STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 12
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic II: Propagation of knowledge to users

PROPAGATION OF KNOWLEDGE ON STATISTICAL QUALITY IN THE CASE OF BUSINESS SURVEYS

Submitted by INSEE, France[1]

Contributed paper

I. INTRODUCTION

1. In statistics, the success of the term *quality* has been consistent for a number of years. We often use this word in publications, conferences, and corridor discussions, without truly realising all that it encompasses. As is typical of rich concepts, we find quality in practice to be polysemic, defying a simplistic, all-encompassing approach. It is therefore subject to individual interpretation. The determining factor is actually the angle from which it is viewed—user, respondent, survey manager, production manager, policy authority, before or after the survey, or in connection with a given use. There are many definitions for quality, though they vary according to the roles of each party involved in the statistics production process.

2. The question of propagating knowledge on statistical quality depends then on the viewpoint we choose. In this brief paper, we will briefly examine various aspects of quality, various users of information on quality, and different kinds of information that are disseminated to the users.

II. WHAT IS QUALITY IN THE FIELD OF BUSINESS STATISTICS?

3. To begin with, let us view the quality from the standpoint of the intended **user** of statistics. Eurostat distinguishes six components to represent quality *of statistics*:

- *Relevance*: statistics must meet users’ needs.
- *Accuracy*: the different types of errors must be estimated.
- *Timeliness and punctuality*: users demand recent data subject to frequent and timely updating.
- *Accessibility and clarity*: the data must be accessible to users, and must be readable.
- *Comparability*: statistics must be comparable in time and space.
- *Coherence*: statistics from different sources must be coherent (using the same definitions, nomenclatures, methodologies, etc.).

4. To understand the concept of quality, we can choose the standpoint of the **respondent**. Data collection is not a neutral operation for the respondent: it requires time and energy; it can require the efforts of several persons, as well as a number of information sources. Therefore, for the user, the lower the burden, the better the survey. More generally, the surveyed unit's perception of the survey is an essential component of the overall process of statistical

[1] Prepared by Pascal Rivière.

production, as individual data constitute our basic input. Such data result from the interaction between the statistics service and the surveyed unit—which can be complex. The quality of this interaction, as perceived by the respondent, must therefore be envisaged in itself.

5. If we view quality from the **production** standpoint, then it is a matter of survey implementation. Despite considerable progress in information technology, conducting a survey is still a complicated, costly undertaking that entails numerous problems. For optimal operation, the survey manager must have access to a series of indicators to determine the status of the task at all times, and to indicate the reliability level of the results, near the end of the survey. The production manager therefore intervenes upstream from conducting the survey (questionnaire, sampling plan), as well as during (management indicators) and after (results reliability analysis) the survey.

6. Data editing is a major part of the quality of this process. One of the main questions it raises is to determine whether or not the questionnaires have been "overedited", because it will impact survey cost and timeliness. The compromise between the cost of checking questionnaires and the cost of unreliable results can be more effectively managed using editing indicators. Of course, more editing increases the final quality of the results; by contrast, it is largely pointless from the statistical perspective to edit many questionnaires. Effective editing and imputation therefore requires a balance between high-quality output and limited manual checking. There is a noteworthy difference between the quality of *statistics* and the quality of the *production process* from the standpoint of the survey manager, for whom a reliable manual editing rate will be a measurement of quality (this, inter alia, improves dissemination lags), while this information is not necessarily useful to judge the reliability of the resulting statistics.

7. If we think of the quality of a set of surveys, we can take the standpoint of **statistics coordination**. In France, it is the national statistical information council (CNIS) that decides, for each survey, whether it will be done or not. Is the survey of interest to the public, to what users, and why? Is the set of tools (questionnaire, sampling design, etc.) sufficient to reach these objectives? These questions can only be answered through meetings of statisticians and professional organisations having some connection to the topic of the survey. If this is a new survey, we must bear in mind that when the decision is taken, the survey has not been conducted, and thus that its «a priori survey quality» must be judged on the basis of only a few factors.

8. Obviously there are other views of «quality» in statistics: **quality of the business register** is a main one. Many papers were written recently in INSEE (especially for the next roundtable on business survey frames), and quality indicators are being defined.

III. WHO ARE THE USERS OF INFORMATION ON STATISTICAL QUALITY?

9. In France, the National Council of Statistical Information (CNIS) is one of the first users of information on statistical quality, as it has to decide, for each survey, whether it can be launched or not. This decision is based on some aspects of quality, for example the relevance (of the survey goals) or the good coordination with other sources of information.

10. Statisticians who manage business surveys have to use information on statistical quality, but for other purposes, knowing the quality of a survey will help them to improve the next one. Generally, many aspects have to be improved and as it is impossible to optimize everything, the main issue is to define a hierarchy between the numerous ways of improvement.

11. In the national statistical institutes (NSIs), people responsible for dissemination of statistics are not interested in the survey process, but they need to have some elements on the quality of the statistics. What is the precision? At what level is it possible to disseminate the results? What are the «poor» variables, that will be difficult to interpret?

12. The researchers within the NSI use the same kind of information on quality, but they need to have something more precise, especially all kinds of metadata. For example, it is important to know if a variable value has been collected or is derived from an imputation.

13. Survey clerks have to be able to use quality indicators on data editing, because it will help them to understand the current situation of the data editing process and to define priorities in their work. Therefore, the indicators have to be as simple as possible, in order to be clearly understood.

14. For the same reason, register clerks will have to use information on quality of the business register, and this information has to be clear and readable.

15. As we can see, many users of the information on statistical data quality belong to the NSI: it gives them a deep understanding of the survey (or the register) which allows them to improve the quality. Then, more generally speaking, if we focus on overall, long-term progress with a view to quality improvement, the techniques applied must be coordinated, standardised, and documented. This is essential in referring to quality. The quality of the work involved may be considered in reference to recognised, documented standards on the final quality of the statistical product. In the absence of a common standard or the possibility of external assessment, individuals may each define quality criteria in their own way (and may each award themselves accolades). That is why a controlled propagation of the knowledge on quality is so important.

IV. WHAT KIND OF INFORMATION IS AVAILABLE FOR BUSINESS SURVEYS?

16. First of all, we can distinguish between general information and survey-specific information. By general information, we mean: guidelines on quality and common methodological documentation.

17. There are some guidelines on quality, particularly the «Business survey charter», in which INSEE undertakes to make surveys that follow a certain number of principles, like accuracy or confidentiality for example.

18. INSEE has a large methodological documentation on the various steps of the survey process, especially in the case of annual business surveys. These methodological materials are about: microediting, partial non response imputation, total non response imputation, macroediting, calculation of main industry, calculation of changes in aggregates, variance calculation, etc. As they have to be used in practice by survey managers, these internal documents do not only describe the process in general: they have to give very precise information, for each function, each database, each variable.

19. The information available for each survey is large, larger than it was for business surveys even 5 years ago. In principle, we must have:

- A «label committee» report: to be launched officially a survey has to be given a label, by the «*Comité du label*». To obtain that label the survey manager has to write a standardised report, in which one must find: the description of the sampling design, the questionnaire, the letter that will be sent to the units, a report of questionnaire testing, and a report of all user's committee meetings. The «label committee» standard report was written in 1997. Since then all the new surveys of the French statistical system have to write a report (following the structure of the standard report) for the label committee, and even the regular surveys, every five years.

- A «statistical data dictionary»: it includes all questions of the questionnaire, all variables of the database (names, definitions, sets of possible values, links with the questionnaire), all kinds of metadata. INSEE has developed a general software named DDS that allows the user to define interactively that kind of dictionary, and to generate automatically an HTML documentation. DDS is available on PC since 1995, and is now used in INSEE for a majority of surveys. The statistical data dictionaries are built for production purposes generally before the survey.

- A detailed description of the results of the survey: when the final database of the survey is achieved, the survey manager has to write a document explaining the contents of the database, so as to make it usable by statisticians, for example statisticians belonging to INSEE's regional offices. In that document (that is sent to all INSEE's departments and regional offices), one can find record layouts, variable definitions, but also, in some cases, information on the quality of the survey, like low quality variables or response rates.

- In the future, a standard quality report: in the case of business surveys, that standard quality report exists (see VI. Example) and has been validated by INSEE. In the long run, the aim is to use that standard report for every business survey. The users will be: CNIS members, survey managers, statisticians, people responsible for dissemination. A part of this document could be used by final users of statistics.

V. HOW DO WE DISSEMINATE AND REUSE THIS INFORMATION?

20. A growing part of the methodological documentation (general or survey-specific) is now available in the intranet site of INSEE's business statistics directorate. The users are statisticians. Part of the survey-specific documentation comes from the «statistical data dictionary» of the survey: it is simplified as the DDS software automatically generates a documentation in a HTML format.

21. The CNIS Internet site (www.cnis.fr) gives summarized information on the characteristics of every survey. It is a part of the information coming from the « label committee » reports (which are not disseminated). CNIS members use it, surveyed units could use it.

22. Methodological and technical meetings are an excellent way of propagating information on quality to the statisticians. It is essential to notice that e-mailing technical papers is never enough to convince and to make people understand: it has to be combined with meetings and oral presentations. In the case of annual business surveys, the coordination committee organizes 5 or 6 meetings per year. The main topics of these meetings are IT issues, methodology, quality, and coordination.

23. Training on business survey methodology in general (this course has been created in 2000) is another way of propagating information on quality: it actually gives all statisticians the same background, a common culture, that simplifies discussions and eases understanding.

24. The business survey charter is systematically given to the enterprises in case of face-to-face interview.

25. The way INSEE will disseminate the standard quality report is not defined today. It will probably be done during the year 2001.

VI. EXAMPLE: THE STANDARD QUALITY REPORT

26. The standard quality report includes three sections: a quality sheet, comments on the quality sheet, and a description of the methods used in the survey.

27. The quality sheet contains all necessary quality indicators on different topics. It must not contain any mention of specific cases, details, or jargon. The sheet must be simple, concise, and readable. It is provided in the section below.

28. The comments function as footnotes to the general terms used in the sheets. The comments are intended to explain the content of the checklist, which is deliberately designed to be in summary form. The "comments" section must also provide an opportunity for self-assessment comments (for example, that a given processing stage was time-consuming, and that the delays might have been reduced through a given action).

29. The description of methods is necessary because survey processing involves a host of activities, that entail a variety of methodologies, organisations, and tools. The objective, for each individual operation, is to determine those applicable: choice of statistics and collection units, sampling method, collection method, keying scheme, data storage method, microediting, macroediting, manual editing criteria. A few lines is sufficient for each of these items, not all of which will be relevant to each survey.

Quality Sheet

Survey

Identification sheet	
Name	
Year	Periodicity
Surveying service	
Identification code	

Description	
Definition of scope	
Size of population	Size of sample
Median response time (estimated)	
Total induced burden (estimated)	

Key dates

First launch meeting
CNIS group meeting
Label committee meeting
Launch of collection
Completion of manager’s work
Final database
Availability of register
First publication of provisional results
First publication of definitive results

A priori quality

Preparation		
Is there a user committee?	Yes □	No □
Was the questionnaire tested?	Yes □	No □
Is the sample co-ordinated?	Yes □	No □
Is the survey part of a larger system of surveys?	Yes □	No □
Was an assessment conducted on co-ordination with administrative sources?	Yes □	No □
Is there a statistical data dictionary?	Yes □	No □

Organisation of production		
Feedback to enterprises	Yes □	No □
Feedback to the register (e.g. out-of-scope items)	Yes □	No □
Provision for reminders?	Yes □	No □
Is there a preliminary letter or cover letter accompanying the questionnaire?	Yes □	No □

Database tools: fundamental features

Are raw figures kept?	Yes □	No □
Is there a code indicating the type of imputation used?	Yes □	No □
Is there an “omissions” modality?	Yes □	No □
Are general or internal tools with documented methodology used for:	Yes □	No □
➲ Sampling	Yes □	No □
▶ If yes ▶▶ Is the sample used as is?	Yes □	No □
▶▶ Is the sample subsequently adjusted ?	Yes □	No □
➲ Edits (microedits or macroedits)?	Yes □	No □
➲ Imputation or reweighting?	Yes □	No □
➲ Survey management?	Yes □	No □

Sampling			
Type of sampling e.g.: exhaustive, with a coverage rate, SSRS , pps, etc.			
Sampling strata used			
Target variables for dissemination ▶ Number ▶ List			
Dissemination strata ▶ Number ▶ List			
Part of population of take-all stratum in total population			%
Average sampling rate (not including take-all stratum)			
Coefficients of variation by target variable and dissemination stratum Distribution of coefficients of variation in dissemination strata			
List of target variables Target variable 1 Target variable 2 Target variable 3, etc.	Maximum	Median	Ninth decile
NB :These composite indicators will be provided when they can be calculated.			

Production-related indicators

Proportion of units in the scope		%
Part of units that were not present at the indicated address		%
Does documentation exist to help the survey clerk make modifications?		
➲ Little to none		□
➲ Only on some variables		□
➲ Complete		□
Proportion of units deemed acceptable by editing, among respondents		%
Effect of edits: checks and modifications of target variables		
List of target variables	Proportion of manually-checked units	Of the units checked, proportion of units modified in respect of the raw value
Target variable 1		
Target variable 2		
Target variable 3, etc.		

Accuracy indicators

Response rate among in-scope units			
Overall response	Distribution of response rates in dissemination strata		
Rate	Maximum	Median	Ninth decile

A posteriori features
Size of sample (a posteriori)
Average sampling rate (a posteriori)

Nonresponse			
Relative weight of total nonresponse imputations, per target variable			
List of target variables	Distribution of weights in dissemination strata		
	Maximum	Median	Ninth decile
Target variable 1			
Target variable 2			
Target variable 3, etc.			
NB : This table will be completed if responses are processed by imputation.			
Proportion of partial nonresponses among respondents			
Coefficients of variation reflecting sampling and nonresponse			
List of target variables	Distribution of coefficients of variation in the dissemination strata		
	Maximum	Median	Ninth decile
Target variable 1			
Target variable 2			
Target variable 3, etc.			
Nonresponse bias in the subpopulation of nonrespondents			
List of target variables	Distribution of biases in the dissemination strata		
	Maximum	Median	Ninth decile
Target variable 1			

Target variable 2 Target variable 3, etc.

Dissemination

Was an availability notice available?	Yes □ no □
Does dissemination documentation exist?	
➲ No, no documentation	□
➲ Yes, simple documentation	□
➲ Yes, a complete data dictionary	□
Number of publications, by type	
➲ "I*nfos Rapides*"	
➲ "*4 Pages"* (or *Insee-Première,* or equivalent)*,*	
➲ CD ROM or "*Insee-Résultats*"	
➲ Articles, studies	
Clarity of publications: rate on a scale of 1-5	

Comparison with other sources

Was comparison attempted?	Yes □	No □
Is comparability provided with:		
➲ Identifiers	Yes □	No □
➲ Definitions of variables	Yes □	No □
➲ Scope	Yes □	No □