CBM 2004:2

Statistiska centralbyrån Statistics Sweden

Design your questions right

How to develop, test, evaluate and improve questionnaires

June 2001
September 2004

Statistiska centralbyrån
Statistics Sweden

Design your questions right

How to develop, test, evaluate and improve questionnaires

June 2001
September 2004

Statistiska centralbyrån
2004

Reference: Birgit Henningsson, Statistics Sweden, Research and Development.
Telephone +46 19 17 63 21, birgit.henningsson@scb.se

Printed in Sweden
SCB-Tryck, Örebro 2004.11 MILJÖMÄRKT Trycksak 341590

CONTENTS **Page**

Preface from the Director General

Design your questions right. How to develop, test, evaluate and improve questionnaires has been produced within the framework of quality improvement work at Statistics Sweden. It reviews in manual format the available and recommended methods for measurement work regarding questionnaire and question design. By investing qualified work in the development of the questionnaire, the foundation is laid for producing statistics of good quality. The manual can be used by anyone responsible for the collection of data, whether working for a national statistical institutes (NSI), a government authority, a non-government organisation (NGO) or a private enterprise.

Svante Öberg

Preface from the authors

This manual outlines various methods for developing questionnaires, and instructions for statistical surveys. It describes a systematic process from the definition of variables to the statement of quality of the survey results. By adhering to a standardised procedure, the development of the questionnaire will be more rational and effective.

The central part relates to methods for testing preliminary versions of a questionnaire under realistic conditions, so that shortcomings can be identified and corrected. This is an aspect that many surveys neglect and for which they pay a high price in the form of lower quality and costlier production.

The presentation is focused on the most common survey methods used to collect data from establishments, municipalities, organisations, individuals and households. Many of these surveys are periodical, whilst others are only carried out once or intermittently.

The primary users of the manual are those persons who develop their own questionnaires and who work with the collection of data. But users of statistics can make good use of the manual as well, when judging whether the material they are analysing has been collected with adequate measurement quality. The manual helps them to check how well the work in each phase of the questionnaire development has been carried out.

The manual is not supposed to be read from beginning to end, but to be used as a reference book. The appropriate chapter can be read to provide ideas and tips. It mirrors Statistics Sweden's understanding of what are currently the best methods for measurement work according to our set of Current Best Methods (CBM). A schematic picture of the measurement work is shown on the inside of the back cover.

This is the first volume regarding measurement. We are grateful for all comments and expect to update and produce a new version in a couple of years.

The main person responsible for the work with the Handbook is **Håkan L Lindström**. The following persons have contributed to different sections: **Gunilla Davidsson, Birgit Henningsson, Anette Björnram** and **Helén Marklund**. In addition, a reading group has helped with comments. The English version of the manual has been finalized in September 2004 by **Birgit Henningsson**, **Chris Denell** and **Sara Hoff**.

1 Why focus on measurement?

1.1 Primary objectives

Measurement development work aims to improve a survey's content, collection methods and questionnaire so that the respondent, with a minimum of effort and with sufficient accuracy, can provide the information that is required for the statistics. The work can also help to formulate the quality declaration of the survey results.

The measurement expert needs to

- have comprehensive practical experience and theoretical knowledge regarding the identification and definition of survey variables
- be able to choose collection methods adapted to the survey situation
- be able to develop questions, response alternatives and instructions by selecting relevant qualitative and quantitative test methods
- be able to measure and judge the accuracy and relevance of the information that is later collected, i.e. to judge whether the questionnaire has worked in practice.

The **immediate objectives** for measurement work are to

- give the survey a **well-defined and measurable content** by becoming familiar with the respondent's circumstances and capacity to answer
- through expert knowledge of the subject-matter area and cognitive and qualitative methods, develop questions, response alternatives and questionnaires adapted to the data collection methods and respondents so that there is the **least possible response bias and response variation**
- through observations, qualitative studies, experiments and evaluations, **produce a standard** or at least an indication **of the measurement uncertainty of the results.**

1.2 Positive side effects

A systematic development and testing of questionnaires and collection methods can make the collection and processing of data simpler and more effective. Some important effects for both the respondents, the survey costs and the production time are:

The **burden on the respondent** is reduced. With a good questionnaire, it takes less time and thought for the respondent to find the right place in the questionnaire, to read the instructions and to answer. By reducing the burden on the respondent, both **unit nonresponse** and **item nonresponse** can be reduced.

Dillman gives an example of a dense and jumbled questionnaire that reduced the number of responses by 3-4 per cent. Akkerboom (1997) reports an extreme case in which the number of responses increased from around 50 to 90 per cent after the testing and revision of a questionnaire.

The need for checking is reduced when the respondent is able to answer in a way that is convenient to him/her.

The volume of checking indicates how large a problem it was to submit the information and also shows to what extent it can be worthwhile to revise the questionnaire and instructions. In general, it is considered that editing requires 30-40 % of the total resources for establishment surveys and 15-20 % for individual surveys (Granquist 1999). A study on Swedish Manufacturing Statistics in 1990 states that roughly 13 % of all information was corrected and that 65 % of all responding establishments had at least one correction (Hedlin 1992). In many cases a direct link between deficiencies in the questionnaire and incorrect information could be shown. For one variable (number of male and female owners), 70 % of the data needed correcting because the instructions were in a footnote, which the respondents didn't notice. For the variable "employed in other activities", a misleading layout led to 50 % of the data needed correction.

The need for **follow-ups by telephone** is lessened if misleading questions, which lead to inconsistent answers and answers of the wrong size order, can be avoided.

1.3 What happens if questionnaire development is neglected?

Measurement errors occur when questions and questionnaires are designed in such a way that the respondent does not understand them correctly, when it is not possible to get the requested answer with good accuracy and when the burden on the respondent is high. This happens often because the producers have not seen the weaknesses and thus have not invested enough resources on testing the questionnaire systematically. But they can also have chosen characteristics that are difficult to measure despite the inaccuracy and restrictions they know this will lead to.

The instructions for quality declarations in Official Statistics of Sweden (SOS) summarise the reasons for measurement errors in the following way:

> **Measurement/observation**
> Sometimes a given record does not agree with the "true" value according to the definition of the variable. There are many reasons for this, e.g that the question does not agree with the respondent's bookkeeping routines, the question is ambiguously formulated, the person has a faulty memory, the respondent is careless (or worse, knowingly misleading), or that the physical measurement methods are marred by deficiencies. In general, it is accepted that measurement errors occur and contribute to the inaccuracy of statistics. They can do this in a systematic way (resulting in distortion) as well as in an ad-hoc way (does not lead to distortion but to increases in inaccuracy).

An empirical study of Christianson and Polfeldt (1996) identifies a number of reasons for measurement errors. Without claiming to be comprehensive, it gives an overview of error risks to be avoided when designing questionnaires. The study classifies measurement errors by cause for a total of 206 selected variables, distributed on 135 products at Statistics Sweden. The reporting is more often based on subjective judgements and indications than on evaluation studies. Below are listed in order of frequency the most common types of errors according to the review carried out in the study:

A	Definition problems. The definition of the statistical terms are not known or are not used by the respondent.
B	Memory errors. Most common in individual or household surveys but can also occur in establishment surveys.
C	Respondents must make estimates. The exact information is not always available or not for the right period.
D	Time period problems. The information relates to another reference period or point in time than that requested, e.g. if an enterprise has a different financial year.
E	Need for calculations. Data for the variable does not exist so the answer must be calculated from other data.
F	Correlated variable used, e.g. when production data is requested but not available, the respondent answers with data on deliveries..
G	Accounting problems. The enterprise does not have the information distributed by the categories or object types requested for the statistics.
H	Classification errors in the background variables, because they are difficult to measure.
I	True value is missing.
J	Inclusion of incorrect components in a variable total. Respondent wrongly adds components that were not requested.
K	Errors not noticed on registration
L	Exclusion of components in a total. Respondent neglects to include one or more components that were requested.

In economic surveys, in particular, the following errors also occur:

- Unit error, when the respondent does not observe in which unit the answer should be given, e.g. he/she reports in SEK thousands, instead of in SEK millions.
- The respondent cannot produce the requested quantitative data and chooses to leave the answer column blank, marks it with a dash or maybe writes 0.
- The wrong exclusion of a component in a variable result and, at the same time, the wrong inclusion of that component in another variable – often, in the "other" or "remaining" item.

The types of errors that appear in a specific survey depend on factors such as subject, level of difficulty, collection method and the motivation of the respondent. These cannot be foreseen solely with help of subject knowledge and measurement experience. The questionnaire should also be tested on actual respondents before it is put into production.

1.4 Development trends

During the 1980s, measurement studies often focused on assessing the effects of measurement methods at user level rather than on identifying the causes of measurement error. A wide range of methods and tools for studying and analysing how the respondent answers questions in a questionnaire is currently in use. Through rather limited testing of a preliminary version of a questionnaire, many causes of errors can be eliminated before the main survey begins. The testing of questionnaires addressed to individuals has been furthest developed. But more experience is also being gathered on testing questionnaires for surveys on establishments.

The area of measurement methods is fast developing. More reliable criteria are being developed for how to choose the best methods, and principles are being formulated for how to use several methods at the same time. In order to successively improve multi-round surveys, systematic approaches are constructed to produce indicators for the occurrence of measurement errors from the production process. Furthermore, as a growing number of collections are computer-assisted, the use of IT support in the collection and measurement processes will rise.

Today the majority of statistical offices (including all the Nordic ones) have a specialist group for the development and testing of questionnaires, usually called e.g. *Questionnaire Design Resource Centre* (QDRC). The group at Statistics Sweden is called the Measurement Laboratory (ML).

References

1. Akkerboom, H. and Dehue, F. (1997) *Examples of whole package tests*. Proceedings of the Workshop on Minimum Standards in Questionnaire Testing held in Örebro 19 - 21 October 1997.
2. Christianson, A. and Polfeldt, T. (1996). *Response quality improvement initiatives at Statistics Sweden*. Proceedings of the 2nd International Conference on Methodological Issues in Official Statistics. Stockholm, September 23-24 1996.
3. Dillman, D.A. *Progress in the design of respondent-friendly self-administered questionnaires*.
4. Granquist, L. (1999) *On improving quality by modern editing*. Report from Statistics Sweden.
5. Hedlin, D. (1992). *Comparison of checked and unchecked data in industry statistics*. Report from Statistics Sweden.

2 Data formation

2.1. Central concepts

Measurement concepts frequently carry various meanings and are often used without being defined. For the sake of clarity, we outline below a number of basic concepts used in this manual.

The **survey unit** is a person, municipality, school, agricultural establishment, etc. about which data is required. The person who answers the questions in an interview or who fills in a questionnaire is called **respondent**. (Other common terms are informant, data provider and responding person).

In a **survey on individuals** a person gives data on him/herself. Surveys that are directed towards enterprises, government authorities, organisations, associations, etc. are here covered by the term **establishment surveys**. These differ from surveys on individuals mainly because there is not one unique respondent, and the "most well-informed respondent" must first be identified among many possible respondents. In **household surveys**, the respondent should be able to answer for e.g. the entire household's income and expenses. In these surveys, there might be some difficulty in deciding who is the best respondent.

We use the concept **question** for the text, which specifies which information should be given and the response alternatives that are offered. The concept is used both when the questionnaire includes actual questions and when a request for information is put in the form "Number of employees at year-end", or "Turnover including VAT during third quarter 2000".

For **fact variables**, which measure, for example, acreage of arable land, delivered quantity, time taken, age, and level of education, there is a **true value.** The variable must be defined precisely by explaining the concept, giving the reference period to which the question applies, and giving the quantity in which it should be measured. With enough effort, a person other than the respondent should come up with the same answer. For an **attitude variable** there is not a true value in the same way - not even with the corresponding definitions. However, it is possible to imagine the existence of an **operative true value.** The question, instructions and response alternatives must then be so clearly formulated that the respondent, at least on average, can come up with the same answer in a hypothetical series of repetitions. Of course, the more complex and hypothetical the question, the harder it is to imagine that there is an operative true value.

For variables where it is hard to imagine a factual or operative true value, it is not possible to decide if one question formulation is better than another or if an answer is more or less correct. For estimates based on such variables, concepts such as confidence level and bias lack actual meaning. A statement on accuracy can no longer be based on sampling theory but must find support in another type of model formulated from another starting point.

We use **questionnaire** as a general term and avoid other common terms such as (question) form, survey and measurement instrument. **The development of a questionnaire** involves formulating instructions, questions and response alternatives, designing a suitable layout for the questionnaire and producing information materials.

With **respondent burden**, we primarily mean the time taken and effort required for the respondent to understand the questions, get hold of the information, fill in the questionnaire and send it in. With incorrectly filled-in questionnaires, the burden on the respondent increases, since the producer of the statistics must get in contact with him/her again for correction and complementary information. When using electronic questionnaires, editing provisions are often included, which means that corrections are imposed. Sensitive questions can also add to the burden on the respondent.

This Handbook primarily deals with the development of questionnaires with standardised questions and fixed (structured) response alternatives - mainly for the production of statistics. The same principles apply, however, to questionnaires for administrative use.

Standardised questions are questions that are formulated in the same way for all respondents. Different questions can be posed to different respondents depending on what sub-group they belong to. The answers to selected questions with accompanying skipping instructions means that all respondents do not have to answer all questions.

In a question with **fixed response alternatives,** the respondent should fill in quantitative data of a given unit in a specific place, put an X or circle around one or several given alternatives, or choose one or several response alternatives given by the interviewer.

With an **open response alternative**, the respondent him/herself must formulate his/her answer. Costs and time to code and register the answers is a deterrent to this alternative. The open response alternative is therefore usually only used for occasional questions and to give the respondent the opportunity to add some final summarising remarks on the survey.

Measurement test methods use, to a large extent, **qualitative methods** (methods which in a detailed way illustrate the data collection method and the questionnaire). Both standardised and non-standardised probes with open answers are used to understand how the respondent interprets the questions and arrives at an answer.

2.2 The formation of data

When choosing the measurement method, constructing the questionnaire or determining how exact a record needs to be, it is important to consider how the respondent can get hold of, process and give the requested information - i.e. how the separate variable values/basic data will be produced. What demands the question makes on the respondent depend on

- which **type of scale** the variable is to be measured in
- how the **basic data formation** is to be carried out
- which **processes** the respondent is required to carry out.

Four types of scales
Type of scale, or **measurement level,** is characterised by how a variable can be measured.

Scale/Level	**Property of scale** **Measurement value can be:**
Ratio scale	Distinguished,distinctive ranked, measured with constant units of measurement and has a zero point
Interval scale	Distinguished, ranked, measured with constant units of measurement
Ordinal scale	Distinguished and ranked
Nominal scale	Distinguished

[The choice of scale is also linked to the processing and type of analysis to be carried out. This discussion lies outside the framework of this manual.]

Giving information on the **ratio scale and interval scale** levels requires the least amount of interpretation for the respondent, because the unit of measurement and the scale stages are well defined. Information of this type can be given on, for example, fuel consumption, temperature, insurance costs. On the other hand, it can require a great deal of preparatory work before the requested information is found or calculated.

Questions for information on the **ordinal scale** level require that the respondent is able to interpret a verbal description of the ends of the scale and decide how large each scale stage (response alternative) should be. For example, the respondents must be able to give a meaning to *Very optimistic, Fairly optimistic, Uncertain, Fairly pessimistic, Very pessimistic* as answers to a question on how they see the future prospects for their establishment or personal finances.

Questions for information on the **nominal scale** level require that the respondent can distinguish between different alternatives. Sometimes the choices are relatively direct and simple, e.g. when reporting town of birth, level of education or civil status. In other cases, it requires calculations and appraisals, e.g. to decide to which socio-economic group a household belongs, which labour force status a person has, or to which industrial sector an establishment belongs.

If it is assumed that the majority of respondents cannot, do not have the time or the inclination to give information on the (theoretically) highest level, it is necessary to consider how variables could best be measured on a lower scale level. For example, "number of trips with local transport" or "number of transports of one type of goods" can be measured in alternative ways over a period. The respondent can:

1. Give the exact number (ratio scale)

2. Mark a size category, e.g. one of (ordinal scale)
   0, 1-5, 6-10, more than 10

3. Choose between verbal descriptions, e.g. (ordinal scale)
   daily
   a few times per week
   a few times each month
   less often
   never

4. Detail that the activity (nominal scale)
   occurs
   does not occur

Three categories of data formation
Three main types of direct data formation can be distinguished, according to how the respondent produces his answers to the questions in the questionnaire:

A. The respondent already has the data recorded and "only" needs to copy it to the form. The respondent is expected to make a correct **copy or a photocopy** - manually or electronically. This can, for example, be information from an establishment's annual report or staff administrative system, costs for a private insurance policy, or information on a person's education.

B. The survey requires that the respondent can do a **measurement or observation and record some information which otherwise would not have been recorded**. For example, to read off a petrol gauge, electricity or water meter the last day of each month or to register travel, travel habits, time use or household costs during a period.

C. The questionnaire asks for **private information, which concerns only the respondent**, e.g. knowledge, plans, memories, attitudes or judgements. In this case, unlike A and B, the respondent can **answer immediately**.

Questions in data formation categories A and B usually concern quantitative information in interval or ratio scales (e.g. prices, costs, volumes, number of employees), and classifications (e.g. level of education, area of activity, type of property). In practice, they can only be used in mail questionnaires or corresponding electronic collections, as the answers cannot be given straight away but must be searched for and identified by the respondent. The great part of data formation in multi-round establishment surveys and even in household surveys falls into categories A or B. Any person who has the requisite authority and knowledge and who has been given sufficient and understandable instructions can give the information.

Questions in category C concern mainly qualitative data on a nominal or ordinal scale. These questions can often only be answered accurately if the respondent is the individual in the sample. If the answers are not to be influenced by other persons, surveys covering this kind of questions have to be interview surveys. Single-round surveys to establishments sometimes cover data formation category C as well, for example questions on recruitment plans, assessments on the economy and expectations. In these cases, mail surveys would probably be more accurate, since the respondent can confirm his/her answers with the relevant person/office.

Processing of own data by the respondent
The producer of the statistics often requests that the respondent not only gives transcripts or direct observations but also carries out appraisals or processes the information. If the processing is to be carried out in the same way by all the respondents, it is necessary to provide precise instructions on the questionnaire of how it should be done. Different types of processing may be required.

1. In the simplest case, the respondent is asked to calculate totals, form ratios and carry out other mathematical or logical actions on available data. Sometimes both the elementary data and the processing results are to be given in the questionnaire, sometimes only the results. (When the producer of the

statistics carries out the processing, the calculated characteristics are called "derived variables".)

2. In more demanding cases, the respondent must him/herself collect information at his/her place of work and carry out the calculations before filling in the information requested in the questionnaire.

3. An "immediate answer" can also be more demanding and require that the respondent can quickly add together and process recollections, estimations, knowledge, etc. "in his/her head". To ensure that all respondents think in the same way, every effort should be made to write clear questions and instructions so that all the respondents have the same frame of reference.

4. In surveys with intelligent electronic questionnaires, the respondent is requested to correct him/herself if the answers are assessed as incorrect. This requires a re-evaluation of answers already given.

Theory and reality

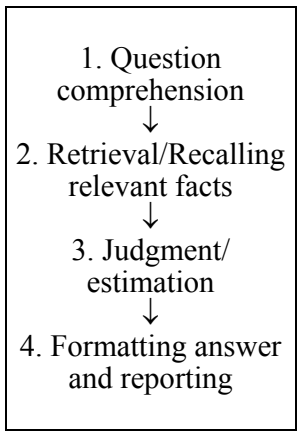
Even if the data is formed in a specific way, this does not necessarily mean that the respondent acts as expected. Even when there is registered information to copy, it can happen that approximate "immediate answers" are given instead. This can happen when the respondent is rushed, does not feel he/she needs to read the instructions or is not sufficiently involved in the field to be able to answer correctly. How the data is actually formed can be studied with the help of specific test methods. An early study within price statistics showed that over 50 % of the establishments gave prices for sales on the Swedish market instead of invoiced prices, as was requested.

2.3 Cognitive model for the response process

A cognitive model is a theory on how a respondent takes in and understands the instructions and questions in a questionnaire, collects the information, appraises it, adapts the answer to the given alternatives and submits it. One cognitive model often referred to consists of the following four stages from when the respondent gets a question to when he/she gives the information:

1. **To understand the words and concepts in the questionnaire and the meaning of the task**. Those who put together a questionnaire like to think that respondents will understand immediately, use the same terminology and give the same meaning to the words and concepts as they do. Experience from questionnaire testing has shown that this is far from always the case, even when the respondent is acting "professionally", i.e. providing information within his occupational role.

2. **To get the information from memory, notes, accounting records, etc.** Information is not always available or structured in the way the question requires. In his/her search for information, the respondent often makes use of his/her own reference framework, which does not necessarily agree with that of the survey.



3. **To process the information and assess what is actually requested.** The information found does not always agree exactly with what is requested but must be redefined, complemented and sorted according to the demands of the question in terms of detail and accuracy.

4. **To formulate the answer according to the questionnaire's formulation and match it to the right framework or response alternative.** In simple cases, this means to report in the right measurement unit and for the right period. In more demanding cases, it can mean deciding what the ends and scale stages are in an ordinal scale, and determining which alternative is correct.

The cognitive model indicates possible difficulties in a questionnaire. Perhaps the respondent does not understand the question, the question puts too great demands on the respondent's memory or the answer demands too much calculation and assessment. It also gives suggestions about how a cognitive questionnaire test could be designed to show the level of difficulty and the reason for the difficulty. In Chapter 6, Phase 3 Cognitive tests, we describe how this is carried out, and how different test tools can be used to point out risks for random and non-random errors.

2.4 Measurement error model for an effective resource allocation

Lacking an understanding of the benefits of systematic questionnaire development, many buyers and users of statistical surveys think that they save money if the questionnaire testing phase is passed over. In doing so, they may underestimate the size of measurement errors that can be caused by faulty questionnaires. Some fairly simple modelling and calculation, supported by empirical studies, may help to convince them that even a slight reallocation of the provided resources from sample size to questionnaire improvement might be of great advantage.

Model reasoning is technical. It is important because it shows how the accuracy of the estimates has an impact at the highest level. It argues that effort is placed on reducing random or non-random errors. The SCB model is shown in the appendix.

3 Phases of measurement - a summary

The development of data collection methods and questionnaires, together with the monitoring of quality, is described in this manual as one process consisting of seven consecutive sub-processes or phases. This is based on the idea that cooperation between subject matter, production and measurement experts are necessary. The classification into phases makes it easier to identify and manage problems in the right order and consequently to avoid extra work and unnecessary costs. The procedure is applicable to questionnaires for:

- all types of respondents
- both single-round and multi-round surveys
- all collection and measurement methods.

Phase 1 Define the survey
Task: To define the survey's variable content, reference period, population and the statistical parameters to be estimated. To decide the data collection method. The client is the responsible person and must make the most important decisions.

Output: A variable list, a tabulation plan and/or an analysis plan.

Phase 2 Questionnaire design
Task: To transfer the variable list into a questionnaire, question by question. The questionnaire should be suited to the data collection method. To formulate the definitions and draft the instructions.

Output: A questionnaire, which is linguistically, logically and technically correct. The layout should be clear and presentable. The questionnaire has not (yet) been tested on actual respondents.

Phase 3 Cognitive tests
Task: To ascertain whether the respondent will understand the words, phrases and concepts used. Whether the respondent can get hold of the information, how he/she will come up with an answer. Whether the respondent thinks that certain questions are sensitive and too private. Even if the questionnaire designer is satisfied with the product, the respondent may not be. Cognitive, generally qualitative tests will show this.

Output: A revised version of the questionnaire, suited to the respondent's ability to understand the questions and instructions, and to his/her capacity and will to answer. The questionnaire should perhaps be tested one more time?

Phase 4 Experimentation
Task: To design and implement an experiment giving quantitative results in order to decide whether it is possible to carry out the survey with acceptable accuracy. Even a well-formulated questionnaire will perhaps not work in a main survey. For example, the burden on the respondent could be too large.

Output: A decision on whether a method is sufficiently good or which method is best according to some quantitative criteria. The survey should perhaps not be implemented if the best alternative is not sufficiently good.

Phase 5 Adjustment for production
Task: To make the further adjustments needed to finalise the test questionnaire for production. For example, a questionnaire can have been tested in a paper version, although the survey is going to be carried out by CATI (Computer-Assisted Telephone Interviews). The preparations for scanning are perhaps not ready. A common problem is that a questionnaire is formulated for one data collection method, although several methods are used in production.

Output: A questionnaire that works both for the respondent and the producer. The reliability can be tested.

Phase 6 Evaluation
Task: To calculate and estimate error indicators. Use these measures to identify sources of errors and to remove them the next time the survey is implemented. The work is carried out partly during the actual data collection, and partly after the collection is completed. The producer has the overview of the work in this phase.

Output: A number of quality measures at different levels of information. For multi-round surveys, also a number of useful indicators for different types of error sources and tools for eliminating or reducing the errors.

Phase 7 Quality declaration
Task: During Phases 1 - 6, a number of observations and indicators on the quality of the data have been produced. Many are localised to the producer and the process. These must be presented in such a way that the information becomes as useful as possible for the client and other users.

Output: Information on measurement quality included in a systematic quality report, covering all relevant aspects of survey quality (e.g. according to the quality declaration template of the Official Statistics of Sweden).

4 Phase 1 - Define the survey

4.1 Introduction

Task: Phase 1 covers work to be done before beginning to formulate concrete questions and generally designing the questionnaire.

This includes
- solving content and relevancy questions
- developing a list of variables and a tabulation or analysis plan.

During the same period decisions should be taken on:
- the data collection method
- the sampling frame, sampling process and sample size.

Before going on to Phase 2, it is important to consider whether the survey is at all feasible, given the availability of data, the burden on the respondents, the cost framework and the time scale.

Distribution of work: To ensure a good basis for the survey's quality, there needs to be close cooperation from the start between the client, the producer and the questionnaire designer. The project manager for the survey usually represents the producer. The term questionnaire designer is used to describe the person who will work with the formulation of the questions and the production of the questionnaire. Sometimes, this is the same person as the project manager; sometimes the project manager works with a measurement specialist. (At Statistics Sweden, there is a special measurement laboratory for this work and a network of persons who work with formulating questions and formatting questionnaires.)

In this phase the client is the driving force by specifying his/her information needs. The task of the questionnaire designer and the producer is primarily to show which resources are available and which technical solutions would work. It is the client who must decide whether the planning work should go forward or not.

It is not sufficient for the client and/or the project manager to have extensive subject-matter knowledge and familiarity with the information that already exists in the survey area. To be able to draft well-functioning questions, the questionnaire designer needs to be involved when the subject-matter problem is re-formulated into a statistical problem. The same applies when choosing the sampling frame and analysing how well it might agree with the survey topic (i.e. the population on which information is to be collected, regardless of whether this population consists of persons, schools, establishments, or milking cows), as well as when deciding which data collection method should be used. How this affects the design of the questionnaire is described below.

4.2 From a general problem to a statistical problem

If a survey is to yield useful results, the client must discuss the basics with the persons who are to carry out the survey and those who are to draft the questions. The problem area should be defined and it should be made clear how it could be covered within the framework of available resources. The client, producer and questionnaire designer should arrive at a common view of how the task is to be

carried out. Similarly, they must plan how the analysis after the data collection should be done.

Much can be gained from careful preparatory work, yielding clear definitions and limitations on what is to be measured by the statistical survey. Transferring generally formulated survey objectives, for example "Swedish eating habits", "Swedish taxpaying attitudes" or "Innovative activities of establishments", into precise questions takes time and consideration. When the client only has a vague idea of the questions, the questionnaire designer is forced to work with several possible interpretations of how the subject-matter problem is to be transferred into measurable statistical characteristics. If the questionnaire designer participates in the work right from Phase 1, there will be significantly less work in the following phases.

4.3 Obtain information on survey units

Over and above the actual questions on the survey's subject matter, it is sometimes necessary to pose technical survey questions in order to calculate the estimates correctly. The questionnaire designer needs to receive the answers to such questions from the responsible person. The following four types of questions need to be considered in each survey:

- The sampling frame nearly always contains a few units that represent **over-coverage**. Questions must be set to identify the units that do not belong to the relevant population. The answers are used to discontinue an interview and/or to sort out afterwards the answers that should not be included.
- Sometimes certain questions are needed to **decide the sampling probability** for the survey units. This can be necessary if the same survey units are in more than one sampling frame or are made up of several sampling units.
- Other questions might be necessary to **define the delineation of the survey unit**, for example which persons are to be included in a housekeeping unit or which organisational unit that the answers in a survey on establishments really refer to.
- In surveys on establishments in particular, it is necessary – in case of further contacts or follow-up surveys - to get information on **which person gave the information** and how he/she can be contacted.

4.4 Choose a data collection method

How questions, response alternatives and questionnaires are constructed depends to a large extent on the data collection method. The method that gives the best measurement quality in a particular survey depends on the scope of the survey, the type of questions and who the respondents are. When designing a survey, the need for statistical precision, as well as costs and timing should be taken into account. This can mean a compromise between the ambition to use the most accurate measurement method and the available resources.

4.4.1 Mail surveys and more modern methods

Paper questionnaires
With **mail surveys**, we mean here that the questionnaire is sent by mail to the respondent for completion and then sent back. In surveys where responding is voluntary (non-mandatory), it is recommended that the questionnaire does not

consist of more than 12-16 pages and does not take more than 30 to 45 minutes to complete.

The measurement advantages of mail surveys are that they permit:

- relatively long explanatory introductions and/or questions
- several, and long, response alternatives
- simple matrix questions (e.g. to set the same questions to all household members, using one answer column per person and one row per question)
- questions on more sensitive subjects than in an interview
- the respondent to take the time needed to give well-prepared answers or to look for the correct information on amounts, number of years, etc.
- such aids as maps, pictures, or symbols that the respondent needs to see to be able to answer the questions.

The measurement limitations of mail surveys imply that the designer should:

- avoid *skipping instructions*, as these increase the risk of an incorrectly filled-in answer. If the respondent has answered both the question with the skipping instructions and the question or questions to be skipped, it is not always clear where the mistake is. It must then be decided if the answer should be counted as item nonresponse or if another method should be used. Whatever the choice, the quality of the results is affected.
- avoid extensive or complicated tables in which the respondent is expected to respond using information from both rows and columns. This often leads to too many incorrect markings.
- limit the number of questions to 12-16 pages and use an “attractive” layout. The risk for both unit nonresponse and item nonresponse increases if the mail questionnaire is seen as too long. There is also a great risk that the respondent's will or ability to give well-prepared and correct answers will be reduced if the questionnaire is too extensive.
- take into consideration that there is a risk that persons in the respondent's surroundings may affect the answers, i.e. questions that must be answered by a specific person should be avoided.
- take for granted that the respondent will look through the questionnaire from beginning to end before answering a question. The answer can therefore be affected not only by questions coming before but also those following behind the actual question, which is different from in an interview survey.

Instructions in mail surveys to establishments must show clearly which position or competence the person answering the questions should have. It cannot be assumed that the receiver of the questionnaire gives it to the person who has the widest knowledge in the subject area. When it is not the "most appropriate person" answering the questions, the accuracy is significantly worsened.

Electronic questionnaires

Electronic questionnaires, i.e. questionnaires which are available on a website or distributed over the Internet or on floppy disks, are common today in administrative applications, such as claims to insurance companies, and in banking, healthcare and private establishments. They are also increasingly used in the collection of data for statistics. Questionnaires can be posted on a website or distributed via a telecommunications network. Electronic distribution over the Internet will probably increase quickly, when security and confidentiality problems of have been solved in a legally tenable way.

When changing from paper to electronic questionnaire, it is not simply a case of transferring the questionnaire to the new technology, but rather to utilise the possibilities presented by the new technology. Electronic questionnaires can be

more or less "intelligent" and can, to a varying extent, have built-in

- checking provisions and error signs
- facilities to correct answers which show error signs and to register other changes
- routines for additions and other calculations
- skipping instructions, which are carried out automatically
- information fields, shown on the screen when clicking on the current question.

An electronic questionnaire can be more complex than a paper questionnaire without raising the burden on the respondent. At the same time, the consistency between the answers can be improved. The questionnaire designer must develop and write the questions in a form that suits the format of the computer screen. But it is probable that those using the electronic questionnaire also wish to make paper printouts to get an overview or to make notes before they answer, and all respondents will not have the facility or inclination to fill in an electronic questionnaire. Though it can be difficult to make sure that the questionnaire on the screen and the questionnaire in paper format are the same, this problem can be solved. For single-round surveys, the programming costs for making use of the options of electronic questionnaires can be unreasonably high.

There are many more collection methods than those discussed here. For collecting a small amount of numerical information from each respondent, Statistics Sweden uses Touchtone Data Entry (TDE) in multi-round surveys on establishments. Figures can be complemented with recorded comments. Techniques for direct retrievals from establishment administrative systems are under development, but have been delayed because the maturity of the data varies substantially and establishments use many different types of ADP system.

4.4.2 Telephone interviews

In surveys directed towards individuals where only the sample person him-/herself can or should answer the questions, an interview is recommended. Mail questionnaires are less suitable, because the answers in these can be influenced or answered by e.g. family members or colleagues. In telephone interviews, the interview time should not be longer than 30 minutes. If a longer interview is necessary, other data collection methods should be chosen – face-to-face interviews or a mail survey. Telephone interviews on subjects that are important to the respondent (i.e. questions on childcare to parents of small children) often work very well even when the interview time is longer.

The measurement advantages of telephone interviews are:

- Computer-assistance techniques have been developed for this method
- There are few or no problems with filter questions or skipping instructions. When an interview is computer-assisted, the skip happens automatically and interview errors are reduced (the interviewer can however mark an incorrect response alternative and in this way cause an incorrect skip).
- It is possible to have several and different follow-up questions to different groups of respondents or to earlier answers, as the skipping instructions do not cause any problem.

The measurement limitations of telephone interviews are:

- Questions must be short and should not contain several information stages. Otherwise the respondent’s short-term memory might fail and it will not be

possible to know how the question has been interpreted or what the answer applies to.
- For the same reason, the response alternatives should not be too many or too long.
- The questions must be written down and the interviewer should read them in such a way as to discourage the respondent from answering too fast, i.e. before he/she has heard the full question or all of the response alternatives.
- Questions that require some thought before answering are less suitable for telephone interviews, as the pace of such an interview is relatively high.

In certain surveys, it is acceptable to use **interviews by proxy** to raise the response rate, i.e. to let someone other than the sample person answer. If the sample person cannot answer the questions him/herself (due to illness, absence, etc.), questions concerning factual personal details can be put to a person who is well acquainted with the sample person and his/her current situation. Attitude questions, however, should be registered as item nonresponse. Interviews by proxy have a better chance of yielding an acceptable result for questions of the type "how things usually are" than for questions on "how things were during a specific period".

4.4.3 Face-to-face interviews

Face-to-face interviews, i.e. when the respondent is interviewed in a personal meeting with an interviewer, are relatively expensive to conduct. The interview time should not be more than one hour. Otherwise, the quality of the answers risks being reduced. Both the respondent’s and the interviewer's concentration lessen in a long interview.

Face-to-face interviews are used primarily when the questionnaire contains many questions, is complicated and/or requires support in the form of e.g. an answer sheet or maps. Other reasons for using face-to-face interviews can be the subject matter or the place of the interview (e.g. interviews at an airport with travellers).

Face-to-face interviews give the questionnaire designer the possibility of producing a questionnaire with advantages from both mail questionnaires and telephone interviews. A disadvantage is that the interviewers can have a relatively strong influence on how the respondent answers, through their way of reading the question and their body language. In particular, the accuracy of answers to attitude questions can be significantly worsened. Face-to-face interviews are usually computer-assisted as well.

4.4.4 Computer-assisted interviews

Because interviews are now mostly computer-assisted, the questionnaire designer has new possibilities. The primary advantages are that built-in controls prevent unreasonable and inconsistent values from being registered and that the interviewer, with the help of automatic skipping instructions, always arrives correctly at the next question, regardless of the amount or complexity of skipping instructions in the questionnaire.

In long interviews, the overview of the survey can be affected. Unfortunately, it is often a major problem (so far) to create tables for computer-assisted interviews. In paper questionnaires, several questions can sometimes be put together into one table and give the interviewer a better overview of the questions. For a computer-assisted interview, it is useful to create an overview diagram on paper, which shows in one block which questions are included in the

questionnaire. This makes it easier for the interviewer to know where he/she is in the interview, particularly if there are several loops (i.e. questions which are only used under certain circumstances). An overview diagram is particularly useful if the interview has to go back a couple of stages in the loop to correct a previous piece of information.

The total time required will vary depending on the type of questionnaire used. If a paper questionnaire is used, a lot of time will be required when the questionnaire comes back completed. For a computerised interview, a lot of time is instead needed to program and thoroughly test the questionnaire (all combinations of possible answers must be tested). Deciding the layout on the screen is also time-consuming.

4.4.5 Several data collection methods in the same survey

Many surveys are ***"mixed mode"*** surveys, which means that more than one method for data collection is used. The aim is often to improve the response rate by giving the respondent more than one way to answer. In other cases, it is necessary to develop questionnaires for different groups of sample persons, for example, in surveys including disabled persons (such as hearing- or sight-impaired persons). In general, it is considered that the number of modes will increase in many surveys concurrently with the computerisation of collection. This is a considerable measurement problem, as a specific formulation will not work equally well, as a rule, in the different collection methods.

A questionnaire is often developed for the main measurement method and then used as far as possible for the others as well. The most commonly occurring mixed mode version is a mail survey followed by telephone interviews for the nonresponse from the mail survey. The questionnaire is most often developed as a mail questionnaire and then the same questionnaire is used in the interviews. The mail questionnaire's biggest advantages (to be able to have many and long response alternatives and long explanatory introductions) are serious disadvantages in telephone interviews. These can therefore not be used fully in a mixed mode survey. If specific instructions are not written for the telephone interviews, every interviewer must be allowed to improvise whenever weaknesses occur. There is then a considerable risk for large variations in their ways of solving problems, and the correlated interviewer variation will be large.

Whatever mixed mode survey is chosen, certain accuracy problems always become apparent. The distribution of answers is frequently different for different data collection methods. In telephone interviews, a larger number of respondents often choose an answer at the beginning or the end of several read-out response alternatives than is the case among those answering the question in a mail survey and seeing all the response alternatives at the same time. In mail surveys, a larger number commonly choose the response alternative "no opinion/don't know" than in telephone interviews. The interviewer does not read out the alternative ”don’t know”. In addition, a majority of people do not like to be seen as lacking knowledge or an opinion in front of an interviewer waiting for an answer.

If one of the data collection methods in a mixed mode survey is a personal visit, the questionnaire designer must think carefully about which question area(s) might be influenced by interviewer bias and take this into account. Examples of such question areas are alcohol consumption, violence in the home, attitudes to tax paying. Some solutions to the problem are special interviewer training and/or that the interviewer hands over a questionnaire with these questions which that

respondent can fill in and give to the interviewer in a sealed envelope or post himself.

4.5 What help is available?

4.5.1 Earlier surveys

If an earlier survey has been carried out in the same subject-matter area, it is very useful to look at the questions and to evaluate how well the questions capture the subject matter. In many survey areas, more information is available on the Internet.

How these surveys worked in the field, i.e. how they were carried out should also be studied. The measurement quality of the survey, evaluations and re-interview studies, if they exist, are particularly interesting. It is also worthwhile to get information on the size of the unit nonresponse and the item nonresponse. Research reports on new methods can be found on the Internet.

4.5.2 Register data

Data from administrative registers is often used to replace various questions in a questionnaire. Register data can sometimes be more reliable than a respondent's answers in a survey. This applies e.g. to events that took place some time back. Information on income has been shown to be more reliable if taken from a register than if given by the sample persons. Using register data also means that the response burden can be reduced, which is sometimes a condition for the survey to be carried out with an acceptable response rate. Sometimes it is preferable not to do a new data collection but to be satisfied with a register study. This might be considered in longitudinal studies and if the respondents are hard to get hold of and reluctant or unable to give information.

It is important to balance what a survey is supposed to cover with what the data in an administrative register can give. But it is not certain that the administrative register really contains all the requested information or satisfactorily describes the facts to be studied. It might be necessary to complement this information with questions in a questionnaire.

There are many studies that identify gaps in coverage and relevance in administrative data - for example non-reported income in taxation registers. The number of crimes reported to the police is significantly less than the number of crimes reported in a survey on crime victims. Another problem is when register data does not have the same reference period or date as the answers in a sample survey. For example, tax assessment data can be two years older than data in a newly carried out household survey.

4.5.3 In-depth interviews and focus groups

In-depth interviews and focus groups with persons representative of the population are efficient. Using these methods, it is possible to expand the total knowledge of the subject area or "check that your ideas are right".

An **in-depth or qualitative interview** aims to lay bare or explain the connection between cause and effect and to the underlying factors. The method is used to identify the areas or variables that are relevant to the respondent. If the in-depth interview focuses on a person's understanding of the concepts, on how he/she interprets certain questions and on how he/she arrives at an answer, the interview

is also called **cognitive.** This means that the interview is carried out within the concepts and methods of cognitive psychology.

A focus group is used for an informal discussion with a group of six to eight persons from the survey target population. The persons should not know each other beforehand. The length of the discussion is preferably 1-1½ hours. It is led by a moderator who, after a short introduction, ensures that the conversation sticks to the subject and that all the participants voice their opinions. The moderator should not intrude his/her own opinions, but might intervene in the discussion if necessary to make things clearer. The discussion is recorded on tape and analysed afterwards. Observers are also used sometimes to capture non-verbal signals and to add to the moderator's observations. Focus groups with business people should preferably be held in the morning to avoid cancellations due to unforeseen events at work.

Example: Early in the planning stage of a quantitative survey aiming to study the general public's view on democracy, some focus group discussions were carried out with persons from the target population. The survey producers wanted to see if the question areas which they themselves had thought important, really were important for "normal people". The focus groups also revealed the preferred choice of words, which is valuable knowledge when drafting the questions in a quantitative survey.

Recruiting persons for focus groups requires hard work to identify and contact appropriate participants. They should be chosen carefully, and not at random, from the survey population. Together they should represent the widest possible spread of the backgrounds relevant to the survey. They should also be talkative, active in social networks and have a good idea of how persons in these networks think and feel.

Focus groups are a good way to get fresh ideas from the world outside statistical survey production. If the group is well composed, matters that might seem strange to "normal people" will be identified. It is therefore important that the participants are not friends or colleagues or members of several focus groups. An advantage with focus groups compared to in-depth interviews is that through the interaction between the participants, associations and new lines of thought appear in a way that does not occur in individual in-depth interviews.

In-depth interviews and focus groups are very useful methods for collecting and getting ideas on relevant question areas, both in terms of content and wording. They also constitute a simple way to gain knowledge on the working methods and procedures in enterprises, municipalities, organisations, etc. All this is very valuable for the questionnaire designer. In certain cases, it has been necessary to put questions to establishments on administrative routines, for example, on which software is used. Otherwise, it is difficult to assess the quality of the given answers or to adequately refine the study domains.

4.5.4 Variable list and tabulation plan

Variable list

Background variables
BV1 Sex
BV2 Age
BV3 Education
BV4 Position

Survey variables
SV1 Traffic regulations
SV2 Resources
SV3 Attitude to XX

The creation of a variable list and a tabulation plan can also yield a reliable overview of the question area in the questionnaire. The variable list is a simple catalogue in which every variable to be included is given a name and a code. A variable can be the source of one or several questions, depending on how difficult it is to measure. The tabulation plan can then be set up quickly, with the help of the variable codes, making it possible to check that all the variables necessary to answer the survey's subject-matter problem are included. The question area can easily be checked against the actual subject problems. The tabulation plan is a good way to ensure that all the necessary background variables and study domains have been included.

Tabulation plan
Background
BV1**BV2*BV3
BV3**BV2*BV5
BV4**BV1*BV5

Traffic
SV1*BV1, BV2
SV2*BV2, BV3
SV3*BV3, BV4

The variable list and the tabulation plan together form an important tool, which should be used in the design of every questionnaire. In the same way, plans for multivariate analyses or for graphic presentations should be put together and checked. It is too late at the stage of the analysis and presentation to discover that variables, which could have contributed largely to the interpretation or comprehension of the results have been overlooked. The possibility to add to the variable list in succeeding phases becomes more and more limited the longer the process goes on, and once the survey has started, it is too late.

4.6 Checklist

Before the questionnaire designer can move onto the actual questionnaire design, the following items should be completed in Phase 1.

CHECKLIST Define the survey content

1. Define the need for information and the proposed questionnaire content. Define the object, the population and the parameters to be estimated.
2. Carry out in-depth interviews with persons from the proposed population of respondents (sometimes even with end-users) and/or carry out focus group meetings.
3. Select and determine the sampling process, the sample size and the data collection method, taking into account quality requirements, financial and other resources and time limits.
4. Review what information in the survey area exists already in the form of register or earlier sample survey data. Try to find out the measurement quality of these and collect good examples within the question area of individual questions, directions and instructions.
5. Decide whether the survey is feasible or not.
6. Put together a variable list and a tabulation plan, i.e. check that everything is included.

The final products of Phase 1 are a variable list and a tabulation plan. There should be decisions on the category of respondents, on how the sample is to be drawn and on the data collection method to be used. It should be clear which information from administrative registers, if any, that is to be used instead of questioning the respondents directly.

For an entirely **new survey**, the work should start with **Phase 1**. It should already have been decided if it is worthwhile to carry out the survey. The less the subject area is known, the more important the preparation phases and the wider their scope.

5 Phase 2 - Questionnaire design

5.1 Introduction

Task: This phase deals with designing questionnaires, question by question. The question area will be translated into concrete questions, adapted to the chosen method for data collection. Definitions and response alternatives shall be formulated, question order and layout decided, and information material and instructions written. Experience shows that certain variables cannot be measured directly. Instead, they must be broken down into several questions. The rules for how the answers should be coded also have to be decided. Examples of such variables are labour force status in the Labour Force Survey and the long-term ill in the Living Conditions Survey. The intended final product is the best possible questionnaire.

- The division of work in this area varies. Many clients produce a preliminary version of the questionnaire and work together with the questionnaire designer to produce a final "drawing board version". This is often a favourable working method, because all the parties are forced to think through the question production and take common responsibility for it. In other cases, the client hands over the entire questionnaire design work to the questionnaire designer. The client must however always approve the final version.

5.2 General tips to the questionnaire designer

The perfect questionnaire, in which all the respondents clearly understand all the questions and can answer them correctly without any problem, will never be constructed. Still, this must be the aim. Even when the first draft of the questionnaire is on the drawing board, it is possible to avoid the most common errors. It should not be necessary to use resources for one or several tests to notice mistakes that could have been picked up before testing. Some prerequisites for a questionnaire to work well are that it has a clear and motivating introduction, is easy to read, understand and fill in, and that the respondent is taken clearly through the questionnaire.

The formulation of a question and its place in the questionnaire determines how the respondent will interpret it and answer it. Below is a list of some basic linguistic and organisational rules of thumb.

Choice of words
- Avoid negatives.
- Consider the tense. Present tense for ongoing activities and imperfect for finished activities.
- Avoid abbreviations, difficult words and technical terms. If they must be used, explain their meaning.
- Use words and formulations that can be interpreted as neutral. Otherwise, the question could be interpreted positively or negatively and be leading.

Be specific
- Specify unit to be used in the answer and desired precision.
- Give reference date/period clearly. Do not change reference date/period between questions if it is not absolutely necessary.
- State which components should be included and which should not *(e.g. "domestic sales", "exclusive VAT", "during 3rd quarter 1999").*

Syntax and comprehensibility
- Use simple and clear language - even when asking technical staff!
- Use short and concrete questions. (In long complex questions, it can be difficult to distinguish which part the respondent should answer).
- Avoid sentences with many and long words.
- Formulate questions so that the respondent can give a clear answer.
- Watch out for abstractions and hypothetical questions
- Ask about one thing at a time - avoid summarising questions.
- Watch out for questions containing conjunctions such as "and", "as well as", "or". Then the question may contain several questions and should be split up.

Order/orientation
- Position the instructions as near the question as possible. Limit the length.
- Follow a logical sequence in the order of the questions.
- Divide the questionnaire into logical blocks.
- Give the blocks titles and divide the questionnaire into sections, also graphically. This applies particularly to so-called omnibus surveys, which often have abrupt transitions between question areas and need intermediate text between them.
- Use graphic signals to show where to find the instructions, important concepts and the logical way through the questionnaire.
- If possible, position questions which can be difficult to answer at the end of the questionnaire. It is important that the respondent does not get stuck on the first questions.
- Position questions, which risk being interpreted as sensitive or as an invasion of privacy, but which must be included, at the end of the questionnaire.

Other
- Avoid mixing questions demanding different types of data formations. (answers to be copied, processed or "directly answered".)
- Avoid changing the direction and position of answer boxes.
- Use the same division of response alternatives as far as possible.
- Never include a question just because "it might be interesting".

5.3 Consider the respondent

Filling in a questionnaire is not a priority
(Allen Gower)

The respondents often answer quickly and many times with insufficient commitment. They usually make no particular effort to understand an unclear or difficult question. Instead, people answer as well as they can to what they believe the question to be. Clear language and unambiguous content is then a must to ensure accurate data.

From the start, think of who will actually be answering the questionnaire and not simply those the questionnaire is meant for. The questionnaire designer needs to understand the level of the respondents' education, their knowledge of the language, if they are used to expressing themselves in written form, etc. and should take this into consideration when drafting the text. The need for definitions or explanations must be adapted to suit the sample persons' situation.

Furthermore, it is important to know if the actual respondents have the knowledge and necessary technical skill to answer the questions. They must understand what information they are expected to give and they must be able to find this information in their experience, memory, accounts, diaries, different activities, etc. It must be made clear to them what type of data is requested and how large the burden on the respondent really is. The examples below describe a case in which the demand had not been adapted to the respondents' situation.

Example 1:
A questionnaire given to the municipalities asked about pre-school activities. During a review afterwards with the respondents in some municipalities, it was asked how the information had been collected. It appeared that, in many municipalities, a person at the central office had been given the task of sending copies of the questionnaire to the different day-care nurseries. This person also collected them, compiled the answers and sent the final questionnaire to Statistics Sweden. It became clear that due to insufficient knowledge about the individual nurseries it was impossible for this person to assess whether the answers from the nurseries were reasonable, or if something might have been misunderstood.

Example 2:
In a test on health issues carried out on children aged 4, it was seen that the terms of illnesses or disorders, which the client and the questionnaire designer had feared would require explanation, did not present any problem at all. However, other, apparently simple, words were difficult to understand. An example of such a word was "giddiness", which the children did not know. It was necessary to change this to "dizzy in your head".

In surveys where **the survey objects are individuals**, the same individuals are usually also the respondents. In such surveys, it is necessary to weigh the need for information against what the respondent can manage to answer with adequate certainty. It is assumed that the persons who will answer the questions have the answers (the information) "in their head", for example life history, consumption, memory of an actual occurrence, knowledge, opinions and attitudes. When the questions concern the entire household, it is assumed that more than one person in the household can answer them. In surveys on, for example, expenses,

income, travel habits and use of time, the respondent is asked to use a diary, tax returns, receipts, insurance papers, tax assessment data, etc. to provide the accurate data.

In ad-hoc surveys where the sample units are enterprises, municipalities, organisations, etc. it is often a problem to know to which person or function the questionnaire should be addressed, even if it is known which skills are required to answer it. A common way to try to find the correct respondent is to address the questionnaire to a person with a specific position, for example managing director, purchasing manager or human resources manager. But regardless of to whom the questionnaire is addressed, the establishment decides internally who should answer the questionnaire. Do not assume that the respondent is automatically a person with good knowledge of the subject matter. In multi-round surveys, efforts are made to establish contact with those persons who are in a good position to give the information. This is particularly important for large establishments, whose answers have a large impact on the statistics.

In surveys where reporting is mandatory, the burden on the respondent can be significantly larger than in surveys with voluntary participation. The need to use a well-formulated questionnaire from a measurement point of view is therefore greater in surveys with mandatory reporting.

The quality of the answers is dependent on the time the respondent must devote to produce the correct information. In surveys to establishments, this is the respondent's working time, and the employer decides how much time can be spent on the task. In surveys in which many respondents are small establishments, there is a risk that a high burden on the respondent will lead to low accuracy of the answers.

The same person can be called upon to answer several different questionnaires sent out by the national statistical institute and by other authorities, industry organisations and research institutes. It is an advantage if the respondent after a time becomes "professional" and familiar with understanding concepts and filling in questionnaires. On the other hand, this increases the risk that the respondent gives the same information if several surveys ask for similar, *but not identical*, information. Coordination and use of standardised questions reduces the burden on the respondent and increases the accuracy of the answer.

5.4 Common mistakes in questionnaires to individuals

It is more difficult to describe how to produce a good questionnaire than to look at questionnaire designs and question formulations that have worked badly. It is important to learn both from your own mistakes and mistakes made by others. This is why we review here some common types of errors that can be avoided already at the drawing board.

Define the questions in time and space

Example
Do you read any evening newspapers?

The question is lacking a reference point in time and can be interpreted in several ways. Some respondents will think that the question refers to whether they read evening newspapers every day and will answer "No" if they usually only read them on Saturdays and Sundays. Other respondents in the same

situation will answer "Yes", because they do read evening newspapers at the weekends. How the question should be formulated depends on the aim. If the aim is to identify those who generally read evening newspapers every day, the question can be formulated:
Do you read an evening newspaper at least five times a week?

Example
How long have you lived here?

Here a reference to space is lacking and there are therefore many possible interpretations. "Here" can be interpreted as the building, the district, the municipality. In the question, a specification of "here" should therefore be included.

The respondent has another frame of reference than the producer

Example
Do you have any long-term illnesses, disorders due to an accident, any disabilities or other weaknesses?

The question formulation can sometimes, as in the example, lead the respondent to leave out chronic illnesses/disorders that should be included. The question contains several words that can give the idea that only serious disorders are asked for. This means that persons with light forms of age-related diabetes, high blood pressure and psoriasis could sometimes answer no to the question. The respondent would have a different frame of reference than the interviewer if he/she rarely or never had problems with the disease, because he/she either had no serious symptoms or the symptoms were held back by well-functioning medicines.

Several questions in one

Example
In a survey on working environment, the respondents are asked to describe how they see their work. They receive the following instructions:

Below are a number of rows with boxes, going from one extreme to another. Describe how things usually are for you by marking an X in each row. So the further left you put an X, the more correct the description to the left is. And the further right you put an X, the more correct the description to the right is. Therefore, the box second to farthest out means that you only partly agree.

	Neither nor	
Far too much to do	☐ ☐ ☐ ☐ ☐ ☐ ☐	Far too little to do
Far too much responsibility	☐ ☐ ☐ ☐ ☐ ☐ ☐	Far too little responsibility
Monotonous work	☐ ☐ ☐ ☐ ☐ ☐ ☐	Varied work
Physically strenuous work	☐ ☐ ☐ ☐ ☐ ☐ ☐	Calm and pleasant work

More questions follow

The first two questions are both really two questions in one. The first deals with to what extent the person has too much to do (and too little responsibility) and

the second to what extent the person has too little to do (and too much responsibility). Those who have neither too little nor too much to do should set an X in the middle box. It has been shown that even those who sometimes had too little and sometimes too much to do set an X in the middle, i.e. they make an average of a working situation, which moves between the two extremes. It is therefore impossible to see how many persons who have just the right amount to do.

The respondent can also be confused when both extremes describe a negative situation. The other rows (questions 3 and 4) go instead from negative to positive extremes.

Below is an example of a question that needs to be written in a clearer way, if the respondent is to be able to answer.

Example
One of the US's largest public opinion measurement institutes, The Harris Poll, asked the question:
Have you often, sometimes, almost never or never had guilty feelings when you have been unfaithful to your wife?

1% answered often, 14% sometimes or almost never and 85% answered that they had never had guilty feelings due to infidelity. Here, there is an implied (filter) question on whether they have been unfaithful or not. Those who answer "No" to this should skip the above question. Errors in question design of this type are often subtle and not as obvious as in the example.

Filter questions work better in interview surveys than in mail questionnaires. In the latter, filter questions and skipping instructions should be avoided, as they are often misunderstood, and instead an extra response alternative can be added (in this case, if they had never been unfaithful to their wife).

Example
The following question was asked child welfare centres:

*Do you today carry out tests/investigations/observations of **all** children in other age groups than 4 year olds to catch children with MBD problems?*

The response alternatives were "Yes" and "No".
Even if this question had really only **one** subject, it is built up into several sub-questions, which can also be answered with "Yes" and "No". How carefully the respondent reads the question varies. If the question contains several parts, which can be answered, it is therefore difficult to know what the answer relates to. The best solution is to divide up the question into several questions.

Leading questions and loaded words

Example
More people have seen the film "Gone with the Wind" than any other film produced in this century. Have you seen it?

The formulation of this question hints that the respondents are in some way going against the norm if they answer no and this is therefore leading. Questions should be neutral or balanced, i.e. if a specific response alternative is justified by a question, other alternatives should also be justified.

A leading question makes it easier to choose one response alternative over another due to the question formulation, for example *"Do you personally think that you are positive towards...?* Here a "yes" answer is indicated by the choice of words in the question, whilst a "no" answer seems to contradict the meaning of the question. The question becomes considerably more balanced if it is written*: "Are you positive or negative towards...?* The question is also leading if it takes advantage of a person's wish to prefer status quo, plays on prestige or uses a well-known person's or organisation's name. Instead of writing, for example "*Leading researchers such as XX believe that...What is your opinion?"* it is better to write *"Certain researchers believe..., while other researchers.... What is your opinion?"*

Example
In a methodology experiment, a survey institute in Sweden posed the following two questions.

The first version was:
"Within the EU, work is going on towards the creation of a currency union, EMU, with a common currency for the countries within the EU. Are you for or against Sweden joining the EMU?

The distribution of answers was: For 38%, Against 48%, Don't Know 13%.

The second version read:
Belgium, Holland, Luxembourg, Italy, Portugal, Spain, Ireland, France, Austria, Germany and Finland will probably join the EMU from the beginning. If this is the case, do you think that Sweden should also join EMU or do you think it should not?

The distribution of answers was: Should join 50%, Should not join 42%, Don't know 8%.

The content of the questions is not exactly the same, nor that of the answers. Version 2 supposes that everyone knows what the EMU is. The questions can therefore not be expected to give the same result. But the difference in the distribution of answers shows clearly how different choices of alternatives can result in a different representation of opinions. Uncertain respondents, in particular, can be influenced and their answers steered in the desired direction.

Vague questions and vague response alternatives

Example
"Do you have a specific doctor that you usually turn to?"

The concept "specific doctor" is too vague. It is not clear whether it means whether you see just Doctor Nilsson, if you have a family doctor or if you regularly visit a specialist doctor, such as an optician or a gynaecologist. The word "usually" excludes those who are listed with a family doctor but who never need to go there.

Example
"How often did you go to church last year?"

The response alternatives are: "Never", "Rarely", "Sometimes", "Regularly". In this example, the response alternatives are too vague. If different people were asked what the words "rarely, now and then, sometimes or often" mean to them,

they would give very diverse answers. The meaning of the words varies, not just between different people, but also depending on what is being asked. There is a big difference between how many times "often" refers to if asking about how often you eat ice cream in the summer or how often you have ear infections.

In the above example, the word "regularly" can also be interpreted in different ways. It can be interpreted that the person has been to church at regular intervals during the year, but also that they have followed their regular behaviour throughout the year and been to church at, for example, Easter, Advent and Christmas.

Instead of these vague response alternatives, it is better to construct alternatives with a time reference, i.e. number of times per day, week, month or year. It is also important to consider that the majority of people understand the middle of the scale to be the "norm value", for example, the number of hours that an average person watches TV during one week. People position themselves according to this, if they cannot provide a more precise answer. With a different scale, the answers would be distributed in a different way.

The question asks too much of the respondent
A question should be drafted in such a way that it is possible for the respondent to come up with an answer. The client can be very interested in getting detailed knowledge in a specific field. The questionnaire designer must formulate the questions in such a way as to get as much information as possible, while still not asking too much of the respondent.

Example 1

53. What type of cheese did you normally eat about 10 years ago? Show how much per day, per week *or* per month?

☐ I rarely or never ate cheese → *Go to question **54***

	slices/day		**slices/week**		**slices/month**
Cheese, 24% fat or more		**or**		**or**	
Low fat cheese, 17% or less		**or**		**or**	
Dessert cheese		**or**		**or**	
	spoons/day		**spoons/week**		**spoons/month**
Soft cheese *(1 dl = 7 tablespoons)*		**or**		**or**	
Cottage cheese *(1 dl = 7 tablesp)*		**or**		**or**	

There is no problem here to understand the question, but there is a problem to provide an answer. Even if the respondent has an answer - can he/she give it? How does the person answer who remembers that, 10 years ago, he/she sometimes made his/her own breakfast (but how often was this?) and then

usually had 4-6 sandwiches with cheese, but when his/her partner made breakfast the filling usually varied?

For each question that the questionnaire designers write, they must ask themselves whether that particular formulation is the best to achieve its aim. The above question is taken from a survey that was directed to persons who had some form of cancer. There was also a control group without the disease. Because it was thought that there is a link between illnesses and eating habits of 10 years, 20 years or even longer ago, it was essential to get the most exact answers possible. In this example, a very large amount of "don't know" answers were received. With another format, a larger number of respondents would probably have been able to give a more exact answer.

One way would be to first ask some questions aimed at getting the respondent to remember as well as possible their life during that period: "Were you studying at the time?", "What were the ages of your children?", "Where did you work?". After this, it should be possible to get closer to the eating habits of the person. Finally, it is possible to set more detailed questions. It is very seldom, however, that questions can be answered as exactly as is hoped for in this example after so many years. The memory is affected by more recent habits.

Be careful with Yes-No questions when referring to attitudes, opinions and values

Example
Two different questions are used to measure the popularity of the prime minister, but they result in different numbers of positive answers.

"Do you think that Prime Minister XX is doing a good job?"

Response alternatives: Yes, No.

"How do you feel that XX is as Prime Minister?"

Response alternatives: Good, Quite good, Quite bad, Bad.

It is human nature that it is easier to agree and say "yes" than to stand up against something, and this is often taken advantage of in opinion surveys. This means that in this example, the first question got a larger number of positive answers than the second question. The second is a better question, from a measurement point-of-view, because it allows the respondent to report an opinion with more nuances.

Example
"Do you think that it would be right to raise taxes so that long-term ill people would be able to have their own room in care facilities?"

The response alternatives are "yes" and "no".

Besides the "yes-no" problem, this question is difficult because it is hypothetical. It is incredibly difficult to measure opinions that would remain the same if the hypothetical situation became reality. Or to measure the probability of future behaviour by asking a hypothetical question. Because hypothetical questions do not oblige anyone to anything, it is much easier to agree with something than to go against it. This applies especially if it is more socially acceptable to agree.

Such an attitude question should be divided up into several questions, which together can show the respondent's attitude. Respondents should, for instance, be able to express that they agree with the idea but think that it can be achieved without a raise in taxes.

Tables and "agree with" questions

Experience has shown that certain tables with several dimensions are rarely completely filled in, however practical and logical they can seem to the questionnaire designer. Many respondents simply fill in the "yes" column in the table. For example, they mark "yes" for those disorders, illnesses or symptoms that they have had, but do not bother to mark "no" or "don't know/don't remember" when these alternatives are correct. The result is a large item nonresponse instead of information about those who have not had certain disorders, illnesses or symptoms. In general it is best to set out each part question as a separate question.

In a special type of table, the "questions" are formulated as statements which the respondents, to varying extents, should agree or disagree with.

Example

	Agree totally	Agree partly	Don't know	Disagree partly	Disagree totally
I risk my health at work	☐	☐	☐	☐	☐
I receive enough support for my work from my boss	☐	☐	☐	☐	☐
plus several other assertions					

To make such a question work well from a measurement point of view is not easy. The table shows several weaknesses. The questionnaire designer must think carefully about what it means to only partly agree or partly disagree. Maybe, in practice, the state of partly agreeing is the same as partly disagreeing. In this example, there is also a completely wrong middle alternative, because the "don't know" alternative is not a part of the scale but lies completely outside as a different possible response alternative.

The assertions included in the table must be very carefully drafted to get unambiguous answers in all response alternatives. Above all, the adjectives "enough", "good" and such like should be avoided if possible. It is hard for the respondent to know what it means to partly agree that they get "enough support", because this is actually a dichotomised variable - either the support is enough or it is not. If the word "enough" is taken away from the example, the respondent can grade the statement more easily.

Several interpretations of the question's content

The following example shows that it is not easy to detect errors in advance - in this case, despite checking the questionnaire and carrying out test interviews with debriefings. However, a cognitive questionnaire test could have detected this type of weakness.

Example
Have you at any time been at home for at least 6 consecutive months to take care of your own children (including a partner's children, adoptive children and foster children)?

Because a trained observer accompanied the interviewer to several face-to-face interviews, it was found that persons, who had already been at home before they had children, answered both "yes" and "no" to this question. Some interpreted it to mean that they should have been home specifically to take care of the children and answered "no", because they were already at home. Others in the same situation answered "yes", because they were at home and took care of their children. The result of such ambiguousness in answers can make it impossible to know what the statistics mean.

The example is instructive because it shows the limitations of drawing board methods and rules of thumb. Things can be missed, even if nothing seems to be wrong on paper. In other words, to see how questions actually work in a survey, it is not possible to replace or leave out the basic qualitative studies.

5.5 Special information for establishment surveys

Many surveys directed to establishments contain questions on plans, ambitions and attitudes to phenomena in society. The same conditions for question design apply here as for the corresponding surveys to individuals.

For surveys directed to enterprises, municipalities, organisations, etc. in which the main questions are quantitative amounts, the conditions governing questionnaire design are different from those governing the majority of questionnaires to individuals. The data asked for is, in general, taken from accounting or activity reports that the enterprise/municipality/organisation is required to produce. The variables are most often quantities - volumes, amounts, numbers, currency, etc. The respondents are requested to fill in the questionnaire with the help of known terms in the instructions (turnover, delivered quantities, number of employees, etc.).

When the questions are to be answered by amounts of varying size, but in a fixed unit, a guiding format for the answer boxes is an important part of the questionnaire's layout if the answer is to be accurate. The risk for insufficient care, rounding off, unit errors and position errors increases with a badly designed questionnaire, such as too small answer boxes. If the questionnaire is to be scanned, the layout must be adapted to specific technical conditions.

For a questionnaire in an establishment survey, it is often a prerequisite that the questionnaire is compressed. The reason is to reduce the amount of paper to be printed, mailed, and registered, especially if the answers are to be scanned. The space for instructions and explanations in the actual questionnaire is therefore very limited. More or less comprehensive information material is therefore sent out with the questionnaire. This is completely contradictory to measurement theory, which recommends spacious questionnaires with instructions printed

near the questions. Experience shows that information and instructions to a question are more often read when they are printed near the question than when they are in a separate brochure. With electronic questionnaires, it is possible to put the instructions so that they are seen when clicking on the question (see more on scanning in section 5.8).

Sometimes the establishment's data from previous survey rounds is printed in advance on the questionnaire. This is to make it easier for the respondent to answer the relevant questions and to give the answers in correct units. It is not possible to say to what extent such pre-printing works as it is supposed to. There is a risk that undetected errors are preserved and that the respondent does not take the time to review if the indicated conditions are correct or if a change has taken place, for example, if the establishment's main activity is the same as the previous year.

5.6 Questions for comparability over time

Many surveys are carried out for the purpose of comparing the results with those of earlier surveys on the same subject. In such cases, it is obvious to consider using the same questions, in the hope to measure a possible change in the most accurate way possible. But there are three problems that the questionnaire designer must pay attention to when deciding whether re-use and comparison are possible.

- Are the questions still relevant? Are the same response alternatives and instructions appropriate?
- Did the questions give reliable results when they were used?
- Has language changed so that it is necessary to choose other words and formulations so that the questions can be understood in the same way as previously?

Both the reality to be described (e.g. forms of employment and savings) and the attitudes towards the relevant phenomena could have changed. Relatively few older questionnaires are systematically developed and the quality of the answers they give need to be reviewed. It must be asked, for example, how the "old" questions worked. How large was the item nonresponse? Did particular groups have a problem answering this or that question? Did the interviewers have to clarify certain questions or explain them?

If the question formulation is still relevant but there were problems in the original survey, the questionnaire designer and client should review the advantages and disadvantages of using the question again. Is a comparison between inaccurate and possibly even irrelevant measurements at two points in time preferable to a comparison between one inaccurate measurement at an earlier time and one accurate and relevant measurement at a later time? When making this decision, the measurement of change is important. One and the same weakness in the question can be of larger or lesser importance for the accuracy of different estimates.

For international comparisons, in particular, it is not sufficient to translate questions word for word. Attention must be paid to cultural disparities and differences in the countries' social and economic situations. Otherwise there is a large risk of comparing apples and oranges. National differences in how the data collection is organised can also be of significance.

5.7 What help is available?

5.7.1 Standards for classification

Within several areas, there exist national and international standards for the classification of important reporting variables, e.g. economic activity, education, status in the labour force, regional divisions. In addition to these established standards, there are also other recommended or generally accepted classifications and distributions: distribution of persons by age groups, distribution of persons by household units, distribution of establishments by size groups, etc. Such standards should only be deviated from for a very good reason, as the survey then loses its comparability with other surveys and the results can be difficult to interpret for users.

5.7.2 Variable list

The variable list, produced in Phase 1, should form the basis for the questionnaire designer when formulating questions, instructions and response alternatives. It covers the agreed question areas and definitions and provides a great help to the questionnaire designer when he/she is drafting the questions. The questions should be checked against the variable list, in order to see that all the survey areas have been covered. Many times, it is not possible to translate a variable directly into one question. It might be necessary to work both with defining the variable's meaning, with making it more concrete, and with developing a group of questions which together measure the phenomenon.

Example
A survey is to identify persons who are disabled. "Disabled" is therefore a concept included in the variable list. When the questionnaire designer writes questions to measure the variable "disabled" he/she needs to know how "disabled" is to be defined in that particular survey. Does it relate only to disabilities originating from not being able to use the legs? Or should other disabilities due to other causes, i.e. heart problems or obesity, be included? Should "disabled" cover problems with using arms and hands? Because the concept "disabled" has different meanings for different people, it is not sufficient to ask the respondents if they are "disabled"; more questions are needed to decide whether and to what extent the person is disabled. It can, for example, be necessary to ask if the respondent can run 100 metres, walk for half an hour at a fast pace, get on and off a bus, get up from a kitchen chair, etc. depending on the definition used for "disabled".

5.7.3 In-depth interviews, focus groups, checking by experts, etc.

In-depth interviews and focus groups are tools that the questionnaire designer can use in Phase 2 as well. They can be used to test a special phrase, a certain question technique or a specific question area, before the final draft questionnaire is ready for a comprehensive test.

A questionnaire designer, working alone, can easily become blind to defects in his/her own work. Informal tests with colleagues or acquaintances can help to find "unnecessary" errors. Before the final drawing board version is finalised, an expert should check the questionnaire. He/she should check that the questionnaire designer is working according to the current best knowledge, that the language is adapted to the intended group of respondents and that the content is correct from a subject-matter point of view. How the questionnaire and data collection will work so that the production will run smoothly should also be looked at.

The concept "expert panel" is used for a systematic review carried out by two or three persons. The review procedure varies in practice, and the checking can be done jointly or independently. In both cases, the producer must ensure that all comments are taken into consideration.

5.8 Layout

How the questionnaire should look in a purely graphic sense, depends on **how** the questionnaire is to be used and **who** is to fill in the answers. If the questionnaire is to be viewed on a PC screen, either via a floppy disk or the Internet, this implies both opportunities and limitations.

Interview questionnaire
Since many years, a layout standard has been available from the Interview Unit of Statistics Sweden for paper questionnaires for interviews. It is important to follow the standard to facilitate the interviewer's work and make it effective. The main rules are:

- Character font is Arial
- Questions and answers to be read out are written in lower-case letters.
- Response alternatives **not** to be read out are written in CAPITAL letters.
- Codes for response alternatives are written to the left of the questions and are ringed by the interviewer.
- *Italics* are used to show instructions.

An example from the Living Conditions Survey:

Question 112 Do you usually go out socially with any of your current colleagues in your spare time? (With colleagues, we mean persons whom you meet nearly every day in your place of work.)

1 YES, TWO OR MORE
2 YES, ONE
3 NO → *Question 114*

For questionnaires in CATI, there are not many options. On the screen, there is space for one question at a time. WIN-CATI, which is now being developed for Statistics Sweden's interviewers, gives somewhat wider possibilities. It will be possible, for example, to show a table question on the screen. But for fonts and such like, the same standard as in DATI is used, i.e. it is not possible to write in bold, in italics or underline, etc. It means that questionnaire designers should, as far as possible, follow the recommendations which already exist and which the interviewers are used to.

Mail survey questionnaires
There are in practice stricter demands on mail survey questionnaires than on interview survey ones. The respondent sees the entire questionnaire immediately instead of hearing one question at a time. No interviewer is there to argue for participation or to cover possible weaknesses in the questionnaire.

It is preferable that the number of pages in a paper questionnaire is one, two or a multiple of four, considering paper volumes, printing, distribution and registration codes. The respondent should see from the beginning that the

questionnaire is easy to manage and well structured. Divisions into logical question blocks and well-considered selection and positioning of information are used to accomplish this.

There is also a range of formal requirements that must be satisfied but that are not reported in detail in this manual. At Statistics Sweden, the logotype of Statistics Sweden must always be included. A letter of introduction on the survey is to provide information about who is responsible for the survey and how it will be used. The letter should give a contact person with telephone number, fax number and e-mail address, give the last date for submission, call attention to the confidentiality protection and to whether it is voluntary or compulsory to submit information.

For mail questionnaires, there exists no standard layout which is used throughout Statistics Sweden or the System of Official Statistics of Sweden, and the variations are considerable. For the purely technical aspects, the Swedish Standards Institute (SIS) has produced a document entitled "Write at the office, Standards and recommendations for the formulation of documents." For information on how a questionnaire should be respondent-friendly, see Jenkins and Dillman (1997).

Design the questionnaire with plenty of space, don't crowd the questions. This facilitates **navigation** - i.e. how the respondent follows the graphic symbols in the questionnaire to answer in the right order. The respondents should not have to think about the direction but should see immediately where the next question is. They should not have to go vertical sometimes and horizontal other times when looking for the spaces in which the data should be written or where the response alternatives are. Difficult navigation increases the risk for item nonresponse.

It is often very difficult to set a question without defining a concept or the scope of the question. The **instructions** that the respondent needs to understand the question correctly should be in the actual question or directly following it. Questionnaires to establishments sometimes require more comprehensive instructions. These are frequently collected together in a separate paper. Many respondents start filling in the questionnaire without consulting the instructions until they get stuck. A better format is to place the questions on the right page of a spread and the instructions on the same level as the questions on the left page. This avoids the risk that the instructions are lost when the respondent comes to fill in the questionnaire.

Skipping instructions are often misunderstood in mail questionnaires. There should only be a few and they should be simple and with a very clear graphic marking for where the respondent is to go. Question tables have been considered demanding to fill in. Groups of cells can easily be passed over and lead to item nonresponse. Both skipping instructions and question tables need to be examined when the questionnaire is tested.

The character font in forms for mail questionnaires is usually **Arial, 10p - bold** for questions and Arial, 10p normal for response alternatives. The introductory letter and comprehensive instructions are written in Times New Roman, which is easier to read. Specific instructions are written in *italics* in Times New Roman.

Questionnaires to establishments, municipalities etc. are often printed in Arial, 9p and Arial, 8p. The questionnaires are compact and, for many, difficult to read. It is possible that such questionnaires once were designed to be filled in with a

typewriter. As there are hardly any typewriters left nowadays, these questionnaires will be filled in by hand and the space is then insufficient.

Certain questionnaires have both the questions and the response alternatives in a **column**. This can give an impression of lightness and be easier to manage in manual data registration.

Example 1
The same questions with one and two columns respectively

Questions about yourself and your household - in one column

1	**In which year were you born?** Year 19 .....
2	**Are you male or female?** 1 ☐ Female 2 ☐ Male
3	**Are you married/cohabiting or single?** 1 ☐ Married/cohabiting 2 ☐ Single

The second questionnaire is written with the question in the left column and the response alternatives in one, or sometimes several, columns on the right. Two (or more) columns save space and allow for more questions on the same page. The division into one question column and one or more response column(s) makes it easier to navigate and to get an overview. The respondent does not need to lower his/her hands each time he/she has written an answer in order to read the text to the next question. Less time is needed to answer if there is a good overview.

Questions about yourself and your household - in two columns

1	**In which year were you born?**	Year 19 .....
2	**Are you male or female**	1 ☐ Female 2 ☐ Male
3	**Are you married/cohabiting or single?**	1 ☐ Married/cohabiting 2 ☐ Single

Example 2

Conditions of small businesses - in two columns

3	**Is the business of your establishment run as a cooperative?**	1 ☐ Yes 2 ☐ No
4	**Does the establishment have a male or a female manager?**	1 ☐ Female 2 ☐ Male
5	**What are the establishment's main activities today?**	....
6	**How large was the yearly turnover, excluding VAT, in 1994?**	*Amount in SEK thousands* .....

Example 3 Several columns

PRODUCTION OF COMMODITIES AND INDUSTRIAL SERVICES IN SWEDEN 1998

1. Income

	Distribution of income of the survey unit		Value 1997, SEK thousands	Value 1998, SEK thousands
01	**Industrial activities**			
02	**Sales of finished products manufactured at another unit within the establishment**			
03	**Other activities**	NACE code		
04	**(Non-industrial**	NACE code:		
05	**activity)**	NACE code:		
06		NACE code:		
07		NACE code:		
08		NACE code:		
09		NACE code:		
10		NACE code:		
11	**Total net turnover including internal deliveries**			
12	**Internal deliveries**	for further processing		
13		of finished products the establishment has produced		
14		of goods or services from non-industrial activities		
15	**Total net turnover excluding internal deliveries**			

The shaded fields on the above questionnaire are pre-printed if the information is available. Character font is Arial 8 p.

Codes are sometimes to the left and sometimes to the right of the check box. If the number is on the left, the text will be close to the check box. This is preferable.

Example 4:

1 ☐ 0-25 km	
2 ☐ 26-50 km	Distance between rows often 3 points.
3 ☐ 51-100 km	
4 ☐ Over 100 km	
5 ☐ Don't know	

There are **templates for questionnaires** that are written with **Crystal Reports**. Traditionally, many questionnaires are written in PageMaker. Questionnaires can also be written in **Word**. If two or more columns are preferred, use *Insert table* under *Table*. The check box is found under *View* - *Tool bar* - *Forms*. The box must not be too small. It is possible to mark and make it larger by selecting a larger character size, for example 12 p (4 x 4 mm).

If much material is to be processed, **scanning** is a good alternative to manual registration of data. A + is inserted in each corner. The boxes must be at least 4 x 4 mm, and should not be too close to each other (see registration value). When scanning, the registration values can be written at the top above a column with reply boxes. It is not necessary to repeat the values in each box although sometimes it is preferable to facilitate checking.

1	2	3	4
☐	☐	☐	☐
☐	☐	☐	☐
☐	☐	☐	☐

In case **colours** are to be used in certain fields, these colours should be pale. Sometimes the answers are faxed, and if bright colours are used, the transmission may result in difficulty to read the answers. Nor are bright colours suitable when scanning. Special attention to the capacity of the scanner must be paid. Be sure to always check how a questionnaire with colours will stand up to being faxed and scanned!

If the respondent has an un-bordered area to write in, i.e. _____, risks for errors in size and unit increase. Using well-designed reply boxes and clearly stating the unit help to avoid mistakes in units and items. Post giro and bank giro forms require the highest level of accuracy, and have a reply box for each digit. The digits are also in groups of three as follows:

☐☐☐ ☐☐☐ ☐☐☐, ☐☐

5.9 Checklist

CHECKLIST Designing questionnaires

1. Find out which national and international standard classifications need to be followed to meet requirements for comparability with other statistics.
2. Formulate the questions using previous experience.
3. Adapt the questionnaire to the collection method and other production conditions (for example how the data will be registered).
4. Improve the first version of the questionnaire by
   - checking by a colleague at the drawing board
   - informal testing of the questionnaire by several colleagues or friends. However, this testing can never replace testing with actual respondents.
   - using an expert panel to go through the questionnaire.
5. Check the language in the questionnaire and the information material.

The final product of Phase 2 should be a questionnaire that is complete and technically correct. To make correct estimates, background variables and required information should be included. The questions should be adapted to the data collection method and the layout for data registration plans.

This does not ensure that the questionnaire is suited to the capacity of the respondent and his/her willingness to reply. Nor is it known how the questionnaire will function in production.

References

Råd & definitioner för Företagsstatistik 1997(Advice & definitions for Structural Business Statistics) 1997. Statistics Sweden
Bergman, L.R. and Wärneryd, B. (1982). *Om datainsamling i survey undersökningar (About data collection and surveys)*. Stockholm: Statistics Sweden and Liber.
Wärneryd, B. (1989:) *Att fråga*. Statistics Sweden
Salant P and Dillman D. A. (1994). *How to conduct your own survey*. John Wiley & Sons, Inc.
Fowler F.J. (1995) *Improving Survey Questions.* SAGE Publications.
Developing and Using Questionnaires (1993) from General Accounting Office, Washington, USA
Survey Measurement and Process Quality. (1997), (edited by Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwartz and Trewin). John Wiley and Sons, Inc.
Jenkins, C.R., and Dillman D.A: *Towards a Theory of Self-Administered Questionnaire Design.* Chapter 9 in SMPQ.
JOS (1985). *Special Issue on Questionnaire Design*. Vol1. No.2 , 1985
JOS *(*1992) *Measurement Errors in Surveys Part I and Part II.* Vol. 8 1 and 3, 1992.
Skriva på kontor, STG handbok 126.. Published by SIS in April 1995.
STG = Swedish general standards institution. SIS = Swedish standards institute

6 Phase 3 - Cognitive tests

If you cannot afford to pilot your study, don't do the study! (Seymor Sudman)

6.1 Contents

Task: To obtain qualitative information on how a drawingboard questionnaire is understood and answered by actual respondents. The questionnaire is then revised to eliminate deficiencies. If many weak points are found, the revision may be so extensive that even the contents must be revised, i.e. Phase 1 must be done again together with Phases 2 and 3.

Cognitive tests (qualitative) are used to find errors in the understanding of the questionnaire, relevance errors, technical errors and difficulties and deficiencies in handling the questionnaire. The tests are mainly directed towards understanding the cognitive process as described in section 2.3 under the heading *Cognitive model for the response process.* They are diagnostic in the sense that they can find the reasons behind errors and give indications on how the questionnaire should be revised to give the best possible exchange of information with the respondents. Cognitive tests primarily make use of observations and in-depth questions on small non-probability samples. In the case of mail questionnaires (especially to establishments) it is necessary to study how the questionnaire reaches the "right" respondent and how this person gathers information and/or calculates the data.

Choice of test method
Cognitive tests with qualitative results and experiments with quantitative results (Phase 4) form the basis for different types of decisions and complement each other. Cognitive tests are significantly cheaper and take less time to carry out than experiments. Experiments that are not preceded by cognitive tests are nearly always a waste of time and resources. Cognitive tests help to discover the factors that affect the contents of the answers. A cognitive test improves the conditions to check problematic factors and formulate better hypotheses on the measurement characteristics of different versions of questionnaires and data collection methods.

In Sweden, establishments are not required by law to reply to pilot surveys. If pilot surveys are to be conducted, permission is needed from the Board of Swedish Industry and Commerce for Better Regulation (NNR). Because of the considerable risk for low participation rates and lack of interest, it is difficult to conduct reliable experiments with establishment surveys. Small qualitative studies with interested respondents are in practice often the only realistic alternative.

6.2 Cognitive tests in summary

Problem analysis
Problem analysis involves evaluating concepts, questions and instructions that can be especially demanding for respondents to respond accurately to. Problem analysis can also assess if some group among the respondents could run into problems with a special section of the questionnaire or in general have more difficulty than others. Hypotheses are formulated on how data is actually formed

and collected. Difficulties that can appear and the reasons behind them can be discovered. Hypotheses can be formulated about

- difficulties in understanding questions and instructions
- difficulties in obtaining information
- effects of the order of the questions, design and layout of the questionnaire
- motivation, conditions and attitudes of respondents towards the survey
- training and actions of the interviewers in the case of interview surveys

Often, certain survey variables are especially important (key variables) and require a high measurement quality. When testing the questionnaire, these variables should be given much attention, even though they may not be among the most difficult to measure. A high level of accuracy in answers that identify the survey unit is also important. Questions used for distribution of survey units in important reporting groups and selection questions must be clearly stated.

Planning and organisation
A test leader is responsible for planning, administration and analysis. The tasks of the test leader include: choosing the test method, defining the tools in the test, giving instructions to the interviewers, documenting how the survey is to be carried out and the results to be achieved. A small number of interviewers will assist the test leader. The test leader and the interviewers recruit the test persons. Since cognitive interviews are in many ways different from standard ones, the interviewers should have received special training to observe, ask in-depth questions and document the viewpoints of the respondents on the questionnaire and their reactions during the testing.

In 1989 the Measurement Laboratory (ML) was established as a network within Statistics Sweden. This network includes social scientists, behavioural scientists, interviewers and statisticians with special know-how about statistical surveys. The Methodology Unit in the Research and Development Department handles the coordination. The folder *Statistics Sweden's Measurement Laboratory - helps you off to the right start with your surveys* describes the activities.

Location for the test
There is a distinction between **field tests** and **laboratory tests.** Field tests are those that take place with the respondents at their workplace. Field tests are usually the best and sometimes the only alternative to test questionnaires and collection procedures of establishment surveys. Cooperation between respondents and others at the establishment must be mapped to understand how the information is collected. Laboratory tests are conducted on premises prepared for measurement studies. Audio and video recording equipment are sometimes available, and sometimes the test leader can monitor how the test is conducted. Testing of questionnaires in surveys on individuals sometimes require a laboratory environment so that the testing is not interrupted by telephone calls or visitors dropping in or influenced by members of the test person's household.

Tools
Cognitive questionnaire testing is done with one respondent at a time. A number of different tools are used, such as different types of in-depth questions and observations. The cognitive tools are described in more detail in section 6.4. The choice depends on the advance knowledge and hypotheses about the questions and the type of difficulties expected with the questionnaire. There are also tools to identify difficulties that the test leader is unable to anticipate. The tests are designed and the tools are adapted to the chosen data collection method and

category of respondents for the planned survey.

Conducting the tests
A test interview begins by the test person reading the information material. Then he/she fills in the questionnaire if it is a mail questionnaire, or replies to the questions if it is an interview survey. Meanwhile, the interviewer watches what is going on but is careful not to disturb or influence the test person. Important observations are noted down and the interview is taped. When the interview is finished, the interviewer conducts a systematic questioning (debriefing) about how the test person felt about the questions. An informative test interview requires that the test person is alert during the whole test and is motivated to express viewpoints. Therefore the interview should not take more than one hour.

When testing establishment questionnaires it would be too time-consuming for the test person actually to collect the requested information. Often this information is not even available at the time of testing. Instead, a study is made of how the information is gathered, which persons are involved, where the information is obtained, what kind of calculations are needed and how much work is needed. (See 6.5 Testing questionnaires for establishments.)

The working methods of the interviewers is evaluated later on
The interviewers participating in the cognitive tests have received special training and been selected because of previous successful experience of standard interviewing. Even when resources are available for a special observer, his/her observations would not say much about how well the group of interviewers as a whole would use the questionnaire and the instructions. These kind of studies are used in larger experiments, Phase 4, or when adapting to production conditions in Phase 5.

Documentation
The interviewers report each test interview separately according to a specific model. The report is very detailed. When the results are compiled, the test leader must carefully differentiate between the spontaneous reactions of the test persons, how they react to in-depth questions, the observations of the interviewers and their conclusions. The evaluation of the test results is done by the test leader in cooperation with the person who will revise the questionnaire. The test protocol does not contain any information that identifies the test persons. The protocol is destroyed when the test and the revision have been completed.

Examples of discoveries through cognitive tests
The examples are from the Bureau of Labor Statistics (BLS) and National Centre for Health Research (NCHR), where cognitive studies have been made with so many persons that quantitative results are useful.

Other classification: *BLS discovered that about 40% of the respondents chose another alternative for their main employment than the one that was correct according to the definition of BLS. As a criterion, BLS had the number of hours worked during a given period, while the respondent regarded him/herself in a more long-term perspective. For example, one person considered him/herself as a student even though he/she worked 30 hours at a fast food restaurant.*
Professional terms differ from everyday language: *The NCHR found that the question 'Have you had serious pains in your abdomen during the last three months?', did not function well for two reasons. Very few of the respondents knew exactly where the abdomen is located. The word "serious" led in some cases to a very high tolerance level for pain.*

One word misunderstood: *BLS first asked if the respondent had read at least one novel during the previous time period. Then the respondent was asked a control question about the titles of the books. About 20% of the said readers of novels replied by listing non-fiction.*
Missing response alternatives: *ML had an assignment to test a questionnaire on alcohol consumption. It proved to be important for large consumers to be able to state that there were certain types of alcohol they did not drink at all. They were not satisfied with a low-consumption alternative and wanted an absolute zero alternative as well.*

6.3 Recruiting test persons

Cognitive tests are done with a number of **test persons,** who in some sense are representative for respondents in the planned survey. The test person is the one who submits information about a **test object**, which can be the test person him/herself, or an enterprise, farm, real estate etc. The number of test persons is small. In the standardised tests at Statistics Sweden, the number has varied between 6 and 15. Even with such a small test it is possible to discover deficiencies in communication with the respondents. However, more test persons would be required to discover some of the more unusual mistakes made by different groups of respondents.

Test objects and test persons **are recruited** in various ways. Telephone directories, membership lists, information from the client, friends and acquaintances are used to find and recruit suitable persons. Travel costs and some form of compensation (*incentives*) are paid to the test persons for their participation. Private persons are usually given lottery tickets, while establishments are preferably offered a sum of money donated in their name to a non-profit organisation of their choice.

A considerable amount of skill is needed to recruit representative test persons so that the test will be informative. Test persons should be "actual" respondents in the coming survey. For surveys on individuals, persons are often chosen to represent groups that are assumed to have some difficulty in handling the questions and submitting the information. Different groups such as those broken down by age can have different cognitive characteristics: questions can be interpreted differently, different factors may be taken into account, or information may be processed in different ways. Sometimes when the differences are assumed to be very great, it is a good idea to do different tests with the groups.

In establishment surveys where both the variable values and their weights in estimates vary considerably, test objects can be recruited. The answers of the test objects make large contributions to the estimates and/or the errors in estimates.

Probability sampling is almost never used, mainly because it seldom gives the distribution of the sampling that is needed. It would be too expensive and time-consuming. Some typical recruiting methods (non-probability sampling) are:

- The sample is drawn with help of personal contacts or from lists that are easy to access, such as a telephone directory or a membership list. *(Convenience sampling)*
- The sample is drawn on the basis of an "expert opinion" that the sample persons are typical or have some set of characteristics that are important in the survey. *(Purposive sampling)*

- It is decided how many units that are to be included in each subgroup, but the interviewer will draw the sample him/herself. *(Quota sampling)*

Another useful but risky method to recruit test persons is to run an advertisement. It is also common to use students or the same test persons in a series of tests. The weak point of the first two methods is that the test persons are not representative for the respondent population. The third method falls short, because after the first test, the test persons begin to regard themselves as experts.

The grade of recruitment is the ratio between the number of test persons that participated in the test and the number of persons that were asked if they would participate. (The concepts response rate and nonresponse rate are not especially meaningful unless the sample is a probability sample.) A recruitment grade of around 50% is often possible for a skilled recruiter. An exceptionally low recruitment grade indicates that the subject for the study is either not interesting or too sensitive, and thus predicts risks for high nonresponse in the coming survey. There is no "refusal processing". It is pointless to persuade reluctant test persons. Experience shows that even if they are persuaded to participate, they will submit very few viewpoints in the questionnaire.

Participants in a qualitative test need to devote more time than respondents in a regular mail survey. They also need to be open about the background of their answers. The conditions are especially different when testing questionnaires for mandatory establishment surveys.

It is reasonable to assume that those who become test persons have a greater interest in the subject area of the survey and a better language ability than those who do not want to participate. In other words, they have fewer problems in understanding words and formulations. If this assumption is true, the capacity of the average respondent to give reliable answers is overestimated. Knäuper and others have shown how deficient cognitive capacity is interrelated with more “don’t know” answers, more item nonresponse etc. Those with poor ability more often misinterpret questions, and there is a risk that this group is not represented in cognitive tests.

6.4 Tools

It is not enough just to observe how the questions are answered to determine if they have been answered correctly or incorrectly. Often, respondents reply quickly and convincingly without having understood the question correctly or being sure of the answer. To determine when and why this occurs, a series of measurement tools are used that are combined in different ways depending on what kind of difficulties are anticipated with the questionnaire. Examples of often-used tools are:

6.4.1 Probes

When the questionnaire is filled in or the interview is carried out, the interviewer asks the respondent a number of **probes** regarding his/her understanding of the questions and how he/she arrived at his/her answers. Probes can have different purposes and are used in various ways. The answers to probes are unstructured.

Common probes are directed towards all test persons. They are formulated with a point of departure in a hypothesis about what the cognitive difficulty of the question is. When the direction is set, different questions can arise:

- questions about how the test person has understood the question and the instructions *(Comprehension probes*).

For example, the Current Population Survey (CPS) asked about employment during the previous week. According to CPS, the definition included the week as from Sunday up to and including Saturday. Probes on how the respondents considered the week to be was as follows:

17% *Sunday up to and including Saturday* ***(correct)***
54% *Monday up to and including Friday*
9% *Monday up to and including Saturday*
6% *Monday up to and including Sunday*
4% *Sunday up to and including Sunday*
10 % *Other way.*

- questions on how the test person arrived at his/her answer (*Information retrieval probes)*. The test person may have answered directly, checked notes or records, calculated the answer, or given a standard answer.

For example, CPS asked in-depth questions on how the test person arrived at the number of hours worked during the previous week: "Did you know immediately how to answer or did you need time to think about it?", "How did you arrive at the answer?"
66% replied they always worked the same number of hours. The remaining persons replied as follows:

40% took an average value for several weeks
36% considered a typical week
24% calculated in another way

-questions that are intended to find out if words, phrases, situations are understood by the test person in the way the questionnaire designer intended *(Frame of reference probes).*
For example "What does 'usually' mean to you?". "When do you feel an illness should be defined as prolonged?"

-questions on how the test person has suited his/her answer to the available alternatives. (*Response category selection probes*)
For example "Why did you choose that particular answer?" "Was there any other response alternative you thought about?"

General (common) **probes** are generally formulated questions. They are used to find difficulties in specific questions that the test leader could not anticipate.
For example "Can you tell more about this?" "Did you find the instructions easy to read?" "Was there any part of the questionnaire that was difficult to follow?"

Random probes are when all interviewers randomly choose a number of questions for general probes. Random probes are not identical with common probes, they vary among the test persons. They are also meant to catch difficulties that the test leader could not anticipate.
For example "Was there anything that was not clear in this question?" "Was it easy to reach your answer?"

Special (specific) **probes** are those that the test interviewer him/herself formulates and uses when, during a test, he/she gets the impression that a test person has difficulties with a particular question. These are important elements in the test and make high demands on the interviewer's ability to make observations and improvise probes.
For example "I noticed you hesitated before answering question 35. Why?"

Different schools prefer to make probes either during an ongoing test **(concurrent use)** or after the completed interview/mail questionnaire **(retrospective use**). The reason for using probes directly when a problem or hesitancy has been noted is that the test person does not risk forgetting what the problem was. On the other hand, the test will be less like the normal interview situation, and there is a risk that the test person, after a few probes, will begin to regard him/herself as an expert. By making the probes after the interview/ questionnaire has been completed, the situation becomes more similar to the actual survey. At ML, probes are used afterwards.

6.4.2 ”Think aloud" method

The "think aloud" method is a useful tool but more difficult to handle than probes alone. It takes longer to carry out and requires additional training of the interviewer. Before the test, the interviewer demonstrates how to think aloud and then lets the test person practice on a short test questionnaire before the actual test begins.

With the think aloud method, the test person continuously makes comments about the filling in of the questionnaire while doing so. These comments are recorded on tape. Since the interviewer is passive during this phase, the test person expresses spontaneous viewpoints and finds difficulties that the test leader could not foresee. This method is best suited for mail questionnaires, but can also be used in interview situations. It is followed up by probes.

6.4.3 Other tools in cognitive tests

Some other established tools which have not yet come into use at Statistics Sweden are:

- **Paraphrasing,** where the test person is asked to re-formulate a question and express it in his/her own words. The result reveals the actual understanding of the context and not just of each word itself.
  For example: *The following question is found in the questionnaire: ".....?" How would you yourself formulate this question?”*

- **Vignettes (**sometimes called "scripts”) are case descriptions of different situations (for example, employment conditions) that the test person is asked to classify according to the response alternatives on the questionnaire. Vignettes are used for borderline cases when it is not certain that the everyday meaning is the same as the formal one. The tool can be useful in Phase 1 as well, together with in-depth interviews or focus groups.
  Example: "A person works a few hours each week in his father's establishment. In return he/she receives room and board at home. In your opinion, is this person employed?"

- ***CARD sorting*** (*free sorting, dimensional sorting, vignette sorting*) is based on how the test person associates and groups together different concepts.

- **Measuring the accuracy of the test person's answers.** (*Confidence rating)* The test person answers by choosing a point on a scale given in advance.

Examples: ”How sure are you that you made the purchase after the first of the year?", "Do you think that the trips you have written into your diary are all the ones you made last month?"

- **Measuring the reply time** *(Response latency)*, i.e. the time that passes from when the question is read to when the test person has replied. The reply time indicates how much effort the test person needs to make to arrive at his/her answer. The indicator is best suited for interviews that can be conducted without interruption and especially for computerised interviews when time measurement is simple. It can also be used with tests of mail questionnaires if the interviewer can see how the questionnaire is filled in. By using intelligent questionnaires we can obtain a measurement of time taken. However, the respondent may have a number of reasons to take breaks.

6.5 Testing questionnaires for establishments

Preparatory work
More preparatory work is required when testing questionnaires for establishments than when testing questionnaires for individuals. If the size of the establishment has considerable bearing on the results, this has to be taken into consideration when choosing test objects. The situation is more complex than in surveys on individuals, and the test planning must be more flexible.

Preparatory contacts need to be made to clarify the relationship between the test object and the respondent, i.e. where the information is found and who will submit it. Information about the workplace is usually available at the main office of the establishment, sometimes about each separate workplace and sometimes only about the establishment as a whole.

In establishment surveys, the statistics producer needs to have an understanding about the position and competence of the respondent, and about how the information is collected and/or calculated by the respondent. Different categories of respondents have different capabilities to give correct answers. It is especially important to make the right choices when asking about coming investments and the need for recruitment, and not about written records. It is not always clear for the establishment who ought to be the respondent, and sometimes the chosen respondent does not have sufficient insight to answer.

CHECKLIST: How the questionnaire reaches the respondent

1. How is a respondent identified/defined at the right level/with the right knowledge?
2. How precisely can the "right" respondent be addressed?
3. Which gatekeepers, i.e. mail clerks, secretaries, receptionists and others mainly answer telephone calls and handle the post?
   How correctly and quickly can these persons identify the "right" respondent and know who shall handle the telephone call/mail delivery?
4. Is the information asked for in the questionnaire really available at the establishment and not at an accounting firm or other workplace?
5. How many and which persons need to cooperate to gather the information?
6. How much work and time does the respondent need to compile and prepare for filling in the questionnaire?

In multi-round surveys, it is an advantage to have the same respondent each time. After this person has learned the concepts and instructions for the questionnaire, the accuracy of the answers should become better. When a regular respondent is replaced by a temporary one in the case of business trips, holidays and illnesses, there is an increased risk for measurement errors, nonresponse or delays. Now and then a permanent change in respondents occurs. Both the level of competence of the respondent and the change of respondent in multi-round surveys are indicators of accuracy and risk for errors in information from an establishment.

Contents are sometimes fixed
The results of the survey will often be used as input to accounts, an index or some planning model, and the variables to be included in the establishment questionnaire are accordingly fixed. Directives and international recommendations can also define the contents. Changes in contents might thus not be possible, even when a test shows that the variables cannot be measured with sufficient accuracy. Another restriction is when changes in the questionnaire require permission from a body outside the national statistical institute (in Sweden the Board of Swedish Industry and Commerce for Better Regulation (NNR).

Field test
Tests of questionnaires for establishments are usually carried out in the field, since this is the only way to create realistic testing conditions. The respondent needs access to accounting and other information - perhaps from the administrative systems of the establishment. Sometimes cooperation by several persons is needed. The necessary compilations and calculations could take days or even weeks to do, and are costly for the establishment, but a questionnaire test at an establishment should not take more than one hour of the establishment's time. Thus it is not possible to monitor and observe all of the preparatory work before the questionnaire is filled in, which could be very extensive. Nor can it be it taken for granted that the information is available at the time of the testing. Instead, the test should focus on how the information would be submitted, or, in multi-round surveys, how the information is usually submitted.

Complex and varied situations
The situation for respondents varies considerably more among establishments than among individuals. The accounting systems and accounting principles, the level of computerisation and the administrative models vary.

Even the organisation of activities in establishments can differ radically. Different questions are answered by different groups of respondents. Different versions of questionnaires are needed to avoid too many skipping instructions. This can be achieved by programming computerised printouts.

Most of the questions concern quantitative measurements. The answers can differ considerably in size, both between the various respondents and between different variables, resulting in risks for "unit errors". This occurs when a respondent thinks of one unit and does not notice that the questionnaire asks for another.

Because the situation for respondents varies, the test must be able to identify and adapt to these variations. More preparation for alternative situations as well as flexibility and ability to make observations is required.

Use of advance information in multi-round surveys
Many surveys are repeated periodically. Helpful information, both documented and informal, may be available with the producer about the accuracy of the results from previous rounds. This information can be used to direct the test with regards to:

1. Questions and sections in the questionnaire that have caused difficulties
2. Questions and sections that are especially important
3. Sub-groups or even individual establishments that have had some special difficulty in submitting correct information
4. Groups of respondents whose information is especially important in the survey

When the same sample is used on several survey occasions, the checked and corrected answers from the previous survey round(s) can be printed on the questionnaire to guide the respondent in the compilation of the new answers.

Time for revision of questionnaire
The client would like to be able to sum up a survey that is carried out monthly or quarterly on a yearly level. The client is also reluctant to let the producer introduce a revised questionnaire at some other point in time than at the turn of the year. Therefore, a decision to make a revision in the questionnaire must be decided well in advance of the turn of the year so that several monthly and quarterly survey rounds can be used to evaluate the old questionnaire and develop a new version.

CHECKLIST: Type of probes for testing questionnaires for establishments

- When are the instructions read, and how are they used and understood? Which instructions are in the "right" place and are read, and which instructions are not read?
- Are there technical terms and concepts in the questionnaire - defined or otherwise - which the establishment does not use or understand correctly?
- Does the questionnaire use concepts and names consistently?
- Does the test person understand the questions correctly?
- How and from which sources is the information collected?
- Is it at all possible to obtain all the information?
- Is the information reported to the degree of detail needed for the statistics?
- Is the periodizing of the statistics the same as that of the establishment?
- In practice, how are recalculations and other preparatory work done?
- How does the respondent estimate answers when information is not available in the accounts, for the desirable period, or on the right organisational level? How uncertain are the estimates?
- How does the respondent orient him/herself in the questionnaire? Are there any questions or information that are overlooked?
- How long does it take to fill in the entire questionnaire and the various parts of the questionnaire?

6.6 Evaluation of test results

An obvious way of confirming accuracy in a test would be to repeat it, preferably with an independent test leader and interviewer. If the repetition leads to the same conclusions as on the first occasion, you have at least strengthened the confidence in the results. As this is hardly possible since it would mean an extra expense for the client, it is necessary to rely on process controls and agreement

with previous experience. There is extensive empirical knowledge available on the capacity of different qualitative methods and how they can be used to complement each other.

To gain insight into the quality of a cognitive test of a questionnaire, the test leader should have reported:

- competence, organisation and experience of the testing organisation
- number of test persons and how they are distributed into "interesting groups"
- criteria for recruiting test persons, recruiting method and extent of participation
- the general design of the test and testing tools
- required time for the test
- detailed information of the test results, so that a user can independently take a stand from the conclusions.

To gain a general understanding of the test results, the following simple questions can be of help:

- how is the number of comments and observations distributed on the test persons
- how are the comments allocated on different parts of the questionnaire
- how are the spontaneous views distributed, and how are they distributed by probes
- what type of questionnaire deficiencies are indicated
- how do the results correspond to the hypotheses and to measurement experience.

An irregular distribution of comments could indicate that a few test persons have been overly critical or disinterested, that the questionnaire was tiring or non-engaging in some sections or that the test process has not worked as planned. If the type of comments varies substantially, it might indicate that the questionnaire has technical flaws or has been cognitively difficult.

Each observation must be evaluated separately, taking measurement experience into consideration, since in-depth questions are open and the answers not standardised. When the test is done with only about ten test persons and they all have been recruited from different backgrounds, they are not expected to notice the same weaknesses in the questionnaire. Nor is such a small number of test persons enough to discover all the major difficulties in the questionnaire. Some defects could pass without being noticed. With 30 to 40 test persons, it is much easier to find both the common problems and problems that are less common.

The fact that a test method reveals flaws in a questionnaire is not sufficient in itself. A revision of the questionnaire has to be made as well. Once the language problems have been exposed (incorrect choice of words, logical errors, missed response alternatives or skipping instructions) it is usually obvious how to improve the question. Changes that concern concepts can be more difficult to arrive at. Just because a faulty question has been modified, the new question will not automatically be without opposition. However, if fewer and less important difficulties are encountered when testing the revised questionnaire, the quality of the test and the revision can be assessed as good. Providing time is sufficient, it is more efficient to do two or more tests with a revision in-between than doing one large test and then using an unverified revision.

The qualitative results only give a first indication of what kind of data quality can be obtained in the survey. An improved questionnaire is best measured by how much the accuracy of the estimates and the efficiency of the production

increased by what was revealed, both in the test and in the follow-up revision. When the changes considerably affect the burden on respondents and it is uncertain what the effects will be on nonresponse and costs, a quantitative experiment is justified.

6.7 Overview, checklist and results

To the statistics producer, the questionnaire may seem perfect on the drawing board, but it is not certain that it is as perfect for the respondent. He/she may not recognise concepts and definitions, instructions may be difficult to understand and hard to find, some information may be difficult to find or require too many recalculations. The layout may appear so vague that the respondent does not know which questions to answer.

Cognitive tests can reveal how a respondent understands concepts, questions and instructions in a questionnaire, how he/she collects information from records or memory, processes and organises it and fits the answers to the response alternatives. They can show the questions that have been misunderstood, how the misunderstanding has arisen, and where the demands have been too high on the ability or desire to process and submit information. The tests can also discover technical errors (e.g. a lack of skipping instructions) and content errors (e.g. insufficient response alternatives) in the questionnaire. The client is then forced to either exclude variables that are difficult to measure, or accept measurements with a large uncertainty. In case the work to remove these errors in the previous phases has not been well done, attention will remain on these errors. The errors that cognitive tests are especially suited to discover might then remain undiscovered. A test-retest procedure is necessary if the first test reveals insufficiencies that require considerable revisions.

CHECKLIST Cognitive tests

1. Identify the sections and variables/questions in the questionnaire and the layout aspects that might be difficult for the respondent to handle.
2. Evaluate the variables by respondent burden and cognitive degree of difficulty. Evaluate the total degree of difficulty of the questionnaire and how it should be tested. Estimate the total burden on respondents.
3. Formulate hypotheses about the cause of the difficulties.
4. Choose which test method and tools to use for the test, e.g. cognitive test questionnaire with special in-depth questions (probes) "think aloud" methods or in-depth interviews.
5. Define the desired characteristics of the test group, decide the number of test objects, premises for the test and define the principles for recruiting test persons.
6. Conduct the test. Test the understanding of and the use of the introduction letter and attached information.
7. Make a detailed documentation of the test showing the hypotheses that have been used and report any unforeseen difficulties.
8. Improve the questionnaire where deficiencies are discovered. When essential changes are required, test the revised questionnaire again (several times if necessary)

Result: A revised version of the questionnaire, suited to the respondents' ability to understand questions and instructions, and to their capacity and desire to answer. After this phase, the questions are adapted to the data collection method,

but the questionnaire is normally not technically ready to use in the production of a main survey.
On the other hand, at this stage it may seem impossible to measure the survey variables with sufficient accuracy, and therefore the survey should be cancelled.

That a questionnaire has been tested and revised by best methods does not automatically mean that it will function flawlessly for all variables. Some variables are quite plainly difficult to measure, no matter how the questions are formulated. In spite of this fact, these variables might be kept because the client wants the information with the accuracy level that can be expected, rather than being without the information at all. The results of the cognitive test will then include a warning about the use of these questions, and suggestions on how to monitor their accuracy in the later phases of the process.

References

Oksenberg, et al. (1991) *New Strategies of Pretesting Survey Questions* JOS 91:3
Knäuper B, et al. (1997) *Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality.* JOS, Vol. 13, No. 2, 1997.

7 Phase 4 - Experimentation

7.1 Contents

Task: When a questionnaire has been qualitatively tested and revised, it normally needs to be tested quantitatively to see if it is good enough to use for statistics production. The questionnaire and method of data collection need to be checked to see if the accuracy matches the needs of the client. Are there any unwanted side effects such as a high rate of nonresponse or high costs? Sometimes it is necessary to compare two or more proposed solutions to see which one is best, considering their effect on a number of factors such as measurement quality, nonresponse, amount of time needed and costs.

7.2 Methodology

The methodology is traditional experiment planning, i.e. probability sampling dimensioned so that hypotheses can be tested by estimating a methodology effect or difference with enough precision to make decisions. Before conducting a single-round survey, **independent** experiments are done. In multi-round surveys, measurement effects can be assessed by **embedded** experiments, i.e. the new method is introduced in a random subsample of the whole survey. As an alternative, a supplementary sample can be used.

An experiment evaluates how the information generally is collected, but cannot reveal the respondents' difficulties with separate questions. Because experiments are considerably more costly and take longer time to do than cognitive tests, it is a waste of time and money if cognitive tests are skipped and questionnaire deficiencies not discovered before this phase. Therefore the questionnaire versions used in experiments and evaluations must be cognitively tested and revised beforehand. Due to restricted resources and the time pressure in statistics production, experiments are used less and less as a basis for decisions.

Independent experiments can be done as preparatory work for a new survey or in parallel with a multi-round survey that needs revising or evaluating. They can be designed without being bound to an old collection method. There is a risk, though, that independent experiments are not taken seriously by respondents, especially those who normally participate in surveys with mandatory reporting.

When experiments are carried out within the framework of the normal survey production, there is also a risk that the organisation gives lower priority to experiments than to "real" surveys. If an experiment requires the service of interviewers, there may be a conflict because the interview capacity is often required in large multi-round surveys. The complexity of the test, e.g. how many factors are being tested, must be weighed against the risk that the test plan may be jeopardised by unforeseen prioritisations.

In multi-round surveys, an experiment can be **embedded** to one or more production rounds. In surveys where participation is obligatory, this is often the realistic alternative and advantageous with regards to response rate, response quality and time frame, since both the producer and the respondent will take it seriously.

However, there is a conflict with embedded experiments. When the results from the experimental part of the sample significantly deviate from the results from

the part of the sample using the usual questionnaire, the results from the experimental part cannot easily be used in the estimation. At worst, the test results must be discarded. The number of sample units that can be used for the estimates in the main survey is thereby reduced correspondingly. In monthly or quarterly surveys, making small samples for the experiment during a series of several months can reduce the risk for losses. Then the experiment can be discontinued as soon as the results are obtained.

Often, a combination of several criteria is used to measure how well a questionnaire and a collection method function. Greater emphasis is usually placed on measuring the effect on other quality indicators than on measurement quality. The most common are:

- level of unit nonresponse
- level of item nonresponse in important survey variables
- level of estimates (How high were the costs, how many business trips etc. were reported?)
- scope of checking and correcting procedures
- amount of time taken to submit information
- cost to collect, register and check the information.

7.3 Examples of decisions after experimentation by Statistics Sweden

Respondent burden. For many years, the Household Budget Surveys have had problems in getting a good response rate. To alleviate this, the producers have for a number of survey rounds tried bookkeeping periods of different lengths and what kind of gifts *(incentives)* to give to the respondents. The effect of different gifts on nonresponse, response rate and quality of answers was studied in an experiment.

Incentives. In the 1985 Household Budget Survey, respondents were given gifts in advance in an embedded experiment. Giving the respondents a mini-calculator in advance yielded a positive effect on the number of completely filled-in accounting books and on the proportion of replies.

The data collection method. When the Living Conditions Survey was started in 1974, comparisons were made on the response rate and response quality of household interviews and individual interviews, respectively. The differences were relatively few. As a result, it was decided to use the more demanding but also more informative method with household interviews.

Change in technique. In 1980, the Labour Force Survey decided to switch from telephone interviews with paper questionnaires to computer-assisted telephone interviews. The change in method was preceded by a large embedded experiment where the interview answers were followed up by re-interviews. The experiment showed good agreement in the spread of replies in most of the variables, with differences in only two variables. The first difference was corrected, while the other one lead to an improvement.

Choice of questionnaire type. The Household Budget Survey chose to use a questionnaire with pre-printed columns after comparing four different versions with regards to the number of answers, cost level and response rate in an independent experiment.

7.4 Overview, checklist and results

CHECKLIST: Experiments
1. Describe the problem
2. Determine the decision-making criteria
3. Formulate hypotheses that can be tested
4. Design and size the experiment
5. Carry out the experiment
6. Test the hypotheses and interpret the results
7. Make decisions

Results
Phase 4 should lead to a number of conclusions about the survey:
- to conduct the survey with the evaluated method for measurement and data collection – or, with the best of the compared methods,
- to revise the survey's questionnaire and design, i.e. to begin again in one of the previous phases
- not to conduct the survey at all.

If the first alternative is chosen, experiments should be conducted so as to obtain the most information possible on accuracy that can be used in the final quality report.

References

Groves Robert M. (1989) *Survey Errors and Survey Costs*. John Wiley and Sons.

8 Phase 5 - Adjustment for production

8.1 Contents

Task: A questionnaire that has been tested for content and measuring quality should only require marginal adjustments. But in full-scale production, there may be technical conditions that were not apparent during the cognitive test (Phase 3) and experimentation stage (Phase 4). There may be instances during the testing itself or problems of various kinds that were not in focus during the cognitive test and experiment. These instances or problems may also be so unusual that they are not discovered with a small number of objects. The questionnaire may need to be further technically adapted before it can be used in production. Since there are more persons and functions involved in production than in a test, it may be necessary to write out detailed instructions on the handling of the questionnaire.

8.2 Production adjustment of questionnaire

There are various reasons why a producer needs to adjust a questionnaire that has been completed with regards to measurement, to the chosen technique for sending, collection and registration.

Because of cost and time factors, a paper version is sometimes used in the test, although a computerised questionnaire is to be used in production. Then the producer must adapt the questionnaire to the system for registering and make a production test, for example to find out if the skipping instructions actually work in practice.

A questionnaire created in Word is sufficient for a cognitive test. The main survey may require the producer to do comprehensive programming in a special program for questionnaire design. This is done in the case of mass printouts of addresses, contact persons, printed information from earlier survey rounds, and/or limitation of the number of questions in each printout depending on the area of activity for each establishment.

Many surveys must use more than one method for data collection to maintain the response rate. When the questionnaire has only been tested for the most important collection method, it needs to be adapted to ensure that it functions well with the other methods as well. Measurement experts and producers do this best cooperatively.

Unfortunately, questionnaires are not always developed systematically but often enter Phase 5 with errors that should have been taken care of earlier. Consequently there are delays, interruptions and/or low accuracy. In addition, all types of shortcomings cannot be discovered with the tools used in this phase.

Phase 5 is directed towards seeing how the questionnaire functions in sending, in contact with the respondents, during collection, registering and measuring, and how much time is spent. Procedures for deregistration and follow-up of non-response should be looked over in this phase, as well as information materials and interview training. However, when checking the contents and format of the questions, only marginal shortcomings are expected to be discovered.

8.3 Adjustment tools

When making difficult single-round surveys and large modifications of multi-round surveys, a "medium-large" test is needed (with several hundred respondents) that covers the whole production chain. Such a test is sometimes called a pilot or test survey. The sample should be large enough and the survey designed so that generalised results can be obtained. The tools used are the same ones as used in cognitive studies, especially observations of the respondents and debriefing (questioning afterwards) of personnel and respondents.

In the case of interview surveys, debriefing with interviewers (and sometimes respondents) gives information on the questions and sections that the interviewer has had difficulties with. It is important to measure the length of time needed for the interview, since time is an important factor when calculating costs. When a lengthy interview time shows that the burden on the respondents has been highly deterrent, the test results help to reveal difficult and time-consuming questions that perhaps can be removed or simplified.

In many surveys, the collection phase is standardised and use production systems and processes that have long been tried and tested. For reasons of time and expense, it is hardly worthwhile in such cases to do a test on the complete production process. Instead, the processes that differ from the norm should be studied, since these could cause unforeseen difficulties. These tests are sometimes called *pilot studies*. (Please note that this term is also used for other types of preliminary surveys).

In one version of such a pilot study, the interviews are made in exactly the same way as will be done later in the main survey. An observer is present and the interview is recorded on tape. Evaluation is made by going through the observations and tape recordings, as well as by debriefing the interviewers. The interviews are mainly done to discover shortcomings in the questionnaire that are too unusual to notice in a small cognitive test, and difficulties to handle information and questions that the better-trained test interviewers did not have.

The pilot interviews are carried out by regular interviewers. Therefore cognitive tools such as probes and "think-aloud" methods cannot be used. The observer sees what happens externally during the interview, but gets no information about the underlying conditions. Information about problems is obtained only if the respondent reveals that there are problems, and if the interviewer spontaneously observes these problems. Therefore, the pilot interviews are not as powerful a tool to identify the respondents' needs for better questionnaires as are the test interviews.

Compared to cognitive tests, trial interviews do not give any information about the reasoning of the respondent, or the causes of their possible errors. On the other hand, trial interviews give more reliable information on the length of time needed for an interview, since they are done under completely "active service" conditions. Just because a question is answered quickly and with reasonable values, it is not a guarantee that the question has been correctly understood or correctly answered. Using observers is more expensive than using a standardised test even when looking at the unit cost per interview. At Statistics Sweden, the Living Conditions Survey (ULF) uses trial interviews each time new questions are inserted into the questionnaires.

A cheaper and simpler method than trial interviews is called *Question rating by interviewers*. Interviewers are trained to be observant of certain types of errors and their causes, but observers are not used. After an agreed-upon number of

interviews have been conducted, the interviewers classify the questions after their level of difficulty according to a table.

8.4 Error signs and correction measures

To ensure that the collection process functions according to plan, and perhaps revise the collection method and the questionnaire, a number of error signs can be monitored during the beginning phase of the collection. This is possible in the case of centrally stored electronic questionnaires, for example in computer-assisted telephone interviews, or when the respondent collects an electronic questionnaire from a website. As long as only a few have visited the website, mistakes can be corrected. However, correcting paper questionnaires that have already been sent out would be difficult to administrate and tremendously expensive.

Limited changes in questionnaires are best made in computer-assisted interview surveys where respondents are contacted successively. Revisions of the questionnaire can then be made from one day to the next. Some of the answers given in the interviews before the improvement may need to be complemented or coded as item nonresponse. At worst, the test results must be discarded. When all the interviewers either work in a central telephone group or can be reached by e-mail, the producer can quickly start up a systematic dialogue. Shortcomings in the questionnaire can be corrected, instructions clarified and the administration strengthened. An electronically accessible compilation of "problems and solutions" will result in a continuously updated collection of advice. The collection process can be improved successively and more consistency obtained.

Overview of possible steps to take for shortcomings:

1. Point-by-point revisions of instructions and variables/questions in centrally stored electronic questionnaires.
2. Sending out additional information in case of serious errors, or even revised questionnaires to respondents of mail questionnaires.
3. Continued (in Phase 6) and in-depth follow-up to obtain:
   - basic material for tests and revisions of questionnaires for the next round of surveys in case of multi-round surveys, and
   - basic material for quality declaration.
4. Sometimes surveys that use questionnaires that have not been tested give such strong error signs that a survey must be discontinued. Only after a new round from Phase 3 or earlier can the survey perhaps be started again.

8.5 Pilot interviews – an example

The following example is taken from a survey regarding the environment in Stockholm. It was a single-round survey conducted with face-to-face interviews and paper questionnaires.

The clients had created the questionnaire themselves together with a response card. Those who had designed the questionnaire had previous experience in constructing questions and were well aware of what to consider when formulating questions. Several persons were involved in designing the questionnaire, and they checked each other's questions on the drawing board. Many of the questions were taken from Statistics Sweden's surveys (especially the Living Conditions Survey) and thus had been tested in other situations. The questionnaire appeared to be well thought-out. Therefore it was decided to rely only on

pilot interviews to see how long the interview would take and how the questions worked in general.

Five interviewers took part in the pilot study and 20 interviews were made.

The average interview time was just over 77 minutes. This was too long, since the average time was not to be more than 60 minutes.

After hearing the interviewers' descriptions about the questions that did not work well, the client decided to delete or shorten some of the questions.

The instructions were improved according to the indications of the interviewers. The questions that did not have a response card were re-written more clearly if possible.

Some of the questions were very difficult to answer directly, for example: "what amount of electricity did you consume last year?" and "how much was your household car used?" The client realised it was necessary in some way to inform the respondents in advance about such questions as the one on electricity consumption. A separate information sheet was made that the interviewers could send out when they scheduled an interview.

The answer alternatives were not in logical order and the layout of the questionnaire was less than optimal. Both the respondents and the interviewers found the interview somewhat confusing. The client then divided the questionnaire into ten different blocks. This resulted in a better overall design.

Summary
The pilot interviews resulted in improvements of the questionnaire and the routines of the interviewers. The combination of a keenly aware client, competent interviewers and a **sufficient amount of time**, saved much of the relevance of the survey. If the general test model had been followed, a cognitive test in Phase 3 would have shown that a number of complicated question tables needed to be deleted since they could not be answered sufficiently well. A CATI questionnaire could also have been developed, resulting in a cheaper alternative.

8.6 Checklist and results

Checklist: Tools to discover error signs
1. Spontaneous comments by respondents - in writing, by telephone etc.
2. Spontaneous observations made by own personnel
3. Debriefing own personnel
4. Debriefing a small number of respondents

Additionally in interview surveys (easiest in a CATI/CAPI environment):

5. Pilot interviews
6. Behaviour coding (including observations in the field) to track "difficult questions"
7. Listening, especially in telephone interviews

Result: A questionnaire that works satisfactorily in a full-scale survey both for the respondent and in the handling, sending, deregistering, data entry and other production stages.

9 Phase 6 - Evaluation

9.1 Contents

Tasks:
In many surveys, shortcomings in the measurement and data collection processes are the largest sources of uncertainty. It is therefore necessary to plan and carry out studies to discover these defects and their origins. Informative measures of accuracy and indicators need to be calculated on how well different processes work. The results lay the foundation of a quality declaration.

When errors are discovered in a multi-round survey, the reasons for these errors need to be identified and eliminated by once again testing and revising the questionnaire. Measurements and indicators on the size of the errors are not enough. We need to know how the errors originated, and this requires cognitive studies.

9.2 Quality assurance

Quality assurance involves measuring if the production process is working according to plan and if the results attain the level of quality that was promised. A deterioration may have occurred, perhaps due to a poorly conducted survey, perhaps because the surveyed facts or the processing conditions have changed so that the questionnaire and/or the design of the survey must be modified. Besides such quantitative measurements that are used in experiments and evaluations, there is a series of quantitative and qualitative signals that indicate risk for low quality response. Some persons may be able to point out that the questionnaire is too complicated and the burden on respondents too high, others may feel the survey contains questions that are too sensitive, while still others think the survey involves too much work. Measurements and indicators are produced by classifying the replies so that it is possible to see if the problem is consistent or concentrated to parts of the survey. Sometimes "risk groups" can be identified that have had more difficulty than others to answer correctly.

When making computer-assisted interviews and using intelligent electronic questionnaires, a counter and timer can be built into the questionnaire, making it possible to measure the number of wrong answers and changed answers. We can also find out the amount of time each section of the questionnaire takes to complete. It is important to get an idea of the burden on the respondents, since these questionnaires demand that the respondent edits the answers that are not accepted.

9.2.1 Debriefing

Debriefing involves questioning on, for example, how a task such as interviewing, checking or coding for a survey was carried out. Those being debriefed have fresh experience of the questionnaire or other aspects of the data collection, they are respondents, checkers, coders or interviewers. The person doing the debriefing prepares an overview in advance and has a list of questions on the elements in the work process that require further information. There is also time to listen to spontaneous viewpoints. Debriefing is often done with one person at a time, since independent viewpoints are preferable. However, to save costs debriefing in groups are also frequent. Compared with focus groups, debriefing is done according to a well-prepared list of questions on a process that all have

participated in. The group is often homogeneous, and everybody knows each other.

When an interviewer is to be debriefed on how the respondents performed, it is important to train him/her on what he/she should look for during the interview. Interviewers sometimes miss the kinds of problems that are observed when coding behaviour, for example. They frequently have difficulty in quantifying their observations, but are good at identifying the elements in the questions and the order of the questions that can lead to misunderstandings. It is frequently difficult for an interviewer to clearly distinguish between his/her own difficulties and those of the respondents.

Respondent debriefing An important advantage with this method is that it is possible to select respondents with unusual values/characteristics. They are very difficult to find, when they are needed as test persons in a cognitive questionnaire test. Debriefing should be done immediately after the interview, while the person still remembers his/her answers. This also saves the cost of having to contact the respondent later. Compared to a cognitive questionnaire, this type of debriefing has more limited goals and a standardised procedure. Because of the sample method, it is necessary to use a regular interviewer, and therefore cognitive tools can only be used to a limited extent. Since it is possible to draw a sub-sample as a probability sample, it is theoretically possible to estimate effects.

9.2.2 Behaviour coding

Behaviour coding involves observing the respondent's behaviour question by question during the ongoing interview. Observations are coded according to a plan that has been made in advance. In an experiment, a "dress rehearsal" or a main survey, behaviour coding yields quantitative estimates on the frequency of difficulties. One valuable use is to measure errors that are too infrequent or errors in the questions that are to be answered by too few respondents to be discovered by a small quantitative test.

Behaviour coding differs from a cognitive test in that observations are limited according to a strict plan, and in-depth questions cannot be used. It is a standardised method with few possibilities to register and follow up error types beyond the coding scheme. Behaviour coding gives information question by question and does not shed light directly on the effect of the order of the questions and the context. Nor does it reveal if the respondent misunderstands the question or estimates the answers, as long as the answers come directly.

Surveys with telephone interviews
A standardised version of behaviour coding is when the interviewer, during an interview, codes the respondent's reactions. This is easiest done during computerised telephone interviews. The interviewers practice at a keyboard registering a small number of types of reactions. These should be easy to identify and classify.

Statistics Canada has introduced a standard with a breakdown of 5-6 codes (all codes are not always used) for different types of behaviour.

CHECKLIST: Behaviour coding of the respondent

- interrupts the question (*Interruption)*
- asks for clarification of the question (*Clarification*)
- asks for the question to be repeated again (*Repetition*)
- troubled by the question (*Uncomfortable)*
- wonders how much time remains (*Time out)*
- doesn't answer exactly but makes an estimation (*Estimation*)

Another breakdown that is used is to code when the respondent

- has doubts or delays in answering
- answers before the question has been asked completely
- misunderstands the question but doesn't realise it him/herself (*Inadequate answer*)
- understands the question but says that the answer is too difficult to remember/describe etc.
- explicitly refuses to answer a particular question.

The number of codes must be limited so that the interviewer can remember them, as well as observe and register the behaviour during the data collection. The coding must not be so lengthy that the rhythm of the interview is disturbed. Only one code per question is allowed. Item nonresponse is registered separately. If no registration is made, the interviewer has understood the question as being correct. According to Statistics Canada, interview costs increase by 0.4% when behaviour coding is used.

One advantage of this form of behaviour coding is that it does not require specialist knowledge. The interviewers can be trained quickly. The same code system can be used from survey to survey. The tool is simple and inexpensive to use, and can also be included in trial interviews to find out which questions are expected to be difficult for the respondents.

A highly simplified version of behaviour coding is used at Zentrum für Umfragungen (ZUMA) in Mannheim and is called problem coding. The interviewer states if he/she has noticed any problems with the question. However, the interviewer does not need to specify the type of problem. After the interview is completed, the interviewer can note the difficulties and later debrief on them.

Surveys with face-to-face interviews

In surveys with face-to-face interviews, the above behaviour coding can of course also be done. An observer can also code the behaviour and reactions of both the interviewer and the respondent, while at the same time the interview is recorded on tape. The respondent's "error signs" can be coded according to a schedule that includes more categories and requires more detailed evaluations than when the interviewer does the coding himself/herself. The interviewer's work is coded by whether the question was

- read correctly
- read with a small discrepancy or
- read with a large discrepancy.

The coding is done afterwards with the aid of protocol and tape recordings. Because an observer is required, this method is comparatively expensive to use. It also requires very well trained personnel, the analysis is more difficult and it takes longer to arrive at the results. Statistics Sweden has previously done a number of such studies on surveys that were already in production. Despite the many valuable observations, it was difficult to show that these lead to any change in the ways to ask survey questions. It is always difficult to attract support for questionnaire revisions after production has started.

9.2.3 Re-interviews

There are two main types of re-interview studies. One type measures the variation in answers and sometimes even the bias. Therefore a relatively large sample is required. The other type measures the quality of the interview, and is made with a small sample. Diagnostic questions are used to find explanations and identify contradictory answers from the interviewer as well as the respondent.

Estimate the variation of answers and bias
The time lag between the first and the second interview is often a compromise: It should be short enough so that the respondent's reality remains unchanged, and long enough so that he/she does not exactly remember his/her answers. Two to three weeks is a common length of this interval. The same data collection method should be used so that no methodological effects are added to the variation in answers. For cost reasons, re-interviews are limited to a randomly drawn subsample. As a rule, the sample size is around several hundred, so that uncertainty can be measured with satisfactory precision.

If the burden on the respondents appears very great, the number of questions can be reduced. The focus will be on the questions that are most important to study; however, these questions must not be used outside their context, which would disturb the comparison (*context effects*). Studies on re-interviews to measure the variation in answers have been done in the Labour Force Survey and the Living Conditions Survey.

Re-interviews with reconciliation are made to estimate the response bias. In this case it is not necessary to use the same measuring method during the second interview. Reconciliation involves finding out the reasons for differences in answers and determining the correct answers. Sometimes this can be done by logical checks and auxiliary information. In other cases it is necessary to contact the respondents. During the first interview, answers can be stored when the collection is done electronically, so that reconciliation can be done directly after the second interview. Re-interviews with reconciliation have mainly been used for evaluation in the Population and Housing Censuses.

Process control
In multi-round interview surveys, re-interviews are used as a process control. A rather small sample (25-50) is drawn for each round of the survey. The sample is drawn so that the interviews are limited to a small number of interviewers. Interviewers with special training who do not know who did the original interview do the re-interviews.

The second interview should discover the shortcomings in the whole collection process, the questionnaire and instructions, and the work methods of the interviewer. Shortcomings in the questionnaire and the instructions can be corrected for the following survey round. Interviewers with a high error

frequency can be given additional training. A simplified form is simply to contact the interviewer to make sure the first interview was actually done.

Statistics Sweden has developed a system for re-interviews with reconciliation, used for e.g. the Labour Force Surveys (when answers are different in re-interviews, the idea is to try to find out the reasons for these differences and determine the correct answers). Since the Labour Force Survey interviews are done in a computer environment, the information from the first interview can be keyed in, collected and immediately compared with the responses in the second interview.

Re-interviews begin as early as five days after the first (original) interview, so that respondents do not forget the details of their employment during the reference week. When the answers differ, the interviewer, by using in-depth questions, finds out which answers are the correct ones and why the difference occurred. The variables "degree of attachment" and "status on the labour market" are the most interesting. The interviewer also asks a few general questions about the survey and how the respondent felt about the first interview. There is a special form to note explanations and observations.

The effects of the errors on the estimates can be measured by using a probability sample and a larger sample size. Summed up over a six-month period with a total of 2 154 respondents, the study showed that 7% of the answers on degree of attachment to the labour market and 5% on status were wrongly classified.

9.2.4 Monitoring

Monitoring involves an interviewer and a "listener" sitting in the same room. The listener sees the computer screen and keyboard, and is able to follow the events in the interview. The listener is an observer and cannot intervene during the interview. The interviewer and the listener both have headsets, and both can hear what the person being interviewed says. Before the interview begins, the interviewer must inform the person being interviewed that another person will be listening to the interview. It is uncommon that anyone reacts negatively.

Since the listener hears both what the interviewer and the person being interviewed say, the listener can study how the questionnaire works for both parties. The listener hears how the interviewer understands the questions, which words are emphasized when read, if words are added or avoided, and which phrases or sentences, questions, terms, concepts, instructions etc. that cause problems in the communication with the respondent. Some examples of listeners' observations are: introduction too long, instructions too difficult or not in logical order, questions that appear to be clear have been understood differently by various respondents.

Monitoring can be used in *dress rehearsals* or in regular production, i.e. in phases where the questionnaire is already complete from a measurement point of view. More knowledge is needed about the way the interviewer handles questions and information, as well as the interaction with the respondents. Monitoring is an effective tool to identify the interviewers' needs for basic and further training. By revising instructions to interviewers, changing the order of the questions and improving the instructions to the respondents, the interview time can be shortened and thus reduce both costs and the burden on the respondents.

Monitoring in one form or another is relatively common when the interviewer is working in a central telephone group. Compared to behaviour coding,

monitoring is more informative regarding misunderstandings and behavioural mistakes. However, monitoring requires more time and expense to carry out and analyse.

9.2.5 Quality reporting by interviewers via e-mail

This method is a highly simplified variation of "observations in the field", and is designed so that reporting can be made at a low cost within 14 days from the start. It is based on cooperation between personnel from the Measurement Laboratory and specially trained interviewers. The interviewers are used to working with personnel from the Measurement Laboratory and reporting to them systematically.

The interviewers receive instructions on which problems are especially important to watch. After having conducted a limited number of interviews (normally five to ten), they report their observations via e-mail. The reporting takes about two hours and focus on which questions cause the greatest problems. The most serious problem is taken up first, and the interviewer continues until the time runs out. Suggestions for improvements are also appreciated. The answers are compiled and a summary overview and analysis is made. The method is used to form a basis for the quality declaration of the survey. If the survey is a multi-round one, improvements can be made in the next collection round.

Quality control via e-mail by the interviewers can reveal which questions the respondents have had difficulty in understanding and answering. However, it cannot reveal when respondents misunderstand the questions but reply quickly anyway. There is not enough time for in-depth questions, so the method cannot be used to reveal the causes of the difficulties in the same way as cognitive tests do. The results are not quantifiable.

9.3 Revision of questionnaires for multi-round surveys

Multi-round surveys require revisions from time to time. There are several reasons:

1. The reality being measured has changed.
   Can be studied by observations, in-depth interviews or focus groups
2. The need for information has changed.
   Can be pointed out by the client, noticed by the producer, or studied systematically by focus groups
3. Words and concepts in the questions have changed their meaning.
   Best studied through cognitive tests
4. Shortcomings in the questionnaire are discovered.
   Studied with cognitive tests, directed with the help of error indicators

The first two reasons are closely related. Sometimes the environment the multi-round survey wants to describe changes. The variables that once gave a relevant description of reality no longer do so several years later. A set of questions suited for measuring economic or social conditions in 1990 has become less relevant by 2000. Phenomena, expectations, attitudes, organisations, goods and services change with time – as do people's attitudes. Classifications such as by industry, education and socio-economic group are revised. Working environment and gender equality rules become stricter, and demands also increase. A situation once regarded as acceptable can have become a working environment problem ten years later.

For example, when industry is re-structured, mobility on the labour market changes, values and habits change, some environmental issues may gain importance, and questionnaires must be adapted to the new situation. Otherwise they will neither meet the information needs of the statistics users nor allow respondents to give relevant answers. For the very reason that the contents in some surveys (mainly those for establishments) are governed by regulations, the questions asked must be relevant for the respondents. The Consumer Price Index is a clear example of a survey that works systematically to change its contents whenever the supply of goods changes.

As a rule of thumb, every multi-round survey should thoroughly review its questionnaire and data collection method at least every fifth year. The revision of a questionnaire in production puts the producer in the same situation as a new survey, and the same methods for studying relevance and content should be applied. The work must begin in Phase 1. The work in Phases 5 and 6 can indicate that a question is no longer relevant for the respondent, for example by increased item nonresponse, "do not know" or "not relevant".

The time plan for a revision is frequently fixed in the year because a monthly or quarterly multi-round product cannot change questionnaire except at year-ends to permit annual summing-ups. Revisions are planned and spread over a calendar year in such a way as to give the opportunity to test the revised questionnaire.

In panel surveys, it is probable that respondents participating for the first time give stronger signals than old-timers about difficulties in the questionnaire. This group should be studied separately, because old-time respondents may have learned how to avoid answers that lead to inquiries and further contacts from the producer. Just because the data passes the edits does not mean that they are accurate. Therefore, putting together new and old respondents may disguise error risks

Suggestions for the improvement of questionnaires in multi-round surveys are sometimes halted by the arguments that “it would disturb the time series" and that "the trend is still the right one". However, there is no general support for these arguments. A question that systematically gives wrong or non-relevant answers will measure the trend for a completely different variable than the one that is interesting. When systematic errors in information occur by excluding a component of a sum, it cannot be expected that just that component follows the same trend as the others. Variations in measurement error or nonresponse as a result of poor questions also give lesser accuracy in estimates of differences and in time series.

9.4 Examples

In this section, some examples are given on how surveys have successfully used error signs to discover measurement problems. The following revision of a question has led to improved quality in various aspects.

Error signs via editing
Editing points out values that appear to be unreasonable, inconsistent or of the wrong size order. Editing directs attention to values that are **outliers**. In a multi-round survey, the **results of editing** are used to trace probable error sources in data collection and questionnaires so that error sources can be eliminated in future rounds. An unclear questionnaire is often the cause of these errors. Much item nonresponse occurs due to various defects in the questionnaire.

Checking cannot reveal systematic measurement errors as long as the data falls within the control limits. These types of errors are called **inliers**. Inliers can occur when there are differences between the definition of the concept in the survey and definitions in the administrative system of the establishment.

External evaluation
In the following example, a lack of agreement was discovered with another statistical product that was used during the checking. The lack of agreement had arisen because the respondent had not been given information on how a quantity was defined. The problem was solved when the producer instead asked in terms of a better-known quantity and did the calculations himself.

Example 9.4.1

Building Production Statistics is an annual sample survey that has been conducted by Statistics Sweden since 1993. Particulars are collected via a mail survey to selected construction companies. During the first few years of the survey, it was noticed that one question was often misunderstood by the respondents. The variable was called "Total production and administrative costs" and was to reflect "all costs during the year". The persons doing the checking work also had access to the Annual Reports specifying the Operating Expenses of the establishments. Operating Expenses for one year are all those costs that have been invoiced during the year, and can include costs for work that was completed the previous year. Costs for work that has not been completed during the year is not to be included. On the other hand, for the establishment, costs are those expenditures the establishment has had during the year, regardless of whether the work has been invoiced or not.

When comparing answers in the Building Questionnaire with Operating Expenses in the Annual Reports, the data was often identical. Operating Expenses should not be identical with the answer on Costs in the Building Production questionnaire when work is still continuing. About 90 per cent of the respondents did not reply to the question on Costs during the year, since they were not familiar with the definition. The remaining 10 per cent either calculated and/or estimated the variable in question. The checking and correction work was very extensive, since it was necessary to look at all the items to determine if the respondent answered with Costs or Operating Expenses. The incorrect items were corrected. This extensive correction work was done manually. It involved deducting last year's work and including the work that had not been completed to arrive at "Costs".

Nowadays, Operating Expenses are collected instead of Costs. A re-calculation is still done for Costs, but today it is done automatically and for all establishments. It is no longer necessary to verify the respondent’s answer, since a familiar variable is now used.

Nonresponse and item nonresponse
The example below signalled considerable unit nonresponse and item nonresponse. Many respondents phoned the producer, and the burden on the respondents was too high. The use of administrative information and calculation algorithms allowed a revision and simplification of questionnaires for small establishments.

Example 9 4.2

During the first two years that the survey Building Production Statistics was conducted, small construction companies had problems in submitting detailed information on their activities as asked for in the mail questionnaire. Therefore, a large part of the checking, re-contacting and correction work was done on small establishments.

A simulation study on material from 1994 showed that the work done on data from small establishments, both by the respondents themselves and by staff from Statistics Sweden, did not have any noticeable effect on the accuracy of the final results. A reduced set of questions for small establishments with a limited amount of data was constructed with the help of figures from 1994. Information that was not collected from small establishments was estimated with a model-dependent estimator on the basis of answers from the other establishments. Estimates from the simulated survey were compared with estimates from the regular survey. No significant differences appeared. As a result, the simplified questionnaire can be used for establishments with up to four employees without any considerable worsening of accuracy.

The simplified questionnaire was used by slightly more than 400 establishments (36 per cent of the sample). As expected, the checking and correction work was reduced, both for the respondents and for the Statistics Sweden's personnel. There was a significant increase in the response rate. Among establishments with zero employees, it increased from 69 per cent to 81 per cent, and among establishments with 1-4 employees from 66 per cent to 76 per cent.

Distribution of the variable
The example below illustrates that on two occasions when **comparing the distribution of the variable**, there was a significant rise in rounding off on one question. The increase signalled that accuracy had worsened as a result of changing the collection method from mail questionnaire to telephone interview.

Example 9.4.3

The questionnaire for the Survey of Housing and Rents asks about loans for one- or two-dwelling buildings. The information was collected in a mail questionnaire in 1991, but in 1993 a telephone interview was conducted. An estimate was made on how many respondents answered in even numbers of hundred thousands of SEK in each collection method. In the 1991 mail questionnaire, 10 per cent replied in even hundred thousands of SEK. In the telephone interviews of 1993, 22 per cent replied so. Questions regarding debts on loans are apparently easier to answer exactly when the respondent has more time and answers in writing.

9.5 Checklists and results

The most common measures of accuracy in measurements can be broadly divided into three groups.

A: Response bias and response variance in estimates

Quantitative methods based on probability sampling must be used to estimate the response bias and response variance. Some of the most common methods are:

Estimates obtained by:

1. Evaluations of bias against another statistical product - sample of a total register.
2. Estimates of measurement errors by embedded or independent experiments.

And in the case of in-depth interviews:

3. Studies of re-interviews - with or without reconciliation, possibly with in-depth questions
4. Measurements of interviewer variance.

Since the 1980s, quantitative studies have become less common. Evaluation studies have been done in connection with recent Population and Housing Censuses. Some surveys have studied re-interviews; studies with or without reconciliation occur.

To take up quantitative methods in this manual would make it unnecessarily comprehensive, especially since the methods are well established and well known. An overview of methods has been done by Groves (1989), and an overview of empirical studies at Statistics Sweden by Lindholm.

B: Indicators and observations from the production process
Different indicators signal whether evaluation, testing and revision are needed for the whole questionnaire or for separate variables.

B.1. **Examples of indicators on problems in the whole questionnaire**:

1. High level of unit nonresponse, especially refusals.
2. Slow rate of return of questionnaires, high share of late replies.
3. Many complaints, spontaneous negative comments, questions or viewpoints on the survey in general from respondents.
4. High number of further contacts and additions due to incompletely filled in questionnaires.
5. Long hours to obtain the information and fill in the questionnaire.

B.2. Examples of indicators on problems with specific questions:

1. High rate of item nonresponse.
2. Many ”do not know” answers, changed answers and unreasonable answers.
3. Irregularities in the distribution of responses in the variable, for example "spikes", significant lopsidedness or many outliers.
4. High frequency of error messages during logical controls and size checks, perhaps even great needs to make corrections and further contacts.
5. When coding automatically, low success rate.
6. Many spontaneous messages from respondents about unclearness and difficulties with specific questions.
7. Many spontaneous observations on difficulties that their own personnel has made.
8. A high level of labour input by the respondents to re-calculate existing information to the information requested.
9. In multi-round surveys - time series breaks or high variation in results between the survey rounds.

C: Special evaluative cognitive and qualitative studies

1. Embedded relevance studies (for example in-depth interviews with certain respondents).
2. Spontaneous observations and debriefings, i.e. the same methods that are used for monitoring quality during the start of the survey.

   In interview surveys:
3. Measurement of interview time and other time taken.
4. Re-interviews for quality control with the help of in-depth questions.
5. Behaviour coding.
6. Monitoring of the quality of the interviewer work.
7. Listening.

Interpretation of signals is on the whole a question of judgment. There are no rules of thumb for how strong a signal needs to be before it leads to further action. Indicators do not say exactly when a disturbance has been caused by an error in the measurement process or by some other factor. The most effective type of further action must be considered and sometimes even tested.

Sometimes, steps other than a general revision of the questionnaire could be the most effective. For example, new information material might be needed for some groups, the interviewers may need additional training, or an extra programme may be needed to assure that all the production processes are working as intended. In multi-round surveys, in-depth contacts with especially error-prone respondents can be more effective than revising the questionnaire. This could be true if an establishment has just changed responding staff and the new respondent(s) need additional instructions for the questionnaire.

Results:
All surveys: Information for the quality declaration of the survey.
Multi-round surveys: Indications of difficulties in collecting information and questionnaires, and suggestions for steps to take directly or after a test.

References

1. Andersson, C., Lindström, H.L., and Polfeldt, T. *Att mäta statistikens kvalitet*. R&D Report 1999:3.
2. Granquist, L. *Granska effektivt*. Report from Statistics Sweden's checking group March 1997.
3. Groves, R. M, *Survey Errors and Survey Costs*. John Wiley & Sons 1989
4. *Evaluation of Population & Housing Census 85. Employment data* Statistical report BE 42 SM 9001
5. *Evaluation of Population & Housing Census 85. Housing data and dwelling data* Statistical report BE 42 SM 9002
6. Journal of Official Statistics , 1992:1, page 63
7. Lindholm P. (1995) *Mättekniska studier vid SCB - en översikt*.

10 Phase 7 - Quality declaration

10.1 Contents

Task: To systematically compile quality information and make it understandable, available and useful for clients and other users.

The producer is responsible for a systematic quality declaration. The measurement expert submits information from his/her studies. The target group consists of the client and other future users.

Result: A quality report according to the instructions for the Official Statistics of Sweden. The quality report includes information on how the questionnaire has been developed and tested, indications from collection and processing, and estimates of or indications on response variance and bias.

10.2 Instructions and prerequisites

The guidelines for quality reporting that were adopted at Statistics Sweden on 11 October 1999[1)] define the new concept for statistical quality for the Official Statistics of Sweden. They are found in the publication MIS 94:3 and subsequent publications. The guidelines consist of five main components, each with a number of sub-components.

Primarily, the measurement work is directed towards improving the contents and accuracy of the main components. The guidelines also contribute towards a more detailed report of uncertainty measures and increase the total accuracy.

Quality concepts for official statistics

Contents
- Statistical target characteristics
  - Units and population
  - Variables
  - Statistical measures
  - Study domains
  - Reference time
- Comprehensiveness

Accuracy
- Overall accuracy
- Sources of inaccuracy
  - Sampling
  - Frame coverage
  - Measurement
  - Nonresponse
  - Data processing
  - Model assumptions
- Presentation of accuracy measures

Timeliness
- Frequency
- Production time
- Punctuality

Comparability and coherence
- Comparability over time
- Comparability between domains
- Coherence with other statistics

Availability and clarity
- Dissemination forms
- Presentation
- Documentation
- Access to micro data
- Information services

[1)] See MIS 2001:1 Quality definition and recommendations for quality declarations of official statistics

Because the quality declaration is user-oriented, it should contain information that helps the user to evaluate the measurement quality of different variables. The user should also be able to determine how to best process the information to obtain acceptable accuracy and overall understanding. This overall understanding requires that the quality declaration can give references to more detailed information and to a contact person with in-depth knowledge. No report is made on the work with questionnaire development and production adaptation unless it has a bearing on the quality of the estimates. Most of this work is documented in methodology reports.

10.2.1 Main component: Contents

The measurement expert contributes to the sample design and the definition of variables with focus groups and/or in-depth interviews to help define the question areas that the respondents know are relevant, but the client/producer did not recognise during his/her work at the drawing board. The work also includes identifying statistical measures that are measurable, i.e. measures for which the respondents have the information on the questions asked. Information on rules and considerations that have steered the choice of variables and measures to assure the quality of the contents should be reported. The user can then evaluate the quality of the preparatory work and understand the reasons why the contents have become what they are.

Example 1

- *The contents of the survey were decided after conducting three focus groups consisting of employers, employment recruiters and employment seekers.*
- *The contents in the survey are completely determined by EU directive no. xxxx.*
- *Questions 7 - 19 are copied from a similar survey conducted in 1995 and have previously been evaluated.*
- *Contacts with establishments show that the majority cannot report transports in volume. Answers can only be received from all establishments if the question concerns the weight of transports.*
- *The questions regarding travel, which the committee wanted to include, were removed to reduce the burden on the respondents.*

10.2.2 Main component: Accuracy

In the 1999 edition of the revised instructions for quality declaration for official statistics, heading **2.2.3 Measurement** reads:

"*Describe the method used for measurement. When a questionnaire is used, it should be reported on in its entirety or in a suitable summary. Describe the measuring difficulties that occurred during the collection of information and their probable consequences on the accuracy of the statistics."*
"If the reported confidence intervals also include uncertainty elements from random measurement errors, make a note of this and describe the situation here or under Sampling. If adjustments for systematic measurement errors have been made, describe these here or under Sampling."

Measurement quality is primarily reported descriptively and concretely, and secondly in the form of advice on the use of the statistics material.

The descriptive part reports on bias, response bias and gross errors in vital estimates. It also reports on specific measures taken to obtain high measurement

quality in the survey. The assessment of whether the quality of the estimates suffices is left to the user.

Without being precise, the advisory part informs the user which estimates are uncertain. It is not necessary to make a quantitative evaluation to find out if a question works well or not. A cognitive study is often sufficient and can also indicate the cause of any specific difficulties. In other cases, behaviour coding or observations can indicate problems with certain questions. The producer has then to take responsibility and evaluate the estimate as “not too bad", and also to make the user aware of this.

10.3 Measurement quality - an example of reporting

The entire questionnaire and a description of the data collection method and its characteristics

- *The survey was originally a mail questionnaire, but TDE (touchtone data entry) was introduced as an alternative in 1997. TDE was used by 51% of the sample in the survey at hand. 22% replied by mail and 12% faxed in their answers. Nonresponse amounted to 15%. A telephone reminder was made 21 days after the survey was sent out. 7% of the sample replied first after the reminder. When registering answers with TDE, there is an immediate checking of size, and the respondent can correct or comment on the information. Methodological studies have shown that answers by TDE are at least as accurate as those on paper questionnaires.*

- *The survey is an interview survey based on paper questionnaires. The average interview time is 67 minutes. To maintain the response rate, telephone interviews were accepted in 11% of the cases. Interviews by proxy with a member of the family account for 4% of all the interviews. In these interviews, questions having to do with knowledge and attitudes (one-fifth of all the questions) are excluded. In telephone interviews, a fatigue effect is seen during the last quarter of the interview, item nonresponse being consistently more common there.*

Systematic and non-systematic errors

Errors can be noticed in several ways. Evaluation, i.e. estimates of accuracy for key variables, and other quantitative measures such as behaviour coding, is desirable. In addition, qualitative information can show obvious weaknesses in the measurements of certain variables (spontaneous remarks from respondents, debriefing of personnel, spontaneous observations upon receipt and processing of incoming material, such as for example listening).

- *In an evaluation survey, it was shown that the number of employed persons in the service sector was underestimated by 3%. The underestimation was especially common for temporarily employed persons in metropolitan areas.*

- *In a study with re-interviews, certain questions were answered differently on the two occasions: If the respondent was "long-term sick", there was a difference of 7% between the two measurement occasions. “Children at home" showed a difference of 2% and “trade union membership” 8%.*

- *Respondents' notes on the questionnaires that were sent in showed they had some difficulty in choosing an alternative for question Qx*

Remarks on some general problems

These problems include how memory effects depend on the length of recollection period and the significance of the variable, if there are personal interests to avoid reporting, if the burden on the respondents is in some way especially high, etc.

An actual example taken from MIS 94:3, page 51.

- *Prices should be invoiced prices, but respondents in some industries report list prices. Discounts should be deducted, but are not always. In a situation when reduced demand leads to increased discounts, the actual price development is over-estimated by an index based on list prices. The opposite is true when discounts decrease.*

Remarks on weaknesses with certain variables

- *Item nonresponse for the question was 7 per cent.*
- *When further contacts were made as a result of checking, it was revealed that 37 per cent of the transporters could only report the weight of the load and did not know the volume.*
- *Spontaneous feedback from the interviewers showed that the question on the sales price for tenant-owned flats was sensitive for many, and led to ambiguous answers.*
- *The task to inform on which floor the flat was often required explanations and help from the interviewer.*

An actual example from the Living Conditions Survey, Appendix 15, page 23.

- *"Questions in the Living Conditions Survey are fairly easy to answer, but it is obvious that different persons understand some of the questions differently. Special caution should be taken when interpreting answers on attitudes, agreement or disagreement, and questions on how often someone exercises or meets with family and friends."*

Steps to obtain good measurement quality

A report should be available on how the questionnaire has been developed through expert checking, cognitive tests and revisions, with information on the standards for questions and classifications used. The account should help the user to determine if the producer has reasonably guarded against mistakes in the formulations, concepts, order of the questions, etc.

- *The questionnaire was originally developed by the client, and was later checked by experts at Statistics Sweden and then revised. The questionaire underwent a cognitive test, which led to re-writing of slightly over one-third of the questions. The introduction was also re-written. These revisions involved such obvious improvements that no new test was done. Since standardised routines for the sampling and the mailing of the questionnaires and the reminders were used, it did not seem necessary to do a production test. However, the cognitive test showed that the questions regarding the previous year's purchase of consumer goods were difficult to answer with certainty, regardless of how these questions were formulated.*

Experiments have been made to choose a method that can justify the choice and give relevant knowledge on the accuracy of the chosen method.

- *Two versions of accounts books were compared in an experiment with a random sample that included 300 units in each group. The results were that the accounts book with pre-printed columns for different kinds of expenses resulted in 15 % more notes than the accounts book without pre-printed columns. The version with the pre-printed columns was chosen. It appeared that the pre-printed columns acted as a reminder and reduced memory errors.*

Process control to notice and take care of disturbances

- *The survey conducts 40 re-interviews each month to identify and correct any errors in the interview. Re-interviews are independent and reconciliation is done.*

- *Twelve per cent of all the answers were corrected after editing. Further contact was made with the establishment for one-fifth of these cases. The follow-up works as training for the respondent, and accuracy is on average better on the next survey occasion.*

- *Behaviour coding in the initial phase of the survey showed that questions 12, 23 and 37 often required explanations. Information and formulation of questions was looked over. Since it was a survey with CATI, the questionnaire could quickly be modified. Before the corrections, 58 interviews had been done, 42 of which had to be coded as item nonresponse.*

Measurement error or other error?

Error classification is not always clear. Perhaps Phase 1 shows that it is not possible to measure the particular variable the client is interested in. Instead, a "similar" variable is chosen that can be measured in Phases 2 or 3. In the latter case there is a lack of relevancy. But if the original impossible question had been asked, the error would have been seen as a measurement error.

When an establishment reports with a split financial year, this can be described as a Content Error (*reference time*). As an alternative, it can be described as a measurement error caused by a time shift error (*telescoping*), since it is still the same variable but with a different reference time.

When a respondent forgets to include a cost that should included in the total, this can be described as item nonresponse in the variable but also as a measurement error in the total.

Sometimes there are different perceptions on the correct way to classify an error. It is important, though, that everything is included and reported clearly in the quality declaration.

References:

Quality definition and recommendations for quality declarations of official statistics. MIS 1994:3
Living conditions appendix 15. Technical report on living condition surveys from 1990-91 and 1992-93. Statistics Sweden 1995 (page 23)

Appendix 1

Measurement error model for more effective resource allocation

Håkan L Lindström

Not understanding the benefit of systematic questionnaire development many buyers and users of statistical surveys think that they save money if the questionnaire testing phase is passed over. Doing so, they underestimate the size of measurement errors caused by faulty questionnaires. Some fairly simple modelling and calculations supported by empirical studies may help to convince them, that already a slight reallocation of the provided resources from sample size to questionnaire improvement will be of great advantage.

By using a statistical model that includes the effects of systematic and random measurement errors, it is possible to calculate how much worse the accuracy is when substandard questionnaires are used. Within the frame of the resources given for the survey, an investment in questionnaire development would considerably reduce the mean square errors of the survey results. In this model, the sample and the measurements are the only error sources allowed to influence the accuracy of the estimates. With empirically based assumptions about costs and how much measurement errors are reduced, we can also show how effective it is to prioritise resources for questionnaire development.

The statistical model assumes a finite population consisting of N objects. Every object has a true value. Object number i (i = 1,2, ...N) has the value x_i. If responses to the questionnaire measures x_i without error, estimates based on x_i-values will have a sampling variation but no measurement bias or variation. When the questionnaire and/or the data collection process leads the respondent to misunderstand the question and answer incorrectly, the response will not be x_i but another value y_i.

The measurement error of the model consists of a systematic component b_i and a random component ε_i. The random error ε_i is 0 in expectation and has variance $\sigma^2_{\varepsilon_i}$.
Then the measured value of object number i is: $Y_i = x_i + b_i + \varepsilon_i$; (1)

Expectation, bias and variance of an estimate of an average $\bar{y}$ can be calculated in two steps - on the probability of the sample (p) and the measurement error model (m). The bias b_i in every object is constant over the measurement error model, and thus the bias in the estimate of an average in the finite population is:

$$E_{pm}(\bar{y}) - \bar{x} = \bar{x} + \bar{b} - \bar{x} = \bar{b} \text{ where } \bar{b} = \sum_U b_i / N . \quad (2)$$

With simple random sampling and with independence between the measurements internally and the true values of the objects and their measurement errors, the mean square error for the estimator $\bar{y}$ based on n observations and measurement errors from a substandard questionnaire is

$$MSE_{pm}(\bar{y}) = \left(\frac{S^2_{x+b,U}}{n} + \frac{\sigma^2_\varepsilon}{n}\right)\left(1 - \frac{n}{N}\right) + \left(\frac{\sigma^2_\varepsilon}{N}\right) + (\bar{b})^2 \ ; \qquad (3)$$

where

$S^2_{x+b,U}$ population variance for true value + systematic measurement error.

$\sigma^2_\varepsilon = \frac{1}{N}\sum_U \sigma^2_{\varepsilon_i}$ the average measurement variance when the measurements are uncorrelated, which can be assumed in a postal questionnaire.

b a bias, that is due to systematically giving the wrong answers such as by being mislead by the formulations and instructions in a questionnaire that has neither been tested nor revised.
N population size
n sample size

To minimise $\bar{b}$ and σ^2_ε , the preliminary questionnaire is tested and revised. Within a given cost frame, the test can be financed by reducing the sample size by n_t so that the sample size of the revised questionnaire is $n - n_t$.

If the test and the revision succeed to completely eliminate both random and systematic measurement errors, the variance for the mean value estimator $\bar{x}$ (based on the tested and revised questionnaire) is unbiased. The mean square error is then:

$$MSE_{pm}(\bar{x}) = Var_p(\bar{x}) = \frac{S^2_{x,U}}{n - n_t}\left(1 - \frac{n - n_t}{N}\right) \qquad (4)$$

When the mean square error (3) for $\bar{y}$ is compared with (4) for $\bar{x}$, we see that as soon as the sample size is not very small, the measures used to reduce bias will most quickly increase accuracy. But also a reduction of random measurement errors can considerably improve the precision of estimates.

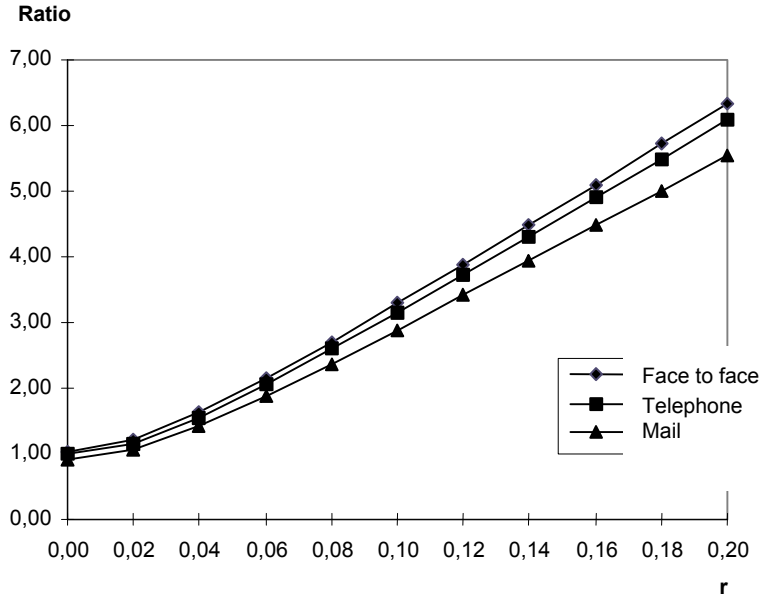
By using empirical data or reasonability assumptions in the formulas, we see how profitable it is to invest in systematic questionnaire development. The following is an example with the conditions:

- A simple standardised questionnaire test costs the same amount as 25 face-to-face interviews, 100 telephone interviews or 250 postal surveys (i.e. n_t = 25, 100 and 250, respectively). The relative costs are based on experience at Statistics Sweden.
- The sample size *n* is set at 1000, a common level in many one-time surveys.
- The components in the mean square error have the relationships $S^2_{x+b,U} + \bar{\sigma}^2_\varepsilon = kS^2_{x.U}$ and $\bar{b} = r * S_{x,U}$.
- The coefficient *k* is placed at 1.1, i.e. a variance increase with 10% as a result of the occurring random measurement errors.
- The coefficient *r* can go from 0.00 to 0.20. The highest value *r* = 0.20 corresponds to10 of 100 answer correctly instead of incorrectly in a question regarding a dichotomised variable where 50% have the quality sought after.

To see if r-values are realistic, one can make simple hypothetical calculations for a dichotomised variable. For example, assume that measurement is made for phenomena that occur in half of the cases. Then $p = 0.5$ and the population variance has the well-known form *p(1-p)* i.e. is here $(0.5)^2$. Then the value r = 0.02 corresponds to 1 additional correct response out of 100 ($\bar{b} = 0.01$ = r*0.50) and r = 0.20 that 10 more of 100 have given correct answers. (It is assumed that there are no random measurement errors but only systematic ones.)

The following diagram shows the square root of the ratio (3)/(4), i.e. the ratio of the mean square error for untested to that of revised questionnaires for the three most common data collection methods; face-to-face interviews, telephone interviews and postal questionnaires.

Square root of the ratio between MSE for untested (3) and tested and revised (4) questionnaires, where k = 1.10 and n=1000.



In the above diagram we see that as soon as a response bias is removed, the gain in accuracy is considerable. Only with postal questionnaires and when there is no bias is there an advantage not to test. Effects in the form of higher response rate, less need for checking and further contacts, and quicker responses that land on the positive side are not taken into account in this model. In practice, standardised tests are always preferable - unless there are small samples of less than several hundred survey objects and a very experienced questionnaire designer.
In many practical situations the effects will be less then what the picture tells. The measurement errors may not be removed totally and there may be non-response bias. Even so, advantages with a revised questionnaire are so great that

it would require very unfavourable conditions for it not to be a better alternative, unless the sample is very small.

This model shows most of all the importance of reducing the systematic measurement errors. But when one is modelling relationships and using variables for class breakdown or selections also random measurement errors may have a highly disturbing effect. In practice, the reduction of both types of measurement errors are concentrated upon rather than giving only one of them priority.

References

- Särndal, C-E.; Swensson, B.; Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag
- Akkerboom Hans and Håkan L Lindström (1997), *Terminology and Abbreviations* for the Workshop - Minimum Standards in Questionnaire Testing
- Nastasha Rees, Forms Consultancy Group, Australian Bureau of Statistics (1997) *Questionnaire Testing Concepts and their Definitions*. Prepared for Workshop on Minimum Standards in Questionnaire Testing