Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Survey Quality

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

by Lars Lyberg

December 2012

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

This product is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
  Finance
  R.H. Coats Bldg., 6th Floor
  150 Tunney's Pasture Driveway
  Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2012

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement (http://www.statcan.gc.ca/reference/licence-eng.html).

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2012
Vol. 38, No. 2, pp. 107-130
Statistics Canada, Catalogue No. 12-001-X

Survey Quality

Lars Lyberg [1]

Abstract

Survey quality is a multi-faceted concept that originates from two different development paths. One path is the total survey error paradigm that rests on four pillars providing principles that guide survey design, survey implementation, survey evaluation, and survey data analysis. We should design surveys so that the mean squared error of an estimate is minimized given budget and other constraints. It is important to take all known error sources into account, to monitor major error sources during implementation, to periodically evaluate major error sources and combinations of these sources after the survey is completed, and to study the effects of errors on the survey analysis. In this context survey quality can be measured by the mean squared error and controlled by observations made during implementation and improved by evaluation studies. The paradigm has both strengths and weaknesses. One strength is that research can be defined by error sources and one weakness is that most total survey error assessments are incomplete in the sense that it is not possible to include the effects of all the error sources. The second path is influenced by ideas from the quality management sciences. These sciences concern business excellence in providing products and services with a focus on customers and competition from other providers. These ideas have had a great influence on many statistical organizations. One effect is the acceptance among data providers that product quality cannot be achieved without a sufficient underlying process quality and process quality cannot be achieved without a good organizational quality. These levels can be controlled and evaluated by service level agreements, customer surveys, paradata analysis using statistical process control, and organizational assessment using business excellence models or other sets of criteria. All levels can be improved by conducting improvement projects chosen by means of priority functions. The ultimate goal of improvement projects is that the processes involved should gradually approach a state where they are error-free. Of course, this might be an unattainable goal, albeit one to strive for. It is not realistic to hope for continuous measurements of the total survey error using the mean squared error. Instead one can hope that continuous quality improvement using management science ideas and statistical methods can minimize biases and other survey process problems so that the variance becomes an approximation of the mean squared error. If that can be achieved we have made the two development paths approximately coincide.

Key Words: Quality management; Total survey error; Quality framework; Mean squared error; Process variability; Statistical process control; Users of survey data.

1. Introduction

This article has been prepared in recognition of Joe Waksberg's unique contributions and leadership in survey methodology. My first encounter with Joe's work was his article on response errors in expenditure surveys written with John Neter (Neter and Waksberg 1964). Among other things that article introduced me to the cognitive phenomenon called telescoping. Later in life I had the opportunity to work with Joe on the first conference and monograph on telephone survey methodology where we were part of the editorial group (Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg 1988). We also collaborated on the preparation of many of the Hansen Lectures that were published in the Journal of Official Statistics (JOS) during my term as its Chief Editor. Joe himself delivered the sixth lecture, which was published in JOS (Waksberg 1998). Joe was a fantastic leader and it is a great honor for me to have been invited to write this article on survey quality, a topic that occupied his mind a lot.

Many of my friends have conveyed their views or sent me materials in preparation of this article. Especially I want to thank Paul Biemer, Dan Kasprzyk, Fritz Scheuren, Dennis Trewin, and Maria Bohata for helping me.

Survey quality is a vague, albeit intuitive, concept with many meanings. In this article I discuss some observations related to the development and treatment of the concept over the last 70 years and for some developments it is possible to trace roots that can be found even farther back. Most of my discussion, however, concerns current issues in government statistical organizations. It is within official statistics that most my survey quality examples take place.

The article is organized as follows: In Section 2 I discuss the total survey error paradigm, including error typologies, treatment of the errors, and survey design taking all error sources into account. In section 3 I discuss quality management philosophies that have had a large impact on survey organizations since the early 1990's. This impact is manifested by methods and approaches like recognition of the user or the client, a discussion of costs and risks in survey research, and the need for organizations to continuously improve. Section 4 provides examples of quality initiatives in survey organizations. Section 5 deals with the difficulties in measuring quality, either

[1] Lars Lyberg, Department of Statistics, Stockholm University, 10691 Stockholm, Sweden. E-mail: Lars.Lyberg@stat.su.se.

directly or indirectly via indicators. How these measures should be communicated to the users or clients is also covered. Section 6, finally, offers some thoughts about how survey practices *must* change to better serve the needs of the users. The last section contains references.

2. The total survey error paradigm

2.1 Some history of survey sampling

There are a number of papers describing the development of early survey sampling methodology. In that early development there is an implicit or explicit recognition of quality issues although they are hidden under labels such as errors and survey usefulness (Deming 1944). The historical overviews provided by, for instance, Kish (1995), Fienberg and Tanur (1996), and O'Muircheartaigh (1997) all emphasize the fact that the period up to 1950 is characterized by a full-bloom development of sampling theory. During the 1920s the International Statistical Institute agreed to promote ideas on representative sampling suggested by Kiear (1897) and Bowley (1913). In 1934 Neyman published his landmark paper on the representative method. Later Fisher's (1935) randomization principle was used in agricultural sampling and Neyman (1938) developed cluster sampling, ratio estimation and two-phase sampling and introduced the concept of confidence interval. Neyman showed that the sampling error could actually be measured by calculating the variance of the estimator. Bill Cochran, Frank Yates, Ed Deming, Morris Hansen and many others further refined the concepts of sampling theory. Hansen led a research group at the U.S. Census Bureau where much of the applied work and new theory development was conducted in those days. One remarkable result of the Census Bureau efforts was the two-volume textbook on sampling theory and methods (Hansen, Hurwitz and Madow 1953). As a matter of fact the advances in sampling theory were so prominent at the time that Stephan (1948) found it worthwhile to write an article about the history of modern sampling methods.

It was early recognized that there could be survey errors other than those attributed to sampling. There are writings on the effects of question wording such as Muscio (1917). Research on questionnaire design was quite extensive in the 1940s. Problems with errors introduced by fieldworkers collecting agricultural data in India were addressed by Mahalanobis (1946), resulting in a method for estimating such errors. The method is called "interpenetration" and can be used to estimate, so called, correlated variances introduced by interviewers, editors, coders and those who supervise these groups. The most prominent error sources were certainly known around 1950. Deming had listed error sources (1944) that constitute the first published typology of survey errors and Hansen and Hurwitz (1946) had discussed subsampling among nonrespondents in an attempt to provide unbiased estimates in a situation with an initial nonresponse. But the methodological emphasis, up to then, had been on developing sampling theory, which is quite understandable. It was very important to be able to show that surveys could be conducted on a sampling basis and in a variety of settings. By 1950 it had been demonstrated quite successfully that this was indeed possible. So it was time to move on to other issues and refinements.

In those early days the use of the word quality was confined to mainly quality control, sometimes as quality control of survey operations. It was common that the quality control was verification and/or estimation of error sizes for various operations. Statistics were known to be plagued by errors other than those stemming from sampling but the process quality issue of how to systematically reduce these errors and biases was still to be developed (Deming 1944; Hansen and Steinberg 1956).

The user 60 years ago was a somewhat obscure player, although not at all ignored by prominent survey methodology developers. For instance, Deming (1950) claimed that until the purpose is stated, there is no right or wrong way of going about a survey. Some other statisticians made similar statements. But the user was really hiding behind terms, such as subject-matter problem, study purpose or the key functions of a statistical system.

Even now survey and quality are vague concepts. As pointed out by Morganstein and Marker (1997) varying definitions of quality undermine improvement work so we should, at least, try to distinguish between different definitions to see what purposes they might serve. One of the most cited definitions is attributed to Joseph Juran, namely quality being a direct function of "fitness for use". It turns out that Deming already in 1944 used the phrase "fitness for purpose", not to define quality, but rather to explain what made a survey product work.

For a long time "good" quality was implicitly equivalent to a small mean squared error (MSE), *i.e.*, data should be accurate and accuracy of an estimate can be measured by MSE, which is the sum of the variance and the squared bias. We have noticed that survey statistics should also be useful, later denoted "relevant". Many of today's quality dimensions were not really an issue at the time. The users, too, were accustomed to the fact that surveys took time to carry out; timeliness was surely on the agenda but not as explicitly as it is today. A census took years to process. The users were accustomed to a technology that could only deliver relatively simple forms of accessibility. Hence, it was natural for users and producers to concentrate on making sure that the statistical problem coincided reasonably well with the subject-matter problem and that MSE was kept on a decent level, where MSE many times was and still is equivalent with just the variance, without a squared bias term added.

Before proceeding any further, let us define "survey". A *survey* is a statistical study designed to measure population characteristics so that population parameters can be estimated. Two examples of parameters are the proportion unemployed at a given time in a population of individuals, and the total revenue of a business or industry sector during a given time period. A survey can be defined as a list of prerequisites (Dalenius 1985a). According to Dalenius a study can be classified as a survey if the following prerequisites are satisfied:

1. The study concerns a set of objects comprising a population;
2. The population under study has one or more measurable properties;
3. The goal of the study is to describe the population by one or more parameters defined in terms of measurable properties, which requires observing (a sample of) the population;
4. To get observational access to the population a frame is needed;
5. A sample of objects is selected from the frame in accordance with a sampling design that specifies a probability mechanism and a sample size n (where n might equal N, the population size);
6. Observations are made on the sample in accordance with a measurement process (*i.e.*, a measurement method and a prescription as to its use);
7. Based on the measurements, an estimation process is applied to compute estimates of the parameters when making inference from the sample to the population under study.

This definition implicitly lists the specific error sources that are present in survey work. For each source there are a number of methods available that minimize the effects but also measure their sizes (Biemer and Lyberg 2003; Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2009).

Deviations from the definition reflect quality flaws. Moreover such deviations are common. In some designs selection probabilities are unknown or the variance estimator chosen might not be the most suitable one, given the sample design applied. Whether such flaws are problematic or not depends on the purpose.

2.2 The components of the total survey error paradigm

The total survey error paradigm is a theoretical framework for optimizing surveys by minimizing the accumulated size of all error sources, given budgetary constraints. In practice this means that we want to minimize the mean squared error for selected survey estimates, namely those that are considered most important by the main stakeholders. The mean squared error is the most common metric for survey work consisting of a sum of variances and squared bias terms from each known error source. Groves and Lyberg (2010) provide a summary of the status of the paradigm in the past and in today's survey practice.

The idea that surveys should be designed taking all error sources into account stems from the early giants in the field. Morris Hansen, Bill Hurwitz, Joe Waksberg, Leon Pritzker, Ed Deming and others at the U.S. Census Bureau, Leslie Kish at the University of Michigan, P.C. Mahalanobis at the Indian Statistical Institute, and Tore Dalenius, Stockholm University were among those who took the lead in survey research, emphasizing errors and optimal design. They worried about the inherent limitations associated with sampling theory since nonsampling errors could make the theory break down. They were very practical and thought a lot about balancing errors and the costs to deal with them. Some of them saw similarities between a factory assembly line (Deming and Geoffrey 1941) and the implementation of some of the survey processes and introduced control methods obtained from industrial applications.

Dalenius (1967) realized that there was as yet no "survey design formula" that could provide an optimal solution to the design problem. The approach taken by Dalenius and also Hansen, Hurwitz and Pritzker (1967) was a strategy of minimizing all biases and going for a minimum-variance scheme so that the variance became an approximation of the MSE. This was supposed to happen through intense verification schemes for ongoing productions and quite extensive evaluation studies for future productions. In 1969 Dalenius, inspired by Hansen, presented a paper on total survey design, where the word "total" reflected the thought about taking all error sources into account. Hansen, Hurwitz, Marks and Mauldin (1951), Hansen, Hurwitz and Bershad (1961), and Hansen, Hurwitz and Pritzker (1964) developed the U.S. Census Bureau Survey Model that reflected contributions from interviewers, coders, editors, and crewleaders and allowed the estimation of those contributions to the total survey error. These estimation schemes were elaborated on by Bailar and Dalenius (1969) and consisted of variations of replication and interpenetration. Bias estimation was assumed to be handled by comparing estimates obtained from the regular operations with those obtained from preferred procedures (that could not be used on a large scale due to financial, administrative or practical reasons). Today this kind of approach is called the "gold standard".

It was stated that good survey design called for reasonably effective control of the total error by careful specifications of the survey procedures, including adequate controls. Hansen, Deming and others did worry about control costs but although statistical process control and acceptance sampling had been implemented in a number of survey organizations, there was very little discussion about continuous process improvement. A lot of the quality work had to do with estimation of error rates, controlling error levels for individual operators and conducting large-scale evaluation studies that usually took a long time. Users were not directly involved in the design process but in the U.S. federal statistical system they had at least some influence on what should be collected and presented. Dalenius (1968) provides more than 200 references on users and user conferences associated with the products of the U.S. Federal statistical system.

While total survey design was first advocated by Hansen, Dalenius and others, users were seldom directly involved in the final determination of survey requirements. Quite often an official, administrator or statistician acted as a subject-matter specialist. Several decades ago this was the way we thought about users. Their opinions counted but they were not really involved in design decisions. Lurking in the back of our heads was the thought that this might not be a perfect model and in the late 1970's Statistics Sweden published an internal booklet called "What to do if a customer shows up on our doorstep".

The basic design approach suggested by Hansen, Dalenius and others contained a number of steps including:

- Specification of an ideal survey goal.
- Analysis of the survey situation regarding financial, methodological and information resources.
- Developing a small number of alternative designs.
- Evaluating the alternatives by reference to associated preliminary assessments of MSE and costs.
- Choosing one of the alternatives or a modification of one of them or deciding not to conduct a survey at all.
- Developing the administrative design including feasibility testing, a process signal system (currently called paradata), a design document, and a Plan B.

Kish (1965) had slightly different views on design. He liked the neo-Bayesian applications in survey sampling and psychometrics advocated by colleagues at the University of Michigan (Ericson 1969; Edwards, Lindman and Savage 1963). For instance, Kish liked the idea that judgment estimates of measurement biases might be combined with sampling variances to construct more realistic estimates of the total survey error. Regarding the optimization problem Kish thought that the multipurpose situation was economically favorable for surveys but that it could be difficult to decide on what to base the design on. If one principal statistic can be identified then that alone can decide the design and if there are a small number of principal statistics a compromise design is possible but if statistics are too disparate a reasonable design might not exist. Kish also emphasized the need for design information obtained from pilot surveys and pretests to facilitate design decisions. Kish noted that survey design and measurement could vary greatly across environments while sampling did less so. That could be one reason that sampling can be easily placed among the traditional statistical theories and methods, while it is more difficult to place the survey process in one specific discipline (Frankel and King 1996 in their interview with Kish).

Kish, like the other giants, emphasized the importance of small biases but appreciated the fact that the reduction of one bias term might increase the total error. Kish was keen on getting a reasonable balance between different error sources and how error structures varied under different design alternatives. Like Hansen and colleagues Kish thought that relevant information should be contemporaneously recorded during implementation (again we see the parallel to paradata). Hansen and colleagues were really concerned about excessive but inadequate controls. They realized that some controls might have to be relaxed due to limited improvements and that degree of improvement in terms of affecting the estimates should be checked out before any relaxation could take place. They also suggested that one might have to compromise relevance to get controllable measurements or abstain from the survey. Both Hansen and colleagues and Kish were vigorously in favor of ending the practice that sampling error is the only survey error measured.

When we look at today's situation we can conclude that we still do not have a design formula for surveys. There is no planning manual to speak of and the literature on design is consequently very small, as is the literature on cost (Groves 1989 is an exception). And no design formula is in sight. Since the advent of the U.S. Census Bureau survey model a number of variants have appeared on the scene, some of them quite complicated (Groves and Lyberg 2010). A common characteristic is the fact that they tend to be incomplete, *i.e.*, they do not take all error sources into account. Most statistical attention is on variance components and especially on measurement error variance. There are a number of other weaknesses associated with the total survey error concept. Most notably a user perspective is missing and a vast majority of users are not in a position to question or even discuss accuracy. The complex error structures and interactions do not invite outside scrutiny and user contacts often tend to concern less technical issues such as timeliness, comparability and costs. Users are not really informed about real levels of accuracy and we know very little about how users perceive information about errors and how to act on that.

As pointed out by Biemer (2001), in his discussion of Platek and Särndal (2001), there is a lack of routine measurements of MSE components in statistical organizations. There are good reasons for this state of affairs. Complexity has already been mentioned and to that we can add factors such as costs, the fact that it is almost impossible to publish such information at the time data are released, and that there is no measure of total error that would take all error sources into account, either because a lack of proper methodology or that some errors defy expression. Groves and Lyberg (2010) list some other weaknesses associated with the total survey error paradigm. For instance, we need to know more about the interplay between variances and biases. It is possible that an increase in simple response variance goes hand in hand with a reduction in response bias, say, when we compare interview mode with self-administrative alternatives. Recently, West and Olson (2010) showed that interviewer variance can occur not only from individual interviewers' effect on the responses within their assignments but also because individual interviewers successfully obtain cooperation from different groups of sample members.

Despite all its limitations, the strengths of the total survey error framework are quite convincing. The framework provides a taxonomic decomposition of errors, it separates variance from bias and observation from nonobservation, and it defines the different steps in the survey process. It serves as a conceptual foundation of the field of survey methodology, where subfields are defined by their associated error structures. Finally, it identifies the gaps in the research literature since any typology will show that some process steps are more "popular" than others. Just compare the respective sizes of the literatures on data collection and data processing.

It seems, however, as if the total survey error framework needs some expansion along lines some of which were identified half a century ago. We need some guidance on trade-offs between measuring error sizes and making processes more error-free. Spencer's (1985) question is: how much should we spend on measuring quality versus quality enhancement? We also need some guidance on how to integrate additional notions into the framework, so that it becomes a total survey quality framework rather than a total survey error framework (Biemer 2010). For instance, if "fitness for use" predominates as a conceptual base, how can we launch research that incorporates error variation associated with different uses? This aspect will be discussed in the next section.

3. Quality management philosophies in survey organizations

During the late 1980's and the early 1990's some statistical organizations were under severe financial pressure and in some cases simultaneously criticized for not being sufficiently attentive to user needs. Governments in Sweden, Australia, New Zealand and Canada as well as the Clinton administration in the U.S. were all keen on improving efficiency and user influence within their respective statistical systems. It was natural for these organizations to look for inspiration in management theories and methods (Drucker 1985) and specifically on what was called quality management (Juran and Gryna 1988). In that newer literature it was possible to study the role of the customer, leadership issues, the notion of continuous quality improvement, and various tools that could help the statistical organization improve. Especially influential to survey practitioners was work by Deming (1986), since he emphasized the role of statistics in quality improvement. He vigorously promoted the idea that improvement work should be led by statisticians, since they are trained in distinguishing between different kinds of process variation. He thought that there were too few statistical leaders advising top management in businesses and he wanted more proactive statisticians to become such leaders. He was especially keen on developing Shewhart's ideas about control charts as a means to distinguish between the different types of variation, namely common and special cause variation. Shewhart's improvement cycle Plan-Do-Check-Act was also part of Deming's thoughts on quality (Shewhart 1939).

Management principles have, of course, existed since ancient times. Juran (1995) provides lots of examples of what was in place in, for instance, the Roman empire. Craftsmanship and a guild system were basic building blocks. There were methods for choosing raw materials and suppliers. Processes were inspected and improved. Workers were trained and motivated and customers got warranties. All these features are found also in today's management systems. The more modern development includes quality frameworks or business excellence models such as Total Quality Management (TQM), International Organization for Standardization (ISO) standards, the Malcolm Baldrige quality award criteria, the European Foundation for Quality Management (EFQM), Six Sigma, Lean Six Sigma, and the Balanced Scorecard. These models are not totally different. They often share a common set of values and common criteria for excellence. Rather they represent a natural development that can be seen in all kinds of work.

Thus, there has been a gradual adoption of quality management models and quality strategies in statistical organizations and a merging with concepts and ideas already used in statistical organizations. My personal timeline for this development is the following (readers are invited to come up with different sets of events and dates):

1875 Taylor introduces what he called scientific management;

1900-1930 Taylor's ideas are used in, for instance, Ford's and Mercedes Benz's assembly lines;

1920's Fisher starts developing theories and methods for experimental design;

1924 Shewhart develops the control chart;

1940 The U.S. War Department develops a guide for analyzing process data;

1944 Deming presents the first typology of survey errors;

1944 Dodge and Romig present theory and tables for acceptance sampling;

1946 Deming goes to Japan;

1950 Ishikawa suggests the fishbone diagram as a tool for identifying factors that have a profound effect on the process outcome;

1954 Juran goes to Japan;

1960 Many businesses embark on a zero defects program;

1960 The U.S. Census Bureau quality control programs are developed;

1961 The U.S. Census Bureau survey model is launched;

1965-1966 Kish and Slobodan Zarkovich start talking about data quality rather than survey errors;

1970's Many statistical organizations provide quality guidelines;

1975 The Total Quality Management (TQM) framework is launched;

1976 The first quality framework in a statistical organization containing more dimensions than relevance and accuracy;

1987-1989 Launching of the ISO 9000, Malcolm Baldrige Award, Six Sigma and EFQM models;

1990's Many statistical organizations start working with quality improvement and excellence models;

1997 The Monograph on Survey Measurement and Process Quality;

1998 Mick Couper introduces the concept "paradata" as a subset of process data;

2001 The Eurostat leadership group on quality organizes the first conference on Quality Management in Official Statistics;

2007 Business architecture ideas enter the survey world.

From the mid 1990's and on quality management philosophies have had an enormous effect on many statistical organizations. The effect is not necessarily higher quality across the board (no one has checked that). But the philosophies have led to an awareness in most organizations of the importance of good contacts with users and clients, and an aspiration in many of them to become "the best" or "world class". Quality is on the agenda.

3.1 The concept of quality

During the last decades it has become obvious that accuracy and relevance are necessary but not sufficient when assessing survey quality. Other dimensions are also important to the users. The development of survey quality frameworks has taken place mainly within official statistics and has been triggered by the rapid technology development and other developments in society. These advanced technologies have created opportunities and user demands regarding potential quality dimensions such as accessibility, timeliness, and coherence that simply were not emphasized before. Decision-making in society has become more complex and global resulting in demands for harmonized and comparable statistics. Thus, there is a need for quality frameworks that can accommodate all these demands. Several frameworks of quality have been developed and they each consist of a number of quality dimensions. Accuracy and relevance are just two of these dimensions.

For instance, the framework developed by OECD (2011) has eight dimensions: relevance, accuracy, timeliness, credibility, accessibility, interpretability, coherence, and cost-efficiency (Table 1). Similar frameworks have been developed by Statistics Canada (Statistics Canada 2002; Brackstone 1999), and Statistics Sweden (Felme, Lyberg and Olsson 1976; Rosén and Elvers 1999). The Federal Statistical System of the U.S. has a strong tradition in emphasizing the accuracy component (U.S. Federal Committee on Statistical Methodology 2001) although it certainly appreciates other dimensions. Perhaps they are viewed as dimensions of a more nonstatistical nature that still need a share of the total survey budget. The International Monetary Fund (IMF) has developed a framework that differs from those of OECD, Australian Bureau of Statistics, Statistics Sweden, and Statistics Canada. IMF's framework consists of a set of prerequisites and five dimensions of quality: integrity, methodological soundness, accuracy and reliability, serviceability, and accessibility (see Weisman, Balyozov and Venter 2010).

Table 1
OECD's quality framework

Dimension	Description
Relevance	Statistics are relevant if users' needs are met.
Accuracy	Closeness between the value finally retained and the true, but unknown, population value.
Credibility	The degree of confidence that users place in data products based on their image of the data provider.
Timeliness	Time length between data availability and the event or phenomenon data describe.
Accessibility	How readily data can be located and accessed from within data holdings.
Interpretability	The ease with which the data user may understand and properly use and analyze the data.
Coherence	Reflects the degree to which data products are logically connected and mutually consistent.
Cost-efficiency	A measure of the costs and provider burden relative to the output.

Without sufficient accuracy, other dimensions are irrelevant but the opposite is also true. Very accurate data can be useless if they are released too late to affect important user decisions or if they are presented in ways that are difficult for the user to access or interpret. Furthermore, quality dimensions are often in conflict. Thus, providing a quality product is a balance act where informed users should be key players. Typical conflicts exist between timeliness and accuracy, since it takes time to get accurate data through, for instance, extensive nonresponse follow-up. Another conflict is the one between comparability and accuracy since application of new and more accurate methodology might disturb comparisons over time (Holt and Jones 1998).

Thus, many organizations have adopted a multi-faceted quality concept consisting not only of accuracy but also other dimensions. We might talk about a quality vector whose components vary slightly between organizations both in number and in contents. There are a number of problems associated with the quality vector approach.

First, the development has not been preceded by user contacts. Producers of statistics have believed that users are interested in a specific set of dimensions even though it is obvious that a vast majority of users think that error structures are too complicated to grasp and assume that the producer should be responsible for delivering the best possible accuracy. In cases where the user or client has specific accuracy requirements a more in-depth dialog can take place between the two. In the rare studies that have investigated user perceptions of information on quality it turns out that users are mostly interested in dimensions that are easily understood, such as timeliness and indicators that are seemingly straight forward, such as response rates. The user wants the producing statistical organization to be credible, which translates into being capable of producing data with small or at least known errors and delivering them in a timely, reliable, and accessible fashion. The thought that it would be possible to produce a total quality measure based on weighted assessments of the different dimensions is not realistic, although Mirotchie (1993) argues to the contrary. In that paper Mirotchie makes a case for a standard set of quality indicators and provides a hypothetical illustration of indexing data quality indicators and computing an actual index (in this illustration the indicators are precision, nonresponse, reliability, timeliness and residuals). Even if a composite indicator in the form of an index were a possible development, the user would like to know which indicators contributed most to an index value. From a user's point of view the least favorable index value could still reflect a situation providing the highest quality. Rarely can a low accuracy be compensated by good ratings on other dimensions, not even in the case of election exit polls where timeliness is imperative. Accuracy is still necessary and there is wide agreement that all reputable organizations should meet accuracy standards (Scheuren 2001; Kalton 2001; Brackstone 2001). Phipps and Fricker (2011) provide an overview of quality frameworks and literature on total survey error. Thus, we can agree that survey quality is a multi-faceted concept involving multiple features of a statistical product or service.

3.2 The quality movement's impact on statistical organizations

Just extending the quality framework from one or two dimensions to several is not sufficient to create a quality environment. In the late 1980's and early 1990's many statistical organizations became interested in quality issues beyond traditional aspects of data quality. Issues concerning customer satisfaction, communicating with customers, competition, process variability, cost of poor quality, waste, business excellence models, core values, best practices, quality assurance, and continuous quality improvement were suddenly part of the everyday activities in many organizations.

Successful organizations know that continuous improvement (Kaizen) is necessary to stay in business and they have developed measures that help them change. This is true also for producers of statistics. Changes that are supposed to improve the statistical product are triggered by user demands, competition from other producers and from producer values that emphasize continuous improvement as part of the general business environment. The measures that can help a statistical organization improve are basically identical to those of other businesses. They can be built on business excellence models such as the European Foundation for

Quality Management (EFQM) (1999). The core values of the EFQM model include results orientation, customer focus, leadership and constancy of purpose, management by process measures and facts, personnel development and involvement, continuous learning, innovation and improvement, development of partnerships, and public responsibility. This model has been adopted by the European Statistical System (ESS) as a tool for national statistical institutes in Europe for achieving organizational quality. The thought is that good product quality, according to the dimensions mentioned (or some other product quality definition) cannot be achieved without good underlying processes used by the organization. It can also be argued that good product quality is achieved most efficiently and reliably by good process quality. If we view quality as a three-level concept it can be visualized as shown in Table 2.

3.2.1 Product quality

The deliverables agreed upon are called the product. It can be one or several estimates, datasets, analyses, registers, standard processes or other survey materials such as frames and questionnaires. Product quality is the traditional quality concept used when informing users or clients about the quality of the product or service. It can be measured and controlled by means of degree of adherence to specifications and requirements for product characteristics adding up to quality dimensions of a framework. Measures of accuracy and margins of error belong here. Also observations whether service levels agreements established with the client have been accomplished are relevant. In line with quality management principles, it is also quite common to conduct user satisfaction surveys to find out what users think about the products and services that are provided.

3.2.2 Process quality

All processes have to be designed so that they deliver what they are supposed to. This means that we have to have some kind of quality assurance perspective when processes are defined. For instance, the process of interviewing implies that a number of elements must be in place for the process to deliver what is expected. Examples of elements are an effective selection of interviewers and a training program, a compensation system as well as supervision and feedback activities. Thus we aim at building quality into the process via the quality assurance. Quality control efforts are only used to check if the process works as intended. It cannot by itself be used to build quality into the process. In Section 4.4 this process view is discussed in more detail. Process quality is measured and controlled via selection, observation and analyses of key process variables, so called process data or paradata (Morganstein and Marker 1997; Couper 1998; Lyberg and Couper 2005). Theory and methods imported from statistical process control can help the producer distinguish between the two types of variation, common and special cause. As long as all variation is contained within the upper and lower control limits associated with the control charts chosen, the process is said to be in statistical control and no process improvements are really possible by trying to adjust individual outcomes. If there are observations falling outside of the control limits, usually set at 3 sigma, then we have indications of special cause variation that should be taken care of so that the variation after adjustment is brought back to common cause variation. The following P-chart illustrates a possible situation:

Process control chart with more extreme values

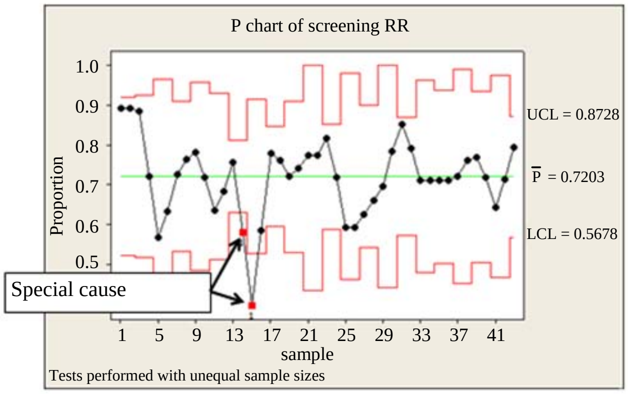

Table 2
Quality as a three-level concept*

Quality level	Main stake-holders	Control instrument	Measures and indicators
Product	User, client	Product specs, SLA, evaluation studies, frameworks, standards	Frameworks, compliance, MSE, user surveys
Process	Survey designer	SPC, charts, acceptance sampling, risk analysis, CBM, SOP, paradata, checklists, verification	Variation via control charts, other paradata analyses, outcomes of periodic evaluation studies
Organization	Agency, owner, society	Excellence models, ISO, CoP, reviews, audits, self-assessments	Scores, strong and weak points, user surveys, staff surveys

* SLA (Service Level Agreement), SPC (Statistical Process Control), CBM (Current Best Methods), SOP (Standard Operating Procedures), and COP (ESS Code of Practice).

Thus, the action sequence is the following. First the roots of the special causes are taken care of so that these variations are eliminated. After that the process displays common cause variation only. If that variation is deemed too large then the process has to change. The kinds of changes necessary are seldom obvious at the outset. Indeed perhaps several are necessary to decrease the process variation. Typically, a process improvement project is needed and the quality management literature has promoted a number of tools that are useful in such projects. Most of these tools are borrowed from statistics (control charts, experiments, regression analysis, Pareto diagrams, scatter plots, stratification) but there are also tools for identifying probable problem root causes (fishbone diagrams, process flow charts, brainstorming). The common thinking is that improvement projects should be "manned" by people working with the process or by people very much familiar with the process in other ways. Sometimes, we talk about forming an improvement team, where also the client or customer participates. In any improvement work suggested changes have to be tested. When Shewhart first developed his control charts he also suggested that improvement work should follow a sequence of operations, Plan-Do-Check-Act. What this sequence tells us is that any process changes suggested should be tested to see if they actually improve the process. If not, another change is made, and testing done again. Deming called this line of thinking the Shewhart cycle but since Deming spent a lot of time promoting it, many eventually called it the Deming cycle. The changes sought after could be decreased process variation, reduced costs, or increased customer satisfaction. The improvement project methodology is described in for instance Joiner (1994), Box and Friends (2006), Breyfogle (2003), and Deming (1986).

Another way of checking the process quality is to use acceptance sampling. Acceptance sampling (Schilling and Neubauer 2009) can be applied in situations where process elements can be grouped in batches. The batches are controlled and based on the outcome of that control it is decided whether the batch should be approved or reworked. Acceptance sampling plans guarantee an average outgoing quality in terms of, say, error rate, but there is no direct quality improvement involved. It is a control instrument that is suitable for operations such as coding, editing and scanning and then only when these processes are not really in statistical control. The method has been heavily criticized by Deming (1986) and others but can be the only control means available in situations where staff turnover is high and there is no time to wait for stable processes.

Global paradata (Scheuren 2001) are "error" rates of different kinds. Examples include nonresponse rates, coding error rates, scanning error rates, listing error rates, *etc*. In some operations the error rates are calculated using verification, which means that the operation is repeated in some way. That is the case for the coding operation. In other operations the calculation can be based on a classification scheme, which is the case for nonresponse rate calculation. These global paradata tell us something about the process. They are process statistics, *i.e.*, summeries of data. A large nonresponse rate indicates problems with the data collection process and a high coding error rate indicates problems with the coding process. From these summaries it is sometimes possible to distinguish common and special cause variation and decide what action to take.

Some standardized processes can be controlled by means of simple checklists. Checklists are very effective when it is crucial that every process step is made and in the right order (Morganstein and Marker 1997). This is the case when airline pilots prepare for take-off. No matter how many times they have taken off, without a checklist the day will come when they forget an item. In statistics production sampling is such a process, albeit with less severe consequences if items are missed. It might very well be the case that a statistical organization has a standardized process for sample selection and a checklist that can be used as a combination of work instruction and control instrument.

There is a kind of checklist that can be used in more creative processes such as the overall survey design process. It is not possible to standardize the survey design process but it is possible to list a number of critical steps that always must be addressed. The list does not tell us how to address them. It just serves as a reminder that an individual step should not be omitted or forgotten. Morganstein and Marker (1997) discuss this kind of checklist and call them (and the simpler checklists) Current Best Methods (CBM). They describe the CBM development process and how the CBMs can be used to decrease the process variation in statistical organizations. For instance, an organization might have seven different imputation methods and systems in its toolbox. It is costly to maintain these seven systems. It is unlikely that they are equally efficient. If they are, it may not be economically feasible to keep them all. In this situation a CBM that describes fewer options to the organization seems like a good idea. This could be accomplished by forming an improvement team consisting of the imputation experts and some clients. CBMs are supposed to be revised when new knowledge is obtained, which implies that there is an expiration date associated with every CBM.

CBMs are of course "best practices" in some sense. Many organizations want best practices implemented and used. Morganstein and Marker offer a process for developing these best practices and keeping them current. It is beneficial for an organization if the variation in process design can be kept at a minimum. It then becomes easier to train people and change the process when it becomes unstable or when new methods are developed. On the other hand, if CBMs and other standards are not vigorously enforced within an organization, they will not be widely used and the investment will not pay off.

3.2.3 Organizational quality

Management is responsible for quality in its widest sense. It is the organization that provides leadership, competence development, tools for good customer relations, investments, and funding. The quality management field has given us business excellence models that can help us evaluate our statistical organizations in the same way other businesses are evaluated. The two main business excellence models are the Baldrige National Quality Program and the European Foundation for Quality Management (EFQM).

These models consist of criteria to be checked when assessing an organization. The Malcolm Baldrige award uses seven main criteria: Leadership, strategic planning, customer and market focus, information and analysis, human resource focus, process management, and business results. Each criterion has a number of subcriteria. For instance, human resource focus consists of work systems, employee education, training and development, and employee well-being and satisfaction. The EFQM model has nine criteria: Leadership, strategy, people, partnerships & resources, processes, products & services, customer results, people results, society results, and key results. These models can be used for self-assessment or external assessment. The organization provides a description of what is in place regarding each criterion and the organization is scored based on that description. Typically self-assessments result in higher scores than external ones. It is very difficult to get a high score from external evaluators since the models are very demanding. For each criterion the organization is asked if it has a good approach in place somewhere in the organization. This is often the case. The next question is how wide-spread this good approach is within the organization. Many organizations lose momentum here, since there is very little truth in the mantra "the good examples are automatically spread throughout an organization". Instead good approaches usually have to be vigorously promoted before they are accepted within the organization. The third question asks whether the approach is periodically evaluated to check if it achieves the results expected. This is where most organizations fail. Their usual strategy is to exhaust an approach until the problems are so great that the approach has to be replaced rather than adjusted. This strategy is, of course, disruptive and expensive and does not score highly in excellence assessments. The maximum number of points that can be obtained using these models is 1000 and very rarely does a winner get more than 450-600 points, which is an indication that there is a lot of room for improvement even in world class organizations.

Some statistical organizations have used business excellence models for assessment. The Czech Statistical Office was announced Czech National Quality Award Winner for 2009 in the Public Sector category based on EFQM. The office got 464 points. Eurostat's leadership group on quality recommended the European national statistical offices to use the EFQM as a model for their quality work and Finland and Sweden are among those that have done so. Since the leadership group released its report in 2001 (see Lyberg, Bergdahl, Blanc, Booleman, Grünewald, Haworth, Japec, Jones, Körner, Linden, Lundholm, Madaleno, Radermacher, Signore, Zilhao, Tzougas and van Brakel 2001) other frameworks and standards have been developed. The European Statistical System has launched its Code of Practice, which consists of a number of principles with associated indicators. Regarding some principles, however, the indicators are more like clarifications. The list of principles resembles other lists that have been developed by the UN and other organizations.

External assessments are probably more reliable than internal ones. There are a number of reasons for that. One is that it is difficult to criticize your peers since you have to interact with them in the future or if your own product or service will be assessed by those peers in the future. Experiences from Statistics Sweden and Statistics Canada show that self-assessments are limited in their capability of identifying serious weaknesses (see Section 5.3).

3.2.4 Some specific consequences for statistical organizations

Most statistical organizations have adopted quality management ideas to varying degrees and with varying success. As pointed out by Colledge and March (1993) it is possible to list a number of obstacles associated with such implementation. For a government agency it can be difficult to motivate its staff through monetary incentives, since there are restrictions on how tax money can be spent. The variety of users and products makes the dialog between the service provider and the user complicated and as mentioned neither the users, or for that matter the providers are totally familiar with all the biases and other quality problems that are present in statistics production. The effect of errors on the uses can vary and are often unknown. To complicate matters further, unlike most other businesses, suppliers are not very enthusiastic. In other businesses suppliers get paid while statistical organizations must motivate theirs, the respondents, who are seldom even given a cash incentive.

On the other hand statistical organizations have a great advantage when it comes to applying quality management principles. A statistical organization knows how to collect and analyse data that can guide improvement efforts. One of the cornerstones in quality management philosophies is that decisions should be based on data and businesses that do not have support from statisticians are often unaware of data quality problems, which can have consequences for their decision-making. By and large, though, a statistical organization is not different from any other business and it is quite possible to apply quality management ideas to improve all aspects of work.

4. Examples of quality initiatives in statistical organizations

In this section we will provide some examples of initiatives that statistical organizations have engaged in as a result of a general interest in quality in society.

4.1 The total survey error

Perhaps the most important thing to notice is that research and development in survey design, implementation, sampling and nonsampling errors, and the effect of errors on the data analysis continue to thrive. Data with small errors is the major goal for reputable organizations, which is indicated by the steady flow of textbooks on data collection, sampling, nonresponse, questionnaire design, measurement errors, and comparative studies. New textbooks are in progress covering gaps such as business surveys, translation of survey materials, and paradata. There are journals such as the *Journal of Official Statistics*, *Survey Methodology*, and *Survey Practice* that are entirely devoted to topics related to statistics production in a wide sense. Numerous other journals such as the *Public Opinion Quarterly*, the *Journal of the American Statistical Association*, and the *Journal of the Royal Statistical Society* devote much space to survey methods. The Wiley series on Survey Methodology and its associated conferences (on panel surveys, telephone survey methods (twice), measurement errors, process quality, business surveys, testing and evaluating questionnaires, computer assisted survey information collection, nonresponse, and comparative surveys) have been very successful and that is the case also for the continuing workshops on nonresponse and total survey error. Thus, there is no shortage of ideas regarding specific error sources and their treatment. Admittedly there are areas that are understudied such as specification errors, data processing errors and the impact of errors on the data analysis but by and large there is a healthy interest in knowing more about survey errors. The challenge lies in communicating this knowledge to people working in statistical organizations and in developing design principles that can be used to improve statistics production. There is a noticeable gap between what is known through research and what is known and applied in the statistical organizations. Thus, staff capacity building seems to be a continuing need, especially since the common idea that good examples spread like ripples within and between organizations is a myth. If that indeed were the case quality would by now be fantastic everywhere. Since it is not, many organizations have developed extensive training programs (Lyberg 2002).

4.2 Risk and risk management

One element of quality management that has entered the survey world is risk and risk management. Eltinge (2011) even talks about Total Survey Risk as an alternative to the total survey error paradigm. The identification and management of risks is an important part of modern internal auditing (Moeller 2005) and is perhaps the only major element that is missing in quality management frameworks such as EFQM. An error source can be seen as more risky than another and should, therefore, be handled with more care and resources than another less risky. For instance, not having an effective system for statistical disclosure control is seen as a very risky situation. Unlawful data disclosure is very rare historically, but when it happens it could potentially destroy future data collection attempts. Certain design decisions can be seen as risky. For instance, if we choose a data collection method that does not fit the survey topic we might get estimates that are so far from the truth that the results are useless. An example might be to study sensitive behaviors using face to face or telephone interviewing instead of a self-administered mode. There are also technical risks that need to be identified and assessed. For instance, the U.S. National Agricultural Statistical Service (Gleaton 2011) like many others has plans for disaster recovery. Groves (2011) and Dillman (1996) both discuss how the production culture and the research culture within a statistical organization might view risks in different ways. Change in statistical organizations is generally slow and there are sometimes good reasons for that. Change might result in failures such as unsuccessful implementation, large costs and decreased comparability. So in some sense both producers and users have a tendency to be hesitant toward changes suggested by researchers and innovators and that might be one reason why change takes a long time. It is very common to have parallel measurements for some time to handle risks associated with implementing a new method or system. According to Groves (2011) the production culture and the users have had the final say about any changes, at least up until now. At the same time innovation is badly needed in many production systems and there are examples of stove-pipe organizations that do not have much time left (to remain unchanged) because the resources to maintain their systems are simply not there. So even though there is resistance against change, lack of resources and competition will make sure that statistical organizations become more process-oriented and efficient. Reducing the number of systems and applications and developing and using more standardization seem to be one road forward.

4.3 The client/customer/user

The advent of quality management ideas in statistical organizations has made the receivers of statistical products and services more visible. Commercial firms have always talked about the client or the customer while government organizations have tended to call them users. In any case the recognition of someone who is supposed to use the endproducts has not been obvious to some providers. Admittedly the user has been a speaking partner since the beginning of the survey industry. In the U.S., conferences for users were quite frequent already 50 years ago (Dalenius 1968; Hansen and Voight 1967). During six months 1965-66, for example, the U.S. Census Bureau organized 23 user conferences across the country and there were also advisory groups. The advisory nature of contacts with users has prevailed in many countries. The user conference format still exists but user input is now complemented by other means such as public discussions and internet forums. Rarely have users been directly involved in the planning and design of surveys. Even when it comes to discussions about the quality of data, producers have acted as stand-in users. The quality frameworks are a good example. The quality dimensions were defined with minimal consultation with users. The literature on how users perceive information about quality is extremely limited (Groves and Lyberg 2010). Also, we do not know if the information on quality that we provide is useful to them (Dalenius 1985b). In fact, an educated guess is that many times it is not. In many surveys the users are many and sometimes unknown and their information and analytical needs cannot be foreseen ahead of time. It is often possible to single out one or a few main users to communicate with, but many of the design and quality problems are so complicated that a vast majority of users expect the service provider to deliver a product with the smallest possible error. Hansen and Voight stated that accuracy should be sufficient to avoid interpretation problems. Today there seems to be consensus among many that what users are interested in are products and services that can be trusted, *i.e.*, the service provider should be credible. It is impossible for most users to check levels of accuracy. Aspects that an average user can discuss are issues such as timeliness, accessibility and relevance. Detailed discussions about technical matters and design trade-off issues including accuracy and comparability are more difficult to have.

During recent decades the user has indeed become more prominent. Some organizations develop service level agreements together with a main user or client, where requirements of the final product or service are listed and can be checked at the time of delivery. Many organizations conducting business surveys have created units that continuously communicate with the largest businesses, since their participation and provision of accurate information is absolutely essential for the estimation process (Willimack, Nichols and Sudman 2002). The large businesses are not users in the strict sense. They are important suppliers often with an interest in the survey results. Another common communication tool is the customer satisfaction survey. The value of such surveys is limited due to the acquiescence phenomenon and problems finding a knowledgeable respondent who is also willing to respond. Also, many customer satisfaction surveys are based on self-selection resulting in zero inferential value. In those surveys the results can only be viewed as lists of issues and concerns that some customers convey. Such information can, of course, be very valuable but is not suitable for estimation purposes. Many survey organizations now conduct user surveys on a continuing basis (Ecochard, Hahn and Junker 2008).

4.4 The process view

Quality management has reemphasized the importance of having a process view in statistics production. To view the production process as a series of actions or steps towards achieving a particular end that satisfies a user, leads to a good product quality. Process quality is an assessment of how far each step meets defined requirements or specifications. One way of controlling the process quality is to collect process data that can vary with each repetition of the process. The interesting process variables to monitor are those that have a large effect on the process's end result. Thus to check a process for stability and variation we need mechanisms for identifying, collecting and analysing these key process variables. The quality management science has given us tools such as the Ishikawa fishbone diagram to identify candidates for key process variables. The statistical process control methodology has given us tools to distinguish between special and common cause variation and how to handle these two variation types. Usually we use control charts originally developed by Shewhart (Deming 1986; Mudryk, Burgess and Xiao 1996) to make those distinctions. Then, again, we use methods from quality management to adjust the process if necessary. Examples include flowcharts, Pareto diagrams, and other simple means for the production team to identify the root causes of problems (Juran 1988).

Process data have been used to check on processes used in statistics production since the 1940's, first within the U.S. Census Bureau and then at Statistics Canada and to some extent also in other agencies. Typical processes that were checked included coding, keying and printing and the process data were mainly error rates. Some of the process checks used at the U.S. Census Bureau were so complicated and expensive that their value was questioned (Lyberg 1981), especially since the associated feedback loops were inefficient and not always aiming for the root causes of the errors. It was common that operators were blamed for system problems and at the time there was no emphasis on continuous quality improvement. The thinking at the time was more directed toward verification and correction.

Morganstein and Marker (1997) developed a generic plan for process continuous improvement that can be used in statistics production. They had worked in many statistical organizations since the 1980's and observed that quality thinking was not really developed in most of them. Their generic plan was built on their first-hand experiences and the general quality management ideas laid out by *e.g.*, Juran (1988), Deming (1986), Box (1990), and Scholtes, Joiner and Streibel (1996). In essence the plan consists of seven steps:

- The critical product characteristics are identified together with the user, both broad and more single effort needs.
- A map of the process flow is developed by a team familiar with the process. The map should include the sequence of process steps, decision points and customers for each step.
- The key process variables are identified among a larger set of process variables.
- The measurement capability is evaluated. It is important that decisions are based on good data, not just data. Available data might be useless. This is an area where statistical organizations should have an advantage over other organizations. One should not reach conclusions about process stability without knowledge about measurement errors. Above all, data should allow quantification of improvement.
- The stability of the process is determined. The variability pattern of the process data is analyzed using control charts and other statistical tools.
- The system capability is determined. If stability is not achieved after special cause variation has been eliminated an improvement effort is called for. System changes must be made when the process variation is so large that it does not meet specifications, such as minimum error rates or production deadlines. Typical methods to reduce variation are the development and implementation of a new training program or the enforcement of a standard operating procedure. The latter can be a process standard, a current best methods standard or a simple checklist.
- The final step of the improvement plan is to establish a system for continuous monitoring of the process. We cannot expect processes to remain stable over time. For many reasons they usually start drifting after some time. A monitoring system helps keeping track of new error structures, new customer requirements, and the potential of improved methods and technology and can suggest process improvements.

The Morganstein and Marker book chapter had a distinct effect on quality work and process thinking in many European statistical organizations. Interest in these issues increased and some organizations started their own quality management system where process improvement was central.

At the 1998 Joint Statistical Meetings Mick Couper presented an invited paper on measuring quality in a CASIC environment. He meant that the new technology generated lots of by-product data that could be used to improve the data collection process. He named those paradata, not in his paper but in his session presentation. This naming caught on very quickly in the survey community and it made sense to define the trilogy data, metadata, and paradata. Thus we had one term for data about the data (metadata) and another for data about the process (paradata). Obviously paradata are process data but for a long time paradata were confined to data about the data collection process, while the term used in many European statistical organizations was "process data" and took all survey processes into account (Aitken, Hörngren, Jones, Lewis and Zilhao 2004). Recently a renewed broadening of the meaning of the concept has taken place. Kennickell, Mulrow and Scheuren (2009) remind us about what they call macro paradata, global process data such as response rates, coverage rates, edit failure rates, and coding error rates that always have been indicators of process quality in statistical organizations. Lyberg and Couper (2005), Kreuter, Couper and Lyberg (2010), and Smith (2011) also use the more inclusive meaning of paradata where other processes than data collection are taken into account. There is a risk that paradata, like quality, becomes an overused concept. There are examples of discussions where all data, apart from the survey estimates, are considered paradata, which, of course, does not make sense.

Paradata is a great naming and they are necessary to judge process quality. However, a word of caution is in place. One should never collect paradata that are not related to process quality and it is important to know how to analyze them. Sometimes statistical process control methods can be used but at other times other analytical techniques are needed. For instance, to be able to control interviewer falsification we might need to look at several processes simultaneously, but theory and methodology for such analysis might not be readily available.

The expanded use of microdata that concern individual records, such as keystroke data and flagged imputed records, is an effect of using new technology. Modern data collection procedures generate enormous amounts of these kinds of paradata but so do systems for computer-assisted manual coding and systems for pure automated coding as well as systems for scanning of data. It makes no sense to confine the concept to data collection.

Quality management has taught us to prevent process problems rather than fix them when they appear, that it is important to distinguish between different types of process variation since they require different actions, that any process intervention or improvement should be based on good data and proper analysis methods, and that even stable processes eventually start drifting, which calls for continuous monitoring.

4.5 Standardization and similar tools

One way of keeping process quality in control is to reduce variation by encouraging the use of standards and similar documents. Colledge and March (1997) discuss four classes of documents.

- A standard is a document that should be adhered to almost without exception. Deviations are not recommended and require approval of senior management. Corrective action should be taken when a standard is not fully met. An organization can become certified according to a standard. This is the case for ISO standards, where a few are relevant to statistical organizations.
- A policy should be applied without exceptions. For instance, an organization can have a policy regarding the use of incentives to boost response rates.
- Several organizations have developed guidelines for different aspects of the statistics production. Typically, guidelines can be skipped if there are “good” reasons to do so.
- A recommended practice is promoted but adherence is not mandatory.

Admittedly, the categories of this classification scheme are not mutually exclusive, especially if we also take language and cultural aspects into account. For instance, in the Swedish language policies and guidelines are very close conceptually. If we consult the unauthorized but consensus based Wikipedia it says that “policies describe standards while guidelines outline best practices for following these guidelines”. This sentence contains three of the categories mentioned by Colledge and March. It is probably best to relate to these different kinds of documents in a similar fashion. They all attempt to improve quality by reducing various types of variation and we should not dwell too much on what they are called.

Although standards have been an important part of survey methodology for a long time they have gained momentum since statistical organizations became interested in quality management. Early standards such as Hansen *et al.* (1967) and U.S. Bureau of the Census (1974) concentrated on discussing the presentation of errors in data. At the U.S. Census Bureau all publications should inform users that data were subject to error, that analysis could be affected by those errors, and that estimated sampling errors are smaller than the total errors. For major surveys the nonsampling errors should be treated in more detail unlike in the past. Many other statistical organizations imported this line of thinking. For instance, the quality frameworks mentioned earlier are expansions including also other quality dimensions than accuracy. The European Statistical System has successively developed and launched what was first called Model Quality Reports and currently just Standard for Quality Reports (Eurostat 2009a). The standard provides recommendations to European National Institutes (notice the conceptual complexity) for preparation of quality reports for a “full” range of statistical processes and their outputs. The standard treats the basic quality dimensions relevance, accuracy, timeliness, accessibility, coherence and comparability.

Let us look at some examples. Regarding measurement error, which is part of the accuracy component, the standard says that the following information should be included in a quality report:

- Identification and general assessment of the main risks in terms of measurement error.
- If available, assessments based on comparisons with external data, reinterviews or experiments.
- Information on failure rates during data editing.
- The efforts made in questionnaire design and testing, information on interviewer training and other work on error reduction.
- Questionnaires used should be annexed in some form.

Regarding timeliness the standard says that the following information should be included:

- For annual or more frequent releases: the average production time for each release of data.
- For annual and more frequent releases: the percentage of releases delivered on time, based on scheduled release dates.
- The reasons for nonpunctual releases explained.

There are also sections on how to communicate information regarding trade-offs between quality dimensions, assessment of user needs and perceptions, performance and cost, respondent burden as well as confidentiality, transparency and security. Even though there is a section on user needs and perceptions, users have obviously not been involved in the preparation of the standard itself. We still know very little about how users perceive and use information about quality. The standard is backed by a much more detailed handbook for quality reports (Eurostat 2009b) and both documents are built around the 15 principles listed in the European Statistics Code of Practice, which is the basic quality framework for the European Statistical System. The Code of Practice principles concern professional independence, mandate for data collection, adequacy of resources, quality commitment, statistical confidentiality, impartiality and objectivity, sound methodology, appropriate statistical procedures, nonexcessive burden on respondents, cost-effectiveness, relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and, finally, accessibility and clarity. Each principle is accompanied by a set of indicators that the individual organization can measure to establish whether it meets the Code or not. Some indicators are vague and very subjective in nature such as "the scope, detail and cost of statistics are commensurate with needs", while others are more specific, such as "a standard daily time for the release of statistics is made public". Peer reviews of compliance to a limited set of the principles have been conducted using an earlier version of the Code and, not surprisingly, many national statistical offices in Europe have problems living up to the Code (Eurostat 2011a). Therefore in order to assist the implementation of the Code a supporting framework has been developed, called the Quality Assurance Framework (QAF) that contains more specific guidance regarding methods and references (Eurostat 2011b). This seems to be a very useful document since its references are mainly summaries of the state-of-the-art in areas such as sampling, questionnaire design, editing and so on, which stimulates conformity to current best practices.

The Code of Practice has many similarities with the UN Fundamental Principles of Official Statistics (de Vries 1999). The latter promotes also the principle of international cooperation and coordination, which is, to a large extent, an element that is missing in today's development of statistics production (Kotz 2005). Even neighbouring countries can have very different approaches and methodological competence levels and the differences are sometimes difficult to explain. Experience shows that global development collaboration is difficult to achieve. We meet, we talk, and we bring back ideas that might fit our own systems. It is harder to agree on common approaches. One global standard that relates to statistics production is the ISO 20252 on market, opinion and social research (International Standards Organization 2006). This is a process standard with around 500 requirements concerning the research activities within an organization. It is a minimum standard for what to do rather than how to do things. It is suitable for organizations that conduct surveys and the organization can apply for certification. In April 2010 more than 300 organizations world-wide had been certified, most of them marketing firms. One national statistical office (Uruguay) was certified in 2009 and Statistics Sweden is planning a certification in 2013 but those are the only national offices that have chosen this path. The standard concerns the organization's system for quality management, management of the executive elements of the research, data collection, data management and processing, and reporting on research projects (Blyth 2012).

The standards of the U.S. Federal Statistical System concentrate on the accuracy component. Although not formally a standard the U.S. Federal Committee on Statistical Methodology (2001) suggests various methods for measuring and reporting sources of error in surveys. In 2002 the U.S. Office of Management and Budget (OMB) issued information quality guidelines (OMB 2002) whose purpose was to ensure and maximize the quality, objectivity, utility, and integrity of information disseminated by federal agencies. OMB (2006a) has also issued standards and guidelines for surveys. They are built in a standard fashion. First comes a standard such as "Response rates must be computed using standard formulas to measure the proportion of the eligible sample that is represented by the responding units in each study, as an indicator of potential nonresponse bias". This standard is then followed by a number of guidelines on how to make the necessary calculations while the final guideline states that "If the overall nonresponse rate exceeds 20%, an analysis of the nonresponse bias should be conducted to see whether data are missing completely at random". As in the case of the ESS standards, the OMB guidelines are complemented by a supporting document (OMB 2006b) that can facilitate adherence to the standards.

Most agencies in the decentralized U.S. Federal Statistical System have documents in place that adapt the OMB guidelines. For instance, the U.S. Census Bureau has its own statistical quality standards that goes into more technical detail compared to the OMB documents. Each standard is described via requirements and sub-requirements and they often provide very specific examples of studies that can be conducted. Examples of other U.S. agencies that have standards related to the quality of information disseminated include the National Center for Health Statistics, National Center for Education Statistics, and the

Energy information Administration. All these standards can be downloaded from the agencies' websites.

Statistics Canada has issued quality guidelines since 1985. They are similar to the ESS guidelines since not just accuracy is emphasized. But they are much more detailed and contain lots of references. A special feature is that for some processes the guidelines prescribe the use of statistical process control. No other agency seems to be doing that. The latest edition of the guidelines is provided in Statistics Canada (2009).

Many other statistical organizations in the world have their own quality standards. They are sometimes described as guidelines or standards and sometimes as business support systems or quality assurance frameworks. In any case, the contents and style vary across organizations but the variation should be manageable. It should be possible to achieve higher degrees of standardization globally, since that has happened in other fields, such as air travel. Apted, Carruthers, Lee, Oehm and Yu (2011) discuss various ways to industrialize the statistical production process at the Australian Bureau of Statistics.

The question is whether international standards would benefit survey quality in general. Some areas where standards would be beneficial include computation of frequently used quality indicators such as error rates and design effects, as well as best practices for translation of survey materials, handling non-native language respondents, and weighting for nonresponse. One must bear in mind that once a standard is issued it has to be continually updated and it is well-known that they can be difficult to enforce. If they are comprehensive, standards can overwhelm the practitioner and, as a result, unless mandated and audited, they are largely ignored.

4.6 Statistical business process models

During recent years concepts like business process models and business architecture have become part of quality work in some statistical organizations. To make production processes more efficient and flexible they can be seen as part of a business architecture model (Reedman and Julien 2010). In statistics production a generic statistical process model is jointly developed by UNECE, Eurostat, and OECD. Any system redesign should be driven by customer demands, risk assessments and new developments. The architectural principles behind this thinking are summarized in Doherty (2010), which discusses architecture renewal at Statistics Canada.

Some of the principles are:

- Decision-making should be corporately optimal, which entails centralization of informatics, methodology support and processing.
- Use of corporate services such as collection, data capture and dissemination should be optimized.
- Reuse should be maximized by having the smallest possible number of distinct business processes and the smallest possible number of computer systems.
- The corporate toolkit should be minimized.
- There should be staff proficiency in tools and systems.
- Rework such as repeated editing should be eliminated.
- The focus should be on the core business and the work with support processes should be outsourced.
- Development should be separated from the on-going operations.
- Electronic data collection should be viewed as the initial mode.
- Structural obstacles, such as overlapping or unclear mandates should be removed.

These principles are very similar to those we identify when we apply quality management principles from the various frameworks and excellence models described previously. The principles represent a move from decentralization to more corporate level thinking. Many statistical organizations realize that stove-pipe thinking is a thing of the past and that a move to more centralization is necessary.

5. Measuring quality

Thus, quality is a multi-faceted concept and measuring it is a complicated task. We have noted that survey quality can be viewed as a three-dimensional concept associated with the final product, the underlying processes that lead to the product, and the organization that provides the means to carry out the processes and deliver the product or service in a successful way. There are basically two ways to measure quality. One is to directly estimate the total survey error or some components thereof. The other is to measure indicators of quality with the hope that they indeed reflect the concept itself.

5.1 Direct estimates of the total survey error

The existing decompositions of the mean squared error described in, for instance, Hansen *et al.* (1964), Fellegi (1964), Anderson, Kasper and Frankel (1979), Biemer and Lyberg (2003), Weisberg (2005), and Groves *et al.* (2009) are all incomplete in the sense that they do not reflect all error sources. It is seldom possible to compute the MSE directly in practical survey situations because this usually requires a parameter estimate that is essentially error free. However, it is possible to obtain a second best estimate of the true parameter value if there are resources available to collect data using some "gold standard" methodology that is not affordable or practical in a normal survey setting. This is the standard evaluation methodology when the true parameter value can be uniquely defined. Gold standard methods are seldom error-free but they can to varying extents provide better estimates, and the difference between the regular estimate and the gold standard estimate can serve as an estimate of the bias, which is the methodology used in census post enumeration surveys (United Nations 2010). Often an evaluation concerns a specific error component such as census undercount, nonresponse bias, interviewer variance or simple response variance, since we want information not on total survey error per se but rather on the components' relative contribution to the total survey error so that root causes of problems can be identified and relevant processes improved. Large evaluation studies are very rare since they are so demanding and their value is sometimes questioned (United Nations 2010). Smaller regular evaluation studies, on the other hand, are necessary to get indications of process and methodological problems.

5.2 Indicators of quality

Continuing reporting of total survey error is a formidable task and no survey organization does that. Instead organizations provide indicators or statements regarding quality. For instance, according to Eurostat's (2009a) handbook for quality reports the following indicators should be measured:

- Coefficient of variation;
- Overcoverage rate;
- Edit failure rate;
- Unit response rate;
- Item response rates;
- Imputation rates;
- Number of mistakes;
- Average size of revisions.

The common theme here is that these paradata summary items are indicators that can be calculated without conducting special studies. The set of indicators that can be calculated directly from the survey data is by definition quite limited and their value questionable. For instance, to include overcoverage but not undercoverage just because only the former can be calculated directly from the available data does not make sense. It is undercoverage that poses the greatest coverage problem in surveys. Admittedly, the handbook prescribes the producer to assess the potential for bias (both sign and magnitude) but it is not clear how this should be accomplished. The producer is urged to include evaluation and quality control results, if such information exists as well. Level of effort measures for processes such as questionnaire design and coder training would be welcomed. There is no standard reporting format for such qualitative and quantitative information. In any case, the key indicator list becomes severely limited when compared to the full list of main error sources and it is hard to see how they are perceived by the users and how they can be used by the producer to improve the process.

The producer needs a more complete list of indicators to be able to measure or assess various levels of quality to make sure that the design implementation is in control or to be able to mount a quality improvement project. The initial survey design must be modified or adapted during the implementation to control costs and maximize quality. Biemer (2010) discusses four strategies for reducing costs and errors in real time, *i.e.*, continuous quality improvement (CQI), responsive design (Groves and Heeringa 2006), Six Sigma (Breyfogle 2003), and adaptive total design and implementation.

When the continuous quality improvement strategy is used, key process variables are identified and so are process characteristics that are critical to quality (CTQ). For each CTQ, real-time, reliable metrics for the cost and quality are developed. The metrics are continuously monitored during the process and intervention is done to ensure that costs and quality are within acceptable limits. The responsive design strategy was developed to reduce nonresponse bias in face to face interviewing. It includes three phases. In the experimental phase a few design options are tested (*e.g.*, regarding incentive level). In the main data collection phase the option chosen in the experimental phase is implemented and the implementation continues until phase capacity is reached. In the nonresponse follow-up phase special methods are implemented to reduce nonresponse bias and control the data collection costs. Such methods include the Hansen-Hurwitz double sampling scheme, increased incentives, and using more experienced interviewers. Again the efforts continue until further reductions of the nonresponse bias are no longer cost-effective. Six Sigma is the most developed business excellence model since it relies so heavily on statistical methods. It contains a large set of techniques and tools that can be used to control and improve processes. Adaptive total design and implementation combines control features of CQI, responsive design and Six Sigma and does that so that it simultaneously monitors multiple error sources. Biemer and Lyberg (2012) give several examples of CTQs and metrics for various survey processes. For instance, regarding the measurement process attributes that are CTQs might include the abilities to identify and repair problematic survey questions, to detect and control response errors, and to minimize interviewer biases and variances. Corresponding metrics might include missing data item by question, refusal rate by size of business, results of replicate measurements, suspicious edits actually changed, and field work results by interviewer. The metrics can be analyzed using statistical process control or analysis-of-variance methodologies. Different related metrics can be displayed together in a dashboard fashion. For instance if one CTQ is the ability to discover interviewer cheating we might want to have a dashboard showing the metrics average interview length by interviewer and the distribution of some sensitive sample characteristic, also by interviewer.

5.3 Self-assessments and audits

The quality management philosophy has introduced the concepts of self-assessment and audit into statistics production. We are anxious to know what users, clients, owners and other stakeholders think about the products and services provided by the statistical organization. There are a number of tools available for this kind of evaluation. We have already mentioned the customer satisfaction survey. Other tools include employee surveys, internal audits and external audits. Customer surveys can shed light on what users think about products and services provided. They can be used to determine user needs and to identify what product characteristics really matter to the users. Another line of questioning might concern the image of the organization and how it compares to the images of other organizations, be they competitors or not. The customer satisfaction survey is very common in society. Often it cannot be used to make inference to the target population of users due to its methodological and conceptual shortcomings. The abundance of satisfaction surveys in society, developed and implemented by people with no formal training in survey methods, contributes to lukewarm receptions in more serious settings resulting in nonresponse and measurement errors. For instance, the 2007 Eurostat User Satisfaction Survey consisted of two separate surveys. One was launched on the Eurostat webpage and the target population consisted of 3,800 registered users. Only those registered users that entered the website during the data collection period were exposed to the survey request and this led to a response rate around 5%. The second survey used email that was sent to a number of main users identified by Eurostat. This more controlled environment generated a response rate of 28%. These surveys also have problems identifying the most suitable respondent. If the "wrong" respondent is chosen within an organization this will most certainly lead to uninformed and misleading results.

The simplest type of self-assessment is the questionnaire or checklist that is filled out by the survey manager. An example is one from Statistics New Zealand. It is a checklist that consists of a number of indicators or assertions such as "information needs are regularly assessed through user consultation", "good and accessible documentation", "indicators of accuracy regularly produced and monitored", and "presentation standards met". The manager is asked to answer yes or no to each assertion and make a comment if deemed necessary. Statistics Sweden had a similar system in place where one of the questions was "has overall quality of your product improved, declined or stayed the same compared to last year?" When results were compiled for these three categories for the entire organization, a very small proportion of the managers reported declining quality, a somewhat larger proportion reported improved quality, while a vast proportion reported status quo. The managers simply did not have the proper means to assess overall quality. Furthermore, vague quantifiers like "regularly", "good", and "meeting standards" invite generous assessments. Also most managers do not want to look bad and status quo becomes a perfect escape route. This system of self-assessment was eventually abandoned by Statistics Sweden. It is possible to increase the value of these assessments by asking additional questions concerning details about how and when quality work was conducted. Some organizations use internal teams that audit important products. Julien and Royce (2007) describe a quality audit of nine products at Statistics Canada, where the purposes were to identify weaknesses and their root causes as well as identifying best practices. Review teams of assistant managers were formed so that each reviewer reviewed three different programs. The main weakness with an approach like this is the internal feature itself. Every reviewer knows that sooner or later it is his or her turn to be reviewed and there is a risk that this fact might hold them back. It is also internal in the sense that users are not explicitly present in the review process. In its general audit program on data quality management, however, Statistics Canada puts great emphasis on its user liaison system (Julien and Born 2006), which is one of the five systems forming the agency's quality assurance framework, the others being corporate planning, methods and standards, dissemination, and program reporting.

A further variant of self-assessment is when it precedes an external audit. Statistics Netherlands (1997) describes how the Department of Statistical Methods is assessed by its staff. The assessment resulted in a listing of weak and strong areas that were later examined by an external team. Typically an external audit uses some kind of benchmark like a set of rules, a standard, or a code of practice for assessment purposes. The audit then results in a number of recommendations for the organization or the individual product or service.

Recently a general system for evaluating the total survey error has been developed at Statistics Sweden. Sweden's Ministry of Finance wants quality evaluation results to be able to monitor quality improvements over time. Survey quality must be assessed for many surveys, administrative registers, and other programs within the agency so there is need for some indicators that can serve as proxies for actual measures of quality. At the same time, the assessment process must be thorough, the reporting simple and the results credible. For each of the error sources specification, frame, nonresponse, measurement, data processing, sampling, model/estimation, and revision eight key products were rated poor, fair, good, very good, and excellent regarding each of five criteria. The criteria were knowledge of risks, communication with users, compliance with standards and best practices, available expertise, and achievement toward risk mitigation and/or improvement plans. The rating guidelines varied by criterion. For knowledge of risks they were:

An Example of the rating guidelines – Knowledge of risks

Poor ●	**Fair** ◓	**Good** ○	**Very Good** ◒	**Excellent** ◉
Internal program documentation does not acknowledge the source of error as a potential factor for product accuracy.	Internal program documentation acknowledges error source as a potential factor in data quality.	Some work has been done to assess the potential impact of the error source on data quality.	Studies have estimated relevant bias and variance components associated with the error source and are well-documented.	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.
	But: No or very little work has been done to assess these risks	But: Evaluations have only considered proxy measures (for example, error rates) of the impact with no evaluations of MSE components	But: Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses	

The evaluation process started with a self-assessment done by each of the eight key products. These reports and other relevant documents were studied by two external reviewers who then met with product owners and their staff to discuss the product processes. After that the reviewers presented detailed assessments and scored each product. The procedure identified important areas to improve within but also across products. In this first evaluation round measurement error turned out to be a problematic area for almost all the key products. As any other approach at measuring or indicating total survey error this one does not really reflect total mean squared error. It requires thorough documentation of processes and improvements made and it is highly dependent on the skills and knowledge of the external reviewers. This study is reported in Biemer, Trewin, Japec, Bergdahl and Pettersson (2012).

5.4 Quality profiles

In continuing surveys there is an opportunity to develop quality profiles. Such documents contain all that is known about the quality of a continuing survey or other statistical product assembled over a number of years. Quality profiles exist for only a few major surveys, all, except one, conducted in the U.S., including the Current Population Survey (Brooks and Bailar 1978), the Survey of Income and Program Participation (Jabine, King and Petroni 1990; Kalton, Winglee and Jabine 1998), the Schools and Staffing Survey (Kalton, Winglee, Krawchuk and Levine 2000), and the American Housing Survey (Chakrabarty and Torres 1996). The exception is the British Household Panel Survey (Lynn 2003). The main problem with a quality profile is that it is not timely, since it compiles results from often time-consuming studies of quality. The goal of the quality profile is to identify areas where knowledge about errors is deficient so that improvements can be made. Kasprzyk and Kalton (2001) and Doyle and Clark (2001) review the use of quality profiles in the U.S.

6. Where do we go from here?

Quality management ideas have been influential in many survey organizations. Concepts such as leadership, quality culture, problem prevention, customer, competition, risk assessment, process thinking, improvement, business excellence, and business architecture are increasingly discussed by leaders of survey organizations, *e.g.*, Trewin (2001), Pink (2010), Fellegi (1996), Brackstone (1999), de Vries (1999), Groves (2011), and Bohata (2011). It seems as if the survey community is moving in a direction where statistics production becomes more streamlined and cost-effective but the pace is slow. Some organizations have started using a quality management model for self-assessment and steering purposes. EFQM is the recommended model for national statistical institutes within the European Statistical System and a couple of institutes, the Czech Republic and Finland, have even applied for their respective national EFQM awards. Some marketing firms are certified according to the ISO 9001 quality management standard and others are certified according to the ISO 20252 standard for market, opinion, and social research. This development ought to result in quality improvements but we cannot be really sure until we start collecting relevant data. One thing is sure, though. Some customers prefer service providers that are certified, have won awards or can show evidence that they are working according to some quality framework or model. Very few customers would think that this is a negative thing.

The margins of error that we associate with estimates are usually too short, since they do not include all sources of variation. Point estimates can be off due to biases. Ideally it would be good if we were able to produce estimates of the total survey error instead of what we produce today. Such a development is, however, not realistic. We are not in a position to produce such estimates, not even occasionally, for reasons that have to do with finances, timing and methodology. That leaves us with indicators of total survey error and its components. Such indicators are of limited value to the users. Users simply do not know what to do with information on nonresponse rates, response variance measured by reinterviews or edit failure rates. On the other hand, such indicators are very useful to the producers of surveys. For instance, reinterview studies can identify fabrication and survey questions with poor response consistency. A majority of users appreciate the service provider's credibility and part of the credibility is the ability to present accurate data. Another important part of credibility is the willingness of the providers to evaluate their own quality and to report the results of such evaluations. Even if these evaluations show problems, it is better for the provider to find the problems than if entities outside the provider's organization find them. Most users do not want to become involved in discussions about errors and trade-offs between errors and for good reasons. It is simply too technical and confusing. If we accept that a good process quality is a prerequisite for a good product quality, we should gradually improve the processes so that they approach ideal bias-free ones. In that way the variance of an estimate becomes a good approximation of the mean squared error.

Despite endless discussions and a myriad of survey quality initiatives, practices have not changed much (Lynn 2004; Pink, Borowik and Lee 2010; Groves 2011; Bohata 2011). Perhaps the lack of competence within survey organizations is one root cause of the slow pace. Many theories and methodologies including statistics, IT, management, communication, and behavioral sciences are needed in survey research. The behavioral sciences are needed to identify the root causes of nonsampling errors. If errors are just quantified no improvement can happen. Current training programs emphasize sampling, non-response, coverage and estimation in the presence of these. Other processes and error sources such as measurement and data processing are not dealt with to the same extent. This leads to a situation where studies on measurement error and data processing error are rare compared to studies on, say, nonresponse. There is a considerable confusion regarding concepts and methods in both the producer and the user camps. Another cause of slow pace might be the consensus philosophy that rules in some organizations when it comes to decision-making regarding changes. This philosophy is one of compromise. Input from many stakeholders is gathered and a decision is usually based on the smallest common denominator, which is never a good standard. Furthermore, arriving at this compromise usually takes a long time and lots of resources. This approach is very far from Plan-Do-Check-Act.

Survey quality is not an absolute entity. Current quality reporting a la one-size-fits-all is not working since fitness for use is defined by each user. Quality dimensions such as timeliness, comparability and accessibility should be decided together with main users while best possible accuracy given various constraints is the responsibility of the service provider.

Have the survey quality discussion and the adoption of quality management strategies resulted in better data? We do not know. Survey quality has not been assessed in a before-after fashion. There is a tendency towards greater standardization and centralization, which should prove cost-efficient but when it comes to data quality some indicators point in the wrong direction. For instance, in many countries nonresponse rates are increasing and error properties of mixed-mode, translation of survey materials, and other design features are not fully known or are different across cultures. There is no design formula, which results in shaky trade-off decisions and problems deciding about intensities with which quality control should be applied. There is a persistent quest for best practices in survey organizations but implementation is difficult and scattered. There is definitely a great need for an upgrade in the competence level across the board. A structured international competence development program for service providers is necessary as is a systematic international collaboration on how to best design and implement surveys. We must serve our users better by providing data with small errors. We can do this by better combining our knowledge about statistics and cognitive phenomena with the principles of quality management. The great positive note is the overwhelming positive attitude toward quality improvement among statistical organizations around the world.

References

Aitken, A., Hörngren, J., Jones, N., Lewis, D. and Zilhao, M. (2004). *Handbook on improving quality by analysis of process variables*. Office for National Statistics, UK.

Anderson, R., Kasper, J. and Frankel, F. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco: Jossey-Bass.

Apted, L., Carruthers, P., Lee, G., Oehm, D. and Yu, F. (2011). Industrialisation of statistical processes, methods and technologies. Paper presented at the International Statistical Institute Meeting, Dublin.

Bailar, B., and Dalenius, T. (1969). Estimating the response variance components of the U.S. Bureau of the Census' Survey Model. *Sankhyā*, B, 341-360.

Biemer, P. (2001). Comment on Platek and Särndal. *Journal of Official Statistics*, 17(1), 25-32.

Biemer, P. (2010). Overview of design issues: Total survey error. In *Handbook of Survey Research*, (Eds., P. Mardsen and J. Wright), Second Edition. Emerald Group Publishing Limited.

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

Biemer, P., and Lyberg, L. (2012). Short course on Total Survey Error. The Joint Program in Survey Methodology (JPSM), April 16-17, Washington, DC.

Biemer, P., Trewin, D., Japec, L., Bergdahl, H. and Pettersson, Å. (2012). A tool for managing product quality. Paper presented at the Q Conference, Athens.

Blyth, B. (2012). ISO 20252; Turning frameworks into best practice. Paper presented at the Q Conference, Athens.

Bohata, M. (2011). Fit-for-purpose statistics for evidence based policy making. Memo, Eurostat.

Bowley, A.L. (1913). Working-class households in reading. *Journal of the Royal Statistical Society*, 76(7), 672-701.

Box, G. (1990). Good quality costs less? How come? *Quality Engineering*, 3, 1, 85-90.

Box, G., and Friends (2006). *Improving Almost Anything: Ideas and Essays*. New-York: John Wiley & Sons, Inc.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25, 2, 139-149.

Brackstone, G. (2001). How important is accuracy? *Proceedings: Symposium 2001*, *Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.

Breyfogle, F. (2003). *Implementing Six Sigma*. Second Edition. New-York: John Wiley & Sons, Inc.

Brooks, C., and Bailar, B. (1978). An error profile: Employment as measured by the Current Population Survey. Working paper 3, Office of Management and Budget, Washington, DC.

Chakrabarty, R., and Torres, G. (1996). American Housing Survey: A Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.

Colledge, M., and March, M. (1993). Quality management: Development of a framework for a statistical agency. *Journal of Business and Economic Statistics*, 11, 157-165.

Colledge, M., and March, M. (1997). Quality policies, standards, guidelines, and recommended practices. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New-York: John Wiley & Sons, Inc.

Couper, M. (1998). Measuring Survey Quality in a CASIC Environment. Paper presented at the Joint Statistical Meetings, American Statistical Association, Dallas, TX.

Dalenius, T. (1967). Nonsampling Errors in Census and Sample Surveys. Report No. 5 in the research project Errors in Surveys. Stockholm University.

Dalenius, T.E. (1968). Official statistics and their uses. *Review of the International Statistical Institute*, 26(2), 121-140.

Dalenius, T. (1969). Designing descriptive sample surveys. In *New Developments in Survey Sampling*, (Eds., N.L. Johnson and H. Smith), New-York: John Wiley & Sons, Inc.

Dalenius, T. (1985a). *Elements of Survey Sampling*. Swedish Agency for Research Cooperation with Developing Countries. Stockholm, Sweden.

Dalenius, T. (1985b). Relevant official statistics. *Journal of Official Statistics*, 1(1), 21-33.

Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359-369.

Deming, E. (1950). *Some Theory of Sampling*. New-York: John Wiley & Sons, Inc.

Deming, E. (1986). *Out of the Crisis*. MIT.

Deming, W.E., and Geoffrey, L. (1941). On sample inspection in the processing of census returns. *Journal of the American Statistical Association*, 36, 215, 351-360.

De Vries, W. (1999). Are we measuring up…? Questions on the performance of national systems. *International Statistical Review*, 67, 1, 63-77.

Dillman, D. (1996). Why innovation is difficult in government surveys (with discussions). *Journal of Official Statistics*, 12, 2, 113-198.

Doherty, K. (2010). How business architecture renewal is changing IT at Statistics Canada. Paper presented at the Meeting on the Management of Statistical Information Systems. Daejeon, South Korea, April 26-29.

Doyle, P., and Clark, C. (2001). Quality profiles and data users. Paper presented at the International Conference on Quality in Official Statistics (Q), Stockholm.

Drucker, P. (1985). *Management*. Harper Colophone.

Ecochard, P., Hahn, M. and Junker, C. (2008). User satisfaction surveys in Eurostat and in the European Statistical System. Paper presented at the Q conference, Rome, Italy.

Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.

Eltinge, J. (2011). Aggregate and systemic components of risk in total survey error models. Paper presented at ITSEW 2011, Quebec, Canada.

Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, Series B, 195-233.

European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.

Eurostat (2009a). ESS Standard for Quality Reports. Eurostat.

Eurostat (2009b). ESS handbook for Quality Reports. Eurostat.

Eurostat (2011a). European statistics Code of Practice. Eurostat.

Eurostat (2011b). Quality assurance framework (QAF). Eurostat.

Fellegi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

Fellegi, I. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2, 165-197.

Felme, S., Lyberg, L. and Olsson, L. (1976). *Kvalitetsskydd av data*. (Protecting Data Quality.) Liber (in Swedish).

Fienberg, S., and Tanur, J. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review*, 64, 237-253.

Fisher, R. (1935). *The Design of Experiments*. New York: Hafner.

Frankel, M., and King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 1, 65-87.

Gleaton, E. (2011). Centralizing LAN services. Memo, National Agricultural Statistics Service, U.S. Department of Agriculture.

Groves, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

Groves, R. (2011). The structure and activities of the U.S. Federal Statistical System: History and recurrent challenges. *The Annals of the American Academy of Political and Social Science*, 631, 163, Sage.

Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls, W. and Waksberg, J. (Eds.) (1988). *Telephone Survey Methodology*. New-York: John Wiley & Sons, Inc.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*, Second Edition. New-York: John Wiley & Sons, Inc.

Groves, R., and Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society*, A, 169, 439-457.

Groves, R., and Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 5, 849-879.

Hansen, M., and Hurwitz, W. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 517-529.

Hansen, M., Hurwitz, W. and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32nd Session, 38, Part 2, 359-374.

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

Hansen, M., Hurwitz, W., Marks, E. and Mauldin, P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46,147-190.

Hansen, M., Hurwitz, W. and Pritzker, L. (1964). The estimation and interpretation of gross differences and simple response variance. In *Contributions to Statistics*, (Ed., C. Rao). Oxford: Pergamon Press, 111-136.

Hansen, M., Hurwitz, W. and Pritzker, L. (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.

Hansen, M., and Steinberg, J. (1956). Control of errors in surveys. *Biometrics*, 462-474.

Hansen, M., and Voigt, R. (1967). Program guidance through the evaluation of uses of official Statistics in the United States Bureau of the Census. Paper presented at the International Statistical institute meeting, Canberra, Australia.

Holt, T., and Jones, T. (1998). Quality work and conflicting policy objectives. *Proceedings of the 84th DGINS Conference*, May 28-29, Stockholm, Sweden. Eurostat.

International Standards Organization (2006). Market, Opinion and Social Research. ISO Standard No. 20252.

Jabine, T., King, K. and Petroni, R. (1990). Survey of Income and Program Participation (SIPP): Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.

Joiner, B. (1994). *Generation Management*. McGraw-Hill.

Julien, C., and Born, A. (2006). Quality management assessment at Statistics Canada. *Proceedings of the Q Conference,* Cardiff, UK.

Julien, C., and Royce, D. (2007). Quality review of key indicators at Statistics Canada. *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, 1113-1120.

Juran, J.M. (1988). *Juran on Planning for Quality*. New York: Free Press.

Juran, J.M. (1995). *A History of Managing for Quality*. ASQC Quality Press.

Juran, J., and Gryna, F. (Eds.) (1988). *Juran's Quality Control Handbook*, 4th Edition. McGraw-Hill.

Kalton, G. (2001). How important is accuracy? *Proceedings: Symposium 2001*, *Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.

Kalton, G., Winglee, M. and Jabine, T. (1998). *SIPP Quality Profile*. U.S. Bureau of the Census, 3rd Edition.

Kalton, G., Winglee, M., Krawchuk, S. and Levine, D. (2000). *Quality Profile for SASS Rounds 1-3: 1987-1995*. Washington, DC: U.S. Department of Education.

Kasprzyk, D., and Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14-15 May 2001, CD-ROM.

Kennickell, A., Mulrow, E. and Scheuren, F. (2009). Paradata or process modeling for inference. Paper presented at the Conference on Modernization of Statistics Production, Stockholm, Sweden.

Kiear, A.N. (1897). The representative method of statistical surveys. *Kristiania Videnskaps-selskabets Skrifter: Historik-filosofiske Klasse*, (in Norwegian), 4, 37-56.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1995). *The Hundred Years' Wars of Survey Sampling*. Centennial Representative Sampling, Rome.

Kotz, S. (2005). Reflections on early history of official statistics and a modest proposal for global coordination. *Journal of Official Statistics*, 21, 2, 139-144.

Kreuter, F., Couper, M. and Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Lyberg, L. (1981). *Control of the Coding Operation in Statistical Investigations: Some Contributions*. Ph.D. dissertation, Stockholm University.

Lyberg, L. (2002). Training of survey statisticians in government agencies-A review. Invited paper presented at the Joint Statistical Meetings, American Statistical Association, New-York.

Lyberg, L., Bergdahl, M., Blanc, M., Booleman, M., Grünewald, W., Haworth, M., Japec, L., Jones, L., Körner, T., Linden, H., Lundholm, G., Madaleno, M., Radermacher, W., Signore, M., Zilhao, M.J., Tzougas, I. and van Brakel, R. (2001). Summary report from the Leadership Group (LEG) on Quality. Eurostat.

Lyberg, L., and Couper, M. (2005). The use of paradata in survey research. Invited paper, International Statistical Institute, Sydney, Australia.

Lynn, P. (Ed.) (2003). *Quality Profile: British Household Panel Survey: Waves 1 to 10: 1991-2000*. Colchester: Institute for Social and Economic Research.

Lynn, P. (2004). Editorial: Measuring and communicating survey quality. *Journal of the Royal Statistical Society*, Series A, 167.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mirotchie, M. (1993). Data quality: A quest for standard indicators. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 729-734.

Moeller, R. (2005). *Brink's Modern Internal Auditing*. Sixth Edition. New-York: John Wiley & Sons, Inc.

Morganstein, D., and Marker, D. (1997). Continuous quality improvement in statistical agencies. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 475-500.

Mudryk, W., Burgess, M.J. and Xiao, P. (1996). Quality control of CATI operations in Statistics Canada, Memo, Statistics Canada.

Muscio, B. (1917). The influence of the form of a question. *The British Journal of Psychology*, 8, 351-389.

Neter, J., and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 305, 18-55.

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*. U.S. Department of Agriculture, Washington, DC.

OECD (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. OECD.

O'Muircheartaigh, C. (1997). Measurement errors in surveys: A historical perspective. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 1-25.

Phipps, P., and Fricker, S. (2011). Quality measures. Memo, Office of Survey Methods Research, U.S. Bureau of Labor Statistics.

Pink, B., Borowik, J. and Lee, G. (2010). The case for an international statistical innovation program-Transforming national and international statistics systems. Paper presented at the Collaboration Leaders Workshop, April 19-23, Sydney, Australia.

Platek, R., and Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17, 1, 1-20 and Discussion, 21-27.

Reedman, L., and Julien, C. (2010). Current and future applications of the generic statistical business process model at Statistics Canada. Paper presented at the Q Conference, Helsinki.

Rosén, B., and Elvers, E. (1999). Quality concept for official statistics. *Encyclopedia of Statistical Sciences*, New-York: John Wiley & Sons, Inc., update Volume 3, 621-629.

Scheuren, F. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.

Schilling, E., and Neubauer, D. (2009). *Acceptance Sampling in Quality Control*, 2nd Ed. Chapman and Hall/CRC.

Scholtes, P., Joiner, B. and Streibel, B. (1996). *The Team Handbook*. Joiner Associates Inc.

Shewhart, W.A. (1939). *Statistical Methods from the Viewpoint of Quality Control*. U.S. Department of Agriculture, Washington, DC, U.S.A.

Smith, T. (2011). Report on the International Workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. NORC/University of Chicago.

Spencer, B. (1985). Optimal data quality. *Journal of the American Statistical Association*, 80, 564-573.

Statistics Canada (2002). Statistics Canada's Quality Assurance Framework, Catalogue No.12-586-XIE, Ottawa.

Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth Edition, Ottawa.

Statistics Netherlands (1997). A self assessment of the Department of Statistical Methods. Research paper No. 9747, Statistics Netherlands.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.

Trewin, D. (2001). The importance of a quality culture. *Proceedings: Symposium 2001*, *Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.

United Nations (2010). *Post Enumeration Surveys: Operational Guidelines*. Department of Economic and Social Affairs, Statistics Division.

U.S. Bureau of the Census (1974). *Standards for Discussion and Presentation of Errors in Data*. U.S. Department of Commerce, Bureau of the Census.

U.S. Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Errors in Surveys*, Statistical Policy Working Paper 31, Washington, DC: U.S. Office of Management and Budget.

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.

U.S. Office of Management and Budget (2006a). *Standards and Guidelines for Statistical Surveys*. U.S. Office for Management and Budget.

U.S. Office of Management and Budget (2006b). Questions and answers when designing surveys for information collection. U.S. Office for management and Budget.

Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-137.

Weisberg, H. (2005). *The Total Survey Error Approach*. The University of Chicago Press.

Weisman, E., Balyozov, Z. and Venter, L. (2010). IMF's data quality assessment framework. Paper presented at the Conference on Data Quality for International Organizations, Helsinkli, May 6-7.

West, B., and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 5, 1004-1026.

Willimack, D., Nichols, E. and Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. In *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge and R. Little), 213-228.

Zarkovich, S. (1966). *Quality of Statistical Data*. Food and Agricultural Organization of the United Nations: Rome, Italy.