DEVELOPMENT OF QUALITY INDICATORS AT STATISTICS SWEDEN

Paul Biemer and Dennis Trewin

January 9, 2012

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 should include metrics that reflect current data quality as well as changes in quality that occur over time. SCB in collaboration with two consultants (Paul Biemer and Dennis Trewin), developed a quality evaluation approach (or model) for this purpose and pilot tested it on eight products: Annual Municipal Accounts (RS), Consumer Price Index (CPI), Foreign Trade of Goods Survey (FTG), Labour Force Survey (LFS), National Accounts (NA), Structural Business Survey (SBS), Business Register (BR), and Total Population Register (TPR). For this initial review the focus was on the Accuracy dimension of survey quality, although some consideration was also given to Relevance. Future reviews will include these and other quality dimensions. For each of these products, Accuracy (or data quality) was assessed for the sources of error that were applicable for each program. These sources of error included: specification error, frame error (including over coverage, under coverage, content error, and missing data), non-response error, measurement error, data processing error, sampling error, model/estimation error, and revision error.

For each product, the quality assessment involved a self-assessment, extensive reviews of relevant documentation, four-hour long quality interview, and product staff reviews of final assessments with feedback. To facilitate the evaluation, a schema was developed whereby each product was scored (using a 10-point scale) according to five criteria. These criteria were essentially the same for each error source and guidelines were developed to impose consistency in the ratings. Overall scores were tallied as a weighted average of the scores for each error source where the weights were 3, 2, or 1 corresponding respectively to high, medium, or low potential risk from an error source. Overall scores were converted to percentages and ranged from 45% to 59% with an average 55% (see Exhibit 3 in the report).

Measurement error stood out as the only error source to be rated as a "high risk" by all products but one. The risk for data processing error was also rated "high" or "medium" by all products. Yet, in terms of their quality ratings, these two error sources scored among the bottom three (only frame error is lower). These findings suggest that future improvement efforts should address the risks to data quality from these two error sources.

The evaluators noted that results of the quality investigations for most products are not well-documented. Most quality evaluations tend to focus on error rates and indirect measures rather than MSE components[1] (i.e., bias and variance measures). Data processing error poses some important risks in areas such as data entry quality control (q.c.), NACE coding, and editing.

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. In addition, the report lays out ten cross-cutting recommendations for improvement that are listed below in no particular order:

1. Improve the integration and coordination of economic statistics from a methods point of view.
2. Improve cooperation between the National Accounts and the statistical areas, particularly in the macro-editing of National Accounts source data.
3. Devote greater attention to the accuracy of NACE coding, especially as a result of moving to self-coding by enterprises.

[1] Mean Squared Error components

4. Increase knowledge of error though evaluation studies, particularly in the areas of data editing and measurement error.
5. Accelerate the research on reduction and compensation for nonresponse, especially in household surveys, with an emphasis on sample representativity rather than high response rates.
6. Foster a closer relationship with the Tax Office to aid in the improvement of the registers and other products.
7. Develop a policy regarding the continuity of statistical series across redesign years with the use of backcasting of time series where important.
8. Improve the relationship between the Information Technology staff (IT) and their client areas.
9. Consider telephone interviewer monitoring for quality control and quality assurance.
10. Develop quality profiles for key products to facilitate future quality evaluations as well as other purposes.

The current quality evaluation model worked well, after some adjustments to the error structure for the Registers and the National Accounts, but could be improved. Improvements in the documentation of quality improvement efforts, the criteria used for scoring the error sources and error structure used in the evaluation of the quality of the National Accounts and other products reliant on a range of source data (e.g. balance of payments) are suggested. In addition, the evaluation model should be extended to include other quality dimensions such as relevance, timeliness, comparability, coherence, and accessibility/clarity. This would capture all of the important attributes associated with total survey quality.

The quality evaluation process should be repeated at regular (for e.g., 12-month) intervals ideally using an external review team who would work closely with the product areas. SCB should also identify other collections for self-assessments that would be facilitated by a knowledgeable internal moderator. Priority should be on the most important collections. Finally, work on ISO standards is important and should be encouraged. It is important that there is agreement within SCB on the standards that apply to each statistical operation.

2 BACKGROUND AND TERMS OF REFERENCE

The government of Sweden stated in SCB'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 reporting in the form of specific indicators that signify any quality improvements that are occurring in pre-specified programs.

Up until 2008 SCB 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 SCB 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 SCB during the year to develop a model that will capture quality changes in the agency's statistical programs.

SCB 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. Currently, six dimensions of total survey quality have been identified – Accuracy, Relevance, Timeliness, Comparability, Coherence, and Accessibility[2]. The director of R&D at SCB has determined that *Accuracy* and *Relevance* should be the immediate focus of the quality improvement initiative and that the agency needs to develop reporting techniques that are more rigorous, transparent and comprehensible for these dimensions. Thus, these two dimensions have been the focus of our efforts in developing a quality evaluation model for use by SCB.

In proposing our approach, we wanted to identify where clear improvements had been made as a result of effort by SCB. We also wanted to have a process which identified the highest priority areas for improvement. Our approach, its applicability to the eight products comprising our review, and its strengths and weaknesses are described Section 3. Section 4 summarises the results of the quality evaluations for the eight products. Section 5 summarises some cross-cutting methodological and other findings. Section 6 proposes a number of improvements in the quality evaluation model. Section 7 discusses next steps and planned future work. Finally, Section 8 provides our recommendations and conclusions.

[2] These quality dimensions differ somewhat from the dimensions that are currently in use by SCB, viz., Content, 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" and consider "Comparability" and "Coherence" as distinct dimensions.

3 PRODUCT QUALITY ASSESSMENT AND MONITORING

3.1 THE SCB QUALITY MODEL

We have developed a proposed SCB quality model for assessing the risks to data quality that exist for a product, knowledge of the risks by both data producers and users, compliance with appropriate standards and best practices, and current and future plans for mitigating the risks. Although the model can be extended to all dimensions of quality, this review focused primarily on Accuracy or data quality. For Accuracy, current risks were assessed separately for each error source that may affect product quality. Error sources may not be the same for all products so they are allowed to differ by product in the evaluation. For example, sampling does not apply to products that employ no sampling. Or if preliminary estimates have little potential risks of disagreeing appreciably with final estimates, revision error would have a low risk. In addition, an error source may be defined slightly differently for some products. As shown in Exhibit 1, three sets of error sources were identified for the eight products considered in this evaluation.

For sample surveys, the survey methods literature defines six essential error sources: specification, frame, sampling, nonresponse, measurement and data processing (Biemer and Lyberg, 2003[3]; see also European Commission, Eurostat, 2009[4]). Two additional error sources were defined for this evaluation: model/estimation error and revision error. A *specification error* arises when the concept implied by the survey question and the concept that should be measured in the survey differ. *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. 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[5]. *Nonresponse error* encompasses both unit and item nonresponse. *Unit nonresponse* occurs when a sampled unit does not respond to any part of a questionnaire. *Item nonresponse* occurs when the questionnaire is only partially completed 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.

Note that, in Exhibit 1, the error sources associated with the two registers – Business and Total Population – is somewhat different than the error sources for the other products. In survey work,

---

[3] Biemer, P. and Lyberg, L. (2003). Introduction to Survey Quality, John Wiley & Sons, Hoboken, New Jersey.

[4] European Commission, Eurostat (2009). Handbook for Quality Reports, Eurostat Methodologies and Working Papers, Downloaded at http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-EHQR_FINAL.pdf on December 20, 2011.

[5] 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.

the primary use of a register is as a sampling frame of some target population. Thus, frame error is expanded to include its subcomponents, viz., overcoverage, undercoverage, duplications, content error, and missing data. Likewise, the error sources associated with the National Accounts are somewhat different from those for the survey products. Frame error is missing from the list even though such errors may be an important issue for the primary data sources that are input to the National Accounts. Instead, they are part of the model/estimation error component reflecting the way they are treated in the estimation process. "Missing data" replaces "nonresponse error" in the list to convey the idea that data may be missing in the estimation process for a number of reasons include scheduled or unscheduled data unavailability, schedule delays, and so forth. For the National Accounts, sampling error in the primary data sources may give rise to inconsistencies in the different components of the National Accounts. However, we have expanded the definition of sampling error to also include the lack of integration in the design of the surveys used for these primary data sources.

Exhibit 1. Sources of Error Considered by Product

Product	**Error Sources**
Survey Products Foreign Trade of Goods Survey (FTG) Labour Force Survey (LFS) Annual Municipal Accounts (RS) Structural Business Survey (SBS) Consumer Price Index (CPI)	Specification error Frame error Nonresponse error Measurement error Data processing error Sampling error Model/estimation error Revision error
Registers Business Register (BR) Total Population Register (TPR)	Specification error Frame: Overcoverage Undercoverage Duplication Missing Data Content Error
Compilations National Accounts (NA)	Specification error Missing Data Content error Sampling error Model/estimation error Revision error

Exhibits 2a-2e provide the quality criteria and quality guidelines that were applied to each error source in Exhibit 1. A two-step rating process was used to assign a rating from 1-10 for each criterion. First, a criterion was graded on a five point qualitative scale corresponding to Poor, Fair, Good, Very Good, and Excellent. These ratings were later refined by choosing between low or high numerical point ratings within each of the five categories. For example, if an error source was assigned a rating of "Good" in step 1 of the evaluation, a numerical rating of either 5 or 6 was later assigned in step 2 to refine this rating.

Each product was also assigned two risk ratings for each error source corresponding to a product's "residual" or ("current") risk and "potential" or ("inherent") risk of error from that source. Residual risk reflects the risk that a serious error might occur from the source despite the current efforts that are in place to mitigate this risk. Potential (or inherent) risk may be thought as the risk of a serious error *prior to* the efforts toward risk mitigation. In other words, it reflects the risk of error from the error source if efforts to maintain current, residual error were to be suspended. For example, a product may have very little risk of nonresponse bias as a result of considerable efforts to maintain high response rates and achieve representativity in the achieved sample. Its residual risk is said to be low. However, remove all of these efforts and nonresponse bias becomes an important risk. As a result, its potential risk is said to be high. Thus, potential risk reflects the effort required to maintain residual risk at its current level. Residual risk does not play an active role in the evaluation nor is it reported in our results. Its sole purpose is to clarify the meaning and facilitate the assessment of potential risk. Potential risk is assessed at three levels: Low, Medium, and High.

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 2a – 2e divided by the highest score attainable, i.e., 50, and then expressed as a percentage. A product's overall score, also expressed as a percentage, is then computed by following formula:

$$\text{Overall Score} = \sum_{\text{all error sources}} \frac{(\text{error-level score}) \times (\text{error source weight})}{50 \times (\text{weight sum})}$$

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.

Exhibits 2a-2e Quality Criteria to be Applied to Each Error Source[6]

Exhibit 2a. Knowledge of Risks				
Poor [1,2] ●	Fair [3,4] ◔	Good [5,6] ○	Very Good [7,8] ◒	Excellent [9,10] ◎
Internal program documentation does not acknowledge the source of error as a potential factor for product accuracy.	Internal program documentation acknowledges error source as a potential factor in data quality. **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. **But:** Evaluations have only considered proxy measures (example, error rates) of the impact with no evaluations of MSE components.	Studies have estimated relevant bias and variance components associated with the error source and are well-documented. **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.

[6] During the evaluation process, we identified a number of ways the guidelines for each criterion rating in Exhibits 2a-2e could be improved. Our current plans are to suggest these and other improvements for use in future quality reviews at Statistics Sweden.

Exhibit 2b. 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 information on the error source.	There are some mentions of the risks of error from this source. **But:** Communications have been largely inadequate considering the potential risks to data quality.	Communications with users and customers have adequately described the risk to many users. **But**: Information conveyed has largely been proxy measures with little communications regarding MSE components.	Communications have shared some of the available information on the relevant MSE components that have been evaluated and assessed or only deal with sampling errors. **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 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 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 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 some important areas. **But:** There are important areas were expertise is lacking.	The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in most important areas. **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 covered.	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 statistical area. **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.	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. There is an excellent working relationship with the statistical area.

Exhibit 2d. Compliance with Standards and Best Practices				
Poor [1,2] ●	Fair [3,4] ◔	Good [5,6] ○	Very Good [7,8] ◒	Excellent [9,10] ◎
There is no evidence that standards and best practices, as they related to this error source, have been applied to the product. Moreover, seriously deficiencies exist that violate standards and best practices as they relate to this error source.	There is evidence that standards and best practices have been applied to the product for this error source. **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.	The relevant standards and best practices have clearly been applied to the product. Either there are no i*mportant* violations or gaps or there may be some important gaps but they are being actively addressed. **But:** Either (a) compliance is not routinely monitor or (b) gaps in compliance exist for some minor areas that are not being addressed.	The relevant standards and best practices have clearly been applied to the product. There are no serious violations of standards and best practices as they relate to this error source **But:** Some key staff may not be aware of the relevant standards and best practices and are not routinely monitoring compliance.	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.

Exhibit 2e. 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 a plan is in place or that any planning has been done for studying or mitigating the risks for this error source.	Some planning has been done for mitigating the risks for this error source. **But:** The plan is in an unfinished state or is poorly written. For example, while the plan might specify key objectives, either there is no provision for measuring progress toward them or the objectives are not measurable.	A written 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. **But:** One of the following deficiencies with the plan exists: a. The plan has not been updated in at least one year. b. There is no evidence that the plan is ever referenced in the work or it is not referenced as often as necessary. c. There are no accountability measures in place to ensure compliance with the plan. d. No metrics are specified for gauging progress toward each objective. e. No resources have yet been allocated.	A well-written 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. None of the deficiencies noted under the "Good" criteria are present. **But:** Progress toward completing the goals and objectives specified in the plan have been only fair or has been inconsistent for some key objectives.	There exist well-documented, short and long-term plans for mitigating the risks to data quality from this error source. The plans are updated periodically as appropriate and are continually referenced in the work. Accountability measures are in place to ensure compliance with the plans. 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.

3.2 APPLICATION TO THE PRODUCTS

The application of this model to the eight products in Exhibit 1 followed a multistep process as follows:

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. Also during this period, the key staff responsible for each product were invited to a meeting that explained the evaluation model and its use. At this meeting, or subsequent to it, the staff used the model to perform a self-assessment of data quality. This review of relevant materials and the self-assessments were essential steps leading to the main data gathering activity – i.e., the quality interview.

THE QUALITY INTERVIEW

Quality interviews were conducted in both Stockholm and Orebro from November 28 – December 5. Each interview took approximately four hours to conduct. The meetings were organised into four parts: (a) descriptions of the processes associated with product design, data collection, data processing, estimation, and reporting, (b) classification into High, Medium, and Low categories of the potential risk associated with each error source, (c) assessment of ratings for each criterion by error source, (d) assessment of the risk of catastrophic error, and (e) a review of the ratings summary including a discussion of the results.

The assessment of the risk of catastrophic error (d) was intended to identify the chance (albeit quite small) that a lapse in quality control or other calamity could result in an error so large as to attract unwanted, external attention to SCB, causing injury to the agency's reputation. This discussion was added to each interview to assess the chance that another high profile, CPI-type error could occur for any of the eight products. It also provided something of a review of the appropriateness of the initial risk assessment.

Detailed minutes were kept of all eight 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

Following the interviews, the minutes were reviewed and the point values assigned to the criteria ratings were refined. The evaluators met to discuss whether the ratings accurately reflected the information uncovered during the quality interviews and whether any adjustments to the ratings were needed. A few small adjustments were made, primarily to correct the scores for inconsistencies with the minutes. Then, for each product, the staff that attended the quality interviews were sent their ratings and the narrative that explained the ratings, and asked to correct any inaccurate or misleading information. On the basis of these reviews and inputs, the narratives s were corrected as suggested. In a few cases, the product ratings were reconsidered, adjusted, and finalised. These final ratings are provided in this report.

SUBSEQUENT QUALITY REVIEWS

The intent is to repeat this process annually to assess the current risks to data quality for each product, efforts that have been made over the years to address these risks, particularly for high and medium risk error sources, and to gauge the effectiveness of these efforts to improve the overall product quality.

3.3 STRENGTHS AND LIMITATIONS OF THE APPROACH

Any effort to evaluate for the accuracy of a system of processes as complex as these eight products will be flawed to some extent. Ideally, the effort should identify the most important threats or risks to the quality of a product and provide a mechanism for accurately measuring, over time, improvements to reduce these as well as new risks. These were our main goals in developing the approach. We believe the model described in this report is capable of achieving this ideal provided that the inputs to the process – in particular, the information needed to accurately assess each criterion – is accurate, complete, timely, and accessible by the evaluators. Ideally, a comprehensive set of documentation should be made available to the evaluators some days prior to each quality interview.

There are two important strengths of the current approach. First, the approach is thorough in that it covers all the important sources of error for each product. Second, the criteria used to assign the ratings for each error source, although still in need of some revision, were effective for identifying and assessing both apparent and hidden risks to data quality. Assuming the information shared prior to and during the quality interview is accurate and complete, we believe the current approach can be used to assign reliable and valid ratings. A weakness of the model is that it currently relies to a large extent on information about quality that is conveniently available prior to the interview and documentation that happens to exist because it was needed for other purposes. During the interview, the key product staff provided much additional information beyond the documentation or that may have been described somewhere in the documentation but was missed by the evaluators. For example, it may have been in Swedish and was not read thoroughly by the evaluators for that reason[7], or it may have been deeply buried in the written documents.

We believe a much better approach would be to prepare documentation, according to an agreed template, that is complete, accurate, and that directly addresses each criterion for each error source. This documentation could be made available to both the evaluators and the key product staff to prepare their assessments prior to the quality interview. If the reviewers are English speakers, it would be necessary to have an English version. These assessments could then be compared and discussed during the interview to arrive at the final, and potentially more accurate, product assessment.

[7] The evaluators used Google Translator to translate the documents – all or in part – from Swedish to English.

4 FINDINGS FOR THE EIGHT STATISTICAL PRODUCTS

Exhibit 3 provides the overall scores for the eight products by error source. To facilitate the exposition of the results, the error sources were consolidated into a single list which appears in first column of the table. The other columns of the table refer to the particular product being evaluated. Note that the interpretation of the error sources may vary across surveys, National Accounts, and registers. The reader is referred to the discussion of the error sources in Section 3.1 for the correct interpretations. The overall scores in the table are expressed in percentages. 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.

4.1 GENERAL OBSERVATIONS

Before discussing each product's detailed ratings, some general observations regarding the results in Exhibit 3 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 avoided for several reasons. First, as noted in Section 3, the aim of the model was to provide a baseline quality level for each product so that future improvements in product quality can be measured against this baseline. Thus, 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 most of these 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 SCB than a medium risk error source for another product if the latter product carries greater importance to SCB or official statistics. Further, although some checks were performed during the evaluations as well as in the days that followed to ensure the criteria were applied consistently, no attempt was made to score the products relative to each other or to force consistency among products. Doing so may have created unintended consequences for using the product scores as baselines for future evaluations.

Finally, the scores assigned to a particular error source for a product have an unknown, inherent level of uncertainty due to some element of subjectivity in the assignment of ratings. 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.

Despite these limitations for inter-product comparisons, we believe the results in Exhibit 3 can serve as a valid baseline for comparing a product's year to year improvements for the product's relevant error sources.

Close inspection of scores in Exhibit 3 yields the following observations:

- Measurement error appears to be the error source with the highest risk; it was rated a high risk for seven out of eight products.
- A close second was data processing error; this error source was at least "medium risk" for all products where this error source was applicable and was high risk for two products.
- By contrast, in terms of their quality ratings, measurement error and data processing error scored among the bottom three (only frame error is lower).
- The highest scores (say, 74 or above) were recorded for specification error, sampling error and nonresponse error; however, these scores were associated with only medium or low risk error sources.

In addition, the data collected in the evaluations leads to these further, general findings:

- Results of quality investigations for most products are not well-documented.
- Most quality evaluations tend to focus on error rates and indirect measures rather than MSE components [8](i.e., bias and variance measures).
- "Available expertise" and "compliance with standards and best practices" are generally rated higher than "knowledge of risks," "communication of these risks to users," and "risk mitigation planning." The latter three criteria appear more challenging to most products.
- Data processing poses some important risks in some areas such as data entry quality control, NACE coding, and editing because of the lack of evaluation of editing methods for most products.

[8] Mean Squared Error Components

Exhibit 3. Product Error-Level and Overall Ratings with Risk Highlighting

(Red Bold = High Risk, Black Bold = Medium Risk, No Bold = Low Risk**)**

	RS	CPI	FTG	LFS	NA	SBS	BR	TPR	Mean Rating
Specification error	**74**	**68**	**62**	66	**56**	**46**	62	44	60
Frame error	36	**42**	**62**	58	N/A	**62**			49
Overcoverage							**48**	**52**	
Undercoverage							**40**	34	
Duplication							46	64	
Nonresponse error/Missing Data	**62**	36	**62**	**66**	64	**74**	40	**60**	58
Measurement error/Content Error	**52**	**40**	**54**	**50**	**58**	**50**	**42**	50	50
Data processing error	**46**	**70**	**46**	**54**	**44**	**52**	N/A	N/A	52
Sampling error	N/A	**54**	N/A	**70**	**44**	**80**	N/A	N/A	62
Model/estimation error	**54**	**64**	**66**	**46**	**44**	**60**	N/A	N/A	56
Revision error	74	N/A	62	N/A	**62**	**58**	N/A	N/A	64
Total	**57**	**55**	**59**	**58**	**51**	**59**	**45**	**52**	55

Those ratings that are high risk (i.e. shown in red) but with a below average score could be regarded as the quality concerns most in need of attention from the SCB Executive. National Accounts is the product with most number of ratings in this category.

In the next section, we discuss the detailed ratings for all eight products individually.

4.2 OBSERVATIONS BY PRODUCT

ANNUAL MUNICIPAL ACCOUNTS

The Annual Municipal Accounts (RS) census faces a number of important risks to data quality. Chief among these are the risks to measurement error. As an example, home health care data is an activity that is highly integrated with other home care which makes it difficult to report as a separate cost. It also can be a source of item nonresponse requiring imputation. There may be other items that may be subject to measurement error and item nonresponse. The RS unit might wish to study the causes of these errors.

Editing error is another area of considerable risk given the large amount of editing that is being done. Due to the complex nature of the editing, editor error is a concern. This concern is heightened by the lack of quality control on the editing. An editor's work is not verified at the editor level although there are opportunities to identify the most egregious errors at the macro editing stage that follows editing. As previously noted, macro editing only identifies suspicious net errors and errors that are large enough to trigger a failure at the aggregate level. Furthermore, the cost-effectiveness of the editing is not being assessed.

We believe that more research should be devoted to understanding the errors associated with the RS data and how errors in this process propagate through the National Accounts to cause biases in the National Accounts estimates.

As a final note, there appears to be an appreciable risk of catastrophic error in RS. Specifically, errors in the disability care estimates in the RS statistics are more vulnerable to criticism because they influence the equalisation system for disability care services (LSS). What a municipality reports on this line as well as RS changes during the editing process directly influences the size of subsidy or fee municipalities receive. It may be important to continue the practice of carefully documenting contacts with municipalities for this row. In addition, more quality control of this row of the spreadsheet in particular would greatly reduce the risk of such an embarrassing error.

Exhibit 4. RS Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**74%**						**M**
	Frame error	**36%**						**L**
	Non-response error	**62%**						**M**
	Measurement error	**52%**						**H**
	Data processing error	**46%**						**M**
	Sampling error							**N/A**
	Model/estimation error	**54%**						**M**
	Revision error	**74%**						**L**
	Total score	**57%**						

For the CPI, the aspects of error risk that most need addressing are (a) the sampling errors in the CPI estimates, (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), an earlier study of about 10 years ago indicated the 95% confidence level error on the annual movements in the CPI as a consequence of the sampling process was plus or minus 0.4% . This is rather large in relative terms when the Riksbank target for the CPI is 2.0% and many payments, including some very large transactions, are indexed to the CPI. Furthermore, the sample size of the Swedish CPI is rather small compared with other countries. It may be appropriate to increase the sample size, at least for some products. It is important that the current study of re-estimating the sampling error is completed before a decision to this effect is taken so that, if the sample size is increased, it is allocated to priced items in an optimal way.

SCB does an excellent job managing substitution bias by the annual updating of weights and reviews of product lists. The index bias problems are elsewhere. With respect to (b), a variety of techniques are used. Clothing and footwear are two of the more problematic areas. Hedonic models were deployed with brand class used as the explanatory variable. This was rather unusual and we were not convinced that the models worked effectively. It was in footwear that the infamous error in the CPI occurred. Perhaps alternative approaches to managing the quality change should be evaluated.

With respect to (c), it may be appropriate to again evaluate alternative methods for estimating owner occupied housing. It is a problematic area but the nature of the housing market and data sources may have changed since the last evaluation.

With respect to (d), it is difficult to assess the size of measurement errors as there is no verification of the field workers except through the data editing processing which will detect the more significant errors. There is some evidence that it may be important. First, studies have shown that quality adjustments are typically prone to error. Second, a study undertaken when introducing hand held computers showed that there were larger than expected differences in the prices collected using the more traditional methods. Third, field controls are limited. Some of these problems should be reduced with the planned use of scanner data.

Exhibit 5. CPI Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**68%**	◒	◒	◎	○	○	**H**
	Frame error	**42%**	◓	◓	○	○	◓	**M**
	Non-response error	**36%**	●	●	◎	◓	◓	**L**
	Measurement error	**40%**	◓	◓	○	◓	◓	**H**
	Data processing error	**70%**	○	○	◎	○	◒	**M**
	Sampling error	**54%**	○	◓	◎	○	◓	**H**
	Model/estimation error	**64%**	◒	◒	○	○	○	**H**
	Revision error							**N/A**
	Total score	**55%**						

The aspects of error risk that most need addressing are (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.

With respect to (a), the asymmetry studies that have been conducted over the years suggest that commodities are being misclassified at high rates. Studies that would show the impact of these errors on the National Accounts and other important uses of the Foreign Trade (FTG) data are needed. With respect to (b), more information needs to be collected to shed light on the causes of the problems. A good starting point might be visits to businesses to observe how they capture this information in the data collection process and to better understand the information that is available to businesses when completing the Intrastat questionnaires. Editing errors (c) can be quite problematic for this survey, especially now that the Service Level Agreement with the National Accounts requires about six days to be cut from the processing schedule. Here it is important to know the extent of the errors, which commodity codes are most prone to error and at what reporting levels, and what is the impact on the National Accounts and other important uses of the data in terms of bias and variance.

With regard to (d), there are some questions as to whether the current method for converting to a statistical value is valid which relies on a Survey of Statistical Values that is conducted approximately on a five yearly basis. Our discussions with the National Accounts analysts revealed some scepticism about the current approaches. Currently, the National Accounts does not use the derived measures of statistical values as a result of this scepticism, even though it is the desired conceptual basis. A rigorous evaluation of the method is sorely needed since using invoice value in the National Accounts is a source of bias in the expenditure based GDP figures.

With respect to (e), Extrastat data are obtained from Customs. There did not seem to be strong monitoring of whether all the data had been received or not. There will be occasions when there will be missing data because of delays in data collection or processing. Furthermore, systems problems at Customs or in the data transfer arrangements may cause data to be lost. It is important that these possibilities are monitored, possibly using a macro-editing approach, as missing data may have a significant impact on the National Accounts or balance of payment statistics.

Finally, the lowest score recorded for the FTG was for data processing reflecting, in part, the lack of knowledge of errors from the editing process as well as the process for keying paper forms. The latter process (keying) maybe violating ISO standards in that there is no quality control for this operation.

Exhibit 6. FTG Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**62%**	○	◒	◒	◒	○	**M**
	Frame error	**62%**	◒	○	◒	○	◒	**M**
	Non-response error	**62%**	◒	○	◒	○	◒	**M**
	Measurement error	**54%**	○	○	○	◒	○	**H**
	Data processing error	**46%**	○	○	○	◓	○	**M**
	Sampling error							**N/A**
	Model/estimation error	**66%**	◒	○	◒	◒	◒	**M**
	Revision error	**62%**	○	○	◒	◒	◒	**L**
	Total score	**59%**						

Two LFS sources of error were determined to be high risk: nonresponse and measurement error. Nonresponse, now at about 25%, is a critical and growing problem in the LFS. Fortunately, register and administrative data are available to provide a rich set of auxiliary variables that can be used for nonresponse adjustment. These include the TPR, Swedish Public Employment Service's Register of Job Seekers, Employment Register, Income and Taxation Register, and the Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA). *Ignorable* nonresponse bias (i.e. nonresponse bias that can be largely eliminated by the use of auxiliary variables) appears to have been well-studied in the LFS. One problem that is difficult to study is the *residual bias*, i.e., the nonresponse bias that remains after adjustment (using the auxiliary variables) due to *nonignorable* nonresponse. To estimate the residual bias, nonresponse follow-up (NRFU) studies would be useful where a sample of the LFS nonrespondents are pursued by field interviewers who conduct face to face interviews. Conducting NRFU studies by telephone have never been successful. This information on the nonrespondents could be quite useful for evaluating the *nonignorable* nonresponse bias in the LFS as well as further examine the effectiveness of the existing adjustment process.

There is a lot of concern in SCB, and outside SCB, about the deteriorating response rates in the LFS. It is worth noting that over 50% of the non-response is due to non-contact and these households may tend to have special characteristics. Rather than reducing non-response, perhaps the focus should be on obtaining a representative sample when following-up non-response. Paradoxically, increasing the response rate may actually increase the nonresponse bias if the additional persons are more typical of existing respondents than nonrespondents. Some prior studies of this at SCB have demonstrated this paradox.

With regard to measurement error, one area of concern is the lack of monitoring of the telephone interviewers. It is standard practice in many NSOs to monitor some random portion (say, 10%) of all interviews in order to reduce interviewer variance, interviewer cheating, and ensure adherence to interviewer guidelines. Altogether eliminating telephone monitoring is not an acceptable way to reduce survey costs. Monitoring is a requirement in the ISO standard. We understand that the current problem is due to systems constraints but that should be rectified in 2012.

There are current plans to study the measurement error in the LFS using methods such as test-retest reinterview, record check studies (especially using the population register), and panel survey evaluation methods such as Markov latent class analysis and quasi-simplex models. We encourage this research activity and recommend that the evaluations focus on the magnitude of the measurement error and its causes. The largest benefit from such studies is to obtain information to inform ways to reduce the measurement error in the LFS.

Finally, some research is needed to evaluate the seasonal adjustment models that are currently in use. By the first quarter of 2012, LFS will be reporting about 1,500 seasonal adjusted series, a very large number. There is a risk that, with so many series, some of these adjustments are adding error and distorting the series. Perhaps the focus should be on adjusting the major aggregate series only.

Exhibit 7. LFS Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**66%**						**L**
	Frame error	**58%**						**L**
	Non-response error	**66%**						**H**
	Measurement error	**50%**						**H**
	Data processing error	**54%**						**M**
	Sampling error	**70%**						**M**
	Model/estimation error	**46%**						**M**
	Revision error							**N/A**
	Total score	**58%**						

The areas most in need of improvement are (a) the processing system, (b) knowledge of risks to accuracy associated with receiving data that was different to the ideal statistical concept, and (c) knowledge of the risks to accuracy associated with the various models used in the National Accounts especially the implicit model associated with the balancing mechanism.

With respect to (a), National Accounts may be at great risk because of the extended use of spreadsheets in their processing system, incomplete documentation and the lack of reliable, adequate IT support. This was identified as the most likely source of a catastrophic error for this product and perhaps for many of the other the products we evaluated. The National Accounts staff have studied the processing systems used by other NSOs but no definite plans are in place to replace the existing system. It is very important that professional IT staff are used in the development of any new system as well as national accountants who are experienced in implementing national accounting systems in other offices.

With respect to (b), there will be specification errors because it is not always possible for data providers to supply data that are consistent with the ideal statistical concept. A good example is the use of invoice value when statistical value is prescribed which means potentially greater adjustment factors are needed to balance the accounts. It is very important that there be analysis of the impact of such specification errors on the National Accounts especially when the risk from specification error is highest. An area of growing concern is the Foreign Trade estimates, where invoiced values are used rather than the conceptually correct statistical values.

With respect to (c), more could be done to analyse the robustness of the models that are used. Some of the imbalances are quite large and, whilst the RAS approach (i.e. bi-proportional adjustment using the marginal row and column totals of supply use tables) is used to eliminate those imbalances by adjusting the unknown values for components of the National Accounts, it depends on assumptions which may or may not be valid. This should be assessed. On the positive side, SCB is very transparent to users about the changes that are made as a result of the balancing process.

There were also concerns about the inconsistencies caused by the primary data sources using different survey frameworks. We have discussed this problem further in Section 5. There is a scope to use a common business framework with many of the primary data sources.

The relationship with primary data sources is not as close as it might be although the move to establish Service Level Agreements is a very positive step. The National Accounts is something of a ‘black box’ to the areas that provide the essential data for the National Accounts estimates. The National Accounts staff could be excellent macro editors if they were given access to preliminary data earlier or if they could work collaboratively with the source data providers during the initial macro editing stages. This would also help improve the consistency between National Accounts estimates and those published by the primary data areas. For example, the Australian Bureau of Statistics (ABS) has made steps to improve the collaboration between the National Accounts and the primary data areas on the preparation of estimates. Such collaboration has proven very effective and similar arrangements should be considered at SCB.

At a meeting with the main users of economic statistics (described in more detail in Section 7), one of the main criticisms was the lack of backcasting of the quarterly National Accounts except for a relatively short period. We were advised by the National Accounts that work on backcasting

was progressing, albeit very slowly, on this so hopefully it will be resolved in the not too distant future. Another criticism was that the National Accounts time series were only available in spreadsheet format. This made them difficult to use. It appeared to be a problem related to the existing National Accounts processing system so may be difficult to fix. However, there may be a private firm with the expertise to undertake the conversion and this possible solution should be investigated.

Exhibit 8. National Accounts Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**56%**	◓	◓	◡	◡	◡	**H**
	Frame error							**N/A**
	Non-response error	**64%**	○	○	◡	○	◡	**L**
	Measurement error	**58%**	○	○	○	○	◡	**H**
	Data processing error	**44%**	○	○	◓	◓	◓	**H**
	Sampling error	**44%**	○	○	○	◓	◓	**H**
	Model/estimation error	**44%**	◓	◓	○	○	◓	**H**
	Revision error	**62%**	◡	◡	◡	◡	◓	**M**
	Total score	**51%**						

Overall, we were quite impressed with the quality of this survey. The aspects of risk that most need addressing are (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.

It seems likely that editing error comes largely from manual editing and poses a high risk relative to other error sources. There are many different data sources and subsystems. Staff members that are unaware of the whole chain could make mistakes because they may also be unaware of the consequences for the other survey processes. A study was conducted comparing data before and after editing. However, its focus was on error rates not on the bias per se so it was limited to that respect. A study is needed that also looks at the effects of editing on bias and variance components so the cost-error trade-off of the editing process could be better understood.

We noted that data entry by keying is not following ISO-standards because there is no validation of the accuracy of keying. The only mechanism for catching keying errors appears to be macro editing which may only trap net errors of the most egregious nature.

Macro editing is undertaken prior to the release of estimates. This is consistent with good practice. After the release, the National Accounts do their own macro editing by confronting the data with other parts of the accounts. They have special knowledge and there would be benefits if this could be done prior to the release of the structural business statistics. This would have the further advantage of greater consistency between the National Accounts and structural business statistics which would please many users. These issues should be discussed before the Service Level Agreement between the two areas is finalised.

There are some systemic differences between preliminary and final estimates. There should be more analysis of the reasons. Assuming the final estimates are more accurate, it may be possible to make some adjustment to the assumptions made at the time of the preliminary estimates in order to eliminate this systematic bias.

Exhibit 9. SBS Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**46%**	◔	◔	◡	○	○	**M**
	Frame error	**62%**	◡	○	○	◡	◡	**M**
	Non-response error	**74%**	◡	◡	◡	◡	◉	**M**
	Measurement error	**50%**	○	◔	○	○	○	**H**
	Data processing error	**52%**	○	◔	◡	○	○	**H**
	Sampling error	**80%**	◡	◡	◉	◡	◡	**M**
	Model/estimation error	**60%**	○	○	◡	◔	◡	**H**
	Revision error	**58%**	○	○	◉	◔	○	**H**
	Total score	**59%**						

BUSINESS REGISTER

For this product, our discussions focused on the Statistical Register not the Public Register. We believe the areas most in need of improvement are (a) the over-coverage caused by not being able to remove inactive and defunct enterprises because of systems limitations, and (b) the inaccuracy of NACE industry coding.

The issue with concern (a) is the lack of capability to address this problem which causes problems to Business Register users. It should be addressed by the development of the new Business Register which we understand is at the project planning phase.

A new register system also provides an opportunity to rethink the processes involved in populating the Business Register and extracting frames for use within SCB. We strongly recommend SCB seize this opportunity to deal with the current major weaknesses of the BR. As an example, we heard several times about the problems caused to the National Accounts by inconsistent frames used by primary data source areas. The redesign of the BR is an opportunity to address this problem. For example, the ABS puts considerable effort into deriving a Common Business Framework (quarterly and annual) from the Business Register. This is to be used by all the relevant collection areas for the selection of their samples. SCB might wish to consider developing such a framework.

As with other collections, we asked about possible catastrophic errors. We were not able to identify anything of this nature – there were a number of checks and balances in place. However, there is an emerging concern which could become quite serious. More legal entities are now doing their own NACE coding. There is less control over this so there are likely to be accuracy problems. Furthermore, there may be incentives to code inaccurately if there are tax advantages. If, for example, too many businesses were coding to manufacturing and this was reflected in the Business Register than this may lead to estimates of growth in manufacturing that are higher than they should be. There is a subsequent risk of policy misinterpretation. It is important that SCB collaborate with the Tax Office to find a way of resolving this problem. It is in the mutual interest of both organisations. This is the thrust of our concern (b). At present, there is little quantification of the accuracy of NACE coding, hence little communication with users and no plans to mitigate this potentially serious risk.

Something else that might be worth considering is the simplification of the Units Model. The current Units Model contains three levels – enterprises, activity units and establishments. It is very difficult to maintain establishments so consideration should be given to reducing to two levels by eliminating establishments. Several other countries have done this.

Exhibit 10. BR Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**62%**	◓	◓	◒	◒	◒	**L**
	Frame error: overcoverage	**48%**	○	○	◓	◓	◒	**M**
	Frame error: undercoverage	**40%**	◓	●	○	○	◓	**M**
	Frame error: duplication	**46%**	●	●	◒	◒	◓	**L**
	Missing data error: item and variable	**40%**	◓	◓	○	○	◓	**L**
	Content error	**42%**	◓	◓	◒	○	◓	**H**
	Total score	**45%**						

The highest risk area for the TPR is overcoverage of the population – i.e., the inclusion of persons on the register who should be excluded because they do not meet the criteria for inclusion. Registered persons who leave Sweden with no intention of returning remain on the register until the tax authorities can verify their status and remove them. Overcoverage may be quite large for some subgroups and can create issues for surveys that use the TPR as a frame. Overall, it is estimated that about 35,000 persons on the register are overcovered; exact figures are difficult to obtain.

In addition, specification error is a medium-level risk for the TPR. There are several issues. One is the difference between an individual's registered address and their current residence. For surveys, the latter is the more important for contacting purposes but the former is on the TPR. The extent of the problem is not well quantified at present. It may be contributing to the relatively large non-contact rate in household surveys. Another issue regards persons having dual citizenship. It is currently impossible to record more than one citizenship on the register, yet dual citizenship is important for some TPR users.

Finally, 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) has not been explored.

To some extent, TPR evaluations cannot proceed independently of the main users of the TPR because what is important is the impact of TRP error on statistics produced by the surveys that use it. Therefore, we believe it is important for the TPR staff to work collaboratively with the users of the TPR on the design of evaluation studies to assess the impact of TPR errors on key estimates produced by surveys. TPR staff could lead some of these projects and would certainly be major players in many of them because they know the structure of the register and are most familiar with the registers strengths and weaknesses. They also can provide important contacts within the Tax Office for evaluation work that requires their cooperation.

Exhibit 11. TPR Ratings Summary by Quality Criteria and Error Sources

	Error Source	Average score	Knowledge of Risks	Communi-cation to Users	Available Expertise	Compliance with standards & best practices	Plan towards mitigation of risks	Risk to data quality
Accuracy(control for error sources)	**Specification error**	**44%**	◓	◓	○	○	●	**L**
	Frame error: overcoverage	**52%**	○	○	◓	○	◓	**H**
	Frame error: undercoverage	**34%**	◓	◓	◓	○	●	**L**
	Frame error: duplication	**64%**	○	○	◒	◒	◓	**L**
	Missing data error: item and variable	**60%**	○	○	◓	○	◒	**M**
	Content error	**50%**	○	○	○	◒	◓	**L**
	Total score	**52%**						

5 SOME CROSS-CUTTING METHODOLOGICAL AND OTHER FINDINGS

During the product interviews and discussions with SCB staff, a number of methodological issues came to our attention. In this section, we provide our thoughts on these issues with hopes that they will generate further activity at SCB. As a caveat, we have not undertaken the background research or detailed consultations with SCB staff regarding these recommendations. Therefore, our comments may seem uninformed to those closest to the issues. Nevertheless, we still note them, in no particular order and without going into much detail, in case they are of interest. We welcome further discussion on these issues or will provide more elaboration if SCB wants to pursue any of the suggestions.

5.1 INTEGRATION OF ECONOMIC STATISTICS

There is more that could be done to improve the coordination of economic statistics from a methods point of view. For example, there is not a common framework used for the sub-annual business collections and the National Accounts advise that this is a contributor to the inconsistency across the different primary sources. Furthermore, there appear to other design differences that may not be necessary.

There seems to be too much reliance on balancing in the National Accounts. The extent of the difference between production and expenditure estimates seemed rather large and this might affect the effectiveness of the balancing method. Research on the sources of the discrepancies between the two methods of estimating GDP should be an ongoing activity to continuously improve the agreement between the production and expenditure based estimation.

The ABS faced similar problems in the 1990s but some serious problems with the National Accounts resulted in investigations to identify the cause. The lack of statistical integration across the statistical collections was an important factor and steps were taken to harmonise designs. In particular, common business frames were introduced. A new frame is provided each quarter (updating for births, deaths and other changes) and all collections are required to use it. It can facilitate sample rotation and the management of the overlap of the sample across different collections. A similar schema could be considered for SCB.

It might be worth looking at this before the design of the new Business Register is finalised. If it is designed correctly, it might provide an opportunity to improve co-ordination of frameworks across collections.

5.2 LACK OF CO-OPERATION BETWEEN THE NATIONAL ACCOUNTS AND STATISTICAL AREAS

There is a reliance on Service Level Agreements at present. These are a positive step although progress towards their implementation seems rather slow. However, even with these arrangements, the National Accounts will remain something of a black box. The relationship between the National Accounts and the primary data source areas seems more estranged than in most developed statistical offices.

The National Accounts are in a great position to be highly effective macro editors. As well as their profound knowledge of economic activity in Sweden they can confront data through the National Accounting framework. However, this should ideally be done before the release of data from primary data sources not after the release. There are several advantages, including: (1) the output

from the primary data sources will be more consistent with those of the National Accounts, (2) the output from the primary data sources will be improved, and (3) there will be a better understanding of the National Accounts requirements which can only lead to improvements in the primary source data especially from the National Accounts perspectives.

The ABS moved in this direction in the mid 1990s and it led to a significant improvement in the alignment between the primary data sources and the National Accounts as well as the quality of the primary source data. It has been widely applauded by the users.

5.3 ACCURACY OF NACE CODING

It appears from what we heard there is likely to be deterioration in the accuracy of NACE coding. To some extent it is because the Tax Office has not as great an interest in the accuracy of NACE coding as SCB. However, the recent steps to allow business to do their own coding could accentuate the problem particularly if there are tax incentives to code themselves to certain industries like manufacturing. The inaccuracies in NACE coding are likely to accumulate over time unless there are special collections to obtain the information on which to reassess NACE coding.

The likely deterioration in industry coding will have important consequences for data quality. First, it will lead to businesses being allocated to the wrong industry. If this is non-random, there will be an upward bias in certain industries and a downward bias in other industries. For example, if more businesses are allocated to manufacturing than actually is the case then there would be an upward growth in the estimate of manufacturing even to the extent that inaccurate economic assessments are being made. The accuracy of the production accounts and production indexes will also be at risk. The NACE coding on the Public Register would also be affected by this deterioration in coding.

The accuracy of samples will also be at risk. Although the Structural Business Survey collects the 'correct' industry, it affects the accuracy of the sample design if businesses are allocated back to a correct industry. This does not affect the largest 500 enterprises as they are completely enumerated. We are not sure whether the corrected NACE code for these largest businesses is fed back to the Business Register.

What can be done about this? We think it is necessary to obtain an industry description to enable SCB to confirm the accuracy of coding. This will require the co-operation of the Tax Office. They need to be convinced. They are more likely to be convinced if there is a risk to the accuracy to the tax base or they start using NACE based statistics to help manage their own activities.

5.4 EVALUATION STUDIES

SCB has world-class capabilities in survey methodology and statistics. It has excellent resources for undertaking evaluation studies of various aspects of statistical quality to better under quality risks, the cost-effectiveness of methods, user needs, etc. There are evaluation studies undertaken but they tend to be of 'error rates' rather than understanding the impact on bias or variance. For example, the cost-effectiveness of editing systems to improve accuracy is largely unknown. There seemed to be scope for a corporately agreed evaluation program focusing on those areas where the studies are likely to lead to significant improvements. Evaluation studies on the cost-effectiveness of data editing seemed to be one area that was potentially rewarding.

As an example, we previously noted that measurement error provides a high risk to seven of the eight products in this evaluation. This suggests that measurement error evaluation should be a

high priority research area at SCB. It might be more effective to design a coordinate research program, combining the talents of methodologists across product areas who would work collaboratively and share results on topics related to measurement error evaluation. As examples, methods for reinterview surveys, administrative record check studies, latent variable error models, and other measurement error evaluation methodologies could be pursued as well as the results from studies that implement these methodologies. The primary goal of the coordinated effort would be reduce measurement error through improved data collection methodologies across all SCB surveys.

Similar efforts should be devoted to understanding and reducing data processing error: another high risk to data quality that scored low overall in our reviews.

5.5 NONRESPONSE IN HOUSEHOLD SURVEYS

We heard several concerns about response rates most notably about the Labour Force Survey and the Household Budget Survey. Non-response rates have deteriorated, as has been the case with many countries, and it may be difficult to greatly improve them. In fact, the efforts to increase may reduce the representativity of the sample. We are not aware of the situation in Sweden but typically response rates are highest in the middle income groups and lower in the lower and upper income groups. Non-response rates (especially non-contact) are particularly high for young adults who are more mobile and harder to contact. If the additional responses, obtained as a result of nonresponse follow-up, are not from the under-represented population groups it will increase rather than decrease non-response bias. This caused us to ask the question of whether non-response in the LFS being managed effectively or not but we did not have enough time to study this in detail.

There is another reason why responses may be more difficult to obtain in the future. With the rapid advent of telephone marketing, people are becoming more anxious about answering the telephone where it is 'unknown caller' or they don't recognise SCB.

We understand a large project on nonresponse is currently being undertaken in SCB. No doubt there will be many interesting and useful findings. In interpreting these findings and deciding what actions to undertake, we suggest that SCB keep in mind that the most important objective is to obtain a representative sample which does not always occur through efforts to increase response rates.

In the U.S., there has been considerable interest in applying two-phase sampling strategies to increase the weighted response rate in ways that minimise nonresponse bias without increasing survey costs. Two-phase sampling involves conducting an initial survey phase where all sample members are pursued. However, later in the survey period, only a subsample of the nonrespondents is pursued – the so-called second phase. Combined with the first phase interviews, the second phase interviews will usually produce a higher *weighted* nonresponse rate than could be achieved with a single phase design. This is due to redoubling the interview effort for a smaller (for e.g., 50% subsample) of the nonrespondents. Also, by reducing the size of the nonresponse follow-up sample, the data collection costs are no greater than the single phase approach. We encourage SCB to investigate this approach for the LFS, possibly conducting the second phase by face to face field methods to maximise response rates.

5.6 RELATIONSHIP WITH THE TAX OFFICE

When discussing major concerns with the Register areas, the biggest concern seemed to be changes in tax forms without prior consultation. Although there seemed to be a generally good working relationship with the Tax Office, and they were a reliable provider of data, it did raise questions about whether the relationship might be strengthened. For example, in Australia there is a Memorandum of Understanding which among other things states that changes to tax forms cannot be made without prior consultation. Proposed changes have often been modified as the result of these consultations. The Memorandum also outlines service level standards. To give the Memorandum of Understanding additional status the heads of the Tax Office and the Statistics Office meet at least once a year to review progress against the different activities listed in the Memorandum.

5.7 POLICY ON CONTINUITY OF STATISTICAL SERIES

It is necessary to redesign collections from time to introduce new methods, new standards or improve efficiency. Although this might improve the collection in many ways, it can be affect continuity and impinge upon the Comparability dimension of survey quality. Users want to be able to bridge the time series before and after the redesign. Discussions suggest there is no policy in SCB on the continuity of series and this may lead to some unfortunate decisions made at the time of the redesign. We suggest SCB's policy specify that every major redesign include some provision for bridging the data series before and after the redesign unless an explicit exception is granted by the Director General. The bridging methodology can take many forms. In some important series like the National Accounts where time series are particularly important to users, the bridging solution should be backcast many years to provide a reasonably continuous series across the break.

5.8 IMPROVING THE RELATIONSHIP BETWEEN IT AND THEIR CLIENT AREAS

There seemed to be considerable frustration about the current arrangements for managing IT applications systems support and development. Whilst there is great sense in centralizing IT from the points of view of technical leadership, consistency of approach (e.g. compliance with agreed IT standards), and more efficient use of resource, it might be possible to maintain these important objectives and provide a more satisfactory experience to client areas if the IT specialists were physically relocated to the client area and for a reasonable period of time before being relocated to another client area. In his way the IT experts will develop a more expert knowledge of the systems they are supporting.

5.9 LACK OF TELEPHONE INTERVIEWER MONITORING

Another area of concern from a data quality perspective is the lack of monitoring of telephone interviewers, particularly in the LFS. Telephone monitoring of telephone interviewers serves multiple purposes. First, monitoring that includes a timely and effective performance feedback loop has been shown to reduce interviewer effects in survey data. It can also be a type of on-the-job training to assist interviewers to continuously improve interviewer performance over time. Through monitoring, survey managers can better understand the strengths and limitations of the questionnaire and interviewing procedures. This information can be very useful for future survey redesigns. Monitoring can detect and deter most forms of interviewer falsification which is a problem for all surveys world-wide. This is particularly an issue for interviewers who work from home. Finally, the lack of telephone monitoring is noncompliant with ISO standards and best practices SCB should consider monitoring at least 5% of all telephone interviews, perhaps more for inexperienced interviewers.

5.10 DEVELOPMENT OF QUALITY PROFILES FOR KEY PRODUCTS

As previously noted, an important deficiency of the current product quality evaluation methodology is its reliance on possibly incomplete documentation, fragments of reports, and anecdotal information for input into the evaluation process. If this information is incomplete, inaccurate, or out-of-date, the reliability and accuracy of the evaluation process could be compromised. One possible solution to this problem is to create a quality profile for each major product that is to be routinely evaluated.

A quality profile is a report that provides a comprehensive picture of the quality of a statistical product, addressing each source of error that is applicable. It reviews and synthesizes all the relevant information that has accumulated over the years for each source of error. The quality profile would provide essential information to quality evaluation process in a consolidated, comprehensive, and accurate manner which would greatly improve equity and reliably of the process. A well-written quality profile would provide essentially all the input required to objectively and accurately apply the criteria developed for the evaluation model.

However, the quality profile has several other important uses that add to its value. For example,

- It describes in some detail the survey design, estimation and data collection procedures for the survey.
- It provides a comprehensive summary of what is known for the survey for all sources of error – both sampling as well as nonsampling error.
- It identifies areas of the survey process where knowledge about survey errors is deficient and may recommend areas in need of improvements to reduce survey error.
- It can also be used to suggest areas where further evaluation and methodological research are needed in order to extend and enhance knowledge of the total mean squared error of key estimates and data series.
- It can be used as a training manual for staff who want to understand what is known about product quality.

We recommend that SCB consider developing quality profiles for the eight products in this review and to use this documentation in the next quality evaluation cycle. The quality profile for an individual product can be built up over time. Even if incomplete, it can still provide a useful source of information.

6 IMPROVEMENTS TO THE QUALITY EVALUATION MODEL

This review was a pilot test of the proposed approach. It worked well but could be improved in several respects. Some suggestions were made in other sections of this report.

Not surprisingly, some concerns have been expressed by the product areas. These are outlined below together with our reactions to these concerns.

1. *Ratings are subjective*. All ratings are inherently subjective. The guidelines that were developed for the assessments go a long way towards achieving consistency and objectivity but in the end, judgments are still involved. We believe that the fairness of the review process is enhanced by giving the product areas an opportunity to comment on their respective ratings.

2. *Ratings are based upon too little information and may not reflect the true situation*. We have also mentioned this concern in this report. Clearly, the better the information provided to the evaluators, the better the assessments will reflect reality. We think this comes down to better documentation which is what we recommend for the future assessments in our report (see section 5.10).

3. *It is not well understood how the evaluation results will be used*. We think these assessments should be treated as one would the results from a pretest. The current trial tested several aspects of the evaluation process – applicability to a wide range of products, appropriateness of the criteria we used, performance of the assessment guidelines, acceptance by staff of the process and results, and so on. However, we should stop short of stating that the scores assigned to each product accurately reflect the real quality of the product. It would take much better documentation or much more than a 4 hour interview to achieve that level of accuracy. We suggest that the validity of the scores be verified before decisions are made to act on them.

4. *The evaluation may have been too ambitious in scope and timing*. We agree that it was ambitious. However, we learned a lot about the effectiveness of the model, how it might be improved, and many strengths and weaknesses of the eight products. Despite the short time for review, we think we have also identified some important areas for improvement by the product areas. Indeed, the feedback we have received on the evaluation process including comments from the product areas that were being reviewed, has been largely positive. If SCB decides to proceed with this approach, we believe that the shortcomings of the current process could be satisfactorily addressed in time for the next assessment.

5. *Better feedback is needed to know what scores to raise and how to raise them*. The product areas also want more detailed feedback as to why they receive the scores they did and what they can do to increase their scores. We have tried to make some suggestions on areas that may need improvement but with better documentation and more time, the quality of the feedback would increase. This evaluation should not be the only feedback products receive from management and others regarding data quality. Rather, this assessment is just one indicator of quality. Further investigation should follow this assessment to verify that improvement is needed and to determine exactly what needs to be done and that the investment in quality improvement will be cost-effective.

7 NEXT STEPS – THE OTHER QUALITY DIMENSIONS

We had plans to look more closely at the Relevance dimension of survey quality during our time at SCB; however, it was decided that thoroughly investigating the Accuracy dimension was a top priority and, consequently, our time devoted toward evaluating Relevance was quite limited. Nevertheless, we managed to achieve two important accomplishments.

First, we held a meeting with some key users – mostly economists from the Riksbank, Nordea, the Ministry of Finance, and the National Institute for Economic Research. Their main use of SCB data was to forecast economic trends so, in that sense, it was a relatively narrow set of users. The discussion was mainly around National Accounts and a summary is provided in the Annex. Although our main purpose for the discussion was to gain some insights into Relevance, the discussions focused mainly on other Dimensions such as Accuracy, Comparability, Coherence, and Accessibility. Interestingly, they were not asking for more macroeconomic accounts statistics. They just wanted the existing statistics to be improved. Given the strong user interest in the work of the SCB, it raises the question of whether it is worth considering the establishment of a macroeconomic statistics user group.

Second, we extended the Accuracy evaluation model to the Relevance dimension and developed evaluation criteria that specifically referenced Relevance. These criteria are shown in Exhibit 12. These criteria could be used for a self-assessment by SCB staff, facilitated by an experienced methodologist such as Heather Bergdahl. It might be treated as a pilot study with one of the purposes obtaining feedback to enable the 'model' to be improved. Also, it would be good to have the input of the key users of these statistics. User needs regarding Relevance may already be known through existing user forums. Otherwise special user group meetings could be scheduled and could cover all the user dimensions of survey quality, not just Relevance.

Exhibit 12. Evaluation Criteria for the Relevance Dimension

	●	◔	○	◕	◎
	Poor (1,2)	Fair (3,4)	Good (5,6)	Very Good (7,8)	Excellent (9,10)
Knowledge of Risks	There is no knowledge of users or user needs. There are no measures of user satisfaction.	Key users are known and arrangements have been made to hold discussions with users on their needs. There are plans under way to conduct user satisfaction surveys.	User satisfaction surveys are conducted on a regular basis that informs this program of statistics. Work has been undertaken to document user needs.	A draft document has been prepared to describe the users, particularly the key users, and outline a plan for meeting the key user needs that have not already been met: an Information Development Plan.	An agreed Information Development Plan has been prepared and distributed to stakeholders.
Communication with Users	There is no effective communication with users.	There is communication but it is essentially passive in nature e.g. information has to be obtained by users searching the SCB website.	Communication is active and electronic media are used effectively to alert users of any new developments.	There is regular communication with government users and documentation of their key needs including any unmet needs.	An active network of government and other users is in place.
Meeting User Needs	There are no staff working on the program that are familiar with the methods required to develop the statistical collections so they will satisfy the unmet needs or there are no plans to meet the unmet user needs.	Staff with the required expertise have been identified or there is a plan in place to acquire the expertise.	Staff with the required expertise could be made available to the program. There are plans to meet the most important unmet needs and these are included in the SCB work program.	Significant progress has been made towards meeting the unmet needs.	The most important unmet needs have been met.

Although the focus of our activities has been primarily on Accuracy and somewhat on Relevance, all the quality dimensions should ultimately be considered. Thus, as we did for Relevance, criteria for evaluating the other quality dimensions – Comparability, Coherence, Timeliness and Accessibility – could also be developed. These criteria will likely be simpler than the model used for Accuracy. Of course, any self-assessment on these dimensions should also reflect the views of users.

The process to "optimise" survey quality will inevitably encounter trade-offs between the dimensions. For example, a survey redesign may improve relevance and accuracy, but comparability may be diminished or even compromised unless a method of bridging the new and old data series is provided. Another common trade-off is between accuracy and timeliness saving calendar time often means reducing efforts to improve data quality. Arriving at the optimal balance of all the quality dimensions requires careful planning and regular feedback from both internal and external users of the data.

8 CONCLUSIONS AND RECOMMENDATIONS

SCB remains a world class statistical organisation. It would not commission studies of this type if it were not interested in continual improvement. We were considerably impressed with the quality of the work done on the eight products in our review. But not surprisingly we have identified a number of areas that require improvement that are discussed in this report.

Our proposed approach was pilot tested on eight products. We believe it worked very well although it can certainly be improved. It seemed to be well-received by the Process Department, Research and Development Department and the statistical areas. We think this is because the process used for reviewing quality was viewed as objective and comprehensive while providing lessons learned that may not have realised with more traditional methods for evaluating quality. It also produces numerical scores both by error source and overall which should increase over time if quality improvements are made or decrease if the opposite occurred. Furthermore, this approach identified the highest priority areas for quality improvement and these areas appear to be consistent with prior beliefs at SCB regarding where improvements are needed. We recommend this approach be adopted for future evaluations after it has been modified slightly from things we learned during this pilot test. As noted above, one key area for improvement is the documentation on data quality which could be greatly enhanced using the quality profile approach.

The instruments we have used should be translated into Swedish and this may be the opportunity to also use language in some parts that may be more familiar to the staff of SCB. We are happy to work together with SCB on this process.

As mentioned above, the process will identify the highest priority areas for improvement. This can be done by the areas responsible for managing the eight products. However, it can also be done corporately and may be a way of identifying projects that might be allocated additional funds. These would be the high risk areas where the rating was less than good. These can be readily identified using graphical presentations in Exhibits 4-11.

We suggest the products be reviewed again in approximately 12 months time using an independent assessor who would work closely with the product areas. SCB should also identify other collections for self-assessment. These do need to be facilitated and Heather Bergdahl is well placed to do this. The priority should be on the most important collections.

We also made some suggestions on changes on methodology and ways of tackling some important statistical problems. Please contact us if you want to discuss these further.

The work on ISO standards is important and should be encouraged. It is important that there is agreement within SCB on the standards to be applied to statistical operations.

Finally, we would like to thank SCB for the opportunity to work on this project.

Although there were some criticisms of Statistics, it was clear that these users were strong supporters of SCB but would like it to improve its performance in some respects. In fact, it appeared that they would regard improved performance as a higher priority than increasing the range of statistics. It was suggested that SCB puts too much emphasis on relevance and not enough on comparability.

Some of the key points made are as follows. Much of the discussion was around the National Accounts.

1. Comparability across time is very important. Most of the participants were involved in economic modelling. As there were no backcast quarterly series except beyond 1993, they were forced to use OECD data instead for longer time series even though the OECD methods were rather crude. They felt that, as a matter of policy, SCB should always provide backcast series when making major revisions to the National Accounts. They suggested SCB should not be too ambitious on the detail that they publish when backcasting.

2. There was also criticism of the delay in the provision of backcast series for the labour force following the 2005 redesign.

3. More generally, there was criticism of the lack of information when methods are changing.

4. The lack of consistency between the National Accounts and the primary data sources was also an area of concern. A special mention was given to the lack of consistency between the National Accounts and the LFS and the impact on measures that combine the two products such as labour productivity.

5. They were very critical of the SCB web site. They found it confusing and it was difficult to find some time series. This was in due to a lot of data being stored in spreadsheet format (Excel). Although they want to be able to download data into Excel, they did not think it should be used for data storage on the web site.

6. There was some discussion on recent delays to the publication of the LFS. Whilst they recognise it may be necessary to do this from time to time it is an inconvenience to users unless more notice is given than in the recent incidences.

7. They were critical of the long delay until the publication of the final annual National Accounts. There seemed to be reasonable contentment with the release times of the quarterly National Accounts.

8. However, they were very supportive of the amount of information SCB provided on errors compared with other national statistical offices.

9. They were critical that a lot of the data provided by SCB to Eurostat did not appear on the Eurostat web site.

10. There was strong support of SCB’s independence from politicians