SURVEY QUALITY ISSUES DURING THE LAST 50 YEARS
Some Observations

Lars Lyberg and Eva Elvers
Statistics Sweden

KEY WORDS: Quality Components, Total Error, MSE Components, Production Process

1. Introduction

Survey quality is a vague concept with many meanings. In this paper we discuss some observations related to the development and treatment of the concept over the last 50 years. Most of our discussion concerns issues dealt with by government organizations and their work with official statistics. This session is devoted to the late Swedish statistics professor Tore Dalenius who was so influential regarding the development of survey methodology in Sweden and abroad during four decades. Our paper is organized in the following way: In Section 2 we give an account of the status of survey methodology around 1950. In Section 3 we discuss various definitions of quality. Section 4 provides some observations regarding the quality development during the period 1950-1980, while Section 5 concerns observations regarding the period 1980 until today. The reason for the 1980 break is that the development after that point is characterized by the influence of a somewhat different set of thoughts than the earlier development. In Section 6 we list specific contributions by Tore Dalenius. In Section 7 we speculate whether there is need for some kind of paradigm shift on how we should view problems related to survey quality. Section 8 contains references. It should be mentioned at the outset that our discussion is by necessity very brief. Thus the reader may rightfully feel that important events and references are missing.

2. The Status of Survey Methodology Around 1950

In the early development of survey methodology 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 the development of sampling theory. During the 1920s the International Statistical Institute agreed to promote ideas on representative sampling suggested by Kiear and Bowley. In 1934 Neyman published his landmark paper on the representative method. Fisher's randomization principle was used in agricultural sampling and Neyman developed cluster sampling, ratio estimation and two-phase sampling and introduced the concept of confidence interval. It was Neyman who showed that a measure of the sampling error could actually be obtained by calculating the variance of the estimator. Cochran, Yates, Deming, Hansen and others further refined the concepts of sampling theory. The latter 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 text-book on sampling theory and methods (Hansen et al 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 early writings on the effects of question wording and research on questionnaire design was quite extensive in the 1940s. Mahalanobis (1946) addressed problems with errors introduced by fieldworkers collecting agricultural data in India, with the method for estimating such errors as a result. The method is called interpenetration and can be used to estimate correlated variances introduced by interviewers, editors, coders, and those who supervise these groups. The most prominent error sources were known around 1950. Deming had listed error sources (1944) and Hansen and Hurwitz (1946) had discussed subsampling among nonrespondents in an attempt to provide unbiased estimates in a situation with initial nonresponse.

At this time quality does not occur as a concept associated with surveys. The use of the word quality is confined to quality control, sometimes as quality control of survey operations. It was clear that statistics were plagued by errors, but today's "quality lingo" was unknown. The user was a somewhat obscure player in those days albeit not ignored by survey methodologists. For instance, Deming (1950) listed 13 factors affecting the usefulness of surveys and claimed that until the purpose is stated, there is no right or wrong way of going about the survey.

3. Definitions of Survey Quality

Both survey and quality are vague concepts. A *survey* is a statistical study designed to measure population characteristics so that population parameters can be estimated and can be defined as a list of prerequisites (Dalenius, 1985). As pointed out by Morganstein and Marker (1997) varying definitions of quality undermine improvement work and we should try to distinguish between different definitions to see what purposes they might serve. One of the most cited definitions is attributed to Juran, namely quality being a direct function of "fitness for use." It turns out that Deming already in 1944 used the phrase "fitness for the purpose," not to define quality but rather to explain what made a survey product work.

For a long time good quality was implicitly equivalent with accurate statistics. Accuracy can be measured by mean squared error (MSE), which is composed of the variance and the squared bias. We have noticed that survey statistics should also be useful, later often denoted "relevant." Many of today's other quality dimensions were not really expected by early users. The users were then accustomed to the fact that surveys took time to carry out and technology did not allow sophisticated forms of accessibility.

During the last decades it has become obvious that accuracy and relevance are necessary but not sufficient when assessing the quality of a survey. Other dimensions are also important to the users. The development of quality frameworks has taken place mainly within official statistics and has been triggered by the rapid technology development and other developments in society. The advanced technology has created opportunities regarding quality dimensions such as accessibility, timeliness, and coherence that simply did not exist before, and these opportunities have affected user demands. Also, decision-making in society has become more complex and global. This has lead to 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.

The framework developed by Eurostat (2000) consists of seven dimensions: relevance (statistics are relevant if users' needs are met), accuracy (closeness between the value obtained and the true, but unknown, population value), timeliness and punctuality (time length between the release time and the end of the reference period; the agreement between the actual release time and the target release time), accessibility and clarity (accessibility refers the physical conditions in which users can obtain data and clarity refers to the information environment), comparability (over time, between geographical areas, and between domains), coherence (the adequacy of statistics to be combined for different uses, especially when they come from different sources), and completeness (the extent to which there are statistics available compared to what should be available). Similar frameworks have been developed by Statistics Canada (Brackstone 1999 and Statistics Canada 2002), and Statistics Sweden (Statistics Sweden 2001). 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). It certainly appreciates other dimensions, but perhaps it sees them more as constraints. The International Monetary Fund (IMF) is developing a somewhat different framework consisting of a set of prerequisites and five dimensions of quality: integrity, methodological soundness, accuracy and reliability, serviceability, and accessibility.

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.

Successful organizations know that continuous improvement is necessary to stay in business and they have developed measures that help them change. This is true also for producers of statistics. 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 EFQM and the Malcolm Balridge Award. The core values of the EFQM model include results orientation, customer focus, leadership and constancy of purpose, management by processes and facts, personnel development and involvement, continuous learning, innovation, and improvement, development of partnerships, and public responsibility.

There are also other quality definitions reported in the literature. Juran and Gryna (1980) distinguish between design quality and quality conformance, concepts that could be used in surveys. An example of design quality would then be the way data are presented. A multicoloured booklet with graphics would be superior to a set of simple tables. Quality conformance is the degree to which the product conforms to its intended

use. One might also say that quality conformance is fitness for use.

As pointed out by Brackstone (1999) and Scheuren (2001), quality has become a buzzword in society. Any definition, sweeping or more distinct, can be challenged. Even the term error is vague. Are we talking about variances and biases as a result of design choices, or are we talking about mistakes in the implementation of the designs? What could we try to agree on?

We believe that there are three kinds or levels of quality in surveys. First, there is product quality, which can be measured along a number of generally accepted dimensions or some other set of characteristics. Second, there is process quality, which tells us something about the processes that lead to the product. How stable is the process? What kind of variation characterizes the process? Third, there is organizational quality, which can be measured against a number of values or criteria. How can the organization make sure that its survey processes are properly managed? It is quite obvious that the levels are closely connected and that good product quality cannot be achieved unless the other two quality levels are up to par.

We might also agree that any measure of quality should be linked to users. As we have implied above, the word quality as in data quality does not appear until rather late in the literature. Zarkovich's book "Quality of statistical data" (1966) is one of the very first to use the term "data quality."

4. The Period 1950-1980

The period 1950-1980 is characterized by the attempts made at disentangling the MSE and other attempts at modelling survey errors. The most famous achievement is the development of the U.S. Census Bureau's survey model (Hansen et al 1964), where the mean squared error (MSE) of the estimator y is decomposed into sampling variance + response variance + covariance + squared bias. The MSE is a useful tool for determining how to allocate resources in a survey so that they will do most good. Using the MSE as a guide, we can begin to address the cost-error tradeoffs.

A number of papers have been written with the purpose of providing guidance on how to estimate MSE components. Examples of such attempts include Fellegi (1964) and Bailar and Dalenius (1969). Perhaps the most important issue related to the Census Bureau survey model is the visualization of the correlated response variance as one specific MSE component, that this component is generally not covered by standard variance estimation formulas, and that the correlated variance contribution to the total error can be quite extensive. Hansen and others knew these facts before they explicitly formulated the model. This knowledge was used to gradually change the U.S. census data collection mode from interviewer enumeration in 1950 to self-administration in subsequent censuses due to the large enumerator effects in 1950.

The basic thought in those days was that quality could be maintained via evaluation studies of MSE components while at the same time keeping individual error sources under control via quality control schemes borrowed from industrial settings. During the 1960s and the early 1970s there is a literature published on administrative applications of statistical quality control. The application areas include those operations that resemble the assembly line in an industry setting, such as keying, coding and editing. This literature was short-lived because it did not really involve the clerks in any improvement actions and the administration around the activity was quite elaborate. One example of this literature is Minton (1972).

The U.S. Census Bureau conducted a number of evaluation studies during these decades which generated a lot of methodology development but like all large-scale evaluation studies conducted in those days the results were published years after the census and was of limited value to the users. Evaluation studies were mostly for the producer of survey statistics. One example is the evaluation of coding performance in the 1970 U.S. Census that was published four years after the census was conducted (U.S. Bureau of the Census, 1974).

Some other observations for this period include:

a. The first attempts at widening the quality concept are made. Statistics Sweden (1979) uses the two main quality aspects relevance and accuracy (alternatively called total error), where relevance is decomposed into contents and timeliness, and where contents has four sub-components, including comparability with other statistics and comparability over time.

b. Pritzker et al (1967) introduced imputation as a means to achieve rectangular data sets, which would make things computationally easier in presence of item nonresponse during days when computation was difficult. Their firm advice was that the imputation degree should be kept on a low level, say 2-5 percent, to avoid adverse effects on estimates.

c. Automation seriously enters the scene. During the 1970s we try to automate coding, editing, data capture and interviewing. Very important studies on recall

errors and on the length of reference periods are conducted by Neter and Waksberg (1965).

d. Triggered by the worldwide 1970 census round, issues in connection with privacy and confidentiality are being addressed by a large number of researchers including Dalenius.

e. Dillman (1978) publishes his book on the Total Design Method, which is basically a prescription on how to achieve decent response rates in mail and telephone surveys based on social exchange theory.

5. The Period 1980 and to This Day

During the last 20 years we have continued to study specific error sources and various means to reduce the errors or compensate for their effects. There have been very few attempts at routinely using the survey models to estimate major MSE components. Platek and Särndal (2001) elaborate on the reasons for this state of affairs. Survey statisticians have often been able to show that some phenomena, such as telescoping, exist or that some survey methods, such as incentives, seem to work. More seldom have they been able to explain why. The statistical theory is simply not sufficient as a sole theory for survey methodology. Theories from psychology, sociology, communication, economy, management, and other disciplines also come into play. These theories can tell us what the root causes of survey quality deficiencies are, and they provide survey design principles aimed at reducing these root causes. These thoughts were, of course, not new to survey designers when the CASM (Cognitive Aspects of Survey Methodology) movement started with a seminar in 1983 resulting in the document by Jabine et al (1984). However, the first huge effort to improve collaboration between survey researchers and cognitive scientists took place already in the late 1970s at a conference sponsored by the British Social Science Research Council. The topic was "retrospective data in surveys." One of the outcomes of the conference, where psychologists and survey researchers met, was the volume by Moss and Goldstein (1979).

Already in the 1950s, however, Morris Hansen at the U.S. Census Bureau brought in scientists from outside statistics who could shed light on causes of errors, and in the late 1960s Tore Dalenius gave courses at Stockholm University on psychological aspects of survey response. Sudman and Bradburn (1974) and Schuman and Presser (1981) also tried to explain causes of errors and offered hypotheses of root causes for problems. What was new with the CASM movement was the organized effort to build this bridge between disciplines. It is now known that the response process has implications for the survey design and especially the choice of data collection mode and the design of questions and the questionnaire. See Tourangeau et al (2000) for more detailed accounts of this movement.

Much of the research on the response process has taken place within the context of surveys of individuals and households. Only more recently has there been an interest in developing similar processes for establishment surveys (see Edwards and Cantor (1991), and Willimack and Nichols (2001)).

During the end of the 1980s the word quality as in data quality became increasingly common. It was definitely recognized that error size is not sufficient for describing the characteristics of a product and the quality movement led by Deming, Juran, Grant, Box, and others had a profound effect on statistical organizations in much the same way it had on other businesses. Many statistical organizations started to apply principles of total quality management (TQM), a concept not embraced by Deming. During these years the Washington Statistical Society organized conferences on Quality Assurance in Government, conferences that focused on TQM, process control, causes of variation, management tools, and teamwork. These efforts were transferred to European statistical organizations in the beginning of the 1990s and were accepted widely.

For instance, in 1999 Statistics Sweden proposed the formation of a Leadership Group (LEG) on Quality to attain improved quality in the European Statistical System (ESS). The ESS comprises Eurostat and the National Statistical Institutes (NSIs) associated with Eurostat, i.e., those organizations that are responsible for producing official statistics in the European Union. The LEG presented its results at an International Conference on Quality in Official Statistics held in Stockholm 2001. A large international audience discussed the recommendations, and other findings on quality in official statistics were also presented.

During the course of its work, the LEG felt the need for the ESS to agree on a common set of values and ideas on how to work with quality-related matters. Some NSIs have developed policy statements for their quality work, but there are no statements pertaining to the entire ESS. The LEG believed that policy might be too strong a notion for such a common set of values and ideas. Instead, the LEG drafted a Quality Declaration consisting of a mission statement and a vision for ESS together with a number of principles or values for quality work in the ESS, which was signed by NSIs belonging to the ESS (see Lyberg et al 2001). After 2001 work has begun on implementing the LEG recommendations across the ESS.

All this has to do with organizational quality as a prerequisite to product quality. Most statistical organizations have produced documents on how they deal with organizational quality, i.e., documents on business plans, strategic plans or protocols. For instance, Statistics New Zealand has produced a number of protocols as a code of practice for the production and release of official statistics. The U.N. has compiled ten Fundamental Principles of Official Statistics (United Nations 1994). Franchet (1999) discusses performance indicators for international statistical organizations. Many other organizations have similar documents in place.

The bottom line is that the concept of quality in statistical organizations has clearly changed during the last decade. The dominating approach is built on the ISO 8402 norm from 1986, which states that quality is ”the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.” Thus accuracy is no longer the sole measure of quality.

In 1987 Statistics Canada released its Quality Guidelines, and Jabine et al (1990) published a quality profile for the U.S. Survey of Income and Program Participation (rather than what was earlier called error profile). The goal of a quality profile is to describe what is known and not known about the survey and its processes and to report any studies on quality dimensions conducted. The strength of a quality profile is that information can be accumulated over time and therefore suited to continuing surveys. It is somewhat disturbing to notice that when it is time for quality reporting we might have relatively little to say about some of the quality dimensions. This is an issue dealt with by Platek and Särndal (2001, with discussions): How much and in what form should quality information be provided to the user? Very little is known about how users as a collective and as individuals view quality information, how much they can understand, and how they use it. There are reasons to believe that apart from a few sophisticated users, most emphasize visible dimensions such as timeliness and comparability, and less emphasis is put on accuracy.

The period starting around 1980 is characterized by impressive technical achievements such as continued development of sampling in general (Särndal et al 1992), multiple imputation (Rubin 1987), disclosure-avoidance techniques (Fienberg and Willenborg 1988), and new ways of modeling measurement errors via latent class analysis, structural equation models, and multilevel analysis (Platek and Särndal, discussion by Biemer 2001). Other achievements come from disciplines outside statistics or from disciplines collaborating with statistics. For instance there is a theory for survey participation (Groves and Couper 1998) based on psychological compliance principles, there is an emerging theory for self-administered questionnaire design (Jenkins and Dillman 1997) based on theories for cognition and visual perception, and there is a consistent literature on the use and effects of using incentives to enhance survey participation (Singer et al 1999).

Now, in 2003, there is a very extensive but fragmented literature on survey quality and related issues. The Wiley series of monographs covering panel surveys, telephone survey methodology, measurement errors in surveys, business survey methods, survey measurement and process quality, computer assisted survey information collection, and survey nonresponse has provided an update of the survey quality field from various aspects. There are thousands of journal articles and unpublished papers written during these 50 years dealing with survey quality or related issues. It is, however, difficult to find a common and by many accepted thread in this extensive literature, especially if we include the literature on management.

6. The Tore Dalenius Effect

Tore Dalenius contributed to the development of survey quality in many ways. His main interests included three areas: (1) optimum stratification in different settings, including multivariate problems (Dalenius and Gurney 1951 and Dalenius 1957), (2) total survey design (Dalenius 1974) and (3) controlling invasion of privacy in sample surveys (Dalenius 1988). His accomplishments regarding sampling are summarized in his 1957 thesis, which is remarkable in the sense that it contains material sufficient for several theses and that is still worth reading. Tore realized early that there was no comprehensive theory for survey planning, so he outlined principles and methods for the planning of sample surveys (Dalenius 1971). Unfortunately the book is in Swedish, but one of the concepts used is “total survey design.” Tore was a proponent of planning a survey with all potential error sources in mind, not just the sampling error. This might sound trivial but what he meant was that the survey should be planned given all these problems rather than accounting for them after the fact. He spoke a lot about optimality, resource allocation, and using resource-oriented approaches, i.e., trying to achieve gains in efficiency by incorporating into the design one or more of the specific administrative and organizational features of the framework within which the survey was to be conducted (Dalenius 1962).

The total survey design idea was moved forward through a major research project sponsored by the Swedish Central Bank and led by Tore. Over 70 research papers were produced in this project and the findings were presented in a report (Dalenius 1974). The ultimate goal was to continue the work and produce a textbook on total survey design with colleagues at the U.S. Census Bureau. That never happened. Instead Tore started working on issues related to privacy and confidentiality.

Tore was completely open to new ideas in the field. As mentioned he led a course on cognitive issues in Sweden around 1969 already. The remarkable thing with that endeavor was that he asked a psychologist to join his research project and describe what cognitive functions were in place when respondents were asked questions of different types. He was an avid fan of Lord, Novick, and other quantitatively oriented behavioral scientists. He tried to incorporate neo-Bayesian ideas in survey sampling when very few others did. He saw the possibilities with automated coding when very few believed in the idea. Tore's research network was truly impressive with champions like Neyman, Mahalanobis, Cochran, Hansen, Hurwitz, Tepping, and Bailar as regular pen pals. He was the founder and the first head of the Survey Research Institute at Statistics Sweden, and he published a lot in Swedish mainly because he wanted to help improve Swedish official statistics.

Tore was not too keen on using the term "quality." He recognized the importance of the user (Dalenius 1985), and he thought that the user should have estimates with small MSEs, i.e., small total errors. He saw other quality dimensions as constraints, and he strongly believed in reducing errors and in designing surveys so that MSE should be as small as possible given a specified research budget.

7. Time for a Paradigm Shift?

The discussions following Platek and Särndal (2001) and Dillman (1996) and also other observations show that there is considerable variation in the survey community regarding the perceptions of quality and what the biggest problems are for future survey work.

a. The user population and user demands have changed during these 50 years. New technology has improved the ways survey data can be accessed and delivered in time. Producers who cannot meet such demands might have to exit the market and consequently new actors conducting surveys and performing secondary analyses have entered the market. For many large organizations quality or customer perceived quality, no matter how it is defined, has become a survival issue (Fellegi 2001).

b. It is not sufficient to rely on MSE as a measure of quality. MSE is normally not sufficient even for accuracy, since the variance and the bias components computed usually do not reflect contributions from all the different error sources. It is a difficult and expensive endeavor to try to estimate MSE and/or its components.

c. As pointed out by Mathiowetz and Groves in the discussion following Platek and Särndal (2001), there is no agreed-upon unified theory or even framework for handling survey quality. The statistical theory is helpful in some respects and theories of human behavior are helpful in other respects, but there are few attempts at integrating the theories.

d. Know-how differs between countries and organizations. Training in survey methodology is fragmented across organizations and countries and there is a variety of training approaches.

e. Most organizations face a situation where delivery demands are increasing while resources are being cut, i.e., they must do more with less. That is not necessarily bad only, but the situation calls for innovation. The attempts at using TQM and similar work philosophies have not been as successful as we thought ten years ago, but some elements have sustained resistance. For instance, viewing survey work as processes, adopting a user perspective, developing current best methods, benchmarking, and promoting teamwork and collaboration are examples of elements that have prevailed.

Perhaps there is need for a revised total error concept for surveys (discussion by Biemer following Platek and Särndal 2001). The revised concept would include comparisons of total MSE and MSE for major components of error, provision of data on the magnitudes of biases and variances, conducting meta-analyses and other study-integration approaches, developing theories for optimal design of survey subprocesses, and documenting approaches so that they can be replicated and eventually adopted as standard practice.

We might want to strive for error-free processes where critical key process variables are checked from time to time to see if the processes are stable. If they are unstable or display a variation that cannot be accepted the processes have to be adjusted through experimentation and application of cognitive methods that can reveal root causes of problems.

8. References

Bailar, B.A., and Dalenius, T. (1969), "Estimating the Response Variance Components of the U:S.Bureau of the Census' Survey Model," *Sankhya*, B, pp. 341-360.

Brackstone, G. (1999), "Managing Data Quality in a Statistical Agency," *Survey Methodology*, 25, 2, pp. 139-149.

Dalenius, T. (1957), *Sampling in Sweden*, Stockholm: Almquist och Wiksell.

Dalenius, T. (1962), "Recent Advances in Sample Survey Theory and Methods," *Annals of Mathematical Statistics*, 33, pp. 325-349.

Dalenius, T. (1974), "Ends and Means of Total Survey Design," *Forskningsprojektet Fel i Undersökningar*, Stockholm: University of Stockholm.

Dalenius, T. (1971), *Principer och metoder för planering av samplingundersökningar*, Intern handbok, nr 4, Statistics Sweden (in Swedish).

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

Dalenius, T. (1985b), "Relevant Official Statistics," *Journal of Official Statistics*, pp.21-33.

Dalenius, T. (1988), *Controlling Invasion of Privacy in Surveys*, Statistics Sweden.

Dalenius, T., and Gurney, M. (1951), "The Problem of Optimum Stratification," *Skandinavisk Aktuarietidskrift,* 34, pp. 133-148.

Eurostat (2000), "Assessment of the Quality in Statistics," Eurostat/A4/Quality/00/General/Standard report, Luxembourg, April 4-5.

Deming, W.E. (1944),"On Errors in Surveys," *American Sociological Review*, 9,4, pp. 359-369.

Deming, W.E. (1950), *Some Theory of survey Sampling*, New York: Wiley.

Dillman, D. (1978), *Mail and Telephone Surveys: The Total Design Method*, New York: Wiley-Interscience.

Edwards, S.M., and Cantor, D. (1991), "Toward a Response Model in Establishment Surveys," in P.Biemer, R. Groves, L.Lyberg, N. Mathiowetz, and S.Sudman (eds), *Measurement Errors in Surveys*, pp. 211-233, New York: Wiley and Sons.

Dillman, D. (1996), "Why Innovation Is Difficult in Government Surveys," *Journal of Official Statistics*, 12,2, pp. 113-198 (with discussions).

Fellegi, I. (1964), "Response Variance and Its Estimation," Journal of the American Statistical Association, 59, pp. 1016-1041.

Fellegi, I. (1996), "Characteristics of an Effective Statistical System," *International Statistical Review,* 64,2.

Fienberg, S.E., and Tanur, J.M. (1996), "Reconsidering the Fundamental Contributions of Fisher and Neyman on Experimentation and Sampling," *International Statistical Review*, 64, pp. 237-253.

Fienberg, S.E., and Willenborg, L.C.R.J. (1998), "Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data," *Journal of Official Statistics*, 14, 4, pp. 337-345.

Franchet, Y. (1999), "Performance Indicators for International Statistical Organisations," *Statistical Journal of the United Nations Economic Commission for Europe*, 16, 4, pp. 241-250.

Hansen, M. and Hurwitz, W. "The Problem of Non-Response in Sample Surveys," *Journal of the American Statistical Association*, 1946, 41, pp. 517-529.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory*. Vols I and II, New York:Wiley.

Groves, R., and Couper, M. (1998), *Nonresponse in Household Interview Surveys*, New York: Wiley-Interscience.

Hansen, M., Hurwitz, W.N., and Pritzker, L. (1964), "The Estimation and Interpretation of Gross Differences and the Simple Response Variance," in C.R. Rao (ed.), *Contributions to Statistics* (presented to P.C. Mahalanobis on the occasion of his 70th birthday), Calcutta: Statistical Publishing Society.

Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (1984), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, Washington, DC: National Academy of Sciences.

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

Jenkins, C., and Dillman, D. (1997), "Towards a Theory of Self-Administered Questionnaire Design," in L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), *Survey Measurement and Process Quality*, pp. 165-196, New York: Wiley and Sons.

Juran, J.M., and Gryna, Jr, F.M. (1980), *Quality Planning and Analysis*, Second edition, McGraw-Hill.

Kish, L. (1995), *The Hundred Years' Wars of Survey*

Sampling, Centennial Representative Sampling, Rome.

Lyberg, L. et al (2001), "Report from the Leadership Group on Quality," Paper presented at the International Conference on Quality in Official Statistics, May, Stockholm.

Mahalanobis, P.C. (1946), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society*, Vol. 109, pp. 325-378.

Minton, G. (1972), "Verification Error in Single Sampling Inspection Plans for Processing Survey Data," *Journal of the American Statistical Association*, 67, 337, pp. 46-54.

Morganstein, D., and Marker, D.A.(1997),"Continuous Quality Improvement in Statistical Agencies," in L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), *Survey Measurement and Process Quality*, pp. 475-500, New York: Wiley and Sons.

Moss, L., and Goldstein, H. (1979), *The Recall Method in Sample Surveys*, London: University of London Institute of Education.

O'Muircheartaigh, C. (1997), "Measurement Errors in Surveys: A Historical Perspective," in L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), *Survey Measurement and Process Quality*, pp. 1-25, New York: Wiley and Sons.

Neter. J., and Waksberg, J. (1965), "Response Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study". U.S. Bureau of the Census Technical Paper 11.

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*, Vol. 97, pp. 558-606.

Platek, R., and Särndal, C.-E. (2001), "Can a Statistician Deliver?" *Journal of Official Statistics*, 17, 1, pp. 1-20 and Discussion, pp. 21-127.

Pritzker, L., Ogus, J., and Hansen, M.H. (1965), "Computer Editing Methods: Some Applications and Results," *Bulletin of the International Statistical Institute,* 41, pp. 442-466.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Scheuren, F. (2001), "How Important Is Accuracy?" *Proceedings of Statistics Canada Symposium*, Statistics Canada.

Schuman, H., and Presser, S. (1981), *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context,* New York: Academic Press.

Singer, E., Van Hoewyk, J., and Gebler, N. (1999), "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys," *Journal of Official Statistics*, 15, pp. 217-230.

Statistics Canada (2002) "Statistics Canada's Quality Assurance Framework 2002". Statistics Canada Catalogue no. 12-586-XIE.

Statistics Sweden (1979) Guidelines for presentation of the quality of statistics. Reports on Statistical Co-ordination 1979:8. Statistics Sweden.

Statistics Sweden (2001) "Quality definition and recommendations for quality declarations of official statistics". Reports on Statistical Co-ordination 2001:1. Statistics Sweden.

Stephan, F.F. (1948), "History of the Uses of Modern Sampling Procedures," *Journal of the American Statistical Association*, 43, pp. 12-39.

Sudman, S., Bradburn, N. and Associates (1974), *Response Effects in Surveys*, Chicago: Aldine.

Särndal, C.-E., Swensson, B., and Wretman, J.(1991), *Model Assisted Survey Sampling*, New York: Springer –Verlag.

Tourangeau, R., Rips, L., and Rasinski, K. (2000), *The Psychology of Survey Response*, Cambridge University Press, Cambridge, UK.

United Nations (1994a), Report of the Special Session of the Statistical Commission, New York, April 11-15, E/1994/20.

U.S. Bureau of the Census (1974), *Coding Performance in the 1970 Census, Evaluation and Research Program PHC(E)-8*, U.S. Government Printing Office.

U.S. Federal Committee on Statistical Methodology (2001), *Measuring and Reporting the Quality of Survey Data*, Statistical Policy Working Paper 31, Washington, DC: U.S. Office of Management and Budget.

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