in the coming year. A methodologist has joined the TPR staff who will focus on overcoverage among other things. In addition, the TPR will play a central role in the population census that will be conducted in 2013. This will provide ample opportunity to address many quality issues in the TPR, particularly overcoverage.

Other issues that need attention in the TPR are specification error and missing data – both medium risk error sources. For specification error, there are several issues. One is the difference between the registered address and the address of an individual's current residence. For surveys and censuses, the latter is the more important for contacting purposes. Moreover, the extent of the problem is not well quantified at present and could be an important cause of nonresponse in household surveys. Another issue affects persons having dual citizenships. Only one indication of citizenship is preserved on the register which is Swedish for Swedish citizens with dual citizenship. For some users, knowing all countries of citizenship would be important.

For missing data, item nonresponse for dwelling unit address is about 5% currently and the impact of this type of missing data on various TPR uses (e.g., the LFS) is yet unexplored.

Another type of missing data is the indicator for persons belonging to the same household. Currently, the cohabitants of a household cannot be determined unless there are children and parents who are registered at the same address. Staff estimate that at least 400,000 persons on the register are unidentified cohabitants. Identification of family and household membership is expected to substantially improve with the 2013 census as well as through the use of a unique identification number that will be assigned to each dwelling unit.

Some noteworthy improvements over the last year include the following:

- A new QD was written for the TPR. Prior to 2012, no QD existed for the TPR.
- A methodologist has been assigned to work with the TPR staff on studies related to data quality improvement and who will focus initially on overcoverage.
- As a consequence, plans have been approved to conduct a study of overcoverage in the TPR in collaboration with subject matter personnel and methodologists.
- Two staff members have joined a working group that includes representatives from the Swedish Tax Agency, the Swedish land register, the Swedish dwelling register, and the Swedish association for local government. This group will meet four times per year to discuss quality issues and plan for quality improvements.
- Plans have been approved to use errors found during the census to correct/enhance TPR content, particularly with regard to household family membership.

The Round 1 to Round 2 changes are shown in Exhibit 11 as well as the current ratings in graphical form. The intrinsic risk for one error source (specification error) was raised from Low to Medium to correct an error in the Round 1 assessment. The specification error issues that affect the TPR were described earlier.

We include the following recommendations for the coming year:

1. 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.

2. 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 who represent the main user groups to understand the effects of overcoverage on key population estimates such as the unemployment rate.
3. Study alternative approaches for classifying a registrant as a non-resident based upon noncontacts, register activity, and so on. If necessary, they could be recorded as a 'likely non-resident' rather than removed from the TPR.
4. It is important to focus on the validity of the information contained in the TPR by conducting studies that attempt to evaluate the validity of the most important variables.

With regard to (2), one idea is to estimate the mean characteristics of overcovered persons. For some uses, overcoverage may not be biasing if it is "completely at random; i.e., if the overcovered persons have characteristics that do not differ appreciably from true residents.

With regard to (4), 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 is 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 validity using this information under plausible assumptions regarding the randomness of the changes.

Finally, as noted in our 2012 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 including users within Statistics Sweden.

***Exhibit 11. TPR Accuracy Ratings for 2012[9]***

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Specification error | 44 | 46 | ◓ | ◓ | ○ | ○ | ◓ | M |
| Frame error: overcoverage | 52 | 56 | ○ | ○ | ○ | ○ | ○ | H |
| Frame error: undercoverage | 38 | 60 | ○ | ○ | ◒ | ◒ | N/A | L |
| Frame error: duplication | 70 | 70 | ○ | ○ | ◒ | ◒ | N/A | L |
| Missing data error: item and variable | 60 | 66 | ○ | ○ | ◒ | ○ | ◒ | M |
| Content error | 50 | 58 | ○ | ○ | ◒ | ◒ | ○ | L |
| **Total score** | **52,2** | **58,0** | | | | | | |

[9] Round 1 was carried out in Nov/Dec 2011. Round 2 was carried out in Nov/Dec 2012. The pink shaded areas indicate areas where improvements have been noted. The blue shaded areas show where deteriorations have been noted. See Annex 2 for more details.

## 4.2 NEW PRODUCT REVIEWS

### *4.2.1 REVIEW OF QUARTERLY GDP DATA QUALITY*

The quarterly National Accounts are a very complex product that relies on many input data sources from both within Statistics Sweden and from external sources. For our review, we could only look at a small number of the data sources that provided the greatest risk to the National Accounts. We also only looked at the production side of the National Accounts. Using the advice of the National Accounts, 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 is used in the quarterly National Accounts. The first two were chosen largely because of the significant contribution they make to the quarterly National Accounts whereas merchanting has been making a significant contribution to recent estimates of change and questions have been asked about the reliability of this data as it indicates increases in GDP are showing a different relationship to the industrial production index than in the past.

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

We believe the areas most in need of improvement, in rough priority order, are (1) a robust processing system for the National Accounts 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.

In addition we strongly support the short term economic statistics project which will integrate those surveys supporting the industrial production index and the services production index. We also support the development of standardized methods for balancing the quarterly National Accounts. Nevertheless, there will always be an element of human judgment involved in the balancing process. This will be necessary to ensure the balancing process does not result in estimates that are implausible. Statistics Sweden’s practice of publishing the discrepancy prior to balancing is an excellent example of transparency in statistics.

On the other hand, we are concerned about the proposal to discontinue the National Accounts research group. It is important to have a group that can research National Accounts although they don’t organizationally need to be part of the National Accounts. For example, in the ABS, National Accounts research is undertaken by a specialist 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 National Accounts and other users of their services is crucial.

Of concern is that the level of experience, and possibly expertise, in National Accounts with the large number of retirements in recent years. We support the steps Sweden is taking to build up this expertise in an area of statistics that is so crucial to the reputation of Statistics Sweden.

With respect to improvement area (1), we strongly encourage Statistics Sweden to engage a specialist in National Accounts processing systems to advise them on the best way forward in terms of a long term processing system for the National Accounts. They should come from a country that has successfully implemented a National Accounts processing system. The expertise does not exist at Statistics Sweden and there is high risk in using internal IT specialists who do not really understand the subtleties of the National Accounts. There are proprietary products available on which are National Accounts processing system can be developed.

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. It may be possible to develop a more sophisticated model that takes this into account. The general approach used by Statistics Sweden is consistent with international practice but we have not heard of other countries having the same ‘bias’ problems in periods of rapid economic change. However, the ABS does find that income tends to be understated and expenditure over-stated by respondents and makes adjustments accordingly. Without these adjustments, value added would be understated.

For Construction, models are used because it is difficult to get reliable estimates directly from surveys. This is not the case for many countries so direct estimates from surveys, of output indicators at least, should be investigated as well as what needs to be done to improve the model based estimates. It is an important sector of the economy and a strong indicator of general economic activity so a review of the activities of other countries would be worthwhile. As Sweden is one of the strongest statistical offices in Europe, this should include a review of the methods outside Europe. The ABS largely relies on survey sources for its annual construction estimates but with some adjustments for owner builders especially for alterations.

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.

With respect to improvement area (4), it is not always easy to understand the impacts on the accuracy of National Accounts of inaccuracies of the source data especially given the complexity of the processes used included the balancing processes. One possibility is to use sensitivity studies where an error is introduced into a particular data source and the impact on GDP is assessed. More specifically, in the formula for the estimate of GDP in (0.1), one could substitute $y_p+\delta$ for $y_p$, where $\delta$ represents a plausible error in the input $y_p$, and then observe the effect on the estimate. This would be done for each of the key data sources (i.e., $y_1,\dots y_P$ ) in turn. This is likely to be an expensive operation so should be seen as a one-off exercise. We do not know whether it is feasible or not and there may be methods for approximating the impacts. The objective is to assess the relative importance of the different input data sources to help focus data development effort.

***Exhibit 12. GDP Quarterly Ratings for 2012***

| Error source | Score | Knowledge of Risks | Communication to Users | Available Expertise | Compliance to standards and best practices | Plans toward mitigation of risk | Risk to data quality |
|---|---|---|---|---|---|---|---|
| Input data source - Index of Service Production, ISP | 58 | ◠ | ○ | ◡ | ◡ | ○ | H |
| Input data source - Index of Industrial Production, IIP | 58 | ◠ | ○ | ◡ | ◡ | ○ | H |
| Input data source - Merchanting Service of global enterprises (also covers royalties, licensing and R&D) | 42 | ◠ | ○ | ○ | ◠ | ○ | H |
| Compilation error (modelling) | 48 | ○ | ○ | ○ | ○ | ◠ | H |
| Compilation error (data processing) | 40 | ◡ | N/A | ◠ | ◠ | ◠ | H |
| Deflation error (including specification error) | 48 | ◠ | ◠ | ◡ | ◡ | ◠ | H |
| Balancing Error | 56 | ○ | ○ | ◡ | ○ | ○ | H |
| Revisions Error | 56 | ◡ | ○ | ◡ | ○ | ◠ | M |
| Total score | 50,5 | | | | | | |

### 4.2.2 *REVIEW OF ANNUAL GDP DATA QUALITY*

As is the case with the quarterly National Accounts, the annual National Accounts are a very complex product that relies on many input data sources from both within Statistics Sweden and from external sources. For our review, we could only look at a small number of the data sources that provided the greatest risk to the annual National Accounts. As with the quarterly National Accounts, we only looked at the production side of the National Accounts. Using the advice of the National Accounts, we selected only one input data source – the structural business statistics which contribute a very high proportion of GDP estimates.

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

We believe the areas most in need of improvement, in priority order are (1) a robust processing system for the National Accounts that includes a 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.

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

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. 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 National Accounts it is not always easy to understand the impacts on the accuracy of National Accounts 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. The volatility of the deflators also has to be taken into account.

***Exhibit 13. GDP Annual Ratings for 2012***

| Error source | Score | Knowledge of Risks | Communication to Users | Available Expertise | Compliance to standards and best | Plans toward mitigation of risk | Risk to data quality |
|---|---|---|---|---|---|---|---|
| Input data source - Structural Business Statistics, SBS | 66 | ◒ | ○ | ◒ | ◒ | ◒ | H |
| Compilation error - modelling | 48 | ○ | ◓ | ○ | ◒ | ◓ | H |
| Compilation error - data processing | 35 | ○ | N/A | ◓ | ◓ | ◓ | H |
| Deflation error (including specification error) | 48 | ◓ | ◓ | ◒ | ◒ | ◓ | H |
| Balancing Error | 50 | ○ | ○ | ○ | ◒ | ◓ | H |
| Revisions Error | 54 | ○ | ◒ | ○ | ○ | ◓ | M |
| Total score | 49,9 | | | | | | |

#### *4.2.3 REVIEW OF THE LIVING CONDITIONS SURVEY (ULF/SILC)*

The Survey of Living Conditions (ULF/SILC) is a long-standing survey dating from the mid-1970's. The survey has undergone a number of expansions, most notably the merging of the Eurostat SILC survey with the older ULF survey. The accumulation of these changes has resulted in a survey design that is quite complicated and unwieldy. One important consequence is that selection probabilities cannot be accurately calculated forcing statisticians assign equal selection probabilities where they are clearly unequal. However, this is only one of the important issues facing methodologists and users of the ULF/SILC).
Some of the problems we identified in our review are listed below:

1. The interview, which is conducted by telephone, averages 36 minutes but can be more than one hour for some situations.
   This is an important consequence of conducting burdensome interviews, particularly by telephone. The survey methods literature suggests that the risk of poor data quality due to satisficing increases with the length of the interview and telephone interviews lasting more than 20 minutes are substantially at risk.
2. 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. An evaluation of the reliability of these data needs be conducted to examine the extent of the problem and either to provide (a) justification for continuing this practice or (b) evidence that it should be discontinued.
3. Response rates are quite low, averaging about 59 percent. They have declined steadily over the years and tend to vary considerably by interview component. The causal factors are very similar to those discussed above for the LFS; however, the nonresponse adjustment approach does not appear similar to the LFS approach. 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, as far as we know.
4. Given the long history of the survey, the questionnaire is sorely in need of refreshing and updating.
   We believe that specification error poses a considerable risk to data quality primarily because an expert review of the survey questions has never been undertaken within the last 20 years.
5. 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.

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

In light of these issues, we have the following recommendations for the ULF/SILC).

1. The ULF/SILC survey in its present form is too complex. Consider redesigning the survey to simplify the sample design and panel structure as well as to shorten/improve the questionnaire.
   - The current interview (for some panels) is too long. Consider shortening the interview to an average of approximately 20 minutes for each component. Maximum interview length should follow best practices for this type of survey which is currently at about 30 minutes.
   - Consider focusing on the SILC component rather than the ULF but with the facility to add questions of special interest to Swedish users. The key user groups should be consulted during the redesign regarding which components of the ULF/SILC should be retained.

2. Appendix 16 from the so-called Appendix series on *The Swedish Survey of Living Conditions – Design and methods* should be updated.
3. The QD is quite confusing and needs to be improved.
4. Document why proper sampling weights cannot be computed in enough detail that the problems can addressed by external reviewers.
5. Evaluate the reliability of data obtained in the children's survey.
6. Conduct item nonresponse rate analysis to identify extent of item nonresponse and items more prone to nonresponse.
7. Conduct unit nonresponse analysis to identify components of the survey most subject to nonresponse. Take steps immediately to reduce unit nonresponse bias.
8. Evaluate coding error for interviewer coded items.

***Exhibit 14. ULF/SILC Ratings for 2012***

| **Error source** | **Score round 2** | **Knowledge of Risks** | **Communication to Users** | **Available Expertise** | **Compliance to standards and best practices** | **Plans toward mitigation of risk** | **Risk to data quality** |
|---|---|---|---|---|---|---|---|
| Specification error | 34 | ◓ | ◓ | ◒ | ◓ | ● | **M** |
| Frame error | 42 | ◓ | ◓ | ◒ | ○ | ◓ | **H** |
| Non-response error | 40 | ○ | ◓ | ○ | ◓ | ◓ | **H** |
| Measurement error | 46 | ◓ | ◓ | ◎ | ◓ | ○ | **H** |
| Data processing error | 42 | ○ | ◓ | ◒ | ○ | ● | **L** |
| Sampling error | 54 | ◒ | ◒ | ◒ | ◓ | ◓ | **M** |
| Model/estimation error | 38 | ○ | ◓ | ◒ | ◓ | ● | **H** |
| Revision error | N/A | N/A | N/A | N/A | N/A | N/A | **N/A** |
| **Total score** | **42,1** | | | | | | |

## 4.3 REVIEW OF USER DIMENSIONS

### 4.3.1 ASSESSMENT OF ACCESSIBILITY & CLARITY AND RELEVANCE/CONTENTS FOR THE CPI

As noted in Section 4.1.4, two user dimensions were included in the CPI review, primarily as a test of these expansions to the ASPIRE process but also to provide feedback to Statistics Sweden regarding current CPI performance on these dimensions. The results of this evaluation are shown in Exhibit 15. The underlying numerical scores with addition comments can be found in Exhibit 4.1 in Annex 4.

***Relevance/Contents.*** The inputs to the CPI and the outputs of the CPI were considered separately. From the point of view of relevance, the inputs were assigned a low intrinsic risk whereas outputs were assigned a high intrinsic risk level based upon their importance to the user community. With inputs, the risk is low because the inputs to the CPI are well defined including through regular discussions at the European level as part of the harmonisation process. With outputs, there is much more choice on the contents of the outputs and how they are disseminated.

There are no suggestions for improvements on improving the relevance of inputs. The current arrangements for ensuring relevance are quite satisfactory.

For the relevance of outputs, we note the regular meetings with the CPI Advisory Board and other major users (i.e. the so called 'power' users. We also note the informal contact with other users through seminars and questions raised with information services staff. There is to be a review of some of the outputs during 2013 and we encourage Stat Sweden to also take the needs of both the power and other users into account for this review.

***Accessibility & Clarity.*** There are four aspects to accessibility and clarity that apply to the CPI. They are (a) ease of data access, (b) documentation (including meta data), (c) availability of quality reports, and (d) user support. All are regarded as high intrinsic risk as all are vital to effective use of the CPI. Unless, some steps were address each of these aspects, effective use of the CPI would be very difficult especially for the power users.

Regarding (a), the web site is the most important way of communicating with users. A User Survey was conducted during 2012 and this should suggest improvements especially in the ease of use of the web site. This should be an area of continuous improvement But Statistics Sweden should also investigate new developments in communication of statistics including communication direct to mobile devices and the use of social media.

Regarding (b) and (c), "The Swedish Consumer Price Index: A handbook of methods" has been published and is available on the SCB web site. An updated QD has been prepared and will be published with the January 2013 CPI release. The documentation should be continuously reviewed particularly as new information becomes available. The User Survey may highlight some areas where documentation can be improved. For (c), an update of the 1999 Study into CPI bias should be considered. This was a valuable and high quality study but some of the findings may have come less relevant over time. Also, the availability of the scanner data may have changed some of the bias estimates.

Regarding (d), there is a help desk to help address questions and a more personal relationship with the power users. There is also a service to support special requests for detailed data. We have no suggestions for further improvement.

### *Exhibit 15. Consumer Price Index (CPI) User Dimension Ratings for 2012*

| | Component | Average score | Knowledge of User Needs | Communication with Users | Available Expertise | Compliance with standards & best practices | Plans towards addressing user needs | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Relevance | Inputs (content, scope, classifications, etc) | 60 | ◒ | ◒ | ○ | ○ | N/A | **L** |
| | Outputs (including microdata and other products) | 66 | ◒ | ◒ | ◒ | ◎ | ◓ | **H** |
| Accessibility & Clarity | Ease of Data Access | 50 | ◒ | ○ | ○ | ○ | ◓ | **H** |
| | Documentation (including metadata) | 50 | ○ | ○ | ○ | ◒ | ◓ | **H** |
| | Availability of Quality Reports | 62 | ○ | ◒ | ◎ | ◒ | ◓ | **H** |
| | User Support | 58 | ○ | ○ | ◒ | ◎ | ◓ | **H** |
| | **Total Score User Quality Dimensions** | 57,4 | | | | | | |

#### 4.3.2 ASSESSMENT OF TIMELINESS & PUNCTUALITY AND COMPARABILITY & COHERENCE FOR THE LFS

As noted in Section 4.1.4, two user dimensions were included in the LFS review, primarily as a test of these expansions to the ASPIRE process but also to provide feedback to Statistics Sweden regarding current LFS performance on these dimensions. The results of this evaluation are shown in Exhibit 16. The underlying numerical scores with addition comments can be found in Exhibit 4.2 in Annex 4.

***Timeliness & Punctuality.*** For the purpose of the report, “timeliness” refers to the official schedule that is in place for reporting the current month’s employment statistics which is 18 to 20 days after the last reference week of the month. “Punctuality” refers to the ability of the LFS to meet this schedule month after month. Both were assigned a high (High) intrinsic risk level based upon their importance to the user community. According to LFS staff, users are generally satisfied with the release schedule for the LFS so the LFS should receive high marks for timeliness. However, since 2010, there have been about five months where the LFS has not met this schedule and has delayed delivery of the estimates up to one week. Thus, punctuality could be improved.

The primary cause of late deliveries is data collection and the difficulties of achieving high response rates as discussed in Section 4.1.4. On some occasions, delivering the estimates on time would mean accepting a response rate that is deemed inadequate by the LFS staff. Thus, there is often a trade-off between punctuality and accuracy. Part of the solution to the punctuality issues will be realized by the solution to poor response rates.

Major users and the press are notified in advance of any delay but are highly critical when this occurs.

***Comparability & Coherence.*** There are three aspects to comparability that apply to the LFS. Comparability (a) across population subgroups defined geographically, (b) across demographic domains (for example, age groups), and (c) across time, in particular, the effects of methodological changes and survey redesigns. With regard to (a) and (b), there is some risk that spurious differences could occur in estimates for geographic and other domains if the data collection methodology differs somewhat across these populations. For example, the LFS is collected both by centralized interviewing (about 35%) while the remainder is collected by decentralized interviewing. The differential effects associated with these two data collection administrations have never been evaluated. Nevertheless, if distribution of these two methods of administration is geographically unbalanced, spurious differences could arise in the estimates for these areas solely as a function of the administrative mode. This assumes that the error distributions for centralized and decentralized interviewing differ markedly. We deemed this to be a medium risk in the absence of any data that would allow a more precise assessment of the risks.

Regarding (c), the LFS has undergone a number of changes and redesigns over its history and considerable effort has been made to understand the effects of these changes on the estimates. Linkage studies that simulate the effects of the design changes on the data series prior to the “break” are one means for evaluating the methodological effects. For example, the effects of the new labour force definitions introduced in 2007 were recently evaluated for all estimates since 1987.

The LFS has maintained good contact with users with regard to these issues and other user-related issues. At present, there are four user groups that meet regularly: (1) the Expert Group on Labour Market statistics, EFAM, (2) User group for labour market statistics at Statistics Sweden, (3) working group SASA, and the (4) NA-LFS Board at Statistics Sweden.

Coherence of the LFS estimates refers to the degree to which the levels or trends for alternate, publicly available estimates of labour force statistics agree as well as the degree of concordance among the LFS and relatable statistics (for example, GDP). In addition to the LFS estimates, there are three other valid sources of employment statistics: (1) quarterly enterprise-based employment statistics, (2) register based employment statistics which is based on tax data (RAMS), and (3) the Swedish Public Employment Service unemployment figures. A comparison of these statistics with the LFS conducted in 2007 found difference in the estimates among these four sources; however, the study was inclusive regarding the sources of the differences although it noted differences in the definitions, data collection and estimation methodologies, and time reference periods as the likely causes. A better understanding of why these sources differ and how much the differences can be attributed to quality issues in the LFS data is sorely needed. Some countries have produced 'Labour Accounts' to provide reconciliation between the different sources of employment statistics.

We have the following recommendations which apply to both Timeliness & Punctuality and Comparability & Coherence:

1. Expand the scope of user needs assessment beyond "power users" to other important user segments in the community.

Currently the "power users" (viz., Riksdag, Ministry of Employment, National Accounts staff, etc.) command the most attention from the LFS as they should. However, there are many other corporate, government, and academic users in the community who represent the majority in terms of number of users whose needs are currently unknown and for whom the LFS has had little contact. Steps should be taken to ensure that the needs of these "smaller" users are not being neglected.

2. Take immediate steps to ensure that the LFS estimates are released on schedule without compromising data quality.

Although punctuality for the LFS has been generally good, recently delays have been problematic and worrisome. Until the problems in the data collection department can be resolved, the risk of future delays is quite high. As previously noted, there is a trade-off between punctuality and data quality. Thus, there is an increasing risk that greater punctuality will lead to comprises in data quality. The temptation to release timely data that is of questionable quality must be resisted at all costs.

***Exhibit 16. Labour Force Survey (LFS) User Dimension Ratings for 2012***

| | Component | Score round 1 | Knowledge of User Needs | Communication with Users | Available Expertise | Compliance with standards & best practices | Plans Towards Addressing User Needs | Risk to data quality |
|---|---|---|---|---|---|---|---|---|
| Comparability/ Coherence | Comparability across geography, populations, and other relevant domains | 52 | ○ | ○ | ◒ | ○ | ◓ | M |
| | Comparability across time (including impacts of redesign) | 74 | ◒ | ◒ | ◒ | ○ | ◒ | H |
| | Coherence with other relevant statistics (including use of standard classifications, | 38 | ◓ | ◓ | ○ | ◓ | ○ | H |
| Timeliness/ Punctuality | Timeliness of release of main aggregrates | 68 | ◒ | ◒ | ○ | ◒ | ○ | H |
| | Timeliness of release of detailed outputs (including microdata) | 68 | ◒ | ◒ | ○ | ◒ | ○ | M |
| | Punctuality | 62 | ◒ | ◒ | ○ | ○ | ○ | H |
| | Total for User Dimensions | 60,4 | | | | | | |

# 5. RECOMMENDATIONS

## 5.1 PROGRESS ON ROUND 1 RECOMMENDATIONS

The Round 1 recommendations were as follows. The current state of play as far as we can assess is shown in italics.

### *5.1.1 Need for 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 commencement of the integration of the surveys supporting the Services Production Index and the Industrial Production Index.

This recommendation remains valid. In particular, it is important that the work on the design of the CBF is completed in time for it to be part of the specifications for the redesigned Business Register.

### *5.1.2 Lack of Co-operation between the National Accounts and Statistical areas*

There is more to be done to improve the relationship but our judgment was that the relationship was good for each of the key input data sources that we considered. The work on the Memoranda of Understanding has continued since Round 1.

It could be said that this recommendation is well on the way to being implemented.

### *5.1.3 Evaluating the Accuracy of NACE Coding*

We were pleased to see that an evaluation study had been undertaken of the accuracy of NACE coding by registered enterprises. However, we have 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 suggest that the Methodology group be asked to assist in the design of a new coding study that uses independent coding. It should also be designed to obtain estimates of coding errors at the different levels of the hierarchy of the NACE coding system.

### *5.1.4 Need for Additional Evaluation Studies*

We are pleased that some new evaluation studies have taken place. 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.

### *5.1.5 Increasing Nonresponse Rates in Household Surveys*

Response rates for household surveys continue to deteriorate which increases the risk of poor quality due to nonresponse bias. This emphasizes the importance of understanding the causes of nonresponse and how to address them. During our stay at Statistics Sweden, we received an

interesting presentation on the initiatives that are underway to study nonresponse issues in household surveys. This presentation provided a comprehensive examination of the multitude of potential causes of poor response rates. This underscores the importance of identifying the most important drivers of nonresponse, or more importantly, nonresponse bias so that a program can be design to effectively reduce the risks of bias.
It is too early for us to offer much advice on the proposed initiatives but we would like to make a few points. First, focusing solely on response rates can increase error risks for other error sources. For example, interviewers may be more apt to falsify all or part of an interview or to accept poor quality responses for the sake of completing the interview. Also, in some cases, increasing the response rate has been shown to increase the nonresponse bias by further increasing the difference between characteristics of respondents and nonrepondents.

Second, call scheduling, particularly with regard to evening and weekend calling, is not optimal at present. We believe the unwillingness of some interviewers to work at these odd hours has greatly contributed to the high level on nonresponse due to noncontact – presently about 50% of the nonresponse rate. Third, there seemed to be a defeatist attitude among the field management team in that they were very pessimistic about being able to improve response rates. This attitude is likely to be passed on to interviewers and can become a self-fulfilling prophecy. Fourth, more effort should be put into measuring the representativeness of the sample, and adjustments for nonrepresentativeness, rather than just rely on levels of nonresponse rates. It may well be that the nonresponse bias problem is not as great as perceived by many users.

High nonresponse rates are likely to remain even with the implementation of the highest priority initiatives from the nonresponse study. In addressing the nonresponse issue, it is important to also address the issue of representative samples and to introduce measures of potential nonresponse bias which are more sophisticated than the simple proxy measures of nonresponse rates for refusals and noncontacts. It is also important to focus on the nonresponse bias in the *adjusted* estimates. Current approaches for adjusting for nonresponse bias could be quite effective at reducing the risks of nonresponse bias even in the presence of falling response rates but not much work has been done to verify this.

#### *5.1.6 Improving the Relationship with the Tax Office*

We were also pleased to note that a decision to amend the standardized tax forms from enterprises was reversed following representations from Statistics Sweden. This shows strength in the relationship although possibly consultation on the form amendment could have taken place earlier. We still think the development of a Memorandum of Understanding worthwhile particularly if the relevant part of the Tax Office moves out of Stockholm as we understand is currently proposed. This should prescribe the minimum number of meetings to be held each year to consider matters of mutual interest.

#### *5.1.7 Establishing a Policy on Continuity of Statistical Series*

We understand there is no policy yet but we think a policy would be worthwhile. As stated last year, 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. In practice this happens to a large extent but there have been some important exceptions.

### *5.1.8 Improving the Relationship between IT and their Client Areas*

There is still not a strong relationship and we sensed frustration with several of the product areas with whom we spoke. We did not speak to the IT area but they are probably also frustrated by the relationships. We were advised that the IT department had recently appointed contact persons with each of the other Departments. This is a step in the right direction. However, it might be worth considering engaging an external expert to study the existing relationship to assess what organizational and other changes are needed to improve the relationship. There has been a strong relationship between the statistical and IT areas at the ABS and Statistics New Zealand. We suggest that Statistics Sweden understand the characteristics of these organizations that have fostered such high levels of satisfaction with IT.

### *5.1.9 Lack of Telephone Interviewing Monitoring*

We were very pleased to see the introduction of telephone monitoring for the LFS. We did have concerns that the interviewers were pre-warned that there was a 50% chance that designated interviews were to be monitored. This may mean their behavior is 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. 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 as well as 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.

### *5.1.10 Development of Quality Profiles for Key Products*

QDs exist for all the products we examined except the National Accounts which utilizes the material they provide to Eurostat in the form of GNI Inventories in lieu of a specific quality declaration document. Furthermore, there have been improvements in all the quality declarations for the products we reviewed last year. Nevertheless there is scope for improvement in most of the QDs and the most important are mentioned in the reports on the individual products. For most, 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.

## 5.2 NEW RECOMMENDATIONS

As can be seen from section 5.1, there has been progress against most of our recommendations of last year. 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 Increase the Focus on Coherence between Relatable Statistics*

We only directly looked at Coherence for LFS but it came up in discussions elsewhere. There was not as much focus as we expected by statistical product areas on the coherence with relatable statistics including the National Accounts. It was also an issue that arose last year with our discussions with users. Although we did not study it directly, we suspect it is not an issue that is mainly due to the use of nonstandard classifications and definitions. Rather, it most likely due to differences in methods and their error properties. In our view, it should be made clear that the statistical product areas are responsible for coherence with related statistics. The main coherence relationships should be specified and the reasons for the lack of coherence studied jointly by the two areas. Reconciliation tables are often a useful device for understanding and specifying the differences. They can also be a valuable device for explaining the differences to users.

After the reasons for the differences are understood, decisions can be made as to whether it is desirable to make any changes to the methods to improve the coherence.

### *5.2.2 Initiate succession planning in some important statistical areas.*

The two statistical product areas where this is of most concern are CPI and National Accounts where a number of very experienced, capable statisticians have retired or are soon to retire. The process for identifying suitable replacements should begin now. We believe the focus should be on moving statisticians with strong technical capabilities (not necessarily price indexes and National Accounts) into these areas. They will pick up the necessary skills quite quickly with the right support and they are much more likely to stay at Statistics Sweden than younger statisticians at an early stage of their career. The retiring statisticians may be able to help with the transition.

# 6. SUMMARY AND CONCLUSIONS

As we stated last year, we believe Stat Sweden remains a world class organisation. In all the products we evaluated for the second time we saw improvements with very few deteriorations. We also saw improvement in documentation particularly through the efforts on quality declarations. Nevertheless there have been a number of areas requiring improvement and these have been identified in this product.

We have reviewed the Accuracy of seven products for the second time. As a result of better information available this time we have corrected the ratings. In the report, we have distinguished the corrections from improvements and Exhibit 4a shows the current ratings, prior year ratings, and the improvements by product. Generally, all products reviewed in Round 1 improved. The average improvement was 4.6 percentage points which is commendable. Because of the corrections, the base ratings will be different to those in our 2012 report. Also, we have changed the risk levels for some error sources as a result of more information. As ratings are weighted according to the level of risk this will also result in changes to the Round 1 ratings.

We reviewed the Accuracy of the National Accounts again but because we used a completely different methodology this year, it is not possible to identify improvements since Round 1. 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 and established a new baseline for each. The National Accounts ratings are summarised in Exhibit 4b.

We reviewed the Accuracy of the ULF/SILC for the first time thereby establishing a baseline for this product. The Accuracy rating for this product is much lower than any product reviewed in either round which suggests that the risks to data quality for this survey is higher than any other product we have ever reviewed. The objectives for the survey are somewhat confused at present, resulting in a very complex and possibly unmanageable design. We are also concerned about the present practice of interviewing children by telephone which is fraught with both data quality and ethical issues. The primary objectives of the survey need to be sorted out so the survey can be redesigned and optimized accordingly. One possibility is to base a redesign on the SILC incorporating the capacity to include topics of special interest to Sweden (for example, using supplementary topical modules). It may also be desirable to retain key aspects of the ULF (for example, some of the longitudinal components) as a research study. However, the most serious problems for the ULF/SILC (for e.g., the sample design, questionnaire design, and data collection methodologies) will require a substantial redesign. These issues are not intractable and can be addressed adequately, given resources and staff availability. Unfortunately, the issues with low response rates will continue as long as the problems for telephone interviewing that were described above for the LFS persist.

We also pilot tested an approach to reviewing the other four Quality Dimensions (other than Accuracy). These so called user dimensions include Relevance/Contents and Accessibility & Clarity which were reviewed for the CPI, and Timeliness & Punctuality and Comparability & Coherence which were reviewed for the LFS.

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 (2012). The recommendations require consideration by the Executive rather than the individual product areas. Most will require some allocation of funding so there may need to be priority decisions made by the Executive.

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 improvements having the greatest opportunity to succeed.

As can be seen from the foregoing, we made a number of changes to our methodology for the Accuracy reviews based on our experience with the Round 1 reviews. One important improvement was to create a checklist 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.

Although they worked well, we did identify some areas for further improvements in the checklists but the changes will be at the margin rather than to the basic approach.

We also created templates for the pilot reviews of the user dimensions. These worked reasonably well but require some adjustments in light of experience. We can assist with these adjustments if it is decided to continue with this approach.

There should be further rounds of quality reviews but there may be advantages in spreading them over more time rather than holding them over such an intense period. Ideally, documenting the results of the evaluation interview should occur on the same day as the interview, if possible, to ensure that important details and nuances revealed in the interview are accurately captured. Scheduling the interviews over a longer time period would facilitate this. For the most important products, an annual review is probably appropriate but less frequent reviews are sufficient for other products. For example, we do not believe it is necessary to review RS on an annual basis. On the other hand, if a product is experiencing significant difficulties (such as the ULF/ SILC), it might be worth reviewing it each year until it is back on track. Furthermore, scheduling the bulk of the review work in December is not ideal because it conflicts with the holiday season. This is not only a considerable distraction for the evaluation process, but it can lead to untimely documentation of the findings and finalization of the report.

The other decision is who should undertake the reviews – external or internal reviewers. External reviewers might be relatively expensive and should only be used for selected reviews. Furthermore, it might be worth establishing a panel of reviewers. It is important to cover both expertise in data collection methodology and the subject matter of the products being reviewed in the review team. Internal reviews should be possible with the development of checklists. These would support self-assessments but self-assessments should still be facilitated to provide support to the product areas as well as ensuing consistency across products. The user dimensions seem particularly suitable for internal review.

Finally we would like to thank Stat Sweden for enabling us to work on this important and interesting project.

# 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, Report to Statistics Sweden.

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

# ANNEX 1 –QUALITY CRITERIA, GUIDELINES AND CHECKLISTS FOR ALL DIMENSIONS OF QUALITY

## *1.1 Quality Criteria, Guidelines and Checklist for Accuracy Applied to Each Error Source, version 2012*

**Exhibit 1.1a. Knowledge of Risks**

| Poor [1,2] | Fair [3,4] | Good [5,6] | Very Good [7,8] | Excellent [9,10] |
|---|---|---|---|---|
| Program documentation does not acknowledge the source of error as a potential factor for product accuracy. | Program documentation acknowledges error source as a potential factor in data quality.<br><br>**But:** No or very little work has been done to assess these risks. | Some work has been done to assess the potential impact of the error source on data quality.<br><br>**But:** Evaluations have only considered proxy measures (example, error rates) of the impact with no evaluations of MSE (bias and variance) components. | Studies have estimated relevant MSE components associated with the error source and are well-documented.<br><br>**But:** Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses. | There is an ongoing program of research to evaluate all the relevant MSE components associated with the error source and their implications for data analysis. The program is well-designed and appropriately focused, and provides the information required to address the risks from this error source. |

**Exhibit 1.1b. Communication with Users**

| Poor [1,2] | Fair [3,4] | Good [5,6] | Very Good [7,8] | Excellent [9,10] |
|---|---|---|---|---|
| Reports, websites, and other communications with data users and customers are devoid of any mention of the error source. | There is acknowledgement of the risks of error from this source.<br><br>**But:** Communications have been largely inadequate considering the importance of these potential risks to data quality. | Communications with users and customers have adequately described the risk to many users.<br><br>**But**: Information conveyed has largely been sampling errors and/or proxy measures with little communications regarding MSE components or the risks have been downplayed leading to a false sense of security. | Communications have shared some of the available information on the relevant MSE components that have been evaluated and the true risks to users have been appropriately conveyed.<br><br>**But:** The information conveyed in could be improved in one or more of these areas: (a) more clarity so that complex ideas are comprehensible to less sophisticated users, (b) improved presentation so data analysts can apply the knowledge more directly in their analyses, or (c) a fuller discussion of the implications of the findings for various types of data analysis so that users are can make informed decisions regarding the results. | Communications regarding the error source have been thorough, cogent, and clear. An appropriate level of detail has been included in the communications so that users should be fully aware of any risks of the error source to data quality and are provided with all the information they need to deal with the risks appropriately in their analyses. |

**Exhibit 1.1c. Available Expertise**

| Poor [1,2] | Fair [3,4] | Good [5,6] | Very Good [7,8] | Excellent [9,10] |
|---|---|---|---|---|
| Among the staff assigned to work on the product, either (a) there are no staff that are familiar with techniques that will be required to deal with the potential risks to accuracy for the product or (b) the expertise of staff that are assigned is sorely inadequate. | The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in at least one important area.<br><br>**But:** For most important areas expertise is still lacking. | The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in most of the important areas.<br><br>**But:** Either (a) there is at least one area that may be critical to accuracy where a higher level of expertise is needed or (b) there are one or more minor areas that could become important in the future that are not well staffed. | The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in all important areas. There is a good working relationship with the key groups involved in activities associated with this error source. Staff are keeping up to date with developments in their areas of expertise<br><br>**But:** There are one or more minor areas that could become important in the future which are not well covered. Current expertise is not adequate to achieve the highest ratings for all evaluation criteria for this error source or the expertise would not be readily available to work on these error sources. | The available expertise required to study this error source and communicate the findings of such studies to data users is more than adequate to achieve the high ratings across all evaluation criteria The relevant experts are actively addressing errors from the source. There is an excellent working relationship with the key groups involved in activities associated with this error source. Staff are keeping up to date with and contributing to developments in their areas of expertise. |

| Exhibit 1.1d. Compliance with Standards and Best Practices | | | | |
|---|---|---|---|---|
| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ○ |
| Staff are mainly unaware of standards and best practices that are relevant for this error source. If some awareness exists, there is no evidence that standards and best practices, as they related to this error source, have been applied to the product. Moreover, serious deficiencies exist that violate standards and best practices as they relate to this error source. | Staff are aware of standards and best practices and there is evidence that these have been applied to the product for this error source.<br><br>**But:** There are still important areas of noncompliance that need to be addressed. These gaps are not currently being addressed or actions to address them have been inadequate. | Staff are well aware of relevant standards and best practices and have clearly been applied them to the product. Important violations or gaps are being actively addressed.<br><br>**But:** Either (a) compliance is not routinely monitored or (b) gaps in compliance exist for some minor areas that are not being addressed. | Staff are well aware of the relevant standards and best practices and have clearly applied them to the product. There are no serious violations of standards and best practices as they relate to this error source<br><br>**But:** Some staff may not keep up to date with latest standards and developments in best practices that are relevant to their work. Compliance may not be routinely monitored. | The product is fully compliant with agreed standards and best practice. The relevant staff are fully aware of the standards and best practices and continually monitor the work to ensure that compliance is maintained. They are actively keeping up to date with and contributing to latest standards and developments in best practices. |
| **Exhibit 1.1e. Achievement Towards Mitigation and/or Improvement Plans** | | | | |
| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ○ |
| There is no evidence that any planning has been done for studying or mitigating the risks for this error source. | An overall plan for error reduction with measurable objectives exists for mitigating the risks for this error source.<br><br>**But:** The plan is not approved by the appropriate level of management. . | A management-approved plan with measurable objectives exists. The plan adequately addresses the work required for mitigating the risks of poor data quality relative to this error source...<br><br>**But:** One of the following deficiencies with the plan exists: a. The overall plan has not been updated in at least one year. b. There is no accountability in place to ensure compliance with the plan. c. No mechanism is specified for gauging progress toward each objective.<br><br>d. No resources have been allocated to implement the plan. | Resources have been allocated to undertake this work. Considerable progress has been made on the plan for mitigating the risks to data. . None of the deficiencies noted under the "Good" criteria are present.<br><br>**But:** Efforts have not yet produced the desired control over the error source that is stipulated in the plan. | Mitigation plans have been fully implemented or well underway. Progress toward all goals and objectives has been excellent. As a result, the level of error in the final estimates due to this error source is being maintained at an acceptable level for the primary purposes of the data. As a result of these efforts, the error source is under control and poses no or very little risk to data quality. Results of the mitigation activities have been fully documented.<br><br>Accountability measures are in place to ensure compliance with the plans. The mitigation plans are reviewed and updated periodically. |

**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<br>☐ No<br>**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<br>☐ No<br>**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<br>☐ No<br>**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<br>☐ No<br>**Very Good** | |
| 5. Existing documentation on the error source is of high quality and explores the implications of errors on data analysis. | ☐ Yes<br>☐ No<br>**Excellent** | |
| 6. There is an ongoing program of research to evaluate the components of the MSE that are relevant for this error source. | ☐ Yes<br>☐ No<br>**Excellent** | |

| Communication | Check Box | Comments |
|---|---|---|
| 1. Users have been informed of the risks from this error source to data quality through reports, websites and other formal means. | ☐ Yes ☐ No **Fair** | |
| 2. These communications have explained the risks in terms of the potential degradation to overall accuracy of the estimates | ☐ Yes ☐ No **Good** | |
| 3. The potential impact on users has 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. | ☐ Yes ☐ No **Good** | |
| 4. Documentation speaks clearly, comprehensively, and with appropriate detail on the size of the MSE components for the target audience. | ☐ Yes ☐ No **Very Good** | |
| 5. The information on data quality conveyed in the communications is sufficiently detailed that less sophisticated data users should be able to know and understand their implications for most uses of the data. | ☐ Yes ☐ No **Excellent** | |
| 6. 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** | |

| 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<br>No<br>**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<br>No<br>**Good** | |
| 3. At least some members of the product staff are adept at communicating risks for this error source to the product area and/or data users clearly and concisely. | Yes<br>No<br>**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<br>No<br>**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 Statistics Sweden statistics. | Yes<br>No<br>**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<br>No<br>**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<br>No<br>**Fair** | |
| 2. Key staff members are aware of best practices in the field that apply as they pertain to this error source | Yes<br>No<br>**Fair** | |
| 3. Current activities for controlling or minimizing data quality risks from this error source comply with all appropriate standards. | Yes<br>No<br>**Good** | |
| 4. There are no serious violations of standards and best practices as they relate to this error source. | Yes<br>No<br>**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<br>No<br>**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<br>No<br>**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<br>☐ No<br>**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<br>☐ No<br>**Fair** | |
| 3. The written plan has been approved by management. | ☐ Yes<br>☐ No<br>**Good** | |
| 4. Progress toward achieving the goals of the risk mitigation plan is regularly reviewed and compliance with the plan is appropriately monitored. The plan is updated appropriately as work progresses and new knowledge is gained regarding the error source. | ☐ Yes<br>☐ No<br>**Very Good** | |
| 5. Mitigation plans have been fully implemented or well underway. Information has been provided to users regarding progress toward risk mitigation. | ☐ Yes<br>☐ No<br>**Excellent** | |
| 6. Quality improvement strategies that have been implemented have been successful at minimizing the risk to data quality from this error source. | ☐ Yes<br>☐ No<br>**Excellent** | |

## *1.2 Quality Criteria, Guidelines and Checklist for User Quality Dimensions Applied to Each Component, version 2012*

**Exhibit 1.2a. Knowledge of User Needs**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ○ |
|---|---|---|---|---|
| Program documentation does not (a) acknowledge the needs of users with regard to this component or (b) provide evidence of the extent to which user needs are being met. | Program documentation acknowledges specific needs that users have regarding this component of this dimension.<br><br>**But:** No or very little work has been done to quantify these needs or to assess user satisfaction with the status quo. | Some work has been done to quantify and document the needs of users and their satisfaction with the status quo.<br><br>**But:** Assessments have only considered some users – for example, major ("power") users of the data. No or very little work has been done to assess the needs or satisfaction of the wider user community | Assessments of needs and satisfaction of the wider user community have been conducted and their results are well-documented.<br><br>**But:** However, these assessments have been limited in scope. For example, the assessments have explored on some uses of the data. In addition the assessments have not gone beyond user feedback mechanisms. For example, staff are unaware of how users are working with the data through reports and publications and how these applications could be improved by work on this component. | There is an ongoing program to assess the needs of the vast majority of the user community. In addition to obtaining direct feedback from users, the user needs assessments have considered how the data from the product is being used and whether problems exist in these uses that can be addressed through improvements of this component. |

**Exhibit 1.2b. Communication with Users**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ○ |
|---|---|---|---|---|
| There is no evidence of two-way, interactive communications with data users and customers regarding this component. | There is evidence of some communications with users.<br><br>**But:** Communications have been largely inadequate considering the diversity of the user community and the importance of this component. For example, communication may have been largely passive relying on users to search websites for information | Communications with users has been active and electronic media have been used to alert some users of new developments.<br><br>**But**: Information conveyed either (a) has not addressed the needs of the wider user community, (b) has not been timely or has occurred too infrequently, or (c) contains important deficiencies that need to be addressed. | Communications with users regarding this component have been comprehensive and have adequately addressed the diversity of user community.<br><br>**But:** The information conveyed in could be improved in one or more of these areas: (a) more clarity so that complex ideas are comprehensible to less sophisticated users, (b) improved presentation so users can apply the information more readily in their work, or (c) greater currency to better reflect recent changes. | Communications regarding the component have been thorough, cogent, and clear. An appropriate level of detail has been included in the communications so that users needs have been fully met. In addition, mechanisms are in place to obtain regular feedback from the user community regarding their current needs and how they might change in the future. Further, there is a process in place to address current and future needs in a timely, effective, and efficient manner. For example, an active network of key users is in place and consider this component on a regular basis. |

**Exhibit 1.2c Available Expertise**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ○ |
|---|---|---|---|---|
| Among the staff assigned to work on the product, either (a) there are no staff that are familiar with techniques that will be required to deal with this component or (b) the expertise of staff that are assigned is sorely inadequate. | The available expertise required to address the issues surrounding this component and communicate with users is adequate in some areas.<br><br>**But:** For many important areas expertise is still lacking. | The available expertise required to address the issues surrounding this component and communicate with users is adequate in most of the important areas.<br><br>**But:** Either (a) there is at least one critical area a higher level of expertise is needed or (b) there are one or more minor areas that could become important in the future that are not well staffed. | The available expertise required to address the issues surrounding this component and communicate with users is adequate in all important areas. There is a good working relationship with the key groups involved in activities associated with this component. Staff are keeping up to date with new developments in the field as they relate to this component.<br><br>**But:** There are one or more minor areas that could become important in the future which are not well covered. Current expertise is not adequate to achieve the highest ratings for all evaluation criteria for this component or the expertise would not be readily available to work on this component. | The available expertise required to address the issues surrounding this component and communicate with users is more than adequate to achieve the high ratings across all evaluation criteria The relevant experts are actively addressing this component. There is an excellent working relationship with the key groups involved in activities associated with this component and staff members are keeping up to date with and are contributing to developments in their areas of expertise. |

| Exhibit 1.2d Compliance with Standards and Best Practices | | | | |
|---|---|---|---|---|
| Poor [1,2] | Fair [3,4] | Good [5,6] | Very Good [7,8] | Excellent [9,10] |
| Staff are mainly unaware of standards and best practices that are relevant for this component. If some awareness exists, there is no evidence that standards and best practices, as they related to this component, have been applied to the product. Moreover, serious deficiencies exist that violate standards and best practices as they relate to this component. | Staff are aware of standards and best practices and there is at least some evidence that these have been applied to the product for this component.<br><br>**But:** There are still important areas of noncompliance that need to be addressed. These gaps are not currently being addressed or actions to address them have been inadequate. | Staff are well aware of relevant standards and best practices and have clearly been applied them to the product. Important violations or gaps are being actively addressed.<br><br>**But:** Either (a) compliance is not routinely monitored or (b) gaps in compliance exist for some minor areas that are not being addressed. | Staff are well aware of the relevant standards and best practices and have clearly applied them to the product. There are no serious violations of standards and best practices as they relate to this component.<br><br>**But:** Some staff may not keep up to date with latest standards and developments in best practices that are relevant to their work. Compliance may not be routinely monitored. | The product is fully compliant with agreed standards and best practice. The relevant staff are fully aware of the standards and best practices and continually monitor the work to ensure that compliance is maintained. They are actively keeping up to date with and contributing to latest standards and to best practices as they relate to this component. |

| Exhibit 1.2e Plans for addressing user needs | | | | |
|---|---|---|---|---|
| Poor [1,2] | Fair [3,4] | Good [5,6] | Very Good [7,8] | Excellent [9,10] |
| There is no evidence that any planning has been done for studying or addressing this issues for this component. | An overall plan exists for addressing the issues for this component. The plans are well-documented and include measurable objectives for addressing the issues for this component.<br><br>**But:** The plan has not been approved by the appropriate level of management. | A management-approved plan with measurable objectives exists. The plan adequately addresses the work required for satisfactorily addressing the issues relative to this component.<br><br>**But:** One of the following deficiencies with the plan exists: a. The overall plan has not been updated in at least one year.<br><br>b. There is no accountability in place to ensure compliance with the plan.<br><br>c. No mechanism is specified for gauging progress toward each objective.<br><br>d. No resources have been allocated to implement the plan. | Resources have been allocated to undertake this work. Considerable progress has been made on the plan. None of the deficiencies noted under the "Good" criteria are present.<br><br>**But:** Efforts have not yet produced the desired level of quality for the component that is stipulated in the plan. | Each activity specified in the written plan has been fully implemented or well underway. Progress toward all goals and objectives has been excellent. As a result of these efforts, the issues associated with this component have all be sufficiently resolved. There is ample evidence that all relevant user groups are quite satisfied with this component. This evidence as well as the results of the quality improvement activities have been fully documented.<br><br>Accountability measures are in place to ensure compliance with the plans. The quality improvement plans are reviewed and updated periodically. |

**Checklist for User Quality Dimensions.** *For each applicable component of a user dimension, 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 explain your response to an item if you wish.*

| **Knowledge of Risks** | **Check Box** | **Comments** |
|---|---|---|
| 1. Documentation exists that acknowledges the specific needs of users with respect to this component. | ☐ Yes<br>☐ No<br>**Fair** | |
| 2. The extent to which user needs are being met has been formally assessed (e.g.,through user satisfaction surveys), quantified, and documented. | ☐ Yes<br>☐ No<br>**Good** | |
| 3. User assessments have been made of a broad cross-section of the user community, not just the "power" or high profile users. | ☐ Yes<br>☐ No<br>**Very Good** | |
| 4. The key results of all actions taken to understanding user needs regarding this component have been well-documented and disseminated appropriately throughout Statistics Sweden. | ☐ Yes<br>☐ No<br>**Very Good** | |
| 5. There is an ongoing program to assess the needs of the vast majority of the user community. | ☐ Yes<br>☐ No<br>**Excellent** | |
| 6. User needs assessments have considered the implications of this component on data analysis, policy analysis, or other user applications and whether there are problems with these uses that can be addressed through improvements of the component. | ☐ Yes<br>☐ No<br>**Excellent** | |

| Communication with Users | Check Box | Comments |
|---|---|---|
| 1. There is some evidence of communication with users regarding this component. | Yes<br>No<br>**Fair** | |
| 2. Communications have broadly covered the whole user community. | Yes<br>No<br>**Good** | |
| 3. Communications with users regarding this component have been interactive using electronic media as well as personal communications (focus groups, etc.) to convey important information and respond to queries about the component. | Yes<br>No<br>**Good** | |
| 4. Information disseminated to users regarding this component have been clear, cogent, and useable. | Yes<br>No<br>**Very Good** | |
| 5. Communications with users have occurred with sufficient frequency to satisfy the needs of users and to reflect the changes as they relate to this component. | Yes<br>No<br>**Excellent** | |
| 6. Mechanisms (for e.g., user communications networks, blogs, etc.) are in place to obtain regular feedback from the user community with regard to this component. | Yes<br>No<br>**Excellent** | |

| Available Expertise | Check Box | Comments |
|---|---|---|
| 1. There is at least one staff member assigned to this product who is familiar with techniques required to deal with improvements as they relate to this component. | Yes<br>No<br>**Fair** | |
| 2. Staff who are familiar with the required techniques have adequate knowledge to address the important issues surrounding this component. | Yes<br>No<br>**Good** | |
| 3. There are no important areas where a higher level of expertise for this component is needed but is not available within the existing product staff. | Yes<br>No<br>**Very Good** | |
| 4. There is a good working relationship with the key groups within Statistics Sweden involved in activities associated with this component. | Yes<br>No<br>**Very Good** | |
| 5. Staff are keeping up to date with new developments in the field as they relate to this component. | Yes<br>No<br>**Very Good** | |
| 6. There are no minor areas that could become important in the future with regard to this component where expertise is lacking. | Yes<br>No<br>**Excellent** | |
| 7. Current expertise is adequate and readily available to achieve the highest ratings for this component. | Yes<br>No<br>**Excellent** | |
| 8. The relevant experts that are actively addressing this component are contributing to developments in this area. | Yes<br>No<br>**Excellent** | |

| **Compliance with Standards and Best Practices** | **Check Box** | **Comments** |
|---|---|---|
| 1. Staff are generally aware of standards as they apply to this component. | Yes<br>No<br>**Fair** | |
| 2. Staff are generally aware of best practices (e.g., the literature concerning this component) that are relevant for this component. | Yes<br>No<br>**Good** | |
| 3. There is some evidence that standards and best practices, as they related to this component, have been applied to the product. | Yes<br>No<br>**Good** | |
| 4. There are no serious deficiencies that violate either the applicable standards or current best practices as they relate to this component nor are there important areas of noncompliance that need to be addressed. Any gaps that exist are minor, are being addressed, and the actions to address them have been adequate. | Yes<br>No<br>**Very Good** | |
| 5. Compliance with standards and best practices as they relate to this component is routinely monitored. | Yes<br>No<br>**Excellent** | |
| 6. With regard to this component, the product is fully compliant with agreed standards and best practice. | Yes<br>No<br>**Excellent** | |
| 7. The relevant staff are actively keeping up to date with and contributing to latest standards and to best practices as they relate to this component. | Yes<br>No<br>**Excellent** | |

| Achievement towards Improvement Plans | Check Box | Comments |
|---|---|---|
| 1. An overall plan exists for addressing the issues for this component. | Yes<br>No<br>**Fair** | |
| 2. A well-documented plan with measurable objectives has been approved by the appropriate level of management and work has begun to implement it. | Yes<br>No<br>**Fair** | |
| 3. The plan adequately addresses the work required for addressing the key issues related to this component. | Yes<br>No<br>**Fair** | |
| 4. The approved plan has either been written or updated within the previous 12 months. | Yes<br>No<br>**Good** | |
| 5. Accountability measures are in place to ensure compliance with the plan and an effective mechanism is in place for gauging progress toward each objective of the plan. | Yes<br>No<br>**Very Good** | |
| 6. Adequate resources have been allocated to implement the plan as it relates to this component. | Yes<br>No<br>**Very Good** | |
| 7. Each activity specified in the plan has either been fully implemented or is well underway. | Yes<br>No<br>**Very Good** | |
| 8. Progress toward all goals and objectives stated in the plan as it relates to this component has been excellent. | Yes<br>No<br>**Excellent** | |
| 9. The issues associated with this component have all be sufficiently resolved and there is sufficient evidence that all relevant user groups are quite satisfied with this component. This evidence as well as the results of the quality improvement activities has been fully documented. | Yes<br>No<br>**Excellent** | |

***Exhibit 2.1 RS Change Ratings between Round 1 and Round 2***

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | *Correction from 2011 rating*<br>*Improvement from 2011 rating*<br>**Comments on changes** |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 74 | N/A | 7 →N/A | 7→N/A | 9 →N/A | 7 →N/A | 7 →N/A | M →N/A[1] | [1]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 | 43 | 60 | 4→5[1] | 1→5[2] | 7 | 5→7[3] | ~~1~~ N/A[4] | L | [1]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.<br>[2]Risk is communicated well in the QD.<br>[3]This error source complies with standards and there is no serious violations of best practices.<br>[4]Given the low risk of frame error, further planning to mitigate risks is unnecessary and is not applicable. |
| Non-response error | 52 | 52 | 5 | 5 | 7 | ~~5~~ 4[1] | ~~9~~ 5[2] | M | [1]Nonresponse is primarily item nonresponse in the sections on educational activities and care/social work in the summary accounts. No study has been done to quantify this risk. This was also true last year so the rating was corrected to reflect this departure from standards and best practices.<br>[2]Likewise, there is no plan for conducting a study of item nonresponse in these areas. However, the rating was elevated to Good to reflect work that has begun to reduce nonresponse bias using editing and the weekly meetings that are held to ensure consistency. |
| Measurement error | 52 | 58 | 3→5[1] | 4→5[2] | 7 | 7 | 5 | ~~H~~ M[3] | [1]Cognitive laboratory evaluation substantially enhanced knowledge of measurement error.<br>[2]Discussion of issues of measurement in the QD was substantially improved.<br>[3]Risk level was corrected to Medium. Any improvements did not change the rating from last year. |
| Data processing error | 46 | 48 | 3→4[1] | 3 | 7 | 5 | 5 [2] | ~~M~~ H[3] | [1]Knowledge of editing error was enhanced somewhat as a result of changes made to the editing process.<br>[2]The value chain analysis that is in process will lift the level of planning to 'very good' in the coming year if some results can be produced that it is having a positive effect on data quality.<br>[3]Risk level was corrected to High to reflect the importance of the error source to data quality. |
| Sampling error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | |
| Model/estimation error | 38 | 38 | ~~5~~ 3[1] | ~~5~~ 3[2] | 7 | 3 | ~~7~~ 3[3] | ~~M~~ H[4] | [1]Most of the modeling is performed by respondents to allocate activity costs and common costs. However there is little knowledge of the errors associated with this process. SCB has a model for allocating common costs that can be used by governments; however, it performance quality is unknown.<br>[2]Users are not well-informed regarding the error risks associated with detailed level data.<br>[3]There is currently no plan approved by management to look into these modeling errors.<br>[4]The risk level was reassessed and changed to High given the importance of modeling to the survey. |
| Revision error | 58 | 58 | ~~7~~ 5[1,2] | ~~7~~ 3[1] | ~~9~~ 8[1] | 7 | ~~7~~ N/A[3] | L | [1]The ratings for Knowledge, communication, and expertise were too high last year based on new knowledge acquired this year. [2]There is no discussion of revision error in the QD and not much is known about it and its affects on NA.<br>[3]Planning to mitigate the risks from this error source appears not to apply given its low importance. |
| **Total score** | **46,7** | **49,6** | | | | | | | |

*Exhibit 2.2 CPI Change Ratings between Round 1 and Round 2*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | *Correction from 2011 rating*<br>*Improvement from 2011 rating*<br>**Comments on changes** |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 68 | 68 | 8 | ~~7~~6[1] | 9 | ~~5~~7[2] | ~~5~~4[3] | **H** | [1]Corrected because the size of the impact has not been conveyed to users.<br>[2]Corrected as there is not a written plan to introduce scanner data which has been approved by management.<br>[3]Corrected as actual practices comply closely with European standards. |
| Frame error | 62 | 62 | ~~4~~7[1] | ~~3~~7[2] | 5 | ~~6~~7[3] | ~~3~~5[4] | **M** | [1]Corrected because of existence of 1999 study which could be repeated in the near future.<br>[2]Corrected as a written plan for improving survey exists.<br>[3]Corrected because of the availability of the 2001 CPI Handbook.<br>[4]Correctedbecause there seem to be no serious violations with European standards with respect to CPI frames. |
| Non-response error | 55 | 55 | ~~1~~3[1] | ~~1~~3[2] | 9 | ~~3~~7[3] | ~~4~~N/A[4] | **L** | [1]Corrected because there is more knowledge than we previously thought.<br>[2]Corrected because we now know more about the steps taken to ensure weights derived from HBS are robust and comply with standards.<br>[3]We have greater understanding of HBS and agree that risk of bias from this source is low. Robustness studies might be worthwhile.<br>[4]Corrected because there is no need for a plan when it is low risk. |
| Measurement error | 58 | 62 | ~~4~~7[1] | ~~4~~5[2] | ~~5~~9[3] | ~~4~~5[4] | 3→5 | **H** | [1]Knowledge corrected because of existence of 1999 study.<br>[2]Communication corrected because we became more aware of the activities with the power user groups.<br>[3]Compliance to best practice corrected because we became aware of the extent of international activity on influencing best practice and because there seems to be conformance with European quality standard.<br>[4]Improvement for mitigation of risk because of plans to introduce scanner data as well as the chain linking procedure for quality change. |
| Data processing error | 70 | 76 | 6→7[1] | 6 | 9 | 6→8[2] | 8 | **M→H** | [1]Improved rating beacause of the new processing system and introduction of shadow system.<br>[2]Improved rating because of the bedding down of the new processing system. |
| Sampling error | 54 | 66 | 5→7[1] | 3→7[1] | 9 | 6 | 4 | **H** | [1]Improved ratings because of the updating of the sample error study. |
| Model/estimation error | 52 | 52 | ~~7~~5[1] | ~~7~~5[2] | 6 | ~~6~~4[3] | 6 | **H** | [1]Knowledge of risks corrected because of better understanding of the models that are used and because quantative studies are somewhat out of date.<br>[2]Communication corrected because checklist suggests communication to users was not as strong as previously thought.<br>[3]Compliance to standards corrected because of better understanding of the standard that is relevant. |
| Revision error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | |
| Total for Accuracy | 60,3 | 63,9 | | | | | | | |

***Exhibit 2.3 FTG Change Ratings between Round 1 and Round 2***

*Correction from 2011 rating*
*Improvement from 2011 rating*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | Comments on changes |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 58 | 58 | 5 | ~~7~~ 5[1] | 7 | 7 | 5 | M | [1]Under the current guidelines, communication should have been "Good" last year, not "Very Good." |
| Frame error | 58 | 58 | ~~7~~ 5[1] | 5 | 7 | 5 | 7 | ~~M~~ L[2] | [1]Corrects error in last years rating for Knowledge of Risks.<br>[2]Also, corrects risk level based upon intrinsic risk of frame error being low. |
| Non-response error | 62 | 66 | 7 | 5→7[1] | 7 | 5 | 7 | M | [1]Communication to users about nonresponse improved as a result of the QD. |
| Measurement error | 54 | 62 | 5→7[1] | 5 | 5→7[2] | 7 | 5 | H | [1]Knowledge of risks gained through writing the QD as well as preparation of the annexes to the SLA with the NA.<br>[2]Working relationship and closer cooperation between the collection unit and the methods group as a result of the SLA. |
| Data processing error | 46 | 60 | 5→7[1] | 5→7[2] | 5→7[3] | 3 | 5→6[4] | ~~M~~ H[5] | [1]Knowledge of risks gained through writing the QD as well as preparation of the documents "Improvements of the work on revisions in the Swedish good" and "Improving macro-editing in Intrastat."<br>[2]Likewise Communication has improved through both of the above mechanisms.<br>[3]Working relationship and closer cooperation between the collection unit and the methods group as a result of the SLA.<br>[4]Some planning is underway for further improvements of editing and coding. Planning and discussions are underway to reduce the misclassification of goods by enterprises.<br>[5]Risk level was re-evaluated and elevated to H based upon the importance of editing to data quality. |
| Sampling error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | |
| Model/estimation error | 66 | 80 | 7→8[1] | 5→7[2] | 7→9[3] | 7→9[4] | 7 | M | [1]Both Knowledge and Communication has improved evidenced by the document "Improvement of the distribution keys for the estimated trade in the Swedish Intrastat."<br>[2]Key staff have made national presentations in connection with the WG Quality Meetings elevating expertise.<br>[3]Swedish Customs adopted SCB's editing system which indicates state of the art systems.<br>[4]Plans are in place to study more sophisticate models for estimation under cut-off using VAT possibly using the Vat Information Exchange System (VIES). |
| Revision error | 62 | 76 | 5→7[1] | 5→7[1] | 7 | 7→9[2] | 7→8[3] | ~~L~~ H[4] | [1]Knowledge and communication of risks improved through writing the QD as well as preparation of the documents "Improvements of the work on revisions in the Swedish goods."<br>[2]Compliance with standards and best practices enhanced through Standardized Toolbox. Above referenced document also provides evidence that best practices are being followed. Progress has been made to rapidly detect and repair causes of large revisions.<br>[3]Plans being developed to identify causes of revision error.<br>[4]The risk level was re-evaluated and elevated to H as a result of the impact on the NA statistics. |
| **Total score** | 57,3 | 65,8 | | | | | | | |

*Exhibit 2.4 LFS Change Ratings between Round 1 and Round 2*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | *Correction from 2011 rating*<br>*Improvement from 2011 rating*<br>*Deterioration from 2011 rating*<br>**Comments on changes** |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 66 | 70 | 7 | 7 | 7 | 7 | 5→7[1] | L | [1]Planning cognitive lab work to reduce specification error. Reinterview survey that is being planned will also help in this regard. |
| Frame error | 58 | 58 | 7 | 7 | 7 | 3 | 5 | L | |
| Non-response error | 56 | 52 | 7→6[1] | 5 | ~~9~~ 5[2] | ~~7~~6→5[3] | 5→5[4] | H | [1]Knowledge of the causes of nonresponse have deteriorated. Although there are theories, the true causes of the increases in both intrinsic and residual risks need to be sorted out.<br>[2]Corrected due to level of expertise in data-collection<br>[3]This is both a correction to the Round 1 ratings and a deterioration. Best practices for telephone panels is to use face to face interviewing for Wave 1 for a number of reasons but foremost is to reduce nonresponse bias. There are other violations of best practices as well.<br>[4]Despite the considerable planning effort, this rating stayed at "Good" because mitigation activities have been slow to materialize while residual risks have climbed to a "critical" or "crisis level. This actually represents somewhat of a deterioration, thus we note it even though there was no change in the rating. |
| Measurement error | 50 | 56 | 5 | 5 | 5 | 3→5[1] | 7→8[2] | H | [1]Monitoring of TIs has commenced and further cognitive testing is being done of questionnaire. However, to achieve compliance with best practices, further examination of measurement error is needed; for example, to better understand the causes and effects of rotation group bias, and removal of the factor that the TIs are to a large extent aware of which calls are being monitored.<br>[2]Plans are in place to conduct reinterview survey; however, more is needed to mitigate measurement errors in the labour force estimates. |
| Data processing error | 54 | 62 | 5 | 3→5[1] | 7 | 7 | 5→7[2] | M | [1]QD documents data editing and provide information on coding error. Improvements planned in conjunction with ISO standards work.<br>[2]Plans to review the automated coding quality are in place. |
| Sampling error | 70 | 78 | 7 | 7→9[1] | 7 | 7→9[2] | 7 | M | [1]QD documents sample design and sampling error.<br>[2]Work on sampling error is well regarded and is consistent with the best in the field. |
| Model/estimation error | 50 | 60 | 5 | 5 | ~~3~~ 5 →6[1] | 3→7[2] | 5→7[3] | M | [1]Error corrected in last year's evaluation of seasonal adjustment expertise.<br>[2]Work on time series adjustment regarded as state of the art. Also work on GREG estimation is very good.<br>[3]Plans in place to revise estimation approach have been approved and implementation is underway. |
| Revision error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | |
| Total for Accuracy | 56,4 | 60,9 | | | | | | | |

*Exhibit 2.5 SBS Change Ratings between Round 1 and Round 2*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | *Correction from 2011 rating*<br>*Improvement from 2011 rating*<br><br>**Comments on changes** |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 50 | 54 | ~~3~~ 5[1] | 3→5[2] | 7 | ~~7~~ 5[3] | 5 | M | [1]Corrected because we have better knowledge including that information contained in the Quality Declaration<br>[2]Improvement because of Quality Declaration<br>[3]The product clearly meets the EU standard. Correction because delays in use of cognitive lab means that best practice is not yet met. |
| Frame error | 64 | 64 | 7 | ~~5~~ 7[1] | ~~5~~ 7[2] | ~~7~~ 6[3] | ~~7~~ 5[4] | M | [1]Correction in communication because of study overcoverage and measurement error using administrative data<br>[2]Corrected because we previously underestimated the degree of expertise that was actually being used for adjusting for frame errors<br>[3]Corrected because of better understanding of EU quality standard.<br>[4]Corrected because improvement plan has not yet been approved by management. |
| Non-response error | 70 | 70 | 7 | ~~7~~ 6[1] | 7 | 7 | ~~9~~ 8[2] | M | Corrected because nonresponse bias is potentially large for some estimates.<br>Corrected becuase there is potential to improve the information provided to users on nonresponse bias. |
| Measurement error | 50 | 52 | 6 | 4→5[1] | 5 | 5 | 5 | H | [1]Improvement because of Quality Declaration. |
| Data processing error | 54 | 60 | 5 | 4→5[1] | 7 | ~~5~~ 6[2] | 5→7[3] | H | [1]Improvement because of the availability of improved documentation for users.<br>[2]Corrected because of better knowledge on steps taken to meet standards.<br>[3]Improvement due to developments in coding system and imputation system. |
| Sampling error | 82 | 84 | 8 | 7→8[1] | 9 | ~~8~~ 9[2] | 8 | M | [1]Improvement due to development of Quality Declaration.<br>[2]Corrected because sample design is regarded as best practice. |
| Model/estimation error | 60 | 60 | 5 | 5 | 8 | 4 | 8 | H | |
| Revision error | 56 | 56 | 6 | 5 | ~~9~~ 8[1] | 4 | 5 | H | [1]Corrected because steps taken to reduce revisions are discussed outside Stat Sweden with other experts. |
| **Total for Accuracy** | 59,6 | 61,4 | | | | | | | |

*Exhibit 2.6 BR Change Ratings between Round 1 and Round 2*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | *Correction from 2011 rating*<br>*Improvement from 2011 rating*<br>*Deterioration from 2011 rating*<br>**Comments on changes** |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 62 | 66 | 4→5[1] | 4→5[2] | 7 | 8 | 8 | L | [1]Improvements due to improved knowledge due to discussions with the Tax Agency on the BR development project.<br>[2]Improvement due to the availability of the Quality Declaration. |
| Frame error- over coverage | 48 | 56 | 6 | 5 | 3→7[1] | ~~3~~ 5[2] | 7 →5[3] | ~~M~~ H[4] | [1]Improvement to Available Expertise because of the discussions leading up to the new BR system including the involvement of the Tax Agency.<br>[2]Corrected regarding Compliance because there does seem to be compliance with Swedish and EU standards.<br>[3]Deterioration because there are no immediate plans in place to address the increasing number of inactive units.<br>[4]The risk level has been corrected to High because we have a better understanding of the processes being used to populate the BR. These suggest the risk of inactive businesses being included is much higher than we previously thought. |
| Frame error- under coverage | 42 | 46 | 3 | ~~2~~ 3[1] | 6 | 6 | 3→5[2] | M | [1]Corrected on the basis of up to date information. Although the problem is communicated, it is not quantified.<br>[2]Improved because there appears to be an attempt to address the undercoverage in the BR system to be developed |
| Frame error - duplication | 55 | 63 | ~~2~~ 5[1] | 2→5[2] | 7 | 8 | N/A[3] | L | [1]Corrected on the basis of new information on the extent of the problem.<br>[2]Improvement due to the availability of the Quality Declaration.<br>[3]No need for a mitiagtion plan when the risk is minor. |
| Missing data error | 48 | 48 | ~~3~~ 5[1] | ~~3~~ 5[2] | 5 | 5 | 4 | L | [1]Corrected because there seems to be more quantification pf the problem than previously thought.<br>[2]Corrected because it appears the risk of missing data have been communicated. Action taken has reduced the size of the problem. |
| Content error | 42 | 46 | 3 | 3 | 7 | 5 | 3→5[1] | H | [1]Althought EU-standards are med, best practice with monitoring reliability of NACE codes is not met. There is work underway to study the accuracy of NACE coding. In addition it will be a consideration for the new BR project. Improvement for these reasons. |
| Total for Accuracy | 47,2 | 52,2 | | | | | | | |

*Exhibit 2.7 TPR Change Ratings between Round 1 and Round 2*

*Correction from 2011 rating*

*Improvement from 2011 rating*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plans towards mitigation of risks | Risk to data quality | Comments on changes |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 44 | 46 | 4 | 4 | 6 | 6 | 2→3[1] | ~~L~~ M[2] | [1]Some plans are being made to investigate these issues, but these plans still need management approval before proceeding.<br>[2]Corrects an error in the Round 1 assessment. |
| Frame error: overcoverage | 52 | 56 | 6 | 6 | 4→5[1] | 6 | 4→5[2] | H | [1]A methodologist has been assigned to the TPR whose focus is overcoverage.<br>[2]Two staff will join a 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. |
| Frame error: undercoverage | 38 | 60 | 3→5[1] | 3→5[2] | 4→7[3] | 5→7[4] | ~~2~~ N/A[5] | L | [1]There has been some progress in understanding the causes of undercoverage; for e.g., the Tax Agency's lag in registering births and immigrations. TPR staff are monitoring this.<br>[2]The QD has a discussion of undercoverage<br>[3]There is now a methodologist to look into these issues. Experts at the Tax Agency are also consulted.<br>[4]The risks appear rather small. Still the staff are actively investigating and monitoring the problem.<br>[5]Given the very low risk of undercoverage bias, planning for risk mitigation does not apply for this area. |
| Frame error: duplication | 70 | 70 | 6 | 6 | 8 | 8 | ~~4~~ N/A[1] | L | [1]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. |
| Missing data error: item and variable | 60 | 66 | 6 | 6 | 4→7[1] | 6 | 8 | M | [1]The collaboration group is working specifically on missing numbers on dwelling. |
| Content error | 50 | 58 | 5 | 5 | 5→7[1] | 7 | 3→5[2] | L | [1]Expertise is elevated based upon the collaboration of the methodologist.<br>[2]Plans have been approved to use errors found during the census to correct TPR content. |
| **Total score** | 52,2 | 58,0 | | | | | | | |

### *Exhibit 3.1 GDP, Quarterly – Numerical Ratings with Comments from Evaluation 2012*

| | Error source | Average score | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality | Comment risk level | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy (control over error sources) | **Input data source - Index of Service Production, ISP** | 58 | 4 | 5 | 7 | 7 | 6 | **H** | 1) service industry kind of activity units in manufacturing enterprises are not included (may affect the extrapolation factors)<br>2) for some industries e.g. real estate because of measurement error<br>3) sampling effects are seen in some smaller industries<br>4) specification for trade due to their smaller margins which can be volatile | There is knowledge of the errors from this data source. You need to be very viligent because the errors can change from one quarter to the next. The shortcomings have been documented but there have been no sensitivity studies to assess the impact of inaccuracies of this data source. In particular, the changes from one quarter to the next are not always well understood. Rating is Fair +. | There is considerable transparency in the contribution of this input data to the volatility of the national accounts, including at media conferences. However, as there have been no studies, users have not been made aware of the quantitative impact of errors in the SPI. Rating is Good. | The nature of the problems are well understood by ISP staff and there are plans to address this. Rating is Very Good. | The EU standard is that source data are regularly assessed and validated. This has been done. With respect to best practice, the ISP is consistent with the best practice for surveys of this type with the possible exception that the enterprise rather than the KAU is used as the survey unit. Rating is Very Good. | The short term economic statistics project will address some of the major problems with the survey. There is an approved plan but it has not yet been implemented. Rating is Good. |
| | **Input data source - Index of Industrial Production, IIP** | 58 | 4 | 5 | 7 | 7 | 6 | **H** | 1) service activity in manufacturing kind of activity units is missing (e.g. merchanting)<br>2) sampling error is potentially high for industries with predominantly smaller enterprises<br>3) measuring "deliveries" (instead of turnover) which could be a specification error<br>4) estimation for below cut-off enterprises<br>5) could be measurement error - enterprises could include more or less than what is required. | There is good knowledge of the errors from this data source. You need to be vilgent because the errors can change from one quarter to the next. The shortcomings have been documented but there has been no sensitivity studies to assess the impact of inaccuracies of this data source, especially from one quarter to the next. Rating is Fair+. | There is considerable transparency with respect to the contribution of this input data to the volatility of the national accounts, including at media conferences. As there have been no studies, users have not been made aware of the potential impact of errors in the IIP. Rating is Good. | The nature of the problem is well understood by IIP staff and there are plans to address this. Rating is Very Good. | The EU standard is that source data are regularly assessed and validated. This has been done. With respect to best practice, the IIP is consistent with the best practice for surveys of this type. Rating is Very Good. | The short term economic statistics project will address the major problems with the survey. There is an approved plan but it has not yet been implemented. Rating is Good. |
| | **Input data source - Merchanting Service of global enterprises (also covers royalties, licensing and R&D)** | 42 | 3 | 5 | 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. | This is a relatively new data source. Questions have been raised about the size of the estimates as they have contributed significantly to divergence of growth rate between IIP and the relative components of GDP. | Users have been made aware of the potential impact, including the contribution of merchanting, through seminars. Rating is Good. | There is expertise within SCB for running surveys of this type although not much experience with this subject matter.Rating is Good. | There is probably more work that needs to de done in this difficult area before being able to say this data source fully complies with EU standards and best practice. Rating is Fair. | There are no current plans to investigate this data source although statistics may be improved if Eurostat proposals on the treatment of global enterprises are implemented. Rating is Good. |
| | **Compilation error (modelling)** | 48 | 5 | 5 | 5 | 6 | 3 | **H** | Models - strong dependency on the work of the analysists.<br>1) intermediate consumption<br>2) construction<br>3) financial services<br>4) real estate<br>5) insurance<br>6) energy<br>7) water supply<br>8) informal economy<br>9) seasonal adjustment | There are numerous models used with that used for intermediate consumption being the most important. These models are documented. Some analysis has been done for intermediate consumption by looking at the size of revisions when annual data is available. There does seem to be a systematic negative bias which is worse when there are downturns in the economy. There has been some evaluation of the construction models. The seasonal adjustment process was revised about 3 years ago to take account of concerns expressed by users. Rating is Good. | Users have access to the available documentation. It is also reported publicly. Rating is Good. | The relevant expertise certainly exists in SCB. One concern is the cessation of the research and analysis group in National Accounts. Rating is Good. | Best practice is probably the relevant consideration. The modelling is probably be consistent with best practice although more work could be done on their evaluation. Several papers have been presented at European level workshops. Rating is Good+. | There have been past studies but little in the way of proposed future studies. Rating is Fair. |
| | **Compilation error (data processing)** | 40 | 7 | N/A | 3 | 3 | 3 | **H** | 1) spreadsheets<br>2) IT-system (objective is to have a more automated process to exclude manual work with input data, also traceability)<br>3) more compilations in SAS<br>4) there have been several false starts at developing a new system. Accessibility to IT resources has been a big issue. | The risks are well understood, some systems changes have been implemented and others have been proposed but IT resources have not yet been dedicated. Rating is Very Good. | Not applicable. | National Accounting systems are very specialised. Although the processing of national accounts of an acknowledged expertise of this group, People with skills in systems development are not available in Statistics Sweden. There will be a need to contract people with experience in implementing National Accounting systems. Such people are available. Rating is Fair. | Best practice is probably the relevant consideration. The processing system is not yet best practice although work has been done on the modules supporting the input data and modelling to reduce the person dependence. Rating is Fair. | Various proposals for systems redevelopments have been planned but there is not yet management approval. Rating is Fair. |
| | **Deflation error (including specification error)** | 48 | 4 | 3 | 7 | 7 | 3 | **H** | 1) possible high sampling errors in some of the producer price indexes<br>2) wage indices have to be used in some cases<br>3) insufficient adjustment for quality in general<br>4) complex products pose difficulties in measuring change over time | There is good knowledge of the risk of errors from this data source but there have been no sensitivity studies to assess the impact of inaccuracies of this data source except for the CPI which is only used to deflate some service industries. Rating is Fair +. | There is good information available on the deflation methods but little quantification of the possible impact on the accuracy of the quarterly national accounts. Rating is Fair. | There is strong expertise within SCB for addressing issues with the deflation process. Rating is Very Good. | The surveys, from which the price indexes are derived, are regularly reviewed consistent with the EU standard. Chain-linking is used to reduce the bias from not having up to date weights. This is all consistent with best practice. Most of the deflators are considered as preferred A or B methods. Rating is Very Good. | There have been past studies but little in the way of proposed future studies. Rating is Fair. |
| | **Balancing Error** | 56 | 5 | 5 | 7 | 6 | 5 | **H** | Dependency on experience of analysts and lack of standardized / formal methods but formal methods are now being developed. | A lot of knowledge would have been built up through practical experience. There is transparency on the processes that have been used. Sensitivity analysis would be possible. | There is transparency with the balancing process. The statistical discepancy (prior to balancing) is published even though the accounts are eventually fully balanced. | There is strong expertise within SCB for addressing issues with the balancing process. | Best practice is probably the relevant consideration. Best practice would be the utilizing of the quarterly supply and use tables. This is planned for the near future. Also the Handbook on the Quarterly Accounts. SCB practices would be consistent with best practice. Rating is Good + | There are some plans for improvement, using supply/use tables, that are currently with management for approval. The supply/use tables are now available. Also, there are regular meetings with the relevant staff during the balancing process. Rating is Good. |
| | **Revisions Error** | 56 | 7 | 5 | 7 | 5 | 4 | **M** | | On average, the revisions are not large and have been quantified. There is a lot of knowledge here. Rating is Very Good. | Revision studies have been published and made available to users. Rating is Good. | There is strong expertise within SCB for addressing issues associated with revisions. Rating is Very Good. | The EU standard on revisions is that they are regularly analysed in order to improve statistical processes. This appears to be addressed. Rating is Good. | Although revisions are produced as part of standard processes, they are not analysed formally although external users would undertake analysis and provide feedback to SCB. Rating is Fair+. |
| | **Total score** | 50,5 | | | | | | | | | | | | |

*Exhibit 3.2 GDP, Annual – Numerical Ratings with Comments from Evaluation 2012*

| | Error source | Average Score | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality | Comment risk level | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy (control over error sources) | **Input data source - Structural Business Statistics, SBS** | 66 | 7 | 5 | 7 | 7 | 7 | **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. | There was a good knowledge of the risks and several studies to quantify the risks. Rating is Very Good. | It does not appear that much of this information has been communicated to users as they tend to be more interested in the quarterly than the annual accounts. However, there is good communication of the errors associated with the SBS by the SBS team. Rating is Good. | SCB has expertise to analyse risks of this type. Rating is Very Good. | The EU standard is that source data are regularly assessed and validated. This is certainly true for SBS which is reviewed annually taking into account the views of National Accounts. Rating is Very Good. | There are annual reviews aimed at improving the SBS for national accounting and other purposes. Rating is Very Good. |
| | **Compilation error - modelling** | 48 | 5 | 4 | 5 | 7 | 3 | **H** | Modelling<br>1) trade margins<br>2) construction<br>3) financial services<br>4) real estate<br>5) insurance<br>6) energy<br>7) informal economy | These models are documented. There has been some evaluation of the construction models and the trade margins. Rating is Good. | Users have access to the available documentation but it does not contain much in the way of quantification. There is less interest fom users compared with the quarterly accounts. Rating is Fair+. | The relevant expertise certainly exists in SCB. One concern is the cessation of the research and analysis group in National Accounts and the loss of experience more generally. Rating is Good. | Best practice is probably the relevant consideration. The modelling would probably be consistent with best practice although more work could be done on their evaluation. Rating is Very Good. | There have been past studies but little in the way of proposed future studies. Rating is Fair. |
| | **Compilation error - data processing** | 35 | 5 | N/A | 3 | 3 | 3 | **H** | Data-processing<br>1) spreadsheets<br>2) IT-system (objective is to have a more automated process to exclude manual work with input data, also traceability)<br>3) more compilations in SAS | The risks are well understood, some systems changes have been implemented and others have been proposed but resources have not been provided. Rating is Good. | Not applicable. | Although National Accounts have the expertise to undertake National Accounts processing, the relevant expertise exist to develop a modern national accounting system is not available in SCB. Rating is Fair. | Best practice is probably the relevant consideration. The processing system is not yet at best practice although there have been work on modules supporting input data and modelling to improve traceability and reduce person dependency. Rating is Fair. | There are plans to develop a new IT system but it has not been approved by management. Rating is Fair. |
| | **Deflation error (including specification error)** | 48 | 4 | 3 | 7 | 7 | 3 | **H** | 1) possible high sampling errors in some producer price indexes<br>2) wage indices are used for collective public consumption and some services<br>3) insufficient adjustment for quality in general<br>4) complex products pose difficulties in measuring change over time<br>5) the models used in constant price estimation for goverment | There is good knowledge of the risk of errors from this data source but there has been no sensitivity studies to assess the impact of inaccuracies of this data source except for the CPI which is only used to deflate some service industries on the production side. Rating is Fair+. | There is good information available on the deflation methods but little quantification of the possible impact. Rating is Fair. | There is strong expertise within SCB for addressing issues with the deflation process. Rating is Very Good. | The surveys, from which the price indexes are derived, are regularly reviewed consistent with the EU standard. Chain-linking is used to reduce the bias from not having up to date weights. This is all consistent with best practice. Rating is Very Good. | There have been past studies but little in the way of proposed future studies. Rating is Fair. |
| | **Balancing Error** | 50 | 5 | 5 | 5 | 7 | 3 | **H** | 1. Objective Editing<br>2. Subjective Editing<br>3. RAS method<br>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. | A lot of knowledge would have been built up through practical experience. There is transparency on the processes that have been used. Sensitivity analysis would be possible. Rating is Good. | There is transparency on the balancing processes used within SCB. Rating is Good. | There is strong expertise within SCB for addressing issues with the balancing process. There has been succession planning but these have failed. Loss of expertise is an increasing concern. Rating is Good. | Best practice is probably the relevant consideration. SCBs balancing practices would be consistent with best practice. Rating is Very Good. | There are no specific plans for improvement although there are regular meetings with the relevant staff during the balancing process. They would like to formalise methods more and reduce the level of person dependency. Rating is Fair. |
| | **Revisions Error** | 54 | 5 | 7 | 5 | 6 | 4 | **M** | 3 yearly estimates are made<br>1) sum of 4 quarters t+ 60 days<br>2) t+9 months (revisions covering largely Government sector)<br>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. | Revisions are produced on a regular basis by SCB but not analysed although there does seem to be knowledge of the cause of the revisions. Rating is Good. | There is transparency with the size of revisions. Users are aware of them and they are discussed from time to time. They are more concerned with revisions to quarterly data. Rating is Very Good. | There is expertise within SCB for handling revisions but concern about the loss of expertise. Rating is Good. | The EU standard on revisions is that they are regularly analysed in order to improve statistical processes. This is done. Rating is Good+. | Revisions are produced as part of standard processes. Rating is Fair+. |
| | **Total score** | 49,9 | | | | | | | | | | | | |

### *Exhibit 3.3 ULF/SILC – Numerical Ratings with Comments from Evaluation 2012*

| | Error source | Average Score | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risks | Risk to data quality | Comment risk level | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with standards & best practices | Plan towards mitigation of risk |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy (control over error sources) | **Specification error** | 34 | 3 | 3 | 7 | 3 | 1 | M | Questions are dated and the risk of specification error should not be ignored or by cognitive experts would focus on this error source.<br>Questionnaire has never been reviewed by a subject matter expert. | QD does not acknowlege Specification error as a risk. It is acknowledged in the checklist although nothing has been done to evaluate the risks. | There has been essentially no communication with users and stakeholders regarding these risks. | There is access to the cognitive lab and their experts necessary to investigate this error source. | Key staff are not aware of the best practices regarding this error source. | There has been no planning to investigate the potential of specification error. This could be addressed as part of the evaluation measurement error via the cognitive laboratory. |
| | **Frame error** | 42 | 3 | 3 | 7 | 5 | 3 | H | Trouble identifying families and hhs using TPR.<br>Sampling is done once a year. New registrants are missed.<br>Overcoverage issues inflate nonresponse rate. | The risk of frame error is acknowledged but little has been done to discover its impact on key estimates. | Likewise, there has been little communication with experts in this regard. | There is a high level of expertise available to investigate these issues, but they are under-utilised. | Standards are being met but no plans to mitigate risks. | Essentially no plans to investigate the potential of frame error (particularly overcoverage) to affect key estimates. |
| | **Non-response error** | 40 | 5 | 3 | 5 | 4 | 3 | H | NR is relatively high and growing.<br>Some knowledge about unit NR bias.<br>Little knowledge about item NR levels and bias<br>No knowledge about item nonresponse | Historical studies have looked at nonresponse bias but these need to be updated given the changing times. | QD does not have an adequate discussion of the risks of nonresponse to data users. Other documentation also does not adequately describe the risks. Studies are needed that evaluate the bias due to nonresponse, not just nonresponse rate analysis. | Statistical expertise is quite good. However, data collection expertise to increase response rates and to deal with the problems in the telephone center is lacking. | There has been some work to increase response rates via contact strategies; however compliance with standards is minimal. | Current plans seem inadequate to reduce the risks of NR bias in the near future. A written plan to address nonresponse in the LCS/ULF has not yet been approved by management. |
| | **Measurement error** | 46 | 3 | 3 | 9 | 3 | 5 | H | Questions are quite complex and wordings need refinement.<br>Design of call monitoring system will miss many careless and deliberate interviewer errors<br>High risk of interviewer variance, especially for child interviews.<br>Child interviews are prone to unreliability and invalidity.<br>Risk of social desirability bias is high for personal questions. | The issues of measurement error are not well-known. This is particularly important for the children interviews as data collected from children is frought with reliability issues. | Documentation of measurement error issues, including the QD, is inadequate. Little mention is made of the risks to data quality of measurement error. | Statistics Sweden has excellent capabilities in this area and staff who have expertise in measurement error evaluation and reduction have been assigned to the survey. | Lack of call-monitoring is the more blatant violation of standards. Compliance with standards will move to a rating of "Good" when call monitoring is implemented. | There are plans to study measurement issues via the cognitive laboratory. |
| | **Data processing error** | 42 | 5 | 3 | 7 | 5 | 1 | L | Primarily in the I/O coding.<br>Interviewer coding of some open ended questions subject to high risk of error<br>5% recoding | Some work has been carried out in the area of I/O coding error evaluation. | The results of studies of I/O coding errors and their risks to data quality have not been shared with users. | There is a high level of expertise available to investigate these issues, but they are under-utilised. | Current practices comply with standards but are not state of the art. | No planning to investigate this error source with is consistent with the L risk level and quite acceptable. |
| | **Sampling error** | 54 | 7 | 7 | 7 | 3 | 3 | M | Sampling design is so complex that staff are not able to compute selection weights. The resulting biases are unknown; however, it is possible that calibration adjustments remove much of this bias. Still, essentially nothing is known about this.<br>Report on simplified survey design | Knowledge of sampling errors is very good. | Samplings errors have been well-communicated. | Expertise in sampling errors is very good considering the expertise of sampling statisticians assigned to the unit. | Given the uncertainties surrounding the current sampling methodology, compliance with standards is only fair. | Essentially no plans to investigate the potential of bias associated with the current weighting approaches. |
| | **Model/estimation error** | 38 | 5 | 3 | 7 | 3 | 1 | H | Calibration modeling for reducing bias and variance are sophisticated and presumably effective. | Fairly good knowledge regarding the benefits of calibration modeling. | These benefits have not be well-communicated, for example, in the QD. | Expertise in calibration estimation is very good considering the expertise of the statisticians assigned to the unit. | Staff are aware of standards regarding weighting but greater knowledge of best practices, particularly as they regard the evaluation of calibration weighting, is needed. | No plans are being made to look at this issue in the future. |
| | **Revision error** | N/A | N/A | N/A | N/A | N/A | N/A | N/A | | | | | | |
| | **Total score** | **42,1** | | | | | | | | | | | | |

### Exhibit 4.1 CPI – Numerical Ratings with Comments for Evaluation of User Quality Dimensions (Relevance/Contents and Accessibility & Clarity) 2012

| | User quality dimensions and components | Average score | Knowledge of Risks | Communication with Users | Available Expertise | Compliance with standards & best practices | Plans towards Addressing User Needs | Risk to data quality | Comment risk level | Knowledge of Risks | Communication with Users | Available Expertise | Compliance with standards & best practices | Plans towards Addressing User Needs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relevance | **Inputs (content, scope, classifications, etc)** | 60 | 7 | 7 | 5 | 5 | N/A | **L** | There are regular discussions with power users and the CPI Advisory Board. Also, the product area is active in Eurostat discussions on HICP. | Most of the focus has been on understanding the needs of the power users and their needs are well understood. There are a small number of important missing items such as financial services but no strong demand to include them. | There is regular communication with the power users and the CPI Advisory Board. There are a large number of less sophisticated users that use the CPI for regulation of contracts and there is only broad knowledge of their needs. | External environment monitoring keeps staff informed on new products and sales outlets. However, it was noted that there would benefits from more product expertise. | Stat Sweden complies with the relevant EU standards and is consistent with good practice. | Given the good situation with input data sources, there is no need for improvement plans. |
| | **Outputs (including microdata and other products)** | 66 | 7 | 7 | 7 | 9 | 3 | **H** | There are regular discussions with power users and the CPI Advisory Board. Also, the product area is active in Eurostat discussions on HICP. Also, a SLA is being developed with National Accounts where there are some differences between their needs and what is currently provided.<br>There is some demand for detailed data but not microdata but special studies at the product level (or higher) are conducted on a commission basis.<br>There are regular discussions with power users and the CPI Advisory Board. Also, the product area is active in Eurostat discussions on HICP. Also, a SLA is being developed with National Accounts where there are some differences between their needs and what is currently provided.<br>There is some demand for detailed data but not microdata but | There are regular user surveys but these focus on the power users. Web site is frequently used and there is little evidence that the information is insufficient. | There is regular communication with power users. A SLA is being developed with the National Accounts. One area of improvement could be to have material on the web site which informs those considering the use of the CPI in contracts of the strengths and weaknesses of it for these purposes. This is done by some NSOs. | Staff have the expertise to understand and address new requirements. | All EU standards are followed and the product is active in discussions at that level. Changes to outputs are monitored by the CPI Advisory Board. | Some parts of the output will be reviewed next year. |
| Accessibility and Clarity | **Ease of Data Access** | 50 | 7 | 5 | 5 | 5 | 3 | **H** | The web site is the most important way of communicating CPI data. There are closer relationships with National Accounts and other CPI users. | User surveys have been recently conducted which include ease of data access. These will be published in December 2012. There have been no cognitive studies on web access. | There is regular contact with the power users. The User Survey results are soon to be published | Given the web site is the main means of communication, there may be a need for more strength for skills in web site design that takes account of cognitive studies | The relevant EU standard is "Dissemination Services use modern information and communications technology and, if appropriate, traditional hard copy". This goes beyond passive web site communication so there are probably some areas for improvement. Even though this is primarily the responsibility of the communications department, the product area could initiate discussions. Some NSOs have been more active in new forms of communication. | There are no definite improvement plans at this time although user surveys are good. |
| | **Documentation (including metadata)** | 50 | 5 | 5 | 5 | 7 | 3 | **H** | The Swedish Consumer Price Index: A handbook of methods has been published is available on the SCB web site.<br>A Quality Declaration has been updated. | A user Survey has recently been conducted which should throw light on areas for improvement | The existence of the two documents demonstrates good communication. | SCB has the expertise to prepare this documentation. Methodology is closely involved. | The relevant EU standard is that "Statistics and the corresponding meta data are presented, and archived, in a form that facilitates proper interpretation and meaningful comparison" and "Meta data are documented according to standardised meta data systems". We did not really discuss this but it is not likely that all the steps are undertaken to warrant a excellent rating. | There are no definite plans for improvement although a User Survey has just been conducted. |
| | **Availability of Quality Reports** | 62 | 5 | 7 | 9 | 7 | 3 | **H** | The Swedish Consumer Price Index: A handbook of methods has been published is available on the SCB web site.<br>A Quality Declaration has been updated. | The recent User Survey should improve knowledge of the needs for information on quality. | The availability of the Quality Report is a significant improvement in this respect. Some of the quality studies are now somewhat out of date. | SCB has been at the forefront of international activity on the quality of CPI statistics. | The relevant European standard is "Users are kept informed about the methodology of statistical processes including the use of administrative data". This is clearly complied with given the existence of the documents mentioned above. However, the contents of the documents could be updated if there were more recent quality studies. | The Quality Declaration is to be published in English. |
| | **User Support** | 58 | 5 | 5 | 7 | 9 | 3 | **H** | There is a help desk to help address questions. There is a more personal relationship with the power users.<br>There is a service to support special requests for detailed data. | Knowledge of the need for improved user support should be able to be derived from the recent User Surveys. | The web site indicates the extent of user support that is available. | There is a help desk and we assume that the staff are knowledgeable about the CPI. The staff dealing with the power users are definitely knowledgeable. | The European standard is "Customer designed analysis are provided when feasible and the public is informed". This is supported. | There is a plan for user support but no plans for improvement at this time. |
| | **Total Score User Quality Dimensions** | 57,4 | | | | | | | | | | | | |

*Exhibit 4.2 LFS – Numerical Ratings with Comments for Evaluation of User Quality Dimensions (Comparability & Coherence and Timeliness & Punctuality) 2012*

| | User quality dimensions and components | Average score | Knowledge of Risks | Communication with Users | Available Expertise | Compliance with standards & best practices | Plans towards Addressing User Needs | Risk to data quality | Comment risk level | Knowledge of Risks | Communication with Users | Available Expertise | Compliance with standards & best practices | Plans towards Addressing User Needs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Comparability and Coherence | **Comparability across geography, populations, and other relevant domains** | 52 | 5 | 5 | 8 | 5 | 3 | **M** | The same methodologies are used across geographic regions and demographic domains which reduces the risk level. There is some potential (medium risk) for the mix of centralized and decentralized telephone interviewing to create spurious differences. | Currently, not much is known about the risks to inter-regional comparability as a result of the imbalance in centralised vs. decentralised telephone interviewing. | Likewise, not much has been communicated to users about these risks. | The expertise for examining these issues are quite good. | Standards are being met but not exceeded. More research is needed to examine the issues. Staff are not particularly knowledgeable about best practices in this area. | No planning to address these issues currently exists. |
| | **Comparability across time (including impacts of redesign)** | 74 | 8 | 8 | 8 | 5 | 8 | **H** | Over its history, there have been substantial changes to the data collection and estimation methodology. These changes have resulted in important shifts in the estimation of labor force characteristics. The risks of such changes to the utility of the LFS estimates is therefore quite high. | Staff understand very well the risks of temporal comparability. During 2012, several reports have been published reviewing the work that has been done to date on temporal comparability and addressing breaks in the data series. | There are four user groups:<br>1) Expert Group on Labour Market statistics, EFAM<br>2) User group for Labour Market statistics at Stat Sweden,<br>3) Working group SASA,<br>4) NA-LFS Board at Stat Sweden | The expertise for examining these issues are quite good. LFS works closely with time series group on seasonal adjustment issues as they relate to comparability. | Standards are being met but not exceeded. More research is needed to examine the issues. Staff are not particularly knowledgeable about best practices in this area. | Plans are in place purblish additional reports on this issue, particularly for series that pre-date 1986. |
| | **Coherence with other relevant statistics (including use of standard classifications, frameworks, etc.)** | 38 | 3 | 3 | 5 | 3 | 5 | **H** | There are several other relevant sources of employment statistics:<br>1) Employer-based employment statistics, quarterly survey<br>2) Register based employment statistics, based on tax data<br>3) Swedish Public Employment Service (unemployment figures)<br>Differences between estimates derived from these sources and the LFS can cause considerable confusion and misinterpretations in the user community. | A comparison of different employment statistics was done in 2007 and is available on Statistics Sweden's website. However, not much is known about why differences occur. Nor has there been much in the way of using these comparisons to understand LFS data quality. | Some discussion of coherence in the QD. However, this section needs to be expanded to include some results from the 2007 study and a discussion of why differences occur and how they should be interpreted vis a vis the LFS. | Expertise for examing Coherence issues is good. However, a collaborative effort with producers of the other sources of estimates may be needed to fully understand the issues. | Some standards are not being met. More research is needed to examine the causes of lack of coherence among the relevant statistics. Staff are not particularly knowledgeable about best practices in this area. | There are some plans to compare the LFS with the NA which have received management approval. |
| Timeliness and Punctuality | **Timeliness of release of main aggregates** | 68 | 8 | 8 | 5 | 8 | 5 | **H** | Timeliness of release of LFS estimates is of critical importance for setting national policy. | The LFS staff interact routinely with the four user groups (Expert Group on Labour Market statistics (EFAM), User group for Labour Market statistics at Stat Sweden, Working group SASA, and NA-LFS Board at Stat Sweden) regarding their issues associated with timeliness. | The LFS staff interact routinely with the four user groups (Expert Group on Labour Market statistics (EFAM), User group for Labour Market statistics at Stat Sweden, Working group SASA, and NA-LFS Board at Stat Sweden) regarding timeliness and the need for more timely data. Needs of other users are captured through central sources. | The staff are well-aware of the need for timeliness. They also know what efforts are needed in order to continue releasing data on the current schedule. However, there is uncertainty as to how to solve the problems in the data collection department that continues to threaten the timeliness and punctuality of the data. | The current release schedule is among the most ambitious in the EU. This suggestion that the LFS is "leading the pack" in Europe. | Although the timeliness of data releases is very good, there should be some planning to ensure that this schedule can continue to be met as well as to improve further as new technologies for better performance are being developed. |
| | **Timeliness of release of detailed outputs (including microdata)** | 68 | 8 | 8 | 5 | 8 | 5 | **M** | Issues under microdata release are essentially the same as for the main aggregates with the latter having somewhat greater risk to user satisfaction. Thus, these two components will be treated simultaneously as one component | | | | | |
| | **Punctuality** | 62 | 8 | 8 | 5 | 5 | 5 | **H** | Standards and best practices are not clearly applied in the data collection area. | The LFS staff understand the issues acciociated with threats to punctuality and delays in the release of labour force statistics. | Information about delays are provided to power-users and is also available on the webpage. However, there are few other mechanisms to obtain regular feedback from the user community regarding timeliness. | The staff are well-aware of the need for timeliness. They also know what efforts are needed in order to continue releasing data on the current schedule. However, there is uncertainty as to how to solve the problems in the data collection department that continues to threaten the timeliness and punctuality of the data. | Punctuality is meeting standards but needs improvement to achieve the level of best practices. | The DG has mandated that there should be no further delays in the release of data. However, there still needs to be planning to ensure that data can be released on time and still maintain acceptable data quality. Such planning has not been done. |
| | **Total for User Dimensions** | 60,4 | | | | | | | | | | | | |