VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 381–415, http://dx.doi.org/10.2478/JOS-2014-0022

A System for Managing the Quality of Official Statistics

Paul Biemer[1], *Dennis Trewin*[2], *Heather Bergdahl*[3], *and Lilli Japec*[4]

This article describes a general framework for improving the quality of statistical programs in organizations that provide a continual flow of statistical products to users and stakeholders. The work stems from a 2011 mandate to Statistics Sweden issued by the Swedish Ministry of Finance to develop a system of quality indicators for tracking developments and changes in product quality and for achieving continual improvements in survey quality across a diverse set of key statistical products. We describe this system, apply it to a number of products at Statistics Sweden, and summarize key results and lessons learned. The implications of this work for monitoring and evaluating product quality in other statistical organizations are also discussed.

Key words: Total survey error; process control; GDP; quality indicators; statistical standards.

1. Introduction

Official statistics include the data and estimates that are published by national statistical offices (NSOs) and other public organizations on the major areas of society and the economy. They provide both quantitative and qualitative information on economic and social development, national productivity, living conditions, health, education, transportation, the environment, and many other areas of national interest. Credibility and confidence in the statistics depends to a large extent on the quality of official statistics. If the quality is suspect, the NSO's reputation as an independent, objective source of trustworthy information could be undermined. Therefore, managing the quality of statistical products is a key objective for all NSOs.

Quality is a vague concept that has become over-used in the literature and a more precise definition is required for the purposes of this article. Here, we define the quality of official statistics in terms of five dimensions that reflect their fitness for use by data users and other constituents. These dimensions, which will be described in more detail subsequently, are: Accuracy, Relevance/Contents, Timeliness & Punctuality, Comparability & Coherence, and Accessibility & Clarity. This article considers all five dimensions but primarily focuses on Accuracy or *data* quality which is considered fundamental to product quality. After providing a brief background for this work, the

[1] RTI International, P.O. Box 12194 Research Triangle Park, NC 27709-2194 North Carolina 27709, U.S.A. Email: ppb@rti.org
[2] Former Australian Statistician, Canberra, Australian Capital Territory, Australia. Email: dennistrewin@grapevine.net.au
[3] Statistics Sweden, SE-70189 Örebro, Sweden. Email: heather.bergdahl@scb.se
[4] Statistics Sweden, P.O. Box 24300, SE-10451 Stockholm, Sweden. Email: lilli.japec@scb.se

article considers a process for continually monitoring, evaluating, and improving quality over time across a diverse set of key data products.

NSOs world-wide are struggling to maintain high quality products as operating budgets continue to decline (see, for example, Struijs et al. 2013; Nealon and Gleaton 2013; Seyb et al. 2013). In fact, the March 2013 issue of the Journal of Official Statistics (JOS 2013) was devoted to cost-effective system architectures for producing high-quality statistics. In 2011, with guidance and support from Statistics Sweden we developed a structured, systematic approach for guiding the quality improvements in the agency's statistical programs and assessing the effects of these improvements on product quality. Referred to as ASPIRE (*A S*ystem for *P*roduct *I*mprovement, *R*eview, and *E*valuation), this approach provides a comprehensive framework for systematically evaluating all dimensions of quality with the primary focus on Accuracy. ASPIRE is quite general and can be applied in essentially any NSO or other statistical organization that supplies a continuous flow of statistical data to a community of users such as economists, researchers, government planners, and policy developers.

ASPIRE comprises an exhaustive inventory of potential risks to data quality for the products being reviewed and evaluates the organization's efforts to understand and mitigate these risks through evaluation studies and process improvements, assigning higher priorities where there are higher risks. The approach imposes a high standard of excellence on products based upon the best practices in the field while objectively and consistently rating products against well-specified quality standards or criteria. The ASPIRE framework provides an integrated approach to quality and risk whilst bringing rigour and heightened objectivity to assessments that might otherwise be based on subjectivity and intuition.

ASPIRE incorporates a number of unique features that may be considered new and innovative in the survey evaluation literature. First, ASPIRE goes beyond assessments that are based solely on compliance with statistical standards. Rather, it encourages continual improvements (both incremental and breakthrough improvements) in areas that represent the highest risks to data quality and thus motivates product excellence. Second, it provides numerical scores by error source, by criterion, and overall error sources and criteria that reflect product and process quality and that can be used for comparisons across time and products. Finally, ASPIRE provides a graphical presentation that can be readily understood by workers, managers, and administrators at all levels. It can communicate a general overview of quality simultaneously across numerous products or be used to "drill down" to view the evaluation details by product, by error source, by criterion level, or by any combinations of the three. Cost optimization is not the goal of ASPIRE; however, it does provides valuable information for cost-benefit analysis.

The first implementation of ASPIRE (referred to as Round 1) was conducted in 2011 for eight key statistical products at Statistics Sweden. This review provided a baseline for measuring improvements for these products in subsequent ASPIRE rounds. In 2012, Round 2 of ASPIRE was conducted for the same eight products while two additional products received an initial review. A third ASPIRE round on these ten products was completed in November 2013. This article presents the theory underlying the ASPIRE methodology, describes the process and its components, and mostly uses the experiences from Round 1 and 2 implementations to illustrate the application of ASPIRE. Further

refinements to the methodology were made in Round 3 but these were relatively minor in nature.

The next section provides an overview of the literature on quality of official statistics and lays the theoretical foundations for ASPIRE. Section 3 describes the ASPIRE approach in some detail including the basic criteria used in the evaluations, scoring system, and methods for ascertaining risks. Section 4 describes how ASPIRE was applied to a number of products at Statistics Sweden in 2011 and 2012 and summarizes some of the key results. Finally, we conclude the article with a discussion of the ASPIRE approach based upon our experience to date and plans for future implementations and evaluations of the methodology.

2. Total Quality

2.1. Product, Process and Organizational Quality

NSOs and other statistical organizations have a long history of addressing various aspects of quality. The concept of quality has evolved over the years to become increasingly complex (Lyberg 2012). Today, we might view quality on three different levels, product, process and organization (Lyberg et al. 1998; Lyberg and Biemer 2008), each with its own set of assessment approaches. These quality levels can only be summarized here; however, Lyberg et al. (1998) describes them in some detail.

Product quality refers to the acceptability of a product (for example, an estimate of the unemployment rate) for its intended uses (for e.g., to monitor job loss/growth in the economy). Improvements in product quality are made by improving the processes generating the product. Thus process quality refers to the ability of survey processes to generate data and other statistical products of high quality. It is important that NSOs possess the knowledge, skills, and appropriate control systems to sustain and improve process quality. Organizational quality refers to the ability of the organization to consistently develop and maintain high quality processes. These three quality levels do not exist independently. Rather, organizational quality is required to achieve quality at the process level which is required for consistent product quality.

As an example, Statistics Sweden's Labour Force Survey (LFS) produces monthly estimates of the unemployment rate whose accuracy can be described in terms of error components that comprise the total mean squared error (MSE) of the estimate – an indicator of product data quality. Reductions in the MSE can only be achieved through process improvements such as more effective follow up of nonrespondents, improved interviewing, better estimation approaches, and so on (i.e., improved process quality). These improvements are possible because the organization possesses the knowledge, skills, and management structure to design and implement improved processes that result in real quality improvements.

The early literature on survey quality focused on product data quality (Accuracy) and the MSE as the primary indicator. Starting with the development of sampling theory in the 1930s and 1940s (Neyman 1934, 1938; Stephan 1948; Hansen et al. 1953) the focus obviously was on minimizing and controlling sampling errors. But it was also recognized early on that other error sources could affect the survey results – for example, the

interviewers and the nonrespondents (Deming 1944). In the 1960s, the importance of minimizing all error sources was stressed by some researchers; particularly, Dalenius (1967), Hansen et al. (1967) and Kish (1962). In order to estimate separate error components, evaluation studies were carried out, especially at the U.S. Census Bureau. Large evaluation studies, however, are expensive and of limited use for improving quality in real time because their findings may lag behind those of the main survey by many months. Standardizing and controlling processes that are known to affect product quality such as sampling, interviewing and coding, therefore became an important part of statistics production. The basic idea is that by continuously improving key survey processes, the overall process approaches an ideal state – that is, one that is stable and repeatable with minimal variation (Biemer and Lyberg 2003). A number of standards, guidelines and recommended practices have been developed over the years spanning from 1970 until today (U.S. Bureau of the Census 1974; Gonzales et al. 1975; U.S. Office of Management and Budget 2002; Eurostat 2005; International Standards Organization 2006; Statistics Canada 2009) all aiming at reducing errors and unnecessary variation. These efforts led to the so-called total survey error approach to survey design (Andersen et al. 1979).

In the late 1970s, the concept of survey quality was broadened via the so-called quality frameworks developed within the survey community (see Subsection 2.2), from encompassing not only Relevance and Accuracy, but also other dimensions of quality. In the 1990s, many survey organizations, influenced by the Total Quality Management (TQM) movement (Groves and Lyberg 2010), started to work on improvement projects. The importance of using process data (later named paradata; see Couper and Lyberg 2005) to evaluate and control process quality was stressed by Morganstein and Marker (1997). To view process quality as key to product quality was a new way of thinking in the survey community but in the private sector the concept of Six Sigma had already started to develop at Motorola in 1985. Also Deming's (1986) emphasis on statistical process control as a means for continuous improvement had large effects on how quality was perceived. Six Sigma (Breyfogle 2003) has become a toolbox for improvement projects, much like TQM, but with a strategic focus and a standardized method for process improvement and control. It turns out that it can also be very useful for improving survey processes.

Outside the survey community in the late 1980s and early 1990s, frameworks for evaluating organizations that strive for excellence were developed, for example, the Baldrige Performance Excellence Program (2013) and the European Foundation for Quality Management (EFQM 2013). These frameworks emphasize customer focus and results, and recognize the importance of leadership, people, partnership and strategy in order for an organization to reach excellence. Other important features of these frameworks are continuous improvement, which they share with Six Sigma and Kaizen, deployment of good practices and external evaluations. Some survey organizations such as the Czech Republic, Statistics Finland and Statistics Sweden have adopted one of these frameworks, namely EFQM.

In the auditing field the Committee of Sponsoring Organizations of the Treadway Commission (COSO) developed a framework to assess and improve internal control systems in the 1990s (COSO 2013) and later a framework to assess and improve enterprise risk management (COSO 2004). Both frameworks stress the importance of risks being

assessed in terms of likelihood and impact. The importance of risk assessment has so far been largely neglected in survey research (Eltinge et al. 2013).

Recently, Kenett and Shmueli (2014) developed a new framework for evaluating the quality of a generic statistical study that includes the dataset, the statistical analysis and the study report which they refer to as InfoQ or Information Quality. InfoQ provides a general framework applicable to data analysis in a broader sense than product quality. Rather it is "the potential of a dataset to achieve a specific goal using a given empirical analysis method." InfoQ framework identifies and examines relationships among the analytic objectives, the data available to achieve those objectives, the analysis of the data, and the ability of the analysis to achieve the objectives. Similar to the quality frameworks for official statistics, InfoQ provides eight dimensions used to deconstruct InfoQ as an approach for assessing it.

As Biemer (2014) notes, InfoQ can be regarded as a general framework that encompasses the survey total quality framework as a special case. Further, the development of InfoQ emphasizes the need for new practical tools for assessing quality in order to inform and caution data users regarding the limitations of a data analysis. In that regard, ASPIRE makes important contributions to data user knowledge and education about survey errors and their potential effects on statistical inference.

Thus, ASPIRE integrates many of the main ideas from the literature and frameworks mentioned above into a tool that will help product managers in survey organizations continually improve product quality. It does not rely solely on evaluations of MSE components for assessing Accuracy; yet it provides a practical, feasible approach to minimizing total survey error. In addition, the process facilitates the communication of quality improvements to stakeholders and users and greatly enhances an organization's ability to set clear goals for continual quality improvement.

As shown in the following, ASPIRE is not only applicable to surveys, but essentially any program that produce statistical products. By "statistical product" we mean virtually any data output that is used for statistical purposes including estimates, data sets, frames, registers, administrative databases, data tables, and indices. A major advantage of ASPIRE's generality in this regard is the consistency of the criteria, guidelines, ratings and definitions across the diverse assortment of statistical products found within NSOs.

2.2. Dimensions of Product Quality

To most statisticians and data analysts, good quality is synonymous with estimates having small mean squared errors (MSEs). The smaller the MSE, the more accurate are the estimates and the better are statistical inferences. As noted above, Deming (1944) recognized that quality should go beyond accurate estimates and should also encompass Relevance (Deming 1944). Over the years, the definition of quality has expanded to encompass other dimensions that are important to data users such as Timeliness, Comparability and Accessibility. This period also saw the development of so-called quality frameworks for official statistics whose use has expanded by new developments in survey methodology, technology and system architectures.

As an example, accessing data sets through the Internet is now common place and, for users, ease of access (i.e., Accessibility) is an important component of quality.

Decision-making in society has become more complex and global resulting in demands for harmonized and comparable statistics across countries and surveys (i.e., Comparability and Coherence). The timeliness of official statistics such as employment figures (i.e., the Timeliness dimension) often drives financial markets. Thus, quality frameworks for official statistics have been established to accommodate all these demands.

Several quality frameworks have been developed – each consisting of a number of quality dimensions. As an example, the quality framework developed by Eurostat (2009) consists of six dimensions: Relevance, Accuracy, Timeliness and Punctuality, Accessibility and Clarity, Comparability, and Coherence. This is essentially the framework adopted for the current report after combining the latter two dimensions into one dimension. Similar frameworks have been developed by, among others, Statistics Canada (Brackstone 1999), Statistics Sweden (Statistiska centralbyrån 2001), the UK Office for National Statistics (ONS 2007), the Organization for Economic Cooperation and Development (OECD 2011) and the International Monetary Fund (IMF 2003).

The work presented in this article emphasizes the Accuracy component of product quality. Biemer and Lyberg (2003) viewed accuracy as the dimension to be optimized in a survey while the other dimensions (the so-called *user dimensions*) can be treated as constraints during the design and implementation phases of production. They argued that sufficient Accuracy is essential for the other quality dimensions to be relevant. However, there are examples where accurate data may lose much of their utility if, for example, they are released too late to affect important decision-making or if they are presented in ways that are difficult for the user to access or interpret. As an example, surveys designed for the surveillance of disease outbreaks must be very timely if diseases are to be effectively contained. Accuracy may be secondary to timeliness in that case or there may be trade-offs involved where accuracy must be compromised to some extent for the sake of timeliness.

ASPIRE can help inform trade-offs among quality dimensions when assessments of these dimensions are incorporated into the evaluation framework. As discussed in Subsection 4.4, extensions of ASPIRE to include the user dimensions have been tested but more work is needed. However, this preliminary work was successful at identifying several important quality trade-offs and providing critical information needed for reconciling conflicting user and producer dimensions of quality.

2.3. Accuracy

For survey products, data accuracy is achieved by minimizing *total survey error* (TSE) which is the totality of error that can arise in the design, collection, processing, and analysis of survey data. (The term, TSE, could be generalized as "total *product* error" to acknowledge that ASPIRE's applications transcend survey products; however, we will use the traditional terminology in this article but note its limitations to describe some of the applications that follow.) A few error sources (such as measurement and data processing errors) are common to almost all surveys; however, other sources of error are dependent upon the survey design, type of data collected, and processing system used to develop the survey products. The ASPIRE system assesses accuracy by first decomposing the total error for a product into a number of error components that hold some appreciable risks to quality for the product. These risks are evaluated in the ASPIRE approach as well as

the steps that have been taken in the design and production stages to contain or mitigate these risks.

To identify the relevant error components, we let $\hat{Y}$ denote a survey estimate (or product) that is subject to errors from a number of sources. One can conceive of an "error-free" version of $\hat{Y}$ denoted by Y which would result if the processes producing $\hat{Y}$ were error free including no sampling error (i.e., a complete census). Thus, the difference, i.e., $\hat{Y} - Y$, i.e., the total survey error, is due to all the errors in the processes that produce $\hat{Y}$, both sampling and nonsampling errors.

The ASPIRE model for surveys decomposes the total survey error into sampling error and seven nonsampling error components, viz., frame error, nonresponse, measurement error, data processing error, modelling/estimation error, revision error, and specification error. *Frame error* (denoted by $\varepsilon_{\text{frame}}$) arises in the process of constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. It includes the inclusion of non-population members (*overcoverage*), exclusions of population members (*undercoverage*), and duplication of population members, which is another type of overcoverage error. Frame error also includes errors in the auxiliary variables associated with the frame units (sometimes referred to as *content error*) as well as missing values for these variables. As examples, information on company size, industry, location, contact name, and address may be missing or erroneous for some enterprises on a business frame or register, thus potentially increasing costs and other errors (for example, sampling and modelling errors)

Nonresponse error ($\varepsilon_{\text{nonresponse}}$) 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* ($\varepsilon_{\text{measurement}}$) includes errors arising from respondents, interviewers, imperfect survey questions and other factors which affect survey responses. *Data processing error* ($\varepsilon_{\text{data processing}}$) includes errors in editing, data entry, coding, computation of weights, and tabulation of the survey data. *Modelling/estimation error* ($\varepsilon_{\text{model/estimation}}$) 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.

Preliminary estimates are published for some key statistics in order to address user needs for timely data. For example, quarterly GDP estimates based on preliminary data are published in order to provide government leaders and other important users with timely, albeit approximate, information on national economic performance. Preliminary estimates may be available one month after the end of a quarter; final estimates may be delayed until the end of the following year or later. Obviously, the utility of the preliminary estimates depends substantially on how close they are to the final, official estimates that are ultimately released. *Revision error* is the difference between a preliminary, published estimate and the final revised estimate and is an important component of the total error for some products.

To see why, let $\hat{Y}_P$ denote the preliminary, published estimate of the parameter Y and let $\hat{Y}$ denote the final estimate. Then the total error in $\hat{Y}_P$ is given by $\hat{Y}_P - Y$ which can be rewritten as $(\hat{Y}_P - \hat{Y}) + (\hat{Y} - Y)$ where $\hat{Y}_P - \hat{Y}$ is the revision error and $\hat{Y} - Y$ is the total error in the final published estimate as described above. Because NSOs are quite interested in reducing the error in all published estimates, not just the revised ones, we focus on both

preliminary and revised estimates in our evaluation of Accuracy. Furthermore, considering revision error as a distinct error source reflects the view that large revisions, regardless of their reasons, are undesirable from the user's perspective and should be avoided. Thus, an important quality goal for any statistical agency is to reduce the size of the revisions which is facilitated by emphasizing revision error whenever it is applicable.

Note, however, that revision error is somewhat unusual because it reflects the combination of all other error sources on the preliminary estimate. For example, the preliminary estimate may differ from the final estimate as a result of late respondents (i.e., nonrespondents at the preliminary deadline) whose characteristics may be estimated or imputed in the preliminary estimate while their reported values are used in the final estimate. Likewise, revisions may correct for other nonsampling errors such as measurement, data processing, or modelling/estimation errors that are identified after the preliminary deadline. In this way, revision error may account for error sources that have already been considered in the assessment of data quality for the revised estimate.

For this review, our primary interest with regard to revision error is on the magnitude of the error – that is, the difference $\hat{Y}_P - \hat{Y}-$ and the steps that could be taken to reduce it and/or its impact on data users. As such, we have not decomposed revision error into its associated subcomponents (nonresponse error, data processing errors, etc.) because these error sources are considered in great detail in the evaluation of the final estimates. Nevertheless, separately decomposing revision error may still be very important in some cases to understand the impact of error sources on revision error that may be distinct from those affecting the final estimates.

For most products, a seventh nonsampling error source – referred to as *specification error* – is also applicable. Specification error arises when the observed variable, y, differs from the desired construct, x – that is, the construct that data analysts and other users prefer. In survey literature, for example Biemer (2011), x is often referred to as a *latent* variable representing the true, unobservable variable and y is often referred to as an indicator of x. As an example, in the European statistics for Foreign Trade of Goods (FTG), the invoice value of goods is collected from enterprises (y) while the statistical value (x) (i.e., the cost of goods at the border of the reporting country excluding costs incurred after crossing the border) is preferred for most statistical uses of the data. Thus, specification error may be defined as the difference between y and x (see, for example, Biemer and Lyberg 2003).

Specification error biases the estimates of population parameters. Let X denote the true population parameter which is a function of x. Then the total survey error (TSE) in a preliminary estimate can be written as

$$\hat{Y}_P - X = (\hat{Y}_P - \hat{Y}) + (\hat{Y} - Y) + (Y - X) \tag{1}$$

where $(\hat{Y}_P - \hat{Y})$ is the revision error, $(\hat{Y} - Y)$ is a combination of errors from multiple sources; specifically $\hat{Y} - Y = \varepsilon_{\text{sampling}} + \varepsilon_{\text{frame}} + \varepsilon_{\text{nonresponse}} + \varepsilon_{\text{measurement}} + \varepsilon_{\text{data processing}} + \varepsilon_{\text{model/estimation}}$, and $(Y - X)$ is the specification error. Likewise, the TSE in the final estimate, $\hat{Y}$, is just the right side of (1) with the revision error term omitted.

Under this model, the total survey error of an estimate includes specification error as well as the other aforementioned sampling and nonsampling errors. Thus, the specification error in the aggregate, $\hat{Y}$, is essentially the difference between the expected value of $\hat{Y}$

conditioned on the concepts implied by the survey instrument (Y) and the population parameter under the preferred or true concept (X). Some would argue that specification error should be part of the Relevance/Contents dimension. However, our view is that it is part of total survey error and, thus, should be considered a component of Accuracy.

3. The ASPIRE Model

The ASPIRE model borrows heavily from the quality assurance literature (see, for example, Juran and Godfrey 1999 and Breyfogle 2003) whose core principle rests on the identification, reduction, and elimination of suboptimal processes as well as the literature on continual improvement or Kaizen (Imai 1986). As a corollary to this principle, Lyberg et al. (1998) argue that improvements in survey processes aimed at reducing error risks (i.e., the probability that important errors will occur) will often produce products with reduced error to the extent that the risks are actually reduced. As an example, data collection processes designed using best practices and state of the art knowledge can achieve lower risks of measurement error and nonresponse, particularly if these processes are routinely monitored for compliance with the design specifications. While continual process improvement is often desirable, it may not always lead to product improvements. For example, some methods for increasing response rates (such as incentives) can actually lead to an increase in the nonresponse bias (see, for example, Keeter et al. 2000; Curtin et al. 2000; Merkle and Edelman 2002).

Thus, an essential ingredient of process improvement is to conduct experiments that directly measure the effects of alternative designs and processes on one or more components of the total error. Such experiments can provide quantitative evidence that the processes implemented actually reduce the errors from the targeted error source compared to the tested alternatives. As an example, the estimation of bias has been used effectively for comparing modes of data collection, alternative incentives, questionnaire design alternatives, and so on. However, this approach may be impracticable for TSE reduction across dozens of surveys generating hundreds of statistical products. It may not even be feasible for a single survey given the many potential sources of error whose effects may interact and vary considerably over the many estimates (products) generated by the survey.

Often the final survey design is a compromise that balances the TSE across many competing objectives; for any particular objective, it may be suboptimal. This "compromised design" phenomenon is not unique to surveys; rather it arises quite often in industrial quality control as well (see, for example, Michalek et al. 2006; Karsak 2004.) Given these complexities, the process improvement principles embodied in ASPIRE provide a feasible and effective approach for achieving product quality improvements across the wide range of products produced by the typical NSO.

ASPIRE is a system for assessing the risks of error from each potential source of error in a product and rating progress that has been made to reduce this risks according to clearly specified evaluation criteria. Its primary goals are to:

(a) identify the current, most important threats or risks to the quality of a product,
(b) apply a structured, comprehensive approach for rating the efforts aimed at reducing these risks, and

(c) identify areas where future efforts are needed to continually improve process and product quality focussing on those high risk error sources where ratings are relatively low.

We believe that product quality will improve to the extent that ASPIRE achieves these three goals. ASPIRE is quite general in that it can be applied to a specific statistical estimate such as the monthly unemployment rate, a range of products produced by a data collection program such as the estimates from a survey of local government agencies, or a frame or register such as the business register or master address frame, or a compilation of a number of statistical inputs such as estimates of gross domestic product (GDP). ASPIRE is also comprehensive in that it considers the errors in official statistics arising from all major error sources from the design of the data collection to final publication or data release.

The ASPIRE model assesses product quality by first decomposing the total error for a product into major error components. It then evaluates the potential (or risks) for these error sources to affect data quality (referred to as "the risks of poor quality") according to five evaluation criteria. Clearly specified and sufficiently detailed guidelines have been developed that are used to evaluate the risks with acceptable inter-rater reliability.

As previously noted, ASPIRE can be customized so that it considers only those error sources that pertain to a specific statistical product. For example, sampling error would not apply to products from the Swedish municipal accounts collection (referred to as RS) which does not employ sampling. As discussed in the next section, the model also accommodates the risk variation across error sources so that a product's overall quality is affected more by the error sources that pose greater error risks. For example, in the RS, revision error was judged as "low risk" because preliminary and final data releases seldom differ appreciably. Moreover, RS data users claim they are seldom affected by such revisions. On the other hand, data processing error is of high risk in the RS due to the amount of editing data receive and the potential for editing error to substantially affect the final estimates.

3.1. Assessing Error Risks

A critical element of the ASPIRE rating system is the assessment of error risk which involves assigning a risk rating to each error source according to its potential impact on product quality. For this purpose, it is important to distinguish between two types of risk: *residual* (or "current") risk and *inherent* (or "potential") risk. *Residual risk* reflects the likelihood that the survey process will produce a serious, impactful error *despite* the current efforts that are in place to reduce or mitigate the risk. *Inherent* risk is the likelihood of such an error *in the absence of* current efforts toward risk mitigation. In other words, inherent risk reflects the expected impact of errors from the error source if efforts to maintain current, residual error were suspended.

As an example, for a survey process that places a high burden on respondents (e.g., lengthy interview or complex data collection protocol), the risk of nonresponse and thus, nonresponse error may be considered inherently high. However, these error risks can be reduced by various data collection strategies such as multiple follow-up attempts, incentives, enhanced interviewer training on techniques for averting refusals, and so on.

Postsurvey adjustments may further reduce the risk of nonresponse bias. Thus, although inherent risk for the survey process is high, the residual may be moderate or low.

One may view the inherent risk rating for an error source as an indicator of the need for measures to control the errors from that source in the process. The greater the inherent risk the greater the need for approaches that will reduce it. The residual risk rating may be regarded as an indicator of the effectiveness of these measures to limit the error from a specific source. It therefore follows that inherent risks should be stable over time. Changes in the survey taking environment that alter the potential for error in the absence of risk mitigation can alter inherent risks, but such environmental changes occur infrequently and usually evolve gradually. On the other hand, residual risks are more transient as they depend upon risk mitigation activities which can change over time or may become less effective. As an example, nonresponse rates may increase over time as contact and refusal aversion strategies that were once effective become less so, thus increasing the residual risk of nonresponse error.

There are some similarities with the ASPIRE approach and those outlined in the program evaluation and risk management literature. Program evaluation consists of methods for collecting and analyzing data in order to address questions about the effectiveness and efficiency of projects, policies and programs (Rossi et al. 2004); for example, an evaluation of the effectiveness of establishing community health centers in low income areas at reducing the need of long hospital stays or expensive emergency room use. Consistent with most program evaluation systems (see McDavid et al. 2013), there is an underlying model and methodology and a performance management system. However, program evaluations often rely on experiments or quasi-experiments that compare the program outcomes with counterfactual outcomes – designs that seldom arise with NSO product evaluations. With respect to risk management, the literature uses the concepts of intrinsic and residual risks, usually uses templates to support the risk analysis, values risks in terms of both impact and likelihood, and relies on a range of risk assessment tools (see Barkley 2004; International Standards Organization 2009). Notwithstanding these commonalities, ASPIRE is the only system to incorporate a total error framework while still remaining accessible to NSO executives who may have very limited knowledge of the complex programs being evaluated.

As shown in the next section, the inherent risk for an error source directly affects a product's overall score because it determines the weight attributed to an error source in computing a product's average rating. While residual risk does not directly affect a product's score, it still plays an important role in the evaluation in two ways. An increase in residual risk from the prior evaluation could suggest that efforts to reduce the inherent risks of error have become less effective. Thus, the product's rating relative to risk mitigation would deteriorate accordingly. In addition, residual risk helps clarify the meaning and facilitate the assessment of inherent risk.

3.2. Evaluation Criteria

In addition to decomposing total error for a product into its component sources and identification of the risks associated with each source, the ASPIRE model evaluates the potential for these error sources to affect data quality according to five evaluation criteria,

viz., Knowledge of Risks, Communication with Users, Available Expertise, Compliance with Standards and Best Practices, and Achievement Towards Risk Mitigation or Improvement Plans. In Round 3, Communication with Users was extended to include data suppliers or providers as well as users. (For example, in the case of the National Accounts, these include departments responsible for key inputs to the GDP calculations such as the foreign trade and business statistics units.) The five criteria are given equal weight; however, differential weights could be used if desired. The guidelines currently used for evaluating these five criteria are shown in Appendix A.

A two-step rating process was used to assign ratings on a 10-point scale for each error source by criterion combination. First, a given criterion is assigned a qualitative rating of Poor, Fair, Good, Very Good, and Excellent based on the check list and subsequent discussions with the product area. Then, in step two, these qualitative ratings are then refined by choosing between low or high numerical point ratings within each of the five categories; for example, Poor (1 or 2), Fair (3 or 4), and so on to complete the 10-point scale. This is further described in the subsequent illustration.

A product's *error-source score* is the sum of its ratings (on a scale of 1 to 10) for the error source across the five criteria divided by the highest possible score attainable (which is 50 for most products) and then expressed as a percentage. A product's overall score, also expressed as a percentage, is then computed by the following formula:

$$\text{Overall Score} = \sum_{\text{all error sources}} \frac{(\text{error-source score}) \times (\text{error-source weight})}{10 \times (\text{number of criteria}) \times (\text{weight sum})} \tag{2}$$

where the "error-source weight" is either 1, 2, or 3 corresponding to an assessment of the source's inherent risk – 1 if low risk, 2 if moderate risk, and 3 if high risk – and "weight sum" is the sum of these "risk" weights over the product's applicable error sources.

The form of the overall score is somewhat arbitrary and other metrics could be used to summarize a product's overall rating. For example, as previously noted, it is possible to weight the five criteria differentially to reflect their relative importance. In addition, Kenett and Shmueli (2014) suggest a metric based upon the weighted geometric mean of scores which also has some desirable properties. Nonetheless, the current metric is intuitive while still providing a useful way to rank and compare products.

4. The Statistics Sweden Experience

As noted above, ASPIRE has been applied to seven key products at Statistics Sweden for three consecutive years (or rounds) and three products for the last two rounds. The quarterly and annual national accounts were considered together in the first round and then considered separately in the last two rounds. Table 1 lists the products and the error sources that were considered in the review for each. These products were considered "key" regarding their importance to the Swedish statistical system. In addition, together they span the breadth of statistical products offered by Statistics Sweden including; business and social surveys, registers, indices, and compilations. As shown in the table, eight products received an initial review in 2011(i.e., Round 1) and a second, follow up review in 2012 (Round 2) although quarterly and annual national accounts were considered separately in the second round. One product received its initial review in 2012. All ten

Table 1. Products and Error Sources Evaluated in Rounds 1, 2, and 3

Product	Round	Error Sources
Survey Products		Specification error
Foreign Trade of Goods (FTG)	1,2,3	Frame error
Labour Force Survey (LFS)	1,2,3	Nonresponse error
Annual Municipal Accounts (RS)	1,2,3	Measurement error
Structural Business Statistics (SBS)	1,2,3	Data processing error
Consumer Price Index (CPI)	1,2,3	Sampling error
Living Conditions Survey (ULF/SILC)	2,3	Model/estimation error
		Revision error
Registers		Specification error
Business Register (BR)	1,2,3	Frame: Overcoverage
Total Population Register (TPR)	1,2,3	Undercoverage
		Duplication
		Missing data
		Content error
Compilations		Input data error (up to four sources)
GDP	1*	Compilation error
GDP by Production Approach	2*,3*	Data Processing error
Annual	2*,3*	Model/Estimation error
Quarterly		Deflation/Reflation error
		Balancing error
		Revision error

* Error sources were modified in Rounds 2 and 3 based upon the error model in Figure 1.

products were reviewed for a third time in November 2013 (Round 3). This section describes some key aspects of these reviews and reports on some of the key findings.

4.1. Implementing ASPIRE

4.1.1. Forming the Evaluation Team

A key issue in forming a program evaluation team is whether to use internal or external evaluators. As summarized in Conley-Tyler (2005), there are important advantages of each approach. Internal evaluators provide some costs advantages and may excel in their intimate knowledge of the specific products and processes to be evaluated. In addition, whereas highly capable external evaluators may be scarce, internal evaluators having high levels of program-specific expertise may be readily available. With regard to costs, Statistics Sweden's experience suggest that cost savings using internal evaluators would be small or nil for broad-based evaluations like ASPIRE once the labor costs devoted to maintaining consistency of ratings across multiple evaluations teams are considered.

On the other hand, external evaluators generally have greater "perceived objectivity" if not greater "real" objectivity – key issues for NSOs intending to make the evaluation results public. Conley-Tyler (2005) notes that external evaluators are more objective and willing to criticise processes, management, and the organization itself. In support of this

claim, Statistics Sweden's prior experiences using internal evaluators engaged in similar activities raised concerns about the objectivity of that approach.

With respect to relevant knowledge of the TSE paradigm, the expertise of external evaluators may be broader and their experiences of having worked in other organisations provide benchmarks for judging quality. Likewise, their knowledge of the total error in official statistics may be greater than that of the internal evaluators. Thus, another advantage to using a small group of external evaluators having broad knowledge to conduct all the evaluations is greater consistency in the ratings across the products.

The advantages of using external evaluators are even stronger for government programs where transparency and objectivity are critical. While transparent evaluation can be achieved by both internal and external evaluators, credibility and legitimation is much greater with external evaluators (Conley-Tyler 2005), particularly if they are recognized experts in both the TSE paradigm and in the functioning of the NSO's statistical programs. This could be the deciding factor for NSOs and other organisations receiving government funding.

In the end, Statistics Sweden opted for external evaluators (Biemer and Trewin) who were aided by two management liaisons (Bergdahl and Japec) who provided internal program context and support for the evaluation.

For each round of ASPIRE, three sets of activities were conducted which may be described as preinterview, interview, and postinterview activities.

4.1.2. Preinterview Activities

a. *Background Reading and Preparation.* Several weeks prior to the onsite evaluation, each of the two external evaluators received an extensive set of materials for each of the products. Central among these was the "quality declaration" (if available) for each product. The quality declaration is a type of quality profile (Biemer and Lyberg, 2003) that documents key aspects of the design, data collection and production process for the product including the major error sources and what is currently known about them, descriptions of previous, current, or planned quality studies, and relevant information related to the user quality dimensions. Questionnaires, training manuals, and reports on recent studies related to quality were also included in the reading materials.
b. *Self-evaluations by Product Teams*. Also during this period, each product team was asked to complete a self-evaluation form that reflects the guidelines the external evaluators used to complete their initial evaluation of the product. In Rounds 2 and 3, the self-evaluations used the checklist format shown in Appendix B.

4.1.3. Quality Interview

A face to face interview lasting about four hours was conducted by the external evaluators with each product team. One important purpose of this interview was to supplement and clarify the information provided in background reading materials and self-evaluations. During these discussions, inherent and residual risks levels (high, medium, and low) were assigned to each applicable error source. Once the risk levels were established, the evaluators separately considered each applicable error source to assign a rating for

each criterion using a simple five-point scale: poor, fair, good, very good, and excellent. At the conclusion of interview, the risk levels and criteria ratings were reviewed and further discussed. Any disputes were clarified and reconciled to the extent possible. Detailed minutes were kept to provide a record of the proceedings. Of particular importance, these minutes captured justifications for the ratings by error source and criterion.

4.1.4. Postinterview Activities

Within a day or two following each interview, the evaluators reviewed the minutes, refined the ratings and resolved any inter-rater discrepancies. Apparent rating inconsistencies within and across products were identified and removed. These ratings and their written justifications were then shared with the product teams who were asked to correct any inaccurate or misleading information and dispute ratings they believe were not justified. This process yields the final ratings and justification narratives. These ratings constitute a major portion of the final report authored by the external evaluators.

Following Round 1, ASPIRE was improved in the following ways:

1. A number of enhancements were made to the rating process. Chief among these was the development of a criterion checklist that could be applied generically across the applicable error sources and products. Items in the checklist were sorted so that the criterion's rating usually followed directly from the last item affirmatively checked. The simple "yes/no" format eliminated much of the subjectivity in the self-evaluation process observed in Round 1. Appendix B shows one such checklist (for Knowledge of Risks).
2. Except for new products, the quality review focused on changes in knowledge, staffing, methodology, processing, planning, mitigation strategies, etc. that may have some implications for data quality. This emphasis reflects the goal of the second and third rounds which are to assess the changes in quality since Round 1.
3. Post-interview, face to face, debriefing meetings were held with product teams that wanted to appeal one of more of their ratings and/or discuss the written rating justifications and recommendations for improvement.
4. In the second round, user dimensions were also evaluated for two products (the Labour Force Survey and the Consumer Price Index) as described in Subsection 4.3.
5. The error sources used in Round 1 for the GDP were substantially revised following in-depth discussions with the National Accounts staff about the GDP production process. This necessitated revamping the criteria used to evaluate GDP data quality. Details regarding this approach are provided in Subsection 4.2.

4.1.5. Illustration – Foreign Trade of Goods (FTG)

In this section, we illustrate how the steps of the process were executed for Statistics Sweden's survey of international trade or the FTG. The FTG collects information on the imports and exports of 9,000 different types of commodities by country of origin and destination for 250 countries resulting in almost two million statistical items being reported each month. The primary uses of the results of the survey include the trade in commodity components of the balance of payments statistics and the expenditure measure of GDP. It consists of two statistical systems: Extrastat (for countries outside the EU) and Intrastat (for EU countries).

In Round 1, measurement error was classified as high inherent risk for the FTG for a number of reasons including possible misclassification of commodities (more so for responses via paper forms than for electronic responses), data concerns regarding net weight (and other quantities) of shipments especially for textiles and chemicals, and errors resulting from the methods used to convert the invoice value to conceptually correct statistical value. At the other extreme, revision error was deemed to be low risk because the size of revisions tended to be relatively small and inconsequential to most users. The other error sources were given medium risk.

In Round 2, these risk ratings were revised based upon further discussions with internal data users such as the National Accounts staff. In particular, revision error was upgraded to high inherent risk after the potential effects of revision error on the GDP estimates were better understood. Likewise, data processing error was raised to high inherent risk after realising the extensive editing that is done in the FTG and the risk it poses to data quality without this editing. Frame error was downgraded to low risk when it was determined that the risk of overcoverage in the FTG frames (viz., the Business Register and National Tax Board VAT register) is much lower than originally thought. Theoretically, changes to inherent risks should only occur when (a) the design of a process undergoes a fundamental change; for example, rather than collecting EU export data directly from enterprises, exports are based upon imports from other EU countries or, (b) as in the case of FTG, the information upon which the current risk level was based is deemed incomplete or erroneous and, thus, the inherent risk level for the product should be corrected.

Note that sampling error is not applicable for the FTG because it employs a cut-off sample that includes all enterprises above a threshold value representing at least 95% of all imports and exports within the EU but there will be modelling error because certain assumptions are made to estimate the contribution of those enterprises below the threshold value.

With regard to quality ratings, processing error received the lowest score which in part reflects the FTG personnel's lack of knowledge at that time about the causes and extent of editing errors which have a high risk of error. In addition, the evaluators had concerns that lack of quality control in the keying of paper forms was a violation of ISO standards. In fact, the number of paper forms that are keyed was quite small (about 10% of all reports) which diminishes any risk of error from this source. Nevertheless, the paper transactions could comprise a sizeable percentage of trade for some commodities and pose an appreciable error risk in those situations.

Notwithstanding these concerns, FTG's overall quality score was among the highest in Round 1. Nonetheless, its rating for measurement error was fairly low and the evaluators provided several recommendations and strong encouragement to take initiatives that would increase that score in the coming year. The evaluators' guidance was apparently followed because important improvements to address measurement error were quite evident in Round 2. For example, communication with data users regarding accuracy, particularly measurement error, substantially improved as a result of enhancements to the quality declaration. In addition, several important studies were completed and documented in reports providing more information on measurement and other error sources.

In addition to these improvements, other quality improvements were made as follows:

- Swedish Customs adopted the FTG editing system for its programs improving the quality of data received by Statistics Sweden.
- Plans are in place to better understand the causes of revision error, its impact on important users such as the National Accounts, and some effective means for reducing it over time.
- An asymmetry study with Finland (i.e., a reconciliation of Swedish imports against Finnish exports and vice versa) was completed which focused on understanding the effects of coding error on trade statistics.
- Work has commenced to replace the current Excel-based macro-editing software with a much improved, flexible and professionally developed system.
- Use of the Statistics Sweden's "Standardized Methods and Toolbox" increased resulting in a number of improved practices.
- A new survey to calibrate the conversion of invoice value to statistical value was scheduled for completion (and, subsequently completed) in 2013.

The current and previous round's ratings are shown in Table 2 in graphical form and the changes are shown in Table 3. Similar tables were developed for Round 3 so that improvements over successive rounds could be shown.

4.2. Error Sources Specific to the Gross Domestic Product

In retrospect, the Round 1 evaluation of the GDP error was somewhat flawed as a result of attempting to force an error structure identical to that used for the surveys. The eight error sources that are applicable to other products cannot easily be applied to GDP considering its unique, extensive and complex error structure. Thus, in Round 2, ASPIRE was modified by tailoring it to more closely reflect the complex GDP error structure. Because of the time constraints, the focus of the Round 2 review was considerably narrower, focusing solely on the estimation of quarterly and annual GDP using the production approach. In addition, the error structure of the GDP estimation process was restructured to more precisely capture the GDP's major error sources. The same approach and error structure can be used as well for GDP compiled from the expenditure approach.

Figure 1 provides a flow diagram that attempts to capture the major activities associated with the estimation of GDP. As shown, the GDP estimation process incorporates two somewhat independent ways for estimating GDP. These are referred to as the production (shown on the left) and the expenditure approaches (shown on the right). Both approaches begin with a number of inputs that must be assembled, processed, and compiled to prepare them for the next step in the process. Each of these inputs is subject to error. The "Compile" stage includes data processing, which may be simply entering the inputs into an Excel spread sheet but may also include some editing as well as modelling/estimation especially when only proxy variables are available. This latter process may involve combining multiple inputs to create derived variables as well as modelling the data to reduce specification and other errors. For producing GDP in current prices, these compiled inputs proceed through an estimation stage which, for the production approach, involves

Table 2. FTG Quality Ratings Matrix for Round 2 with Round 1 vs. Round 2 Scores by Error Source

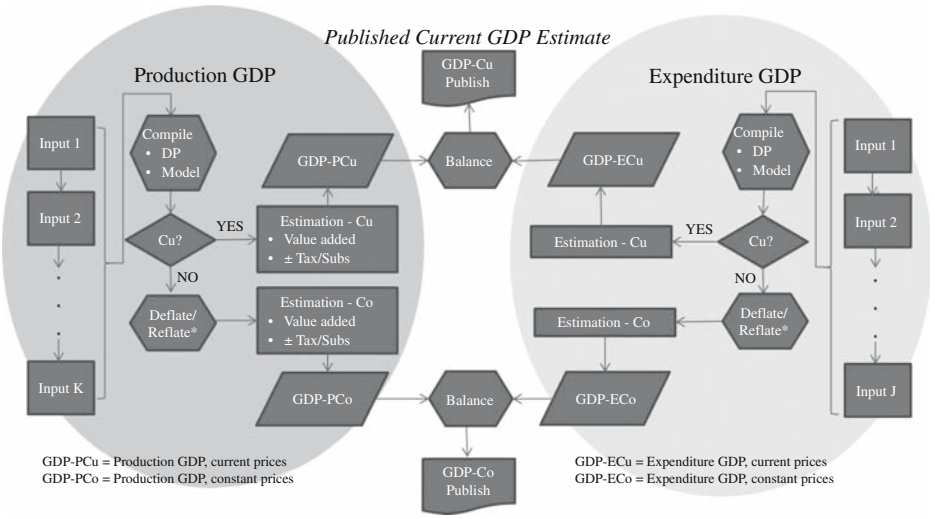
Error source	Score round 1	Score round 2	Knowledge of Risks	Communi cation to Users	Available Expertise	Compliance with Standards & Best Practices	Plan Towards Mitigation of Risks	Risk to Data Quality
Specification error	58	58	○	○	◡	◡	○	M
Frame error	58	58	○	○	◡	○	◡	L
Non-response error	62	66	◡	◡	◡	○	◡	M
Measurement error	54	62	◡	○	◡	◡	○	H
Data processing	46	60	◡	◡	◡	◓	○	H
Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Model/estimation	66	80	◡	◡	◎	◎	◡	M
Revisionerror	62	76	◡	◡	◡	◎	◡	H
Total score	57,3	65,8						

Scores					Levels of Risk			Changes from round 1	
●	◓	○	◡	◎	H	M	L		
Poor	Fair	Good	Very good	Excellent	High	Medium	Low	Improvements	Deteriorations

Table 3. FTG Round 1 to Round 2 Rating Changes and Corrections with Annotations

Error source	Score round 1	Score round 2	Knowledge of Risks	Communication to Users	Available Expertise	Compliance with Standards & Best Practices	Plans Towards Mitigation of Risks	Risk to Data Quality	*Correction from 2011 rating* / *Improvement from 2011 rating* Comments on changes
Specification error	58	58	5	~~7~~5[1]	7	7	5	M	[1]Under the current guidelines, communication should have been "Good" last year, not "Very Good."
Frame error	58	58	~~7~~5[1]	5	7	5	7	~~M~~L[2]	[1]Corrects error in last years rating for Knowledge of Risks. [2]Also, corrects risk level based upon intrinsic risk of frame error being low.
Non-response error	62	66	7	5→7[1]	7	5	7	M	[1]Communication to users about nonresponse improved as a result of the QD.
Measurement error	54	62	5→7[1]	5	5→7[2]	7	5	H	[1]Knowledge of risks gained through writing the QD as well as preparation of the annexes to the SLA with the NA. [2]Working relationship and closer cooperation between the col lection unit and the methods group as a result of the SLA.
Data processing error	46	60	5→7[1]	5→7[2]	5→7[3]	3	5→6[4]	~~M~~H[5]	[1]Knowledge of risks gained through writing the QD as well as preparation of the documents "Improvements of the work on revisions in the Swedish good" and "Improving macro-editing in Intrastat." [2]Likewise Communication has improved through both of the above mechanisms. [3]Working relationship and closer cooperation between the col lection unit and the methods group as a result of the SLA. [4]Some planning is underway for further improvements of editing and coding. Planning and discussions are underway to reduce the misclassification of goods by enterprises. [5]Risk level was re-evaluated and elevated to H based upon the importance of editing to data quality.
Sampling error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Model/estimation error	66	80	7→8[1]	5→7[2]	7→9[3]	7→9[4]	7	M	[1]Both Knowledge and Communication has improved evidenced by the document "Improvement of the distribution keys for the estimated trade in the Swedish Intrastat." [2]Key staff have made national presentations in connection with the WG Quality Meetings elevating expertise. [3]Swedish Customs adopted SCB's editing system which indicates state of the art systems. [4]Plans are in place to study more sophisticate models for estimation under cutoff using VAT possibly using the Vat Information Exchange System (VIES).
Revision error	62	76	5→7[1]	5→7[1]	7	7→9[2]	7→8[3]	~~L~~H[4]	[1]Knowledge and communication of risks improved through writing the QD as well as preparation of the documents "Improvements of the work on revisions in the Swedish goods." [2]Compliance with standards and best practices enhanced through Standardized Toolbox. Above referenced document also provides evidence that best practices are being followed. Progress has been made to rapidly detect and repair causes of large revisions. [3]Plans being developed to identify causes of revision error. [4]The risk level was re-evaluated and elevated to H as a result of the impact on the NA statistics.
Total score	57,3	65,8							

Note: *(Shaded cells denote either improvements (light) or deteriorations (dark) in ratings since Round 1. Corrections denoted by strikeouts with correct rating inserted. Footnotes describe reasons for the changes.*)



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

Fig. 1. High-Level Process Flow Diagram for Estimating Current and Constant Price GDP by Production and Expenditure Approaches

adding taxes and deducting subsidies (subs). For constant prices, the current prices must be "deflated" using the appropriate price indices before adjustments for taxes and subsidies.

Both the production and expenditure approaches will produce interim estimates of GDP (both current and constant prices) which must then be "balanced" or forced into agreement as the economic theory dictates (see, for example, Lequiller and Blades 2006). This balancing process produces the preliminary estimates of GDP for both current (denoted by Cu in the exhibit) and constant (denoted by Co) prices. The latter differs from the former primarily by a deflation/reflation process that adjusts prices to a common base-year. The preliminary estimates are subsequently revised when additional data become available. Thus, the error sources associated with the GDP estimation process are as shown in Table 1, bottom panel.

In the evaluation of production GDP, considerable attention was given to the error in the inputs and their effects on the error in the GDP estimates. Priority was given to inputs that posed the greatest risk to GDP error. These were determined by the evaluators in collaboration with the National Accounts staff.

4.3. Overall Results for All Products for Round 2

This section further illustrates some important uses of ASPIRE to compare the scores of all ten products in Round 2. Table 4 provides the overall scores for the six survey products and two registers and Table 5 provides the overall scores for the National Accounts only because the structure of their error sources is quite different from the other products. To facilitate the exposition of the results, the error sources were consolidated into a single list which appears in the first column of Table 4. The other columns of the table refer to the particular product being evaluated. For each product, the bold figures correspond to "High Risk" error sources, italic corresponds to "Medium Risk," and non-bold corresponds to

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

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

In this table,individual and mean ratings can be compared across products (columns) and by error source (rows) as well as. Note, for example, that the LCS and Measurement error/Content Error have the lowest average ratings. The FTG shows the greatest improvement from Round 1 to Round 2.

BOLD = HIGH RISK

ITALICS = MEDIUM RISK

REGULAR FONT = LOW RISK

N/A = NOT APPLICABLE

Table 5. Product Error-Level, Overall Level, and Error Source-Level Rating with Risk-Levels Highlighted for the National Accounts

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

BOLD = HIGH RISK
ITALICS = MEDIUM RISK
REGULAR FONT = LOW RISK
N/A = NOT APPLICABLE
N.E. = NOT EVALUATED

"Low Risk" error sources. The same applies to the second table for the two National Accounts products. Note that the interpretation of the error sources (see Subsection 2.3) and criteria may vary between surveys and registers.

Before discussing the results in Tables 4 and 5, a few cautions should be stated. There is a natural tendency to compare the overall scores across the products or to rank the products by their total score. The interpretation of such comparisons may not be straightforward for several reasons. First, the total score for a product reflects a weighting of the error sources by the risk levels which can vary considerably across products. Products with many high risk error sources, such as the National Accounts, may be at somewhat of a disadvantage in such comparisons because they must perform well in many high risk areas in order to achieve a high score. Second, the assessment of low, medium, or high risk is done within a product, not across products. Thus, it is possible that a high risk error source for one product could be of less importance to Statistics Sweden than a medium risk error source for another product if the latter product carries greater importance to Statistics Sweden. (For example, measurement error for the ULF/SILC may be somewhat lower priority than it is for the CPI.) Finally, the scores assigned to a particular error source for a product have an unknown level of uncertainty due to a number of factors. We believe rating consistency and reliability considerably improved with the development of the checklist as discussed above. Still, a difference of 2 or 3 points in the overall product scores may not be meaningful because an independent reassessment of the product could reasonably produce a new score that differs from the current score by that margin. Note further that, because of the very different approach taken in Round 2 for the National Accounts, comparisons to Round 1 for the GDP ratings are not meaningful.

Close inspection of scores in Tables 4 and 5 yield the following general observations:

- The average score for all products in Round 2 was 57 compared to 54 in Round 1 – a 5.6 percent improvement in the ratings. However, among products evaluated in both

Table 6. User Dimensions and their Components

Timeliness & Punctuality	Accessibility & Clarity
• Timeliness of release of main aggregates • Timeliness of release of detailed outputs (including microdata) • Punctuality	• Ease of data access • Documentation (including metadata) • Availability of Quality Reports • User support
Comparability & Coherence	**Relevance/Contents**
• Comparability across geography, populations, and other relevant domains • Comparability across time (including impacts of redesign) • Coherence with other relevant statistics (including use of standard classifications, frameworks, etc.)	• Inputs (content, scope, classification, etc.) • Outputs (including microdata and other products)

rounds, the improvement was about 8.5 percent. The introduction of ASPIRE undoubtedly led to some of these improvements as the ratings for all seven products that were reviewed in Round 1 improved in the current round. A significant influence was the development of Quality Declarations consistent with one of the strong recommendations of the evaluators.

- In both rounds, measurement error had the highest average inherent risk of any error source. It also ranked near the bottom in percent mitigated risk, defined as the total points earned divided by the maximum points achievable for an error source expressed as a percentage.
- By contrast, sampling error ranked the highest in percent mitigated risk, earning roughly 70% of the maximum points achievable in both rounds. Revision error is also highly ranked although it only applies to three products in Table 5 and the two National Accounts products.
- "Available expertise" and "compliance with standards and best practices" are generally rated higher than "knowledge of risks," "communication (of these risks) with users," and "achievement towards risk mitigation or improvement plans." The latter three criteria appear more challenging to most products.

ASPIRE identified many areas where improvements to data quality are needed with the highest priorities assigned to areas having high risks and low ratings. In addition, a number of "cross-cutting" recommendations were made. These are recommendations that affect multiple products such as: better documentation of quality and use of quality profiles, more evaluations of measurement errors, improved IT-client relationships, better succession planning in some areas, and so on. Costs varied considerably among the recommendations and limited resources constrained the scope of the improvements that Statistics Sweden could pursue. Because some improvement projects, particularly those that cut across product areas, required substantial allocations or reallocations of funding, decisions regarding which projects and activities to pursue in the

future should be left to management. Nevertheless, product areas may have some capacity to implement the most important improvements and this has happened to some extent.

The results of all three rounds of ASPIRE can be found in Biemer and Trewin (2012, 2013, and 2014). These reports are available by request from the authors.

4.4. Assessing the User Dimensions

As noted previously, the ASPIRE system was expanded in Round 2 to incorporate a process for evaluating the four user dimensions of quality. These are Accessibility & Clarity, Comparability & Coherence, Relevance/Contents, and Timeliness & Punctuality. The primary goal of this application was to develop a process for assessing the user quality dimensions. The system was tested on two products: the LFS and the CPI. The evaluation framework is completely consistent with the Accuracy framework; that is, each dimension was first decomposed into mutually exclusive components (analogous to the error sources defined for Accuracy) which, for the most part, are those described in the ESS Quality Assurance Framework (ESS 2011). Quality for each component was assessed according to five criteria that are similar to the five Accuracy criteria; viz., Knowledge of User Needs, Communication with Users, Available Expertise (to address user needs), Compliance with Standards and Best Practices, and Plans toward Addressing User Needs and were applied to each of the components under a dimension.

The components associated with each user dimension appear in Table 6. As was done for Accuracy, checklists were developed for each criterion and were generic across dimensions and components within dimensions.

The LFS was evaluated for Timeliness & Punctuality and Comparability & Coherence and the CPI was evaluated for Relevance/Contents and Accessibility & Clarity. The assessment process, which proceeded much like the process for Accuracy, seemed to work well for their initial application. However, some needed improvements were identified. For example, the checklists and criteria could be enhanced to better capture the risks of poor quality associated with each dimension. Also, direct communication with the users of these statistics is recommended to provide information on quality from the broader user community. In this trial evaluation, we largely relied on the advice of product staff on their interaction with users.

5. Discussion

Although this article has focused on the application of ASPIRE to ten Statistics Sweden products, it can be applied much more generally. As we have demonstrated, it can be used for survey products, administrative data products, registers and 'compilation' products such as the National Accounts. It can also be applied in other government statistical offices as well as in private sector or university statistical products. By design, it performs best for products that recur regularly and that are reviewed repeatedly so that improvements (or deteriorations) in quality can be assessed across time. While one-time ASPIRE reviews could provide useful insights regarding a product's current quality-level, multiple reviews would be more effective if the

objective is quality improvement. We believe that annual reviews are sufficiently frequent to track improvements for most programs. Less frequent (say biennial) reviews may be sufficient for lower risk programs or programs whose improvement efforts require more than one year to generate measureable results.

Any method for evaluating the quality of products as complex as those considered in this article will have its limitations. Estimating the total MSE (or even its key components) for a product such as the CPI or quarterly GDP is virtually impossible because the data required are largely unobtainable. Further, any data that can be collected on nonsampling errors are themselves subject to nonsampling errors. For example, a survey of nonrespondents to estimate the nonresponse bias in the LCS/ULF is also subject to nonresponse. The ASPIRE approach does not provide direct measures of the total MSE of a product. However, ASPIRE's ratings are negatively correlated with the risks of poor data quality; specifically, improved quality ratings reflect lower error risks. In addition, ASPIRE ratings are positively affected when MSE components have been estimated. For example, the rating for Knowledge of Risks is elevated when the bias from the error source has been estimated. Likewise, the rating for Communication with Users is elevated if those estimates have been documented and disseminated.

As noted in Section 3, the primary goals of ASPIRE are to identify the current, most important threats or risks to the quality of a product, apply a structured, comprehensive approach for rating the efforts aimed at reducing these risks, and identify areas where future efforts are needed to continually improve process and product quality focussing on those high risk error sources where ratings are relatively low. We believe that product quality will improve to the extent that ASPIRE achieves these three goals. A key requirement for this is that inputs to process – in particular, the information needed to accurately assess each criterion – are accurate, complete, timely, and accessible by the evaluators. Thus, continuing to update and improve the documentation of quality is an important activity to ensure ASPIRE's success.

Based upon this work, we believe ASPIRE succeeds in four areas. First, the approach is comprehensive in that it (a) covers all the important sources of error for a product and (b) uses criteria that span all the important risks to product quality. Second, the checklists used to assign the ratings under each criterion seem quite effective at identifying and assessing both manifest and hidden risks to data quality. To the extent that the documentation and other information shared during the ASPIRE process is both accurate and complete, the current approach assigns reliable ratings that reflect true data quality risks. Third, ASPIRE successfully identifies areas where, from an organizational perspective, improvements are needed and have very high priority. It further prioritizes these needs when it is not possible or sensible to undertake all quality improvements. For example, areas having highest risk and lowest ratings, assuming other factors are equal, should be assigned highest priority for improvement. Of course, the overall importance of the product relative to other products also should be taken into account as well as the resource requirements and the likely success of the improvement effort.

Finally, if implemented appropriately, the ASPIRE framework should generally increase organizational transparency and accountability both internally and externally.

Within the organization, this will enhance communication across products and quality improvement projects thus fostering greater collaboration and sharing of quality improvement ideas and results. Externally, this transparency will lead to greater organizational credibility and product confidence. In addition, providing this detailed information on data quality issues to external users can generate external pressure on the organization to make swifter and greater progress on quality improvements.

One weakness of the model is that it is, at best, a proxy measure for product quality because it makes no attempt to estimate the TSE and its components. However, quantitative assessments of TSE are reflected in the ratings and can also be used to supplement the information obtained from our approach. Another potential weakness of the approach is that it can be somewhat subjective in that it relies heavily on the knowledge, skill, and impartiality of the evaluators. However, we believe it would be undesirable to remove all the subjectivity from the process because that would be akin to automating the review process. A purely objective process may not optimally utilize the expertise of the evaluators nor allow for more complex judgments to be applied to the process. It is important, however, that any subjectivity in the ratings does not lead to inequities and inconsistencies across reviews. A number of safeguards have been put in place to prevent these potential adverse effects including the quality guidelines, checklists, the rating revision process, and the ratings appeal process.

With respect to possible future research, there are several thrusts. First, further testing and evaluation of the ASPIRE approach should focus on its long-term effects on product quality. For example, there could be some assessment of value of improvements projects that have been launched following recommendations from the ASPIRE process. Key users should be informed of the improvements completed and still underway and consulted to obtain their views on whether quality has been improved. Thus, the evaluation could determine whether quality improvements have increased under ASPIRE and whether ASPIRE is worth the investment of resources. The evaluation might also assess whether actual improvements correlate well with the changes in ratings for individual products and the quantitative information on error components that might be available for some products. Finally, staff within the organization should be consulted in the evaluation to elicit their opinions regarding the benefits and issues associated with ASPIRE.

Second, research could be conducted to further reduce inter-rater variation as well as intra-rater bias. Cognitive laboratories might be used for this purpose. Third, further work could extend the ASPIRE approach to the user dimensions. Whilst external evaluators are preferred, a satisfactory evaluation of the user dimensions could rely primarily on internal evaluators by using the structured approach we propose for obtaining feedback from both internal and external users across the range of quality dimensions.

Finally, we hope to see ASPIRE or a similar approach be implemented in other NSOs to see if similar quality improvements can be realized in other countries and organizations. For the sake of cross-country comparisons, settling on a unified approach that is applicable across diverse NSOs and cultures would offer clear advantages.

Appendix A – Evaluation Criteria and Guidelines for Accuracy

Exhibit 1.1a. Knowledge of Risks

Poor [1,2] ●	Fair [3,4] ◓	Good [5,6] ○	Very Good [7,8] ◒	Excellent [9,10] ◎
Program documentation does not acknowledge the source of error as a potential factor for product accuracy.	Program documentation acknowledges error source as a potential factor in data quality. **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 (bias and variance) components.	Studies have estimated relevant MSE 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.

Exhibit 1.1b. Communication with Users

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

Exhibit 1.1c. Available Expertise

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

Exhibit 1.1d. Compliance with Standards and Best Practices

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

Exhibit 1.1e. Achievement Towards Mitigation and/or Improvement Plans

Poor [1,2] ●	Fair [3,4] ◓	Good [5,6] ○	Very Good [7,8] ◒	Excellent [9,10] ◎
There is no evidence that any planning has been done for studying or mitigating the risks for this error source.	An overall plan for error reduction with measurable objectives exists for mitigating the risks for this error source. **But:** The plan is not approved by the appropriate level of management.	A management-approved plan with measurable objectives exists. The plan adequately addresses the work required for mitigating the risks of poor data quality relative to this error source. . . **But:** One of the following deficiencies with the plan exists: a. The overall plan has not been updated in at least one year. b. There is no accountability in place to ensure compliance with the plan. c. No mechanism is specified for gauging progress toward each objective. d. No resources have been allocated to implement the plan.	Resources have been allocated to undertake this work. Considerable progress has been made on the plan for mitigating the risks to data. None of the deficiencies noted under the "Good" criteria are present. **But:** Efforts have not yet produced the desired control over the error source that is stipulated in the plan.	Mitigation plans have been fully implemented or well underway. Progress toward all goals and objectives has been excellent. As a result, the level of error in the final estimates due to this error source is being maintained at an acceptable level for the primary purposes of the data. As a result of these efforts, the error source is under control and poses no or very little risk to data quality. Results of the mitigation activities have been fully documented. Accountability measures are in place to ensure compliance with the plans. The mitigation plans are reviewed and updated periodically.

Appendix B – Example of a Criterion Checklist (Knowledge of Risks)

For each applicable error source, indicate either compliance or noncompliance with an item in the checklist by marking "Yes" or "No," respectively. In order to achieve a higher rating for a criterion, all items for that higher rating must be checked. You may use the "Comments" field to provide comments you deem necessary to explain your response to an item.

Knowledge of Risks	Check Box	Comments
1. Documentation exists that acknowledges this error source as a potential risk.	☐ Yes ☐ No **Fair**	
2. The documentation indicates that some work has been carried out to evaluate the effects of the error source on the key estimates from the survey.	☐ Yes ☐ No **Good**	
3. Reports exist that gauge the impact of the source of error on data quality using proxy measures (e.g., error rates, missing data rates, qualitative measures of error, etc.)	☐ Yes ☐ No **Good**	
4. At least one component of the total MSE (bias and variance) of key estimates that is most relevant for the error source has been estimated and is documented.	☐ Yes ☐ No **Very Good**	
5. Existing documentation on the error source is of high quality and explores the implications of errors on data analysis.	☐ Yes ☐ No **Excellent**	
6. There is an ongoing program of research to evaluate the components of the MSE that are relevant for this error source.	☐ Yes ☐ No **Excellent**	

6. References

Andersen, R., J. Kaspar, and M. Frankel. 1979. *Total Survey Error*. San Francisco: Jossey-Bass Publishers.

Baldrige Performance Excellence Program 2013. *The 2013–2014 Criteria for Performance Excellence*. Available at: http://www.nist.gov/baldrige/ (accessed August 3, 2013).

Barkley, B.T. 2004. *Project Risk Management*. New York: McGraw Hill Professional.

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

Biemer, P. 2014. "Comment on 'On Information Quality' by Kenett and Shmueli." *Journal of the Royal Statistical Society,* Series A. Vol. 177, Part 1: 27–29.

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

Biemer, P. and D. Trewin. 2012. *Development of Quality Indicators at Statistics Sweden*. Report to Statistics Sweden, January 2012.

Biemer, P. and D. Trewin. 2013. *A Second Application of the ASPIRE Quality Evaluation System for Statistics Sweden*. Report to Statistics Sweden, January 2013.

Biemer, P. and D. Trewin. 2014. *A Third Application of ASPIRE for Statistics Sweden*. Report to Statistics Sweden, January 2014.

Brackstone, G. 1999. "Managing Data Quality in a Statistical Agency." *Survey Methodology* 25: 139–149.

Breyfogle, F. 2003. *Implementing Six Sigma*, 2nd edition. New York: John Wiley & Sons.

Conley-Tyler, M. 2005. "A Fundamental Choice: Internal or External Evaluation?" *Evaluation Journal of Australasia* 4: 3–11.

COSO, 2004. *Enterprise Risk Management – Integrated Framework*. Available at: http://www.coso.org/documents/coso_erm_executivesummary.pdf (accessed August 3, 2013).

COSO, 2013. *Internal Control – Integrated Framework, 2013*. Available at: http://www.coso.org/documents/coso%202013%20icfr%20executive_summary.pdf (accessed August 3, 2013).

Couper, M. and L. Lyberg. 2005. "The Use of Paradata in Survey Research." In Proceedings of the 55th Session of the International Statistical Institute, Sydney, Australia, April 7, 2005. Available at: http://isi.cbs.nl/iamamember/CD6-Sydney2005/ISI_Final_Proceedings.htm (accessed June 26, 2014).

Curtin, R., S. Presser, and E. Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64: 413–428.

Dalenius, T. 1967. *Nonsampling Errors in Census and Sample Surveys*. Report no. 5 in the research project Errors in Surveys, Stockholm University.

Deming, E. 1944. "On Errors in Surveys." *American Sociological Review* 9: 359–369.

Deming, E. 1986. *Out of the Crisis*. Cambridge, MA: MIT Press.

EFQM, 2013. "An Overview of the Excellence Model." Available at: https://www.google.com/url?q=http://www2.efqm.org/en/PdfResources/EFQM%2520Excellence%2520Model%25202013%2520EN%2520extract.pdf&sa=U&ei=9BasU4nkHsqTqAbUhIGQCg&ved=0CAUQFjAA&client=internal-uds-cse&usg=AFQjCNHthJnhRPIS1t6cfa4Ka9ePXOLRxg (accessed June 26, 2014).

Eltinge, J., P. Biemer, and A. Holmberg. 2013. "A Potential Framework for Integration of Architecture and Methodology to Improve Statistical Production Systems." *Journal of Official Statistics* 29: 125–145. DOI: http://dx.doi.org/10.2478/jos-2013-0007.

European Statistical System (ESS) 2011. "Quality Assurance Framework of the European Statistical System, Version 1.1." Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF (accessed August 9, 2013).

Eurostat 2005. "European Statistics Code of Practice, Revised Edition." Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF (accessed June 26, 2014).

Eurostat 2009. *Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009, Eurostat General/Standard report*, Luxembourg, April 4–5. Available at: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009R0223 (accessed June 18, 2014).

Gonzales, M.E., J.L. Ogus, G. Shapiro, and B.J. Tepping. 1975. "Standards for Discussion and Presentation of Errors in Surveys and Census Data." *Journal of American Statistical Association* 70: 5–23.

Groves, R.M. and L.E. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74: 849–879. DOI:http://dx.doi.org/10.1093/poq/nfq065.

Hansen, M., W. Hurwitz, and W. Madow. 1953. *Sample Survey Methods and Theory, Volumes I and II*. New York: John Wiley & Sons.

Hansen, M., W. Hurwitz, and L. Pritzker. 1967. *Standardization of Procedures for the Evaluation of Data: Measurement Errors and Statistical Standards in the Bureau of the Census*. Paper presented at the 36th session of the International Statistical Institute.

Imai, M. 1986. *Kaisen: the Key to Japan's Competitive Success*. New York: McGraw-Hill Education.

International Monetary Fund (IMF) 2003. *Data Quality Assessment Framework and Data Quality Program*. Available at: http://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm (accessed June 21, 2013).

International Standards Organization 2006. *Market, Opinion and Social Research ISO Standard No. 20252*. Available at: www.standards.org/standards/listing/iso_20252 (accessed August 8, 2014).

International Standards Organization 2009. *Risk Mangement: Principles and Guidelines for Implementation*, ISO/DIS 31000 Standard No. 31000. Available at: www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43170 (accessed August 8, 2014).

Journal of Official Statistics 2013. *Special Issue on Systems and Architectures for High-Quality Statistics Production*, edited by B. Lorenc, I. Jansson, P. Biemer, J. Eltinge, and A. Holmberg, Vol. 1, March, 2013.

Juran, J. and B. Godfrey. 1999. *Juran's Quality Handbook*. New York: McGraw-Hill.

Karsak, E.E. 2004. "Fuzzy Multiple Objective Decision Making Approach to Prioritize Design Requirements in Quality Function Deployment." *International Journal of Production Research* 42: 3957–3974.

Keeter, S., C. Miller, A. Kohut, R. Groves, and S. Presser. 2000. "Consequences of Reducing Nonresponse in a Large National Telephone Survey." *Public Opinion Quarterly* 64: 125–148. DOI: http://dx.doi.org/10.1086/317759.

Kenett, R.S. and G. Shmueli. 2014. "On Information Quality." *Journal of the Royal Statistical Society,* Series A 177: 3–38. DOI:http://dx.doi.org/10.1111/rssa.12007.

Kish, L. 1962. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association* 57: 92–115.

Lequiller, F and D. Blades. 2006. *Understanding National Accounts*. Paris: OECD 2006. Available at: http://www.eastafritac.org/images/uploads/documents_storage/Understanding_National_Accounts_-_OECD.pdf (accessed June 21, 2013).

Lyberg, L. and P. Biemer. 2008. "Quality Assurance and Quality Control in Surveys." In *International Handbook on Survey Methodology*, edited by J. Hox, E. de Leeuw, and D. Dillman, 421–441. Mahwah, NJ: Lawrence Erlbaum Associates.

Lyberg, L., L. Japec, and P. Biemer. 1998. "Quality Improvement in Surveys – A Process Perspective." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 23–31.

Lyberg, L. 2012. "Survey Quality." *Survey Methodology* 38: 107–130.

McDavid, J., I. Huse, and L. Hawthorn. 2013. *Program Evaluation and Performance Measurement: An Introduction to Practice, Second Edition*. New York: Sage Publications.

Michalek, J.J., O. Ceryan, P.Y. Papalambros, and Y. Koren. 2006. "Balancing Marketing and Manufacturing Objectives in Product Line Design." *ASME Journal of Mechanical Design* 128: 1196–1204. DOI: http://dx.doi.org/10.1115/1.2336252.

Merkle, D. and M. Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and R. Little, 243–257. New York: John Wiley and Sons.

Morganstein, D. and D. Marker. 1997. "Continuous Quality Improvement in Statistical Agencies." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 475–500. New York: Wiley and Sons.

Nealon, J. and E. Gleaton. 2013. "Consolidation and Standardization of Survey Operations at a Decentralized Federal Statistical Agency." *Journal of Official Statistic* 29: 5–28. DOI: http://dx.doi.org/10.2478/jos-2013-0002.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–606.

Neyman, J. 1938. *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, DC: U.S. Department of Agriculture.

Organisation for Economic Cooperation and Development (OECD) 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Available at: http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en (accessed June 21, 2013).

Office of National Statistics (ONS) 2007. *Guidelines for Measuring Statistical Quality, Version 3.1*. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html (accessed June 21, 2013).

Rossi, P.H., W.M. Lipsey, and H.E. Freeman. 2004. *Evaluation: A Systematic Approach*. 7th ed. Thousand Oaks, CA: Sage Publishers.

Seyb, A., R. McKenzie, and A. Skerrett. 2013. "Innovative Production Systems at New Zealand: Overcoming the Design and Build Bottleneck." *Journal of Official Statistics* 29: 73–97. DOI: http://dx.doi.org/10.2478/jos-2013-0005.

Statistics Canada 2009. *Statistics Canada Quality Guidelines, Fifth Edition*. Available at: http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-539-X&CHROPG=1&lang=eng (accessed March 10, 2014).

Statistiska centralbyrån 2001. *Quality Definition and Recommendations for Quality Declarations of Official Statistics*. Available at: http://www.scb.se/Grupp/Hitta_statistik/Forsta_Statistik/Metod/_Dokument/MIS2001_1.pdf (accessed June 18, 2014).

Stephan, F.F. 1948. "History of the Uses of Modern Sampling Procedures." *Journal of the American Statistical Association* 43: 12–39.

Struijs, P., A. Camstra, R. Renssen, and B. Braaksma. 2013. "Redesign of Statistics Production within an Architectural Framework: The Dutch Experience." *Journal of Official Statistics* 29: 49–71. DOI: http://dx.doi.org/10.2478/jos-2013-0004.

U.S. Bureau of the Census 1974. "Technical Paper 32: Standards for Discussion and Presentation of Errors in Data. U.S. Department of Commerce." U.S. Government Printing Office, Technical Paper 32, Department of Commerce.

U.S. Office of Management and Budget 2002. "Guidelines for Ensuring, and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies." *Federal Register*, 67, 36, February 22.

Received August 2013
Revised April 2014
Accepted June 2014