Session 6

REVIEWING AND REPORTING QUALITY IN SURVEY DATA

Quality Declarations at Statistics Sweden - Principles and Practises

Claes Andersson, Håkan L. Lindström and Lars Lyberg

Statistics Sweden

Abstract: The general principles for quality definition and quality declaration at Statistics Sweden are presented. Their development over the last two decades is discussed in the light of an increasing concern for the users of statistics. Some background is given to explain how variation in ambitions, techniques and resources have changed the possibilities to study and measure quality. For the major quality component accuracy the statements of the quality level of the subcomponents are presented one by one. We mention general approaches to promote measurement and presentation of product quality. Finally we give some examples of products with good quality declarations.

Key words: Quality declaration, accuracy.

The Development of the Quality Concept

The production of high quality statistics has always been a concern at Statistics Sweden (SCB). Quality, as we see it, is a vector of components. Together these components describe the quality but they cannot all be measured in a quantitative way and therefore they cannot be added to each other to form a measure of the total quality.

Different sets of principles have been applied for different groups of products. In 1979 some general principles were formulated for quality definition and quality declaration at Statistics Sweden. There was a strong recommendation from the chief statistician that these should be followed by all surveys. At that time the concept focused on the effects on estimates of those procedures that influence the mean squared error, i.e., strategy, data collection method, and measurement, nonresponse, data entry, editing, coding and estimation models. The general recommendation was supported by more detailed definitions and recommendations in specific areas. Special decisions were made regarding mandatory reporting of nonresponse, variance calculations and questionnaire testing, i.e., three specific areas. The basic principles were quite general and should be applied on censuses and sample surveys based on administrative records or survey data.

But other aspects of quality were not totally forgotten. Also the content of the survey, the comparability with other surveys and time aspects were included in the general principles. The main reason why these aspects were less elaborated upon is obviously that the work with the quality declarations was led by statisticians while subject matter specialists were less interested in quality issues at that time. Another reason may be that the production of

official statistics at that time was highly concentrated to Statistics Sweden. Most of our statistics were then centrally financed by the government. Other clients were less important. Ideas on user or customer satisfaction were virtually non-existent.

Eventually Statistics Sweden has become more and more dependent on the market. The responsibility for official statistics has been transferred to about two dozen agencies since 1994 and Statistics Sweden is one of them, albeit the largest. Now only about half the volume of our work is funded by the government and the other half by other clients. In most cases we are competing for the contracts. This has forced us to turn to a broader definition of quality, highly influenced by ideas from total quality management. Current important principles are:

(i) **The focus is on the user**. A product's quality is determined by the user's opinion of the product and its usefulness. The user's opinion should direct the approach of the development work.

(ii) Quality encompasses **all characteristics of a product** (commodity or service) influencing how well the product satisfies the user's needs and expectations.

With this definition quality has a **descriptive meaning from the producer's point of view**. This is why we use the concept **content** and do not speak about relevance. The **user makes the judgement** and decides if the quality of the product is good or bad in relation to his/her intended use of it. It is essential for the producer to be aware of the quality judgements of existing and potential users, since these judgements provide a basis for development leading to higher quality, which in turn leads to increased user satisfaction.

One consequence of this point of view is that the user have to take more responsibility for the level of quality and that resources are set aside to reach it. Before 1994 the producer felt this responsibility and could argue with the government about the resources needed. The user could expect the producer to provide good accuracy within the limit of given resources. One can hope that the new situation will promote more intense discussions.

The development of the present quality concept was preceded by long discussions within SCB in order to make it both useful and generally accepted. Quality can be seen as a vector consisting of four main components: **Content**, **Accuracy**, **Accessibility,** and **Timeliness and periodicity**, which are broken down into 23 subcomponents. There was a fairly common acceptance of the subcomponents *per se* but opinions varied on the best grouping into main components. Especially the subcomponents on time and comparability were under intense discussion. The subcomponents are given below:

Content:	**Timeliness and periodicity:**
Statistical entity: • Type of unit and population • Variables • Types of statistical measure • Study domains Comparability with other statistics	• Time of reference • Length of production time • Punctuality • Periodicity • Comparability over time
Accuracy:	**Accessibility:**
Overall accuracy Sources of uncertainty: • Coverage • Strategy (sampling and estimation) • Measurement • Nonresponse • Processing • Model assumptions Presentation of uncertainty measures	• Forms of dissemination • Presentation • Documentation • Access to microdata • Information services

All producers of official statistics now have to follow these principles.

It is obvious that the quality information will be different for different subcomponents. Especially it is difficult to produce information on the accuracy subcomponents in the way users want. The users want to know the quality of the product and are less interested in the process leading to the product. Still we often have to compromise since process information is what the producer can offer and also has to offer. The information on accuracy that is offered can be classified in the following levels:

1. Quantified quality of the product, like evaluation results, variance calculations, response variation.
2. Quantified process indicators, like nonresponse and editing rates.
3. Generalized knowledge on error tendencies from "comparable" surveys.
4. Process descriptions like coding or editing rules.
5. Common sense conclusions/Vague knowledge - for example about the presence of a black market economy, car accidents not accounted for, etc.
6. No knowledge whatsoever of the quality.

Other than saying that level 1 is what everyone should strive for and level 6 is unacceptable, there is no absolute ordering between the other levels. In practise we get more information about the process quality that is useful

mainly for the producer than we get information about the product quality needed by the user. The computerization of survey processes will easily provide us with more and more process information. This is much cheaper and easier to come by compared to information on product quality.

The widening of the quality concept mirrors how the users are given more and more influence on official statistics in Sweden and, as a consequence, how "new" quality components increase in importance. Accessibility is a component of great importance for many users. They tend to prefer to produce their own statistics and ask mainly for edited files or easily accessible data in databases. The pressure on fast production is also growing and so is the demand for decreased costs. Demand for a high-level quality regarding one component is often in conflict with the possibility to maintain quality regarding other components. The demand for fast production will make it harder to obtain an acceptable response rate.

One important issue is how to use of the quality declaration. Statistics Sweden is no longer the only producer of official statistics from the collected data and is consequently no longer in a position to advise on its use. A quality declaration cannot be written to suit all situations since data files and statistics will be disseminated in many ways and also further processed by different users. We have to provide the potential users with information that makes it possible for them to derive the quality of their own statistical products. It must also be made obvious that it is their responsibility to declare the quality of their products.

In the near future we anticipate an increasing demand for statistics to be standardized in order to make comparisons and mergings with other statistics possible, especially with those produced by international organizations. International comparability may for example be in conflict with the "the most relevant" content in a survey or in a country.

Organizational and Business Aspects

Statistics Sweden has become a hybrid. We are an agency responsible for our share of statistics funded by government together with half a dozen other agencies with identical responsibilities. Most of the surveys that Statistics Sweden is responsible for concern areas where there is no obvious subject matter agency. One example is the national accounts, another is the labour force survey. An example of a responsibility shift is the following: the responsibility for environment statistics has been transferred to our agency for environmental protection. The shift in responsibility means that the agency gets the funding for statistics production and can do the work itself, or let an outside firm do it. Statistics Sweden could be one of those firms.

Thus Statistics Sweden is also a statistical firm. Most of the production work has so far gone back to our agency following extensive negotiations regarding cost and quality. But we are in a very competitive environment. It is very tempting for all these agencies to try to do some work themselves or perhaps let several firms, including Statistics Sweden, do different parts of the work. On top of this Statistics Sweden shall oversee this new system. Basically, one

specific department at Statistics Sweden, R&D, where the authors work, has this task. We are supposed to report to government via our Director General how all the agencies, including our own, perform. We are also responsible for general methodological development that should benefit the entire Swedish statistical system. This also means that we should train and consult with these other agencies, as we always have done within our own agency.

The organizational structure and the different roles of Statistics Sweden might seem very complicated. To some extent that is true but so far the transition has been really smooth. We can notice an increased interest in official statistics among agencies and we can definitely notice an increased interest in quality and methodology, which must be good for the status of statistics in Sweden.

So how does all this reflect on product quality? We do not know yet. There is little money for evaluations and methodological studies. We believe that product quality must be achieved through improved and stable processes. As has been pointed out there is no shortage of process data and the collection of such data and trying to develop standardized procedures will help improve quality. Insofar these other agencies choose to let Statistics Sweden do the production work this approach is valid also for them. If they work on their own it is our job to see to it that the products are up to par when it comes to quality declarations, publishing, and proper use of accepted methodology. But we are in no position to tell them how to do things.

The Development of Quality Assurance - Variation in Resources

The view on how to tackle quality problems has varied at Statistics Sweden during the last decades, due to variation in financial and methodological resources. In the beginning of the period finances were in good shape and a number of evaluation studies took place. For a number of years, special funds were available for quality studies. There were both smaller studies shedding light on specific error sources in specific surveys, and larger ones, most notably the evaluations of the population censuses. These studies led to improvements in the methodology used but rarely to profound process improvements. The improvements had a tendency to stay within the surveys evaluated. For instance, evaluations of the coding process in the censuses led first to the use of independent verification and then to automated coding but very little of these achievements spilled over to other surveys.

Then there was a period of redesign of surveys. Typically, a group of methodologists studied the design and came up with quite remarkable suggestions on how to change the design so that it became more efficient. In some cases the improvements led to significant reductions in sample size with reductions in costs and respondent burden as consequences. This approach was very demanding, since each attempt drew a lot of methodological resources. Perhaps it was possible to conduct one or two such efforts per year and knowing that the number of surveys is around 100, it is

easy to see that the approach does not seem too efficient. It was basically abandoned since it exhausted the methodological staff.

There was a feeling that procedures that were common to many surveys, like questionnaire development, coding, editing, nonresponse reduction and adjustment, estimation, and analysis should be done similarly across the organization. Therefore a lot of current best practices were developed during the 80s, most notably in the areas of questionnaire development, editing, automated coding, nonresponse, and estimation. These practices, however, had difficulties becoming known across the organization. Partly this state of affairs was due to a lack of financial pressure and a lack of competition. Meanwhile, of course, a lot of developments and improvements took place within the individual surveys but the common slogan that visualizing good examples would do the trick simply was not true. There was no systematic benchmarking within our organization.

During this period two general measures were developed to keep track of quality: the nonresponse barometer and the yearly quality report based on self-evaluations performed by the survey managers. Both of these efforts are described below.

Statistics Sweden bought into the total quality management concept in 1993. The reason was that our position had become more vulnerable in connection with the creation of the new statistical system. The customer became more visible and we realized that improvements must involve those who work on the processes. All of a sudden Statistics Sweden had to compete for work which called for some changes. One such change is to create current best methods of the kind just mentioned but in such a way that these methods are readily accepted by those involved in everyday survey work. There is a great need for such standardizations since, if they are applied consistently across the organization, they will reduce variation and save resources. Currently two such standardization projects are underway, one on nonresponse reduction and one on editing. Next year two new ones will start, one on questionnaire development and one on time series analysis. The project groups are set up such that implementation becomes more or less "automatic." Implementation will be assured through management follow-up and the fact that process owners have participated in the development.

Also, we are developing a system for quality assurance based on process thinking. Product quality is generated through process quality by means of checklists. For a number of survey operations, there are checklists that each survey manager should use to make sure that all process steps have been taken.

We are fairly convinced that there will be little room for large quality studies in the future. We believe that the route to take is working on processes, standardize them, measure key process variables on a continuing basis and use checklists to assure product quality.

The Accuracy Components

In many surveys at Statistics Sweden the presentation of accuracy can be made in a rather standardized way. This is due to the existence of a number of administrative registers which have been transformed to sampling frames. Whether we make a complete enumeration or a sample survey by using these registers as sampling frames, they give us possibilities and set limits to what we can do.

In the following paragraphs we present the subcomponents of the major quality component Accuracy.

Overall accuracy

The final goal for the quality declaration is to present the overall accuracy. This is, however, seldom accomplished. The existence of frames makes it almost always possible to check that estimates from the survey agree with known parameters computed from the register, though.

Some surveys present overviews of their current knowledge of errors referring to a series of experiments, observations and analysis.

Coverage

Most surveys can easily express coverage rates in relation to the sampling frame when there are sampling units like individuals, organizations, farms, enterprises, etc. From contacts with the authorities producing administrative records and from experiences of earlier surveys one often has a very good understanding of the number of units that has not been included in the frame and those who have not been excluded in time. Typically there are only a few percent over- and undercoverage in the surveys conducted by Statistics Sweden.

For some surveys the demand for very quick presentation is strong. Sometimes preliminary results have to be published before all the data have been collected. If this is to be classified as undercoverage or nonresponse error depends on the " time cut-off" rules.

A more important problem often appears when the units are events like road accidents, crimes and some types of economic activities. Nonobservation or underreporting results in some estimates being much too low. It is very rare that we have a good knowledge of the size of this error.

Strategy

The concept strategy includes sampling plan, sample size and estimation plan. Most surveys at Statistics Sweden use some stratification by register variables. The sample is usually simple random or systematic within each stratum. The sample is more often than not allocated to promote good quality in study domains. Allocation to reach best possible precision in population

estimates is rare. The chief statistician has declared that the calculation of precision must not be neglected in any sample survey.

Formulas for estimation and variance estimation are usually straightforward. Some surveys have found it useful to develop generalized variance functions to reduce the amount of calculation.

At Statistics Sweden several computer programs for the estimation in sample surveys have been developed. One of the recent programs, CLAN, is designed to estimate several different rational functions $f(\mathbf{t})$ of different totals, $\mathbf{t}$, (for instance means, ratios, ratios of ratios, etc.) and their standard errors in the same run. Since CLAN was written in the SAS macro language it works on PCs as well as on mainframe computers.

A large number of estimators, including the use of auxiliary information and calibration can be handled. The user may combine the choice of estimators with the specification of complex sets of domains in a very flexible manner.

So far four strategies have been implemented in CLAN. The strategies imply stratification of elements and clusters and the sample selection with SRSWOR. The majority of surveys conducted at Statistics Sweden, including a number of surveys that use pps-sampling, various types of network sampling and two-phase sampling schemes for stratification, can in different ways be brought back to these four strategies.

Measurement and data collection

Usually only process information or vague knowledge is available. Sometimes the producer will mention variables that are hard to measure, sometimes statements about the direction and size of the bias are made. A low occurrence of item nonresponse or absence of complaints by respondents (or the opposite) is sometimes mentioned as an indication of good measurement quality. Comparisons with other statistics are used as indicators of reasonable results. A number of surveys have conducted reinterview studies or other evaluations of their questionnaires.

A growing number of surveys have had their questionnaire pretested - usually by the Measurement, Evaluation and Development Laboratory (ML) of Statistics Sweden. Even if we can be rather confident that pretesting means an improvement of the questionnaire it is not designed to measure errors.

Processing

The data collection process may cause problems of different kinds. For instance, when data are collected from different administrative registers, these data are not always very well adapted to the needs we have. Actions have to be taken in order to edit and control the data.

Traditional data entry is becoming less frequent at Statistics Sweden depending on the use of computerized collection techniques, CAPI, CATI, TDE, etc. and scanning of questionnaires.

Automated coding of occupation and education is done on a regular basis.

Each moment in the data processing should have some effect on the final quality of both the micro data and the estimates computed from these data.

Nonresponse

A great effort has been made to develop and standardize the presentation of nonresponse rates. Since 1985 there is an overview, "The Nonresponse Barometer" that presents time series on response rates in all sample surveys and some censuses at Statistics Sweden. At first only a few important surveys on individuals and households were included and their rates had to be accepted without standardization. The reports covering the last few years include almost all surveys. The main features of their sampling and data collection plans are mentioned. Design changes that may have an influence on the response rate are mentioned.

Model assumptions

Statistical results sometimes rely on complex calculation schemes. These schemes may presume a model relationship among the input statistics if the calculated results are to be valid. This is the case for a lot of statistics on the environment and on the public economy. Errors in the model specifications may generate important errors. The models must be explained and robustness to specification errors explained.

Presentation of uncertainty measures.

Uncertainty of statistical estimates must be reported for all Swedish official statistics according to a set of recommendations by Statistics Sweden.

Evaluations

The results of censuses and sample surveys suffer from a number of errors. Editing, control of coding, etc. reduce the errors but cannot eliminate them. It is important that users of statistics have a possibility to judge how statistics can be used and what conclusions can be drawn from the published information. Also the producer should have an interest in knowing the actual quality of the published figures.

One possible way to get knowledge of the sizes of the errors in the final estimates is to carry out evaluation studies. These studies are not carried out on a regular basis and not in every statistical product at Statistics Sweden. The reason is of course limited financial and human resources as previously mentioned. An evaluation study is often considered complicated and it is often thought that available resources can be of better use in the main survey.

However, evaluation studies are carried out in the most important surveys and censuses as part of the quality control and as a basis for development of improved methods of data collection, editing, coding, etc.

Traditionally, evaluation studies at Statistics Sweden have been mostly *producer-* oriented rather than *user-* oriented and it is the major component *accuracy* and its subcomponent *measurement* that have been of interest. Perhaps this is natural since, in general, qualified statisticians have been responsible for these studies.

Random and systematic errors can occur both in estimated parameters and in the measurement of variable values. Systematic errors in variable values might eliminate each other when aggregated (if you are fortunate) and make the net bias in the estimates due to measurement error small, while the random part of measurement errors in general will increase the random errors in the estimates.

Often not only the net effect of the systematic errors are of interest, but also the gross effect. The reason for this is that in a survey, as well as in a census, the collected data are to be used not only to estimate the parameters that have been studied in the evaluation study but also to estimate other parameters that perhaps no one thought of at the time of the evaluation, for example estimates in totally different domains of study. Another reason is that the results from a census is usually used as a sampling frame where the variables are used to define different strata. Serious errors in the stratification might then ruin a sample survey. Further, when independent researchers outside Statistics Sweden use the data for different kinds of analysis, often by sophisticated methods, the researcher has to know the quality of the data he/she is using in order to draw valid conclusions.

Evaluation of the component accuracy is usually concentrated on the size of errors in the statistical estimates, for instance in estimated totals, ratios or mean values. In cases where the aim of the studies is to measure the systematic error in different estimates, evaluations can be conducted in different ways depending on level of ambition and available resources.

Crude measures and indicators of systematic errors can be obtained by comparisons between estimates from different surveys where related parameters are estimated, or by the study of correlated background variables whose values are known for the whole population. These types of measurements can often be made without further data collection but they give limited information about the measurement process.

Only in exceptional cases can good estimates of measurement errors be obtained without collecting additional data about the units in the main survey. If the aim of the study is to measure the reliability, you "only" need to repeat the main measurement process under conditions that are as identical as possible with the main survey. Such evaluation studies have been conducted rather frequently at Statistics Sweden. They are not to expensive and usually they give valuable information to the producer of the statistics. The results are seldom disseminated outside the agency but rather published in internal memos.

When you want to know something about the bias in published estimates, you need data with "true" values for at least a subsample of the units. The word "true" is used in an operational way here (in some cases there might not even be a true value). In practice it means that we are using a measurement process that is considered significantly better than the ordinary one.

"True" values are most often determined by matching and reconciliation. The statements given in the main survey are compared with the corresponding statements given in the subsample, where the questions or the wording of the questions need not be identical. If there is no discrepancy between the statements they are considered true. In other cases the respondent is asked to confirm the "true" statement, (the original, the new, or perhaps a third one).

This technique is rather expensive and is mainly used to evaluate the large registers and censuses.

Some discussions have concerned how to present the result from the evaluations. Let $\hat{\theta}$ be the estimate of a parameter based on the units in the evaluation sample and on the ordinary statements and let $\tilde{\theta}$ be the estimate of the same parameter based on the "true" measurements in the evaluation sample. Sometimes (1), $100 \times (\hat{\theta} - \tilde{\theta}) / \tilde{\theta}$ and sometimes (2), $100 \times (\tilde{\theta} - \hat{\theta}) / \hat{\theta}$, are used as measurement of relative bias. In (1) the deviation is shown as % of "true" (unbiased) value, while (2) shows the deviation as % of the "official" estimate. Of course you can always get (1) from (2) and vice versa but (1) is the *producer's* measurement as it tells the producer the deviation from the goal while (2) tells the *user* something about the error in the published figures. Measurement (1) was for example used in the evaluation of the Register of Employment while (2) has been used in the evaluations of the population censuses.

An example: The evaluation of the 1990 Census of Population and Housing

The evaluation program contained the following studies:

Evaluation of

- household data for dwelling households
- housing data for occupied dwellings
- employment data
- education data

The evaluation covered about 17 000 units which were sampled from persons in Sweden who were in the ages of 16 to 74 years and registered in Sweden on the 1st of November, 1990.

"True" values were determined by matching and reconciliation in the same way as described above. The statements given in the census were compered with corresponding statements given in the Labour Force Survey of November 1990. For about 16% of the units no "true" value could be

determined depending partly on missing values in one or both of the two and partly on discrepancies that could not be reconciliated.

The results from the evaluation study are published in the official statistical series. Both net errors and gross errors are estimated. Estimates are given for many different combinations of sex, age, region, type of dwelling, type of household, etc.

The Nonresponse Barometer

Since 1985 a yearly report on the nonresponse in some surveys at Statistics Sweden has been published. The aim of the report is,

- to show the amount of nonresponse in a number of surveys at Statistics Sweden,
- to give a picture of the "response climate" (i.e., do individuals, businesses, and other institutions become more or less willing to answer survey questions?),
- to be one (of several) instruments to compare different statistical products over time.

The aim of the barometer is not to describe the quality component named nonresponse, but rather to give a description of the size of the nonresponse for different surveys and over time. The *effects* of the nonresponse, for example nonresponse bias, are not handled here. An estimate of these effects is given in the yearly quality survey "The Quality Report" described below.

Only unit nonresponse is treated. Item nonresponse, that is where a unit has participated in the survey but has not given answers to every question, is not.

Measurement of the nonresponse rates are given both as weighted and unweighted figures. Weighting is done in many different ways depending on which measure that is considered best adapted to its purpose. In surveys related to businesses, the estimated number of employees or the turn-around (according to the information in the register) in the nonresponse businesses is used as well as the estimated number of nonresponse businesses in the population. When the units are individuals the weighting means that the estimated number of nonrespondents in the population is used.

The total nonresponse is classified according to cause (*refusals*, *no contact* and *other*), where this is possible (mostly in surveys of individuals).

The response climate is measured by asking the person responsible for the survey to give an estimate of the changes in response climate from last year and the change from five years ago. Four alternatives are possible: "*Better*, *Neither better nor worse*, *Worse*, *No opinion/irrelevant.*"

Another measure of the response climate is obtained by asking the interviewers' supervisors about their opinions. Then we get more information and it is also possible to say something about regional differences.

Until now the presentations of the results have been made by the different departments at Statistics Sweden. In the future the presentation of the nonresponse results will probably be more difficult depending on the new distribution of the responsibility for official statistics.

The Quality Report

The Quality Report is produced in order to provide a basis for an analysis of the development of the quality in the statistical products at Statistics Sweden. The report has been published yearly since 1988.

A questionnaire is administrated to every person responsible for a statistical product at Statistics Sweden. The questionnaire consists of three parts. In part 1, the respondent is asked to state general external factors of importance for the quality changes of his/her product, also the measures (if any) carried out to handle them are to be reported. In part 2 the respondent is asked to give estimates of how the quality level of their product has changed since last year. Estimates are to be given for each of the 23 quality components described above. Changes are measured by a five degree scale, *"much worse," "slightly worse," "unchanged," "slightly better,"* and *"much better,"* or *"not relevant."* In part 3 the respondent is asked to report remaining quality problems that are judged to be of special importance. Also planned improvements of importance for the quality shall be reported.

The questionnaire is sometimes filled in by a team, involving several persons working with the product. The questionnaire is examined and approved by the respondent's manager. In each department the responsible statistician examines all the questionnaires, make necessary completions and the data processing.

The respondents are instructed to fill in the form with the intended user's perspective in mind but it is important to note that the users normally have no possibility to share their opinions since they are not explicitly ask to do so.

The measurement process should be continually improved. Especially in part 2 there are currently substantial possibilities for subjective judgement when a change is reported. However, in order to improve the quality of the responses, the respondent is asked to give a motivation and a comment whenever a change is reported.

Needless to say, generally the alternative "unchanged" dominates the statements in part 2, (70%-90%). When a change is reported the alternatives "slightly better" and "much better" dominate. Future plans include a redesign of the entire process.

Endnote

A number of factors affect quality and the possibilities to declare quality. Cuts in funding generally means that there is less room for nonresponse follow-up. Money is also the reason why there are fewer evaluation studies these days. Demands for faster production also contributes to less time for nonresponse follow-up but also to less planning time in general. The new technology means that it is easier to measure process variables which to some degree can compensate for the problems mentioned.

A continuing quality problem is that so much of statistics are based on the use of administrative registers. Normally the producer of statistics has very little influence on the collection of such data and normally does not know much about coverage and measurement error problems unless special studies can be designed.

International organizations get more and more influence on contents and methods. It is more and more common that specific surveys deliver to systems of different kinds, like index systems or accounts systems. Concepts such as comparability, additivity, and completeness might get new meanings and country comparisons, for instance, might take precedence before local needs. Thus, international cooperation seems necessary.

Data Quality at the Energy Information Administration: The Quest for a Summary Measure

Renee Miller
Energy Information Administration

At one point not too long ago, I found myself in a conversation with a relative who, like many taxpayers, was not convinced he was always getting his money's worth. He was familiar with the Energy Information Administration's (EIA) data and said that the data were useful. I thought I was off the hook, but then he said something like, "How do you know if the data are any good? Do you have some kind of a measure?"

Skeptical relatives have not been our only questioners. "How do we know if the data are any good" is a question with which we at EIA have been grappling for years. It recently emerged during our Business Re-engineering efforts. The Business Re-engineering team[1] was chartered to rethink three core business processes: data operations, data integration, and product preparation and dissemination.

At several points the team discussed making EIA data more timely to better meet the needs of our customers. During these discussions someone would raise the concern about balancing timeliness and quality. We thought a summary measure of data quality would be helpful in this situation and we discussed what that measure might be. The discussion of a summary measure of data quality led to thinking about how information on data quality is presented to the public. The issues of how we ensure data quality, whether there is a summary measure of data quality, and how we report information on data quality to the public are intertwined.

This paper begins with a few words about EIA, then discusses our attempts to measure data quality. It continues with a discussion of additional activities to ensure data quality. It then presents a proposal for a summary measure of data quality and goes on to describe some recent developments. The paper ends with some thoughts on where we go from here.

EIA in a Nutshell

EIA is almost twenty years old. Congress established the agency in 1977 to be an independent source of energy information. It combined data gathering functions formerly performed by the Bureau of Mines, the Federal Energy Administration, and the Federal Power Commission. Besides combining data gathering functions, the new agency was also a combination of people. Some came from the three predecessor agencies. Some of us are from other statistical agencies such as the Bureau of the Census, Bureau of Labor Statistics and National Center for Health Statistics. Others came from academia and many other places.

---

[1]Members of the Business Re-engineering Team included: Project Director, Chuck Heath; Core group members: Ray Boyer, Clyde Boykins, Ann Ducca, Sue Harris, Mike Lehr, Dorine Andrews, and Lori Gillespie; Champions: Chuck Allen, George Baker, Noel Balthasar, Yvonne Bishop, Ken Brown, Bill Dorsey, Lamar Gowland, Mike Lehr, Nancy Leach, Bob Manicke, Renee Miller, Ken Vagts, Howard Walton, and John Weiner.

The activities described in this paper reflect traditions carried over from many of these agencies, plus new ones that we developed as we strive to become "team EIA."

There will be references to EIA's re-engineering efforts. These efforts stemmed from the realization that with declining budget and staff levels EIA could not improve (or even maintain) its level of customer service by doing "business as usual." In August of 1995 EIA's Quality Council chartered a team, and in April of 1996 the team delivered a blueprint for a re-engineered EIA. We are in the process of implementing parts of the blueprint.

Attempts to Measure Data Quality

During the development of the Business Re-engineering blueprint, the team developed measures for various processes. The idea was that instead of having a long line of staff checking and rechecking work, we would have measures that would indicate whether the processes were functioning effectively. One measure that eluded us, but that we kept coming back to, was a summary measure of data quality. In searching for a measure, we reviewed some approaches tried or suggested previously.

These approaches included validation studies, data comparisons, supply/disposition balances, and elements of data quality. Revision error, response rates and sampling error were also revisited as described below.

Validation Studies

In its early days, EIA conducted validation studies. Reflecting their extensiveness, they were called "cradle to grave" examinations of the data. They included a search for deficiencies in the universe list, an audit of company records to determine if they corresponded to what was reported, a check for transcription errors by comparing hardcopy to the automated data file, and many other activities. Reference [1] provides an example of a study pertaining to data collections on coal production.

As a result there was some information pertaining to each source of nonsampling error (coverage, measurement, nonresponse, and processing). Sometimes the information was quantitative. What was lacking was a way of adding it all up to get total survey error, because sometimes the errors were offsetting.

Overall, the studies were not popular with either the survey respondents or with the survey managers. Furthermore, they were expensive. They were stopped when our budget was reduced in the early 1980's.

Data Comparisons

EIA staff also compared the data series of interest called the reference series with other data collected by EIA or other organizations. In the early days comparative sources were plentiful. At the aggregate level, we computed each comparative series as a percentage of the reference series.

An early study focussed on the data on imports of crude oil. There were three comparative series

with data for three years, which we considered as nine independent estimates. We had nine ratios of the comparative series to the reference series and computed the mean, the standard deviation of the mean and a 95 percent confidence interval. The 95 percent confidence interval was 99.2 - 100.8 for imports of crude oil based on data for 1977, 1978 and 1979. We then concluded that the EIA reference estimate was accurate to within 1 percent[2].

As might be expected, these estimates of accuracy were not well received. Sometimes the comparative estimates had well-known problems. There was often no indication that they had been validated and most of the time there was no documentation on how the comparative series were obtained. Although we continue to perform comparisons and present the results to the public, we stopped coming to conclusions about data quality based on them.

In the eighties, we presented the results of data comparisons in a series of reports that became known as the "State-of-the-Data" reports [2]. Staff in the Office of Statistical Standards prepared these reports with input from survey staff. These comparisons were performed at the respondent level as well as at the aggregate level.

Currently survey staff members prepare annual feature articles comparing EIA data with other sources. The articles appear in EIA's *Petroleum Supply Monthly* and *Petroleum Marketing Monthly*, which have a wider distribution than the earlier "State-of-the-Data" reports. Examples of comparisons the user can find in these reports include data on imports of crude oil and petroleum products from EIA and the Bureau of the Census and prices of petroleum products from EIA and the Bureau of Labor Statistics.

These articles provide a vehicle to let users know that some observed differences in the data series stem from the different definitions or universes used in the data collections. In addition to the feature articles, results of comparisons have been presented at conferences such as the annual meetings of the American Statistical Association and Washington Statistical Society meetings. Some EIA programs, such as end-use consumption and electric power, routinely include comparisons in Appendices to their data publications [3, 4, 5, and 6].

Comparisons often raise more questions than they resolve. While it is comforting when data collected from different sources correspond well, there is still the possibility that they are not correct. Comparisons have been most useful when data are collected from the same respondents and we can match records. In these situations we can identify the individual respondents with differing responses and follow up to find out why. We have gained information about how respondents are interpreting our definitions and instructions through these follow-ups.

Supply/Disposition Balances

In addition to data comparisons, EIA staff members look for symptoms of problems in the published data by examining supply/disposition balances. The expectation is that supply should equal disposition. Both components, in turn, consist of several parts. Production, imports, and

[2] An Assessment of the Accuracy of Principal Data Series of the Energy Information Administration, DOE/EIA-0292, June 1981, page 23.

withdrawals from storage make up supply. Consumption, exports, and additions to storage make up disposition. Because we obtain the data that comprise the supply/disposition figures from different surveys, a balancing item is needed for supply to equal disposition.

EIA staff members have been using the balancing item as a warning signal. If, for example, the balancing item increased sharply from one year to the next, it could be an indication of errors in one or more of the components. A small balancing item, however, does not conclusively show that all the figures are of good quality. There could be offsetting problems; e.g., production, a component of supply, could be overstated and imports, another component of supply, understated. Therefore, it is difficult to use the balancing item as a measure of data quality.

While looking for symptoms of problems in the data through comparisons and balances is likely to identify major problem areas, it does not give us a systematic way of quantifying errors in the published data.

Elements of Quality

In August 1991, then EIA Administrator, Dr. Calvin Kent took a different approach to assessing data quality. In a presentation at the annual meeting of the American Statistical Association entitled, "Quality of Energy Data," he discussed four elements of quality: timeliness, consistency, continuity, and customer satisfaction. In the past few years we have made progress in measuring two of the elements: timeliness and customer satisfaction.

We measured timeliness as the number of days between the last day of the reference period and the "released for printing date" shown inside the front cover of the publication [7]. We have compiled data on timeliness for annual publications for the years 1990 through 1994 and for monthly and quarterly publications for 1993 to 1995. In the future we will use the date the publication returns from the printer to better reflect the date the customer receives it.

To measure customer satisfaction, for the past two years we have been surveying EIA's telephone customers. In February of 1996, EIA volunteers surveyed 264 telephone customers. The volunteers asked customers about their satisfaction in two broad areas: customer service and information quality. The first area included: ease of access, courtesy, familiarity with the information, understanding the customer request, and promptness in responding. The second area included: availability, relevance, accuracy, comprehensiveness, and timeliness.

About 73 percent of the respondents said they were either satisfied or very satisfied with the timeliness of the information. By contrast, 90 percent of respondents said they were satisfied or very satisfied with the accuracy of EIA data. Several interviewers noted, however, that a few respondents said they had no way of knowing whether the data were accurate or not. These respondents, nevertheless gave us a high rating because they said they had no reason to believe the data were not accurate.

With respect to timeliness, during 1995 there were widespread efforts to make data available earlier through electronic dissemination. However, in 1996 EIA received about the same overall rating on timeliness as in 1995. We think that one reason that the ratings did not change was that the

customer survey was conducted with telephone customers, a group that may not be fully aware of the electronic data.

EIA does not have measures for the two remaining elements: consistency and continuity. By consistency we meant--how do EIA data compare with other similar series and are our data internally consistent? (An example of a data inconsistency would be more domestic electricity sold than generated). For continuity our questions were: Do we measure the same thing over time? How does EIA handle revisions to the data? How does EIA handle breaks in its time series resulting from industry changes and modifications to our survey forms?

The four measures, in contrast to the work performed in the validation studies, do not attempt to measure total survey error. Rather, they are related to the fitness for use of our statistical products.

The Usual Suspects: Revision Error, Response Rates, and Sampling Error

Revision error, the difference between preliminary and final estimates, has been suggested as a measure of data quality. This measure has been criticized, however, because it does not address the issue of the quality of the final estimates. In addition, suppose there are no revisions. Does that mean there is no error?

Nevertheless, we have found the computation of revision error to be useful in improving our preliminary estimates. We present information on revision error to the public in annual feature articles to EIA's *Petroleum Supply Monthly* and *Natural Gas Monthly*. Other program areas, such as electric power and petroleum marketing, include the information in appendices to their publications [6, 8]. We show the preliminary estimate, the final estimate, and the percent difference. In the feature articles, we provide explanations of the differences, if available.

In addition, we have been tracking revision error as part of our organizational performance measurement system. EIA developed the system while participating as a pilot project under the Government Performance and Results Act of 1993 [7].

We also compute and publish information on response rates and sampling error, generally in the explanatory notes section of our publications. While these are important measures, they do not tell the whole story with respect to data quality.

Other Activities to Ensure Quality

Besides the activities just described, there are other activities performed throughout the agency to ensure the quality of the data. While we have found these activities useful, we also found that they did not lend themselves to measuring data quality. The activities include editing of the data and the development of statistical standards. In addition, we conducted audits to check compliance with standards. Furthermore, we have performed site visits with selected respondents. This section summarizes these activities.

Edits

Prior to publication, survey staff members edit the data using consistency checks, comparisons with previous reported values, and other more complex methods [9, 10]. They follow up by phone with respondents who have reported seemingly anomalous values. Because of the wealth of historical data for the weekly and monthly surveys, time series methods have been used to predict the current value and to construct tolerance limits for the new data. Examples of these methods are featured in Statistical Policy Working Paper 18, "Data Editing in Federal Statistical Agencies" [11].

Edits tell us about the quality of the reported data to some extent. We have found, however, that there are errors that edits cannot detect such as a respondent consistently reporting residential deliveries as commercial deliveries. Therefore, we have not been able to translate information from edits to a measure of quality for the published data.

Standards and Audits

EIA has developed a manual that contains copies of the agency's statistical standards [12]. In the foreword to the manual, we state that standards "help ensure data quality, remove ambiguities, avoid duplication of effort, and improve responsiveness to our data users." The standards cover both data collection and processing, and data presentation.

Using an analogy from the health area, standards are much like the advice to maintain a low-fat diet and to exercise regularly. While a low-fat diet and exercise purport to contribute to our long-term well-being, they do not ensure that we are disease free on a daily basis. There is a similar situation for statistical standards which is why adherence to standards has not been accepted as a measure of data quality.

As noted in the paper, "Quality in Federal Surveys: Do Standards Really Matter?" [13], the relationship between standards and data quality is tenuous. Nevertheless the paper notes that standards were helpful in establishing the credibility of EIA data, along with rigorous programs of enforcement, evaluation, and education.

EIA has conducted audits as part of its enforcement program. The initial round was broad in scope and mainly concentrated on standards and documentation review along with data processing issues. These audits checked each system for compliance with each standard [14].

The next round focussed on the quality control activities to determine whether they were adequate to control nonsampling error [15]. For each source of nonsampling error, we developed checklists of activities that could control these errors. The EIA standards manual and Statistical Policy Working Paper 15, "Quality in Establishment Surveys" [16] were used to develop the checklists. In addition, results of data evaluations were used to determine whether any identified anomalies resulted from a failure in a quality control procedure.

Site Visits

In the early 1990's, we started a program of site visits. During the visits we spoke to respondents about how they are interpreting selected items on the forms. In addition, we asked whether their records correspond with the items we are requesting. Unlike the validation studies of prior years we have not asked respondents to produce records for verification. We visited about a dozen respondents, covering coal and natural gas production, consumption and distribution data. While we have obtained useful data from these visits, the sample size has been too small to draw inferences. Furthermore, we did not ask all respondents the same questions.

Proposal Developed During Business Re-engineering to Measure Data Quality

During our re-engineering efforts the issue of a summary measure of data quality arose several times. And several times we concluded it could not be done. With the activities previously described as background, following is a rating scheme we tried to develop.[3]

The Business Re-engineering team ultimately decided it would not be workable because it would require much time and judgement. It is being presented because we learned something from the experience. The team was a diverse group consisting of managers, statisticians, analysts, computer specialists and interdisciplinary staff. We all had different reference points. The attempt to develop a summary measure proved helpful in making us all realize what was involved.

Dimensions of Data Quality

We began by listing some of the dimensions of data quality:

- sampling error
- measurement error (the difference between the value collected during the survey and the true value. It includes both reporting error and specification error)
- coverage
- nonresponse
- methodological consistency (this is the same as "continuity" in Dr. Kent's scheme described earlier. It pertains to breaks in the data series and whether the changes and their impact on the data are documented)

These dimensions differ in the ease with which we can quantify them. Sampling error, on the one hand, can be computed directly from the survey data. Methodological consistency, on the other hand, cannot be directly computed. Ideally a series should be stable over time; i.e., not have any

[3]Dwight French, Office of Energy Markets and End Use, participated as a subject matter expert in the re-engineering effort and worked with this author on the rating scheme for data quality.

breaks. Sometimes due to changes in the industry, it is inevitable that a data collection is modified. Does that series get penalized for having breaks?

Some other dimensions sound like they should be easily quantified, such as measurement error. EIA does not have information for each survey on an ongoing basis. This is the type of information that we obtained from the validation studies which have been discontinued.

Since we had some information for each survey, we thought we could gather it together. We would then rate each survey on each dimension using a 1 to 5 scale (where 5 is very satisfactory and 1 is very unsatisfactory). We would have 2 scales: level of knowledge about the category and level of quality. In this way we would obtain information on how much we know about data quality as well as information about the quality of the data.

Using nonresponse as an example, a survey might get a score of 5 on knowledge if documentation was available on the response rate, on our follow-up and imputation procedures, and key information was presented in the publications. A survey might get a 5 on the quality scale for nonresponse if the response rate was 98 percent in terms of both number of respondents and volumes reported.

We would then combine component scores into an overall score for a survey, program, or EIA as a whole. To ensure consistency we started to develop guidelines on what represents a "5" versus a "4" and so on. It got very complicated quickly. Furthermore, for measurement error and methodological consistency, we found it difficult to develop "quality" measures; therefore, we only had "knowledge" measures.

Perceived Complications

Because of the perceived complications and other issues, the Business Re-engineering team decided not to pursue this procedure. One issue was who would do the ratings. Another concern was that we could not really ensure consistency. Furthermore, there was the perception that a lot of time would be involved in performing the ratings. The general feeling was that even if we could be precise enough to ensure consistency, we would not be giving the user much more information than is provided in the explanatory notes section of our publications.

This comment raised the issue of whether the approach we should take should focus on the descriptive explanatory material, perhaps standardizing it. All of our data publications contain explanatory material. We have a standard on publication of energy statistics which is based on a directive in the *Statistical Policy Handbook* [17]. The EIA publication standard specifies that we describe the survey design and provides a checklist of activities to include. It also specifies that we point out the limitations of the data. The detail we provide on the limitations of the data and on features of the survey design varies across EIA.

Recent Developments

There are two recent developments at EIA that could affect the approach we take to presenting information on data quality. One is quite specific, the development of a quality profile for the Residential Energy Consumption Survey. The other is more global, electronic dissemination.

Quality Profiles

Last spring, EIA published its first quality profile, an extensive profile of the Residential Energy Consumption Survey. It was prepared by Thomas B. Jabine [18] in a joint effort between the Offices of Energy Markets and End Use and Statistical Standards. As described in the report the purpose of the *Residential Energy Consumption Survey Quality Profile* (*Quality Profile*) is "to present, in a convenient form, a report on what has been learned about the quality of RECS data since the survey began."

The report provides an overview of the survey and presents information about three major sources of nonsampling error: coverage error, nonresponse, and measurement error. It also discusses the contributions to nonsampling error of data processing and imputation procedures. In addition, it looks at the effects of estimation procedures on data quality. Furthermore, the report presents results of studies that have compared RECS data with data from other surveys, and describes relevant research currently in progress.

The *Quality Profile* has been very well-received. Several members of our energy advisory committee said it was a good model of how we should document our surveys [19]. They pointed out that customer satisfaction depends on data quality and that a quality profile would give users all the information they would need to determine data quality. They suggested that we do more profiles. Unfortunately, due to our budget constraints and reduction in staff levels, that does not appear likely.

Electronic Data Dissemination

As mentioned earlier, there was a concentrated effort to make data available electronically at EIA. Electronic dissemination has produced new possibilities. One is that the user would click on a data value and see a standardized description that explained it [19].

We have taken a couple of steps in that direction. EIA has developed an Electronic Styles and Standards Manual [20]. It requires that when a publication is released electronically that it is released in its entirety so that the explanatory material is included. For products released as files, we are required to provide data sources and caveats concerning the data.

Another step is the development of a succinct set of notes for data from the Commercial Building Energy Consumption Survey that will be released on the internet [21].Topics in the notes include: survey methodology, target population, sample design, changes in the survey from the previous

cycle, sampling rates, data collection procedures, response rates, minimizing nonresponse to the survey, and a general discussion of sampling and nonsampling errors.

Where Do We Go from Here?

The issue of a summary measure of data quality does not appear to be going away. EIA is moving toward a performance-based budget. In addition, implementation of the business re-engineering blueprint includes a pilot test to integrate survey operation activities. As part of this effort, staff members are developing measures to monitor the process overall. Ideally we would like to include a measure of data quality, apart from revision error.

We realize it is not likely that we will find the perfect measure. While we have not found a summary measure of data quality, there is agreement that providing users information on what we know about the quality of the data is crucial.

We have been giving our users explanatory material for years. Yet during our customer satisfaction survey, some have told us that they have no way of judging the quality of our data. They think EIA has good quality data, but they say they do not know for sure. Something seems amiss here.

Perhaps the future direction should be to make the information on data quality easily accessible and understandable. We could cover in a concise way the dimensions of data quality that we identified: sampling error, measurement error, coverage, nonresponse and methodological consistency. The notes developed for commercial consumption data that will be released on the internet are in the direction of this goal. They were based on work done at the National Center for Education Statistics.

Building on each other's work perhaps we can attain information on data quality that is so clear and accessible that users, themselves, will be able to answer the question, "How do we know if the data are any good?"

References

1. Management Engineers, Incorporated. "Coal Production Data Systems Validation: Final Report." Reston, Virginia, September, 1982.

2. Energy Information Administration. *An Assessment of the Quality of Selected EIA Data Series.* DOE/EIA-0292 (83), (85), (86), and (87). Washington D.C.

3. Energy Information Administration. *Household Energy Consumption and Expenditures 1993.* DOE/EIA-0321(93). Washington D.C. October 1995.

4. Energy Information Administration. *Commercial Buildings Energy Consumption and Expenditures 1992.* DOE/EIA-0318(92). Washington D.C. April 1995.

5. Energy Information Administration. *Manufacturing Consumption of Energy 1991.* DOE/EIA-0512(91). Washington, D.C. December 1994.

6. Energy Information Administration. *Electric Power Monthly March 1996 With Data for December 1995.* DOE/EIA-0226(96/03). Washington D.C. March 18, 1996.

7. Kirkendall, N.J. "Organizational Performance Measurement in the Energy Information Administration." Presented at Bureau of the Census Annual Research Conference, 1996.

8. Energy Information Administration. *Petroleum Marketing Monthly April 1996 With Data for January 1996.* DOE/EIA-0380(96/04). Washington D.C. April 2, 1996.

9. Swann, T.C. "Electronic Data Collection in the Petroleum Supply Reporting System." Presented at the meeting of the American Statistical Association Committee on Energy Statistics. Washington, D.C. April 28-29, 1988.

10. Weir, P., Emery, R., Walker, J. "The Graphical Editing Analysis Query System." Presented at Conference of European Statisticians, Work Session on Statistical Data Editing. Statistical Commission and Economic Commission for Europe. Athens, Greece. November 6-9, 1995.

11. Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology. *Data Editing in Federal Statistical Agencies, Statistical Policy Working Paper 18,* page 35. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. Washington D.C. May 1990.

12. Energy Information Administration. *The Energy Information Administration Standards Manual.* DOE/EIA-0521.

13. Freedman, S.R. "Quality in Federal Surveys: Do Standards Really Matter?" Presented at the Annual Meeting of the American Statistical Association, 1990.

14. Dandekar, R. "Future Data Quality Audits." Presented at the meeting of the American Statistical Association Committee on Energy Statistics, Washington D.C. November 2-3, 1989.

15. Energy Information Administration, Office of Statistical Standards. "Electric Power Data Evaluation." Unpublished report, 1994.

16. Subcommittee on Measurement of Quality in Establishment Surveys. Federal Committee on Statistical Methodology. *Quality in Establishment Surveys. Statistical Policy Working Paper 15.* Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. Washington D.C. July 1988.

17. U.S. Department of Commerce. Office of Federal Statistical Policy and Standards. *Statistical Policy Handbook.* Washington, D.C. May 1978.

18. Energy Information Administration. *Energy Consumption Series: Residential Energy Consumption Survey Quality Profile.* DOE/EIA-0555(96)/1. Washington, D.C. March 1996.

19. American Statistical Association Committee on Energy Statistics. Transcript of the Meeting. April 26, 1996, pages 93 - 118.

20. Energy Information Administration. "EIA Style and Standards for Electronic Products." Unpublished report. Washington D.C. January 1996.

21. Energy Information Administration, Office of Energy Markets and End Use. "Methodology and Technical Notes." Unpublished notes, prepared for Commercial Buildings Energy Consumption Survey data to be released on the internet.

REVIEWING AND REPORTING QUALITY

JAY WAITE, DISCUSSANT

I really appreciated the opportunity to discuss these two papers. It was especially refreshing that the two papers were actually on the same subject and that the papers presented were faithful to the papers as written.

Both papers were broadly similar and both discussed work of two statistical agencies in searching for what I fear is an unreachable goal. They are both striving for a linear measure of the quality of a survey.

In fact, quality like beauty, is most often in the eye of the beholder.

In the final analysis, quality is defined by the customer. Even if massive resources are expended to precisely measure the primary components of quality, different customers with different uses in mind will want to weight the components differently. To some, timing is critical and without it nothing else matters. To others, accuracy is the last and final determinant of a quality data product.

It seems that as statistical organizations, either public or private, we are fundamentally in the business of producing and delivering information. In this context, we need to be clear about what we are producing, for whom we are producing it, and what are our customers' key determinants of quality. This is harder than it appears, especially for public- sector agencies.

Before trying to get a producer's view of a linear measure of quality, we should look toward our customers.

Who are our customers, or at least who would we like our customers to be? What do these people want from our data? (This may be quite different from what we think they should want.)

How can we get them what they want with the aspects of quality that they value most?

What will it cost us to give them that?

What are they willing to pay us for it?

The Swedish paper talks about the four faces of quality. They defined quality in the context of four areas.

1. Statistical Entity
2. Accuracy
3. Timeliness
4. Accessibility

These aspects of quality are often conflicting and while with greater or lesser degrees of success, we may be able to measure each of them. It doesn't really make sense to me to just add the up somehow either weighted or unweighted to get a total quality measure.

Consider an example of the restaurant business. Suppose we wanted to compare two restaurants on the same quality scale. Let's begin with the four categories defined in the Swedish paper. First, we would need to make a transformation of language in order to talk about restaurants. Let us define the four areas of a quality restaurant as follows:

Statistical entity	------->	Menu
Accuracy	------->	Taste/pleasing to the palate
Timeliness	------->	Quick service
Accessibility	------->	Location

Now consider two restaurants.

Restaurant one is a Five Star restaurant with food to die for. It is a the Greenbriar resort in West Virginia.

The second restaurant is a wonderful little McDonalds restaurant just four blocks from my house.

Now let's try to measure these two restaurants on quality.

First: Menu

The Greenbriar has a vast menu and if they don't have it, they will get it for you. It is fair to say that on the menu scale, the Greenbriar is a clear 10.

The local McDonalds? Well, if you want hamburgers, chicken, or fish and french fries cooked in a tub of grease, this is the place for you. A rack of lamb with some fine wine might be a problem though.

Second: Taste/ Pleasing to the palate

Here again, the Greenbriar is at the top of the chart. If you have ever eaten there, you know what I mean.

The local McDonalds tastes good too, but in a different way and to a decidedly different clientele.

Third: Quick Service

Well, here the Greenbriar is not so good. While it is true that you can get literally anything you want at the Greenbriar, it is not so clear that you can get it quickly.

The local McDonalds, on the other hand, specializes in getting you its admittedly limited menu quickly. In fact, its name is synonymous with fast food.

Fourth: Location

Well, here there is no comparison. The Greenbriar is four hours away by train. The McDonalds--well you know the McDonalds.

Which of these two restaurants should score the highest on the summary measure of quality scale? Clearly it depends on who is voting, and what they are looking for in a restaurant experience. It obviously doesn't make any sense to try to get a numeric score for these two restaurants and then compare them somehow. They are both quality restaurants, but in a different way. The same is true of surveys.

This ambiguity does not mean that we shouldn't try to measure quality for both internal management improvement decisions and for advertising purposes.

We clearly must measure all we can and seek to improve all we can, but we should not fall prey to the fiction that somehow we can put a universal metric of quality on all surveys.

THE ENERGY PAPER

On this paper as well, many measures of quality are proposed--both internal and external. Clearly, resources available to measure quality are shrinking. Does this imply that interest in quality is shrinking or only that the measures that we are producing are not what our customers feel they need?

If it's the former, then we are in trouble. If it is the latter, we may be able to get the needed resources if we measure and improve the aspects of quality that our customers and sponsors think are important.

One statement in the energy paper struck me. The authors state that users don't seem to know very much about the quality of your surveys now and yet they are basically happy with the existing quality. I found myself wondering what would be accomplished if they succeed in answering the questions about the quality and their customers find that their happiness has been misplaced.

Maybe a better question for statisticians to attempt to answer is "Are our data good enough that they are the data of choice for most of our important users? How can we improve them for this purpose?"

All this gets us back to the three really important questions we should be asking ourselves about quality.

1. Who are our customers?
2. What do they want?
3. What are they willing to pay for it?

My final advice:

Seek to understand as much about the quality of your survey as you can.
Seek to understand your customers needs and wants.
Make sure that you are measuring improvement in the areas of quality that are most important to your customers.
And finally, don't waste your time searching for the Holy Grail.

DISCUSSION:
Reviewing and Reporting Quality in Survey Data

Richard A. Kulka
Research Triangle Institute

Obviously, there have been many papers presented in the past few years on the quality of survey data, including the preparation of extensive "quality profiles" for selected major surveys in U.S. Statistical agencies, such as the Survey of Income and Program Participation (SIPP) at the Census Bureau, the Schools and Staffing Survey (SASS) at NCES, and the Residential Energy Consumption Survey (RECS) at EIA. There is also a large volume of methodological work on various aspects or dimensions of data quality, both as addenda and footnotes to major reports and as full-blown publications in their own right. While both of the statistical agencies described in these papers have contributed substantially to this literature and tradition, these two papers and the new philosophies to approaching data quality that they describe and represent are both remarkably congruent with one another and substantially different than the current mainstream in a few major respects.

In essence, these papers might be jointly entitled "Approaches to Survey Data Quality in an Era of Declining Resources for Statistical Data Collection and in the Face of Increased Competition for the Collection and Analysis of Statistical Information." Both address the meaning of data quality in the face of dwindling resources for statistical data, and both explicitly recognize competition for those resources and the need to focus more directly on quality from a customer, client, or user's perspective, as contrasted with that of the developer or producer of statistical information and products.

In effect, both papers argue, at least implicitly, that the market for statistical information is such that we must increasingly view data quality largely from a customer's or user's perspective. That is, the ultimate standard for quality is in the eye of the user. While this perspective is in evidence to some degree throughout the U.S. statistical system, and a recent focus on customer satisfaction and reinventing government (e.g., GRPA and other legislation) essentially requires that all government agencies give at least some serious attention to this perspective, both EIA and Statistics Sweden have been specially zealous in pursuing these concepts, in both cases apparently due to significant changes in their business environments or climates (i.e., budget reductions and the need to compete with other agencies in providing their services and products). More specifically, the approaches being taken by each agency derive from a "Total Quality Management" paradigm. In one form or another, this perspective has drawn significant attention throughout the statistical establishment, and, assuming that such attention will not diminish significantly over the next few years, it is important to explore more thoroughly some of the key *implications* of this particular view of data quality. Even if one does not fully agree with this particular point of view, it is still important that we recognize some of the important implications of taking its basic tenets seriously.

First, consider two of the basic distinctions made in these two particular papers. One, already alluded to, is an emphasis on quality as defined by the customer or data user rather than a

producer-centered view of data quality. As the paper by our colleagues at Statistics Sweden notes, the vast majority of approaches we have typically taken to define and assess data quality are decidedly producer-focused rather than client-focused in their basic orientations. A second dimension highlighted in this session is an emphasis on process versus product quality, a distinction also emphasized in particular in the paper developed by Statistics Sweden.

The distinction between a user-based versus producer-based definition of statistical data quality is stated most clearly by Claes Andersson and his colleagues at Statistics Canada:

> A product's quality is determined by the user's opinion of the product and its usefulness. The user's opinion should direct the approach of the development work. . . . Quality encompasses **all characteristics of a product** (commodity or service) influencing how well the product satisfies the user's needs and expectations. . . . The **user makes the judgement** and decides if the quality of the product is good or bad in relation to his/her intended use of it.

Thus, as noted by Anderson et al., it is essential for the producer to be fully aware of the quality judgments of both existing and potential users, since these provide the very basis for efforts to improve quality .

But that is not all. Another consequence of this way of defining quality is that users must then take greater responsibility for the level of quality and for ensuring that sufficient resources are available to reach this level, a responsibility previously vested in most instances with the producers of statistical data. As a result, this point of view also entails yet another important responsibility implied by the EIA paper--a responsibility to provide the information necessary for users to make such judgments in an informed and effective manner. However, this clear and obvious need also presents a potential dilemma. As noted by Renee Miller, EIA has:

> been providing our users with explanatory material for years. Yet during our customer satisfaction survey, some have told us that they have no way of judging the quality of our data. They think that EIA has good quality data, but they say they do not know for sure. Something seems amiss here.

Indeed. And, is it not in effect a major new responsibility of producers of statistical data and products to ensure that these customers and users have adequate information to make fully informed judgments of quality in relation to their critical uses and needs?

In turn, this need for information touches on the other basic distinction alluded to earlier--process versus product quality. As our colleagues at Statistics Sweden point out, users are fundamentally most interested in the quality of the product and less so in the process leading to the product. However, process information is often the only thing producers have to offer. Thus, there is a natural tension between the information generally available and what the user ideally wants or needs. Ironically, the major new surge throughout our industry to computerize survey and related information collection processes offers considerably more and higher quality

information on process quality (of substantial utility to producers of statistical data) but very little more information on product quality (of greatest interest to users), except by inference.

There is a general faith that product quality can be and is achieved through increased process quality. Thus, we work hard on improving processes (e.g., by developing standardized procedures and checklists, by continuously measuring key process variables, etc.) to achieve product quality. That process quality will automatically result in product quality cannot always be assumed, however. For example, in her paper, Miller correctly notes that adherence to standards is not synonymous with achieving data quality, just maintaining a low-fat diet and exercising regularly does not ensure that one will be disease free. Thus, the relationship between standards and standard processes and data quality is not an exact one, although standards are most clearly quite helpful from another perspective—for establishing credibility among our users. However, as the EIA customer satisfaction data described by Miller suggest, the relationship between standards and credibility may still be a very tenuous one.

Perhaps the key implication then of this new perspective on data quality—a user or customer-based perspective—is that one must take quite seriously a fundamental responsibility to provide the information necessary for users to make well-informed judgments on the quality of our statistical services and products. And, these two papers provide several good examples of how difficult it may be in practice to shift our basic paradigms in this direction. First, consider the EIA observation that, in spite of their having provided users with detailed explanatory information (derived from a producer-based persepective) for several years, customer surveys indicated that users still felt that they had no good way of judging the quality of EIA data.

Similarly, Andersson and his colleagues highlight two additional examples from Statistics Sweden in describing their evaluation studies, which have mostly been producer rather than user-oriented. For example, they note that while statisticians undertaking such studies (producers) typically focus on the *net* effects of systematic bias, from a broader, user-based perspective *gross* effects are also of legitimate interest for a number reasons. In the same context, they note that measures of relative bias derived from evaluations can be presented either as deviations between observed and "true" values as: (1) % of "true" (unbiased) value (of interest to *producers* as deviation from goal), or (2) % of the "official" estimate (of greater interest to *users* as an indication of error in published figures).

Another example is evident in the description by Andersson et al. of The Quality Report published annually since 1988 by Statistics Sweden to provide a basis for an analysis of the development of quality in their statistical products. These reports are based on questionnaires administered to every person responsible for a statistical product at the SCB, who are asked to respond from the intended user's perspective. However, users themselves are not routinely asked to share their opinions directly.

A final example that best illustrates perhaps how far we may still need to go to fully meet the demands and responsibilities associated with a user-based orientation to data quality is derived from Miller's discussion of EIA's *Quality Profile* for the Residential Energy Consumption Survey (RECS):

The *Quality Profile* has been very well-received. Several members of our energy advisory committee thought it a good model of how we should document our surveys. They pointed out that customer satisfaction is a function of data quality and that a quality profile would give users all the information they would need to determine data quality. It was suggested that we do more profiles. Unfortunately due to our budget constarints and reduction in staff levels that does not appear likely.

In an era where we purport to be adopting and take quite seriously a true user or client-based approach to data quality, can we really afford *not* to provide such information both routinely and in considerable depth?

In closing, I wish to thank the authors from SCB and EIA for presenting these very stimulating papers. Read in combination, they are extremely informative and thought-provoking and may well serve as precisely the right type of stimulus to ensure that we clearly recognize the full potential, implications and responsibilities that accompany these new and still somewhat controversial ways of conceiving of data quality.