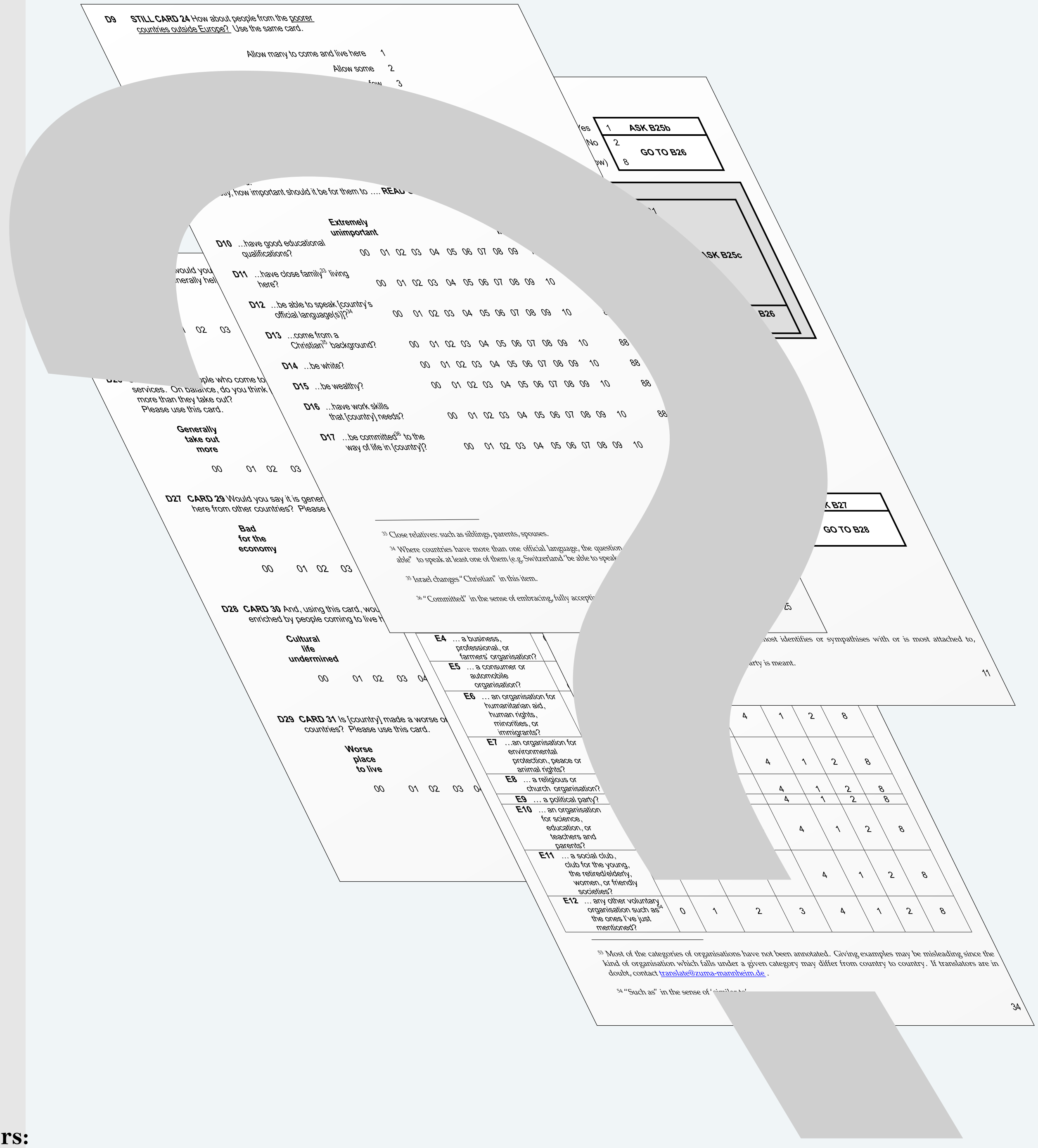
Handbook of Recommended Practices for Questionnarie Development and Testing in the European Statistical System

European Commission Grant Agreement 200410300002



Authors:

G. Brancato, S. Macchia, M. Murgia, M. Signore, G. Simeoni - **Italian National Institute of Statistics, ISTAT**

K. Blanke, T. Körner, A. Nimmergut - **Federal Statistical Office Germany, FSO**

P. Lima, R. Paulino - **National Statistical Institute of Portugal, INE**

J.H.P. Hoffmeyer-Zlotnik - **German Center for Survey Research and Methodology, ZUMA**

Version 1

Acknowledgements

We are grateful to the experts from the network countries who supported us in all relevant stages of the work: Anja Ahola, Dirkjan Beukenhorst, Trine Dale, Gustav Haraldsen. We also thank all colleagues from European and overseas NSIs who helped us in understanding the current practices and in the review of the draft version of the handbook.

Executive summary

Executive Summary

Questionnaires constitute the basis of every survey-based statistical measurement. They are by far the most important measurement instruments statisticians use to grasp the phenomena to be measured. Errors due to an insufficient questionnaire can hardly be compensated at later stages of the data collection process. Therefore, having systematic questionnaire design and testing procedures in place is vital for data quality, particularly for a minimisation of the measurement error.

Against this background, the Directors General of the members of the European Statistical System (ESS) stressed the importance of questionnaire design and testing in the European Statistics Code of Practice, endorsed in February 2005. Principle 8 of the Code states that "appropriate statistical procedures, implemented from data collection to data validation, must underpin quality statistics." One of the indicators referring to this principle requires that "questionnaires are systematically tested prior to the data collection."

Taking the Code of Practice as a starting point, this Recommended Practice Manual aims at further specifying the requirements of the Code of Practice. The Recommended Practice Manual is structured into two parts. 1) The present Executive Summary of the Recommended Practices for Questionnaire Development and Testing in the European Statistical System summarises the requirements for questionnaire design and testing in the European Statistical System (ESS). At the same time it briefly presents tools and methods considered as appropriate. It finally contains recommendations on how to develop strategies towards questionnaire design and testing tailored to the requirements of specific statistics. 2) The Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System describes the methods and tools in detail and gives practical hints and recommendations for their application. Furthermore, it presents the theoretical background of the question-answer process and contains suggestions for further readings.

The Recommended Practices were developed with the financial support of Eurostat.

1. Scope of the Recommended Practices

The European Statistics Code of Practice determines the requirements for questionnaire design and testing on a very general level. The Recommended Practices presented here further specify these requirements. In analogy to the Code of Practice, the Recommended Practices apply to all "Community statistics as defined in Council regulation (EC) No 322/97 […], produced and disseminated by national statistical authorities and the Community's statistical authority (Eurostat)." All data collection instruments in these statistics have to provide valid and reliable results, i.e. make sure that survey questions

- are understood and answered correctly by the respondents,
- can be administered properly by interviewers (if relevant) and
- do not adversely affect survey cooperation.

However, it is recommended to apply the Recommended Practices also in all further data collections carried out by National Statistical Institutes at the national and regional levels.

Questionnaire design, according to the Code of Practice, has to make sure that European Statistics "accurately and reliably portray reality" (Principle 12). Hence, the wording, structure and layout of all questionnaires must lead to valid and reliable results. The accuracy of the measurement clearly is the key requirement of the code. Nevertheless, the Code of Practice names a number of further requirements which have some impact on questionnaire design. These requirements include the limitation of the response burden (necessitating e.g. a respondent-friendly design), the automatisation of routine clerical operations like data capture (necessitating a computer-assisted or OCR-ready questionnaire) as well as the use of the full productivity potential of information and communications technology in data collection and processing (like in questionnaires using CAPI, CATI, and CASI technology). Taking all these requirements into account makes questionnaire design a complex scientific task.

The process of questionnaire design includes various successive steps: the development of a conceptual framework, writing and sequencing the questions, making proper use of visual design elements as well as implementing electronic questionnaires technically. In order to achieve cross-national comparability in European or international surveys, two further tasks are necessary. The translations of the questions or questionnaires have to be functionally equivalent, i.e. the respondents in different countries must have the same understanding of the questions. The demographic as well as socio-economic variables have to be harmonised through commonly accepted instruments. Therefore, possible approaches towards the translation of questions and a number of tested and accepted measurement instruments for demographic and socio-economic variables are briefly outlined in the following paragraphs.

Regarding *questionnaire testing*, two basic requirements of the Code of Practice have to be distinguished. In all surveys of European statistics, questionnaires have to be tested 1) in a systematic way and 2) prior to the data collection. This relates to both paper-and-pencil as well as computer-assisted modes of data collection, carried out either in a self-administered or interviewer-assisted way. When surveys are conducted using multiple modes of data collection, all questionnaire versions should be tested.

1) The first requirement indicates that, in any European Statistics, questionnaire testing has to be carried out using *systematic methods*. This means that the methodology used has to be sound, needs to be applied in a specific order, and needs to be appropriate for the specific requirements of each individual survey. A choice has to be made from a wide range of methods used under field conditions (field testing) as well as under laboratory conditions (pre-field testing). For example, testing a questionnaire by using only informal methods, or by using respondent debriefings as the only method, would clearly not meet the requirement of carrying out the test in a systematic way. In any case, a consistent testing strategy has to be developed individually for each new and ongoing survey.
2) The second requirement states that every questionnaire has to be (systematically) tested *prior to being used for collecting data* for European statistics. This requirement covers all existing questionnaires (given that a systematic test did not yet take place or it is evident that some questions need improvement) as well as new questionnaires. It implies that a questionnaire should undergo systematic testing if one or more of the following circumstances apply:
    - legislative changes mandate a new survey,
    - new questions, which were formerly not tested by the statistical institute, have to be asked,
    - questions in existing surveys are being modified, even if apparently minor changes are made,
    - the data collection instrument has to be changed (e.g. use of computer-assisted interviewing) or an additional data collection mode is introduced (e.g. web surveys in parallel to mail surveys), or
    - poor data quality has been indicated by a review of nonresponse rates and biases, validations against other surveys or re-interview studies, deficiencies in internal consistency or other evidence.

If the necessary tests are not conducted, or if sufficient evidence from existing research is not presented, the Code of Practice implies that European statistics must not include the questions in the survey. If the tests are conducted and the results show that a question performs poorly or adversely affects responses to other survey questions or survey response rates, European statistics must not use the survey question without modifying and testing it further.

2. Recommended Practices for Questionnaire Design

The questionnaire in the first instance is a measurement instrument. Its main purpose is to operationalise the user's information demand into a format which allows a statistical measurement. The concepts of "reality" must be operationalised in a way that enables the subject-matter specialists and users to carry out the necessary analyses, that the questionnaire designer can implement into the questionnaire, and that the respondents can understand and answer properly. Hence, the design of questionnaires must primarily take into account the statistical requirements of data users. In order to provide a valid and reliable measurement, the wording, structure, and layout must make allowance for the nature and characteristics of the respondent population. Further requirements include the organisational setting of the statistical office and the technical requirements for data processing. Additionally, the questionnaire should impose the lowest possible response burden and remain both respondent- and interviewer-friendly.

The questionnaire should ask relevant questions and permit data to be collected efficiently and with minimum error, while facilitating the coding and capture of data and minimising the amount of editing and imputation that is required.

The following recommendations support an efficient implementation of the questionnaire design process.

■ Literature review

Every questionnaire design should start with a review of the existing literature as well as other data sources (including available registers and non-official statistics) or surveys with similar research topics.

■ Specification of the survey objectives

Before starting to draft the questionnaire, the survey objectives should be specified in co-operation with important users and stakeholders. At the same time, decisions should be made on basic concepts such as the target population, on sampling design, and also on the available resources and possibly the the data collection mode to be preferred. Especially in the case of new surveys or major redesigns of surveys, an intensive user-focused consultation should be conducted in order to identify the concepts required. Besides informal communication with experts and users, expert group meetings involving key users should be held in order to identify the concrete information demand to be satisfied. These user requirements should be documented, and reconciled (in case of trade-offs).

■ Conceptualisation and operationalisation

Once the survey concepts have been established, they need to be translated into observable variables. The complexity of the theoretical concepts – even in apparently simple cases – requires a strict selection of empirical traits (often referred to as indicators) which can be observed in a survey. These indicators are deemed to be a suitable (though never direct or perfect) representation of the concept. For example, if the concept under consideration is poverty, one possible operationalisation is income, again subdivided into various subdimensions. Although this translation is at least in part based on conventions within the scientific community, some methods like dimension/indicator analysis, semantic analysis, facet design, content sampling, symbolic interactionism and concept mapping can facilitate this crucial step.

■ Exploring concepts: focus groups and in-depth interviews

Especially in new surveys it should be explored whether the concepts and indicators the users are seeking are compatible with those the respondents have in mind. A number of qualitative methods are available in order to get an idea of how respondents think about what survey designers have conceived. Such methods include for instance focus groups and in-depth interviewing.

Focus groups are composed of a small number of target population members guided by a moderator. The objective is to learn how potential respondents use the terms related to the topic, how they understand general concepts or specific terminology, and how they perceive the questions in terms of sensitivity or difficulty. Additionally, focus groups could also be a useful method for pre-field testing of questionnaires.

In-depth or qualitative interviews in a similar way focus on the respondents' understanding of the concepts, on how the respondents interpret certain questions and on how they arrive at their answers. In contrast to focus groups, in-depth interviews are not based on a group discussion.

When specifying the survey concepts and variables, standard definitions of concepts, variables, classifications, statistical units and populations should be used (if available).

■ Definition of a list of variables and a tabulation plan

Once the objectives and concepts are defined, a concrete tabulation plan and a list of variables should be laid down specifying the expected output of the survey. The variable and value list is to be seen as a simple list of variable names and values as well as of the corresponding definitions. With regard to the variable list, it is recommended to draw a distinction between background variables (e.g. demographic variables), variables used to measure the survey concepts, and technical variables (e.g. required for weighting).

For the specification of the list of variables, it can be useful to use questionnaire schemes in order to illustrate the contents and scope of the questionnaire in a systematic way. One suitable approach are “entity/relationship schemes” (or “conceptual schemes”). These schemes provide an overview on units, items and relationships covered by the survey.

- **Decision on the data collection mode**

The selection of an appropriate data collection mode must take into account the number, the contents and the scope of the survey variables. Also, aspects like the sensitivity of the questions or possible problems of information retrieval should be considered. Additionally, possible preferences of the target population should be taken into consideration. It is important to note that the questionnaire should not be designed without a previous decision on the data collection mode. Standard data collection modes are listed in table below.

Data collection modes

Technology	**Type of administration**	
	Interviewer administration	**Self-administration**
Computer-assisted data collection	CAPI, CATI	CASI: WBS (or CAWI), DBM, EMS and TDE
Paper and Pencil (PAPI)	PAPI face-to-face interview	PAPI (mail surveys)

CAPI: Computer-Assisted Personal Interviewing, CATI: Computer-Assisted Telephone Interviewing, CASI: Computer-Assisted Self-Interviewing, WBS: Web Based Survey, EMS: E-Mail Survey, TDE: Touch-tone Data Entry, PAPI: Paper-and-Pencil Interviewing, DBM: Disk by Mail, CAWI: Computer-Assisted Web-Interviewing.

- **Writing and sequencing the questions**

Questionnaires must translate the variables to be measured into a language that respondents understand. It is essential that the words used in questionnaires have the same meaning for all the respondents and at the same time correspond to the concepts to be measured. Additionally, one should only ask questions on which the respondents can actually retrieve the necessary information.

In order to be understood correctly, questions should be worded using *simple* and *unequivocal* terms which have the same meaning for all members of the target population. Notions which respondents might not be familiar with should be defined. *Long* or *complex questions* as well as *hypothetical questions*, *loaded questions*, and questions with *double negations* should be avoided.

The questionnaire designer has to be particularly careful when choosing the *stimuli* of the questions. The stimuli should measure the variables of interest appropriately. Never more than one stimulus should be used in a question. Every question should be explicit about the *reference period*, which is best placed before the stimulus of the question as such. In the case of business surveys, questions should be chosen which are compatible with the reference periods and the response categories of the establishments' record-keeping practices.

The answer categories must cover all possible responses and correspond as closely as possible to the variation in the population. Furthermore they must be disjoint, i.e. there should be no overlaps leading to ambiguous results.

The respondents and/or interviewers should be provided with clear *instructions*. Instructions should however only be used if necessary and should be kept reasonably short. In order not to get overlooked they have to be placed close to the question to which they refer.

Also, the *questionnaire structure* should encourage respondents to answer the questions as accurately as possible. To this end, the questionnaire must focus on the topic of the survey, be as brief as possible, flow smoothly from one question to the next, facilitate respondents' recall and direct them to the appropriate information source. Respondents should have the feeling that filling in the questionnaire is an interesting and diverting task. In the introduction to the questionnaire, the title or subject of the survey should be provided, the sponsor should be identified, the purpose of the survey should be explained, and the respondent should be asked for his or her cooperation. Also, the statistical office carrying out the survey, the confidentiality protection measures, any record linkage plans, and any data sharing arrangements that are in place should be (briefly) indicated.

The *sequence* of the questions should be self-evident to the respondents, so that the questionnaire follows a logical stream. Questions should be arranged into logical groups. Questions on the same topic should be grouped in the same section. With regard to section ordering, one should proceed from less to more complex topics. The use of checks should be carefully evaluated against the increase of difficulties in controlling the interview. The opening questions should be applicable to all respondents, be easy and interesting to complete, and establish that the respondent is a member of the survey population. At the end of the questionnaire, space for additional comments by respondents should be provided and an expression of appreciation to the respondent should be included.

Apart from creating a smooth logical flow of the questions, one should be aware of the fact that the *order of questions* could have a strong impact on the results. Therefore, the questions must be scrutinised with respect to their context. The respondent's interpretation of a certain question may be influenced by earlier questions. To facilitate the measurement of sensitive questions, the topic should be reached gradually and in a way that makes the respondent feel at ease, and the wording should be as neutral as possible. In general, sensitive questions should not be placed at the beginning of the questionnaire. In order to avoid context effects, bias due to the sequence of questions should be addressed during the questionnaire testing.

■ Visual design elements

Besides the wording and structure of the questionnaire, visual design elements require some consideration. Not only verbal, but also non-verbal stimuli guide the respondents and interviewers when filling in the questionnaire. Using the right non-verbal stimuli (like graphical symbols and layout) is vital in order to achieve valid and reliable estimates and to enhance the usability for both respondents and interviewers. When looking at visual design elements, the emotional, functional and reflective levels have to be distinguished.

At the *emotional level* (determined by the connection of the brain with the visceral nervous system), the respondent and the interviewer produce positive or negative reactions (emotions) without reasoning. In order to be understood correctly, it is important that a questionnaire produces immediate positive reactions. It is therefore much more than just a cosmetic detail that questionnaires should be visually pleasing and look easy to complete. For the same reason, it is vital to give a positive first impression in the cover letter and on the front cover, and to make the questionnaire appear professional and businesslike as well as (if interviewer-administered) interviewer-friendly.

The *functional level* determines the usability of the questionnaire, i.e. whether the information is cognitively processed by the respondent as intended by the survey designer: The structure of tasks as well as the cognitive processes required of the respondent and the interviewer should be designed in a way as to enable them to give the correct answer. Therefore, questionnaire designers have to be aware of how non-verbal information is processed. There are a number of general findings which should be taken into account when constructing a questionnaire. (1) The structure of tasks should not be too complex (e.g. one question should be asked at a time, the number of items per question should be reduced, and complex matrices should be avoided etc.). (2) The questionnaire should make use of the "natural" mappings of the respondents, so that it is self-evident how to navigate through the questionnaire. For example, the questions should start where the respondent and interviewer expect them to start, i.e. in the upper left corner of the page or the screen. Elements that belong together in the questionnaire should be grouped together (like the question, the respective answer space and response options, as well as the instructions on that question). The visual design should also make clear where the respondents and interviewers are expected to enter their responses (e.g. by the use of a figure/ground contrast). Skip instructions should always use strong visual symbols (like arrows). (3) Where the questionnaire cannot make reference to a pre-established conceptual model (mapping) in the mind of the respondent or interviewer, one should standardise the verbal as well as the design elements for similar questions throughout the questionnaire.

Finally, the respondents take a conscious and reflected decision when participating in a survey or answering survey questions. This level is referred to as the *reflective level*. At the reflective level, respondents and interviewers attribute meaning also to non-verbal features of the questionnaire and to their activity of providing answers. Those involved in questionnaire design should be aware that this can influence the readiness to participate in a survey as well as the responses to the questions.

■ Electronic questionnaire design

A number of special considerations apply regarding the design of electronic questionnaires, as used for CATI and CAPI surveys. Electronic questionnaires of course have the same main objectives as PAPI questionnaires, namely to collect information according to the survey information needs. Given the special opportunities due to the electronic implementation, however, the design of the screen layout, the management of the various types of errors and the wording have to be tailored to the needs of both the respondent and the interviewer. The screen layout setting must enable the interviewer immediately to understand what to read and where to find it. At the same time, electronic questionnaires enable us to detect and reconcile inconsistencies already during the interview. The electronic questionnaire must be designed in such a way as to solve the greatest number of inconsistencies while paying attention to the fluency of the interview and not frustrating the respondent. The error management should make the interviewer understand sooner what kind of error happened and which questions were involved in it. The customisation of texts is to help the operator in reminding the respondent of the information previously collected.

An electronic questionnaire should allow the possibility of modifying responses previously given, without missing any relevant information (i.e. to allow a change of route when backing up). If the questionnaire contains ad hoc modules and some of them are “stand alone”, it should allow the possibility of completing them in any order. In the case of CATI surveys, the call scheduler must be planned in such a way as to give the sample units the same probability of being contacted without changing their inclusion probability.

From a technical point of view, an electronic questionnaire must measure what it is intended to measure (*effective),* be easy to modify when changes occur and easy to correct in case of errors (*flexible*), be structured in modules which may be used for other studies or survey waves, be easily adaptable to different hardware or software platforms (*portable*) and be *efficient* in terms of response time for screen replacement.

Two testing methods are particularly useful when the data collection mode is computer-assisted, hence implying the implementation of a software for the interview management: the functionality and the usability tests. The former consists in the assessment that the electronic questionnaire performs according to the given specifications and is robust with respect to unexpected events. In the latter, the interest shifts towards the users of the system, thus testing if they use the system in a correct and efficient way.

■ Cross-national harmonisation of questions

In order to achieve cross-national comparability in international and European surveys, the questionnaires used in different countries have to be functionally equivalent, i.e. must actually measure the same concepts in different cultures and languages. Therefore, in cross-national research, the translation of questionnaires plays an important role. Secondly, the harmonisation of demographic and socio-economic variables is a further prerequisite to cross-national comparability.

Functionally equivalent translations require quite sophisticated techniques for translation, taking into account the syntactic, semantic and pragmatic levels of the source language questionnaire. Normally, cross-national research starts with the agreement that one language (mostly English) is used as reference language. A drafting group formulates the questions of the questionnaire. Native speakers of the reference language (English) in this drafting group will be not only language experts, but also experts for the cultural background of the concepts and questions. During the actual translation process ideally a bilingual but "unicultural" member translates the questionnaire. It is important to note that cultural differences are ignored if an item is only translated from one language into another without analysing the cultural background of the national question wording in the (English) master copy. Normally, two independent draft translations should be made per language. A pretest should be an integral part of the translation process.

Besides appropriate translation techniques, it is important to use harmonised demographic and socio-economic variables. Context variables or background variables are variables that contain information necessary to define homogeneous subgroups, to establish causal relations between attitudes and societal facts, and to define differences between scores on scales. In cross-national research, standardised instruments or indices exist only for a very small group of variables including "occupation", "education", and "status". For the variables "income", "family", and "ethnicity", there is preparatory work in progress.

4. Recommended Practices for Questionnaire Testing

Questionnaire testing is critical for identifying problems for both respondents and interviewers with regard to, e.g. question wording and content, order/context effects, and visual design. Problems with question wording include, for example, confusion with the overall meaning of the question as well as misinterpretation of individual terms or concepts. Problems with skip instructions may result in missing data and frustration of the interviewers and/or respondents. Visual design especially concerns self-administered questionnaires; however, poor visual design can easily lead to confusion and inappropriate measurements also in interviewer-administered surveys.

Questionnaire testing is a broad term that incorporates many different methods or combinations of methods. This section summarises and briefly describes those methods that should be used to test and evaluate questionnaires in the ESS. These methods have specific strengths and weaknesses that make them valuable for different types of problems. Consequently, they are useful at different stages of questionnaire development. In order to get a sound understanding of the response process, to identify problem questions and to suggest adequate improvements, the use of a combination of questionnaire testing methods is indispensable. In most cases, the use of one single method will not be sufficient. The appropriate combination of these methods determines their suitability to meeting the objectives of questionnaire testing.

In the Recommended Practices, we distinguish two major categories of questionnaire testing methods – pre-field and field methods. This distinction is an analytical one, but nevertheless reflects some principal and operational differences.

Pre-field methods are normally (although not necessarily) applied under laboratory conditions. This means that the interviews are not carried out in exactly the same way as later on in the field. “Laboratory conditions” refer to an observational environment which may totally or partially differ from the actual field conditions. The interviews may not be carried out in the same environment as in the field (e.g. in a cognitive laboratory). Only small parts of the questionnaire might be included. Additional questions might be added with regard to how the respondents perceive the questions. Consequently, the interview flow might deviate from the later fieldwork quite substantially. They are generally used in a preliminary stage of the questionnaire testing process and mostly qualitative in nature. Pre-field methods are particularly suitable to collect information on how respondents proceed when answering the questions. Often, the focus is on single questions rather than the whole questionnaire. They include expert group reviews and cognitive interviews such as think aloud interviews, probing, respondent debriefings, confidence ratings, paraphrasing, sorting, vignette techniques, and analyses of response latencies. One method sometimes described in the literature as a pre-field method is the use of focus groups, that we treated in the design of the questionnaire, since it is more closely related to a preliminary analysis and the development of concepts than to the actual testing of a draft questionnaire. Nevertheless, focus groups might also play a role in pre-field testing.

Field methods are those used to evaluate questionnaires tested under field conditions. This means that the interview is carried out in a way very similar to the subsequent fieldwork (regarding setting, lengths, choice and order of questions, etc.), and the majority of the conditions mirror the real survey situations. The test might be conducted during the data collection phase, e.g. in the context of a pilot study, in conjunction with the actual data collection, or in parallel to ongoing or recurring surveys. Therefore, field testing often includes bigger sample sizes and allows quantitative analyses. The focus is more on the complete questionnaire instead of individual questions. Field methods include behaviour coding, interviewer debriefings, respondent debriefings, follow-up interviews, experiments.

It is important to note that, in practice, some of the methods could be used either under laboratory or under field conditions. In our classification we refer to what can be seen as the major area of application.

Other techniques, based on the analysis of item non-response, editing and imputation rates, or response distributions, are commonly performed on the data coming from the testing methods as well as on data from the real data collection phase. In the last case, these analyses result particularly useful especially for ongoing surveys. These are referred in the handbook as post-evaluation methods.

■ Pre-field methods

Expert groups are the only testing method discussed here which does not involve the respondents. For this reason, when using expert groups, some other testing method involving the respondents should be applied additionally. Expert groups are composed of survey methodologists or questionnaire design experts, but also subject-matter experts. These senior survey researchers should have experience with potential problems of questions and/or questionnaires from other surveys. The objective of expert groups is to evaluate the questionnaire for potential problems for either interviewers or respondents. Usually these reviews are conducted early in the questionnaire development process. There are two ways of conducting expert groups: (1) by structured discussion on each question, with a standardised coding scheme or (2) in the form of a not formally structured discussion on each question, but without a standardised coding list.

Observational interviews are frequently used in order to identify problems in the wording, question order, visual design etc. of self-administered questionnaires. They also provide reliable estimates of the time needed to complete the questionnaire. During the interview, which is ideally carried out in a laboratory and video-recorded, the respondent's behaviour (e.g. whether all the questions and instructions are read before answering) and observed cognitive processes (e.g. counting on fingers or writing calculations on the page) are watched closely in order detect questionnaire features which might produce measurement errors. In combination with observational interviews, follow-up probes are often used in order to elicit information about the reasons and motives for the respondent's behaviour.

Cognitive interviews are typically used after a questionnaire was constructed based on focus groups and has been improved in expert groups. The objective of cognitive interviews is to obtain qualitative information on how the questions are understood and answered by actual respondents. The idea is that an understanding of cognitive processes supports the development of design rules that govern the choice of the response categories, the question ordering and visual design etc. They consist of one-on-one in-depth interviews in which respondents describe their thoughts while answering the survey questions or after having answered the questions. Cognitive interviews are usually carried out in somewhat controlled rooms like labs or other suitable rooms and are recorded on video or tape for further analyses. The interviewer as well as the test person should play an active role. However, laboratory interviews may in some cases be impractical or unsuitable. In business surveys, for example, cognitive interviews are usually conducted on-site at the offices or location of the business or institutional respondent, where the respondent has access to the necessary records.

Cognitive interviews are usually carried out for a subset of particularly important and/or difficult questions. Depending on the requirements of a survey, various techniques can be used:

- In a *think aloud* interview, the test person is asked to speak all thoughts aloud which lead to an answer. Think aloud sessions are used to identify difficulties in question comprehension, misperceptions of the response task, types of recall strategies used and reactions to sensitive questions. They can be carried out either concurrently or retrospectively. However, their effectiveness in testing depends heavily on the ability and willingness of the test persons to articulate their thoughts.
- *Probing* examines an answer with the help of additional questions that are asked by the interviewer to achieve additional information. Probing questions are used when the information provided by the test person is incomplete or has revealed potential problems with a specific question. The researcher asks, for instance, how respondents made their choice among the response options or how they interpreted an instruction. It is important for the formulation of probes that one should already have an idea or hypothesis about what the cognitive difficulty of a question is. Probes can be categorised in different ways. Either they are grouped according to the time they are asked (follow-up or retrospective) or they are grouped according to whether they aim at getting information about how the question was understood (comprehension) versus how the test person arrived at a particular

answer (information retrieval). In business surveys, for example, in which data retrieval often involves business records, probing techniques are used which ask the respondents to describe those records and their contents or to show the records to the researcher.

- *Paraphrasing* means that, after answering, the test persons have to repeat the questions in their own words. This permits to examine whether the test person understands the question and interprets it in the manner intended. Paraphrasing may also reveal better wordings for questions, for example, if different test persons consistently use the same terminology. There are two main approaches towards paraphrasing: either it is tried to remember the question word by word or the question's context is repeated in own words. Usually it is easier to assess whether the test person understood a question if the person repeats it in his or her own words. Paraphrasing is especially useful to detect complex and/or confusing questions.
- *Confidence ratings* ask the test person to assess the degree of reliability of their answers with the help of a given scale. Confidence ratings attempt to identify questions that test persons find difficult to answer by having them rate their level of confidence in the answer they have provided. The theory is that low confidence ratings are often the result of a lack of knowledge or a difficult recall task.
- *Sorting* aims at providing information on how test persons categorise terms or understand the underlying concepts. The test persons are given a set of objects that are arranged in no particular order and are asked to sort them according to their own or given criteria. Sorting procedures provide an efficient method for assessing perceived similarities between a large number of stimulus objects within a particular domain (for example visual patterns, social institutions, political policies, consumer products etc.). An analysis of the resulting subjective groups reveals information on how test persons cognitively represent and organise their knowledge in that domain but it may also provide a means for assessing differences between respondent populations in such cognitive representations.
- *Vignette classifications* can be regarded as a special type of sorting. Test persons are given certain descriptions as vignettes to determine by themselves if these have to be included into the answering process or not, or to decide if they relate to a particular survey question or concept. A hypothetical vignette may describe a certain type of behaviour or activity. The test person might be asked: "How should the person described in this scenario respond to the following question?"
- *Response latency* is the time measured between question and answer either by a stop watch or qualitative estimation (called qualitative timing). Response latency methods have been used to identify problems with the interpretation of questions, memory retrieval and response selection. Response latencies can be measured particularly easily in all types of computer-assisted data collection.

■ Field methods

Testing the questions in the field is a crucial stage in the development and assessment of survey questions along with the development and testing of associated procedures and instructions. No matter how much development work has been done, the questionnaire has to be tested under field conditions, and in particular the questions have to be tested in conjunction with the mode in which they will be used. Field methods can be applied in a field test, in pilot surveys or in ongoing surveys. Various techniques are used for field testing:

- *Behaviour coding* consists of systematic coding of the interaction between the interviewer and the respondent from live or taped interviews in order to evaluate the quality of the questionnaire. The main objective of behaviour coding is to provide information on the quality of the questionnaire as a standardised measurement instrument. In most cases, behaviour coding is conducted in the field, but it can also be used in the laboratory. By the assistance of a coding list the reactions of interviewers and respondents are classified and registered systematically, so that frequencies of reactions on each question can be provided.
- *Interviewer debriefings* consist of an organised discussion of the questionnaire between interviewers who conducted the fieldwork, and the designers/researchers. The aim is to obtain useful feedback from the interviewers on the performance of the questionnaire to get a better understanding of questionnaire weaknesses. This technique combines standardised interviewer debriefing questionnaires and focus group style interviews to gather information from interviewers about either a previously used survey instrument or a draft instrument. Although the information is indirect and possibly filtered, it reveals some of the difficulties experienced by respondents. Interviewer

debriefings are generally carried out after field tests and/or after ongoing surveys, rather than during earlier stages of the survey development process.

- *Respondent debriefings* involve incorporating structured follow-up questions at the end of a field test interview or focus group style discussions with other interviewers to elicit quantitative and qualitative information about the respondents' interpretations of survey questions. The primary objective is to determine whether concepts and questions are understood by respondents in the same way as intended by the questionnaire designers. A critical aspect of a successful respondent debriefing is that question designers and researchers must have a clear idea about potential problems so that good debriefing questions can be developed. Ideas about potential problems can stem from pre-field techniques conducted prior to the field test, from an analysis of data of a previous survey, from a careful review of questionnaires, or from an observation by the interviewers.
- *Follow-up interviews* are intensive semi-structured interviews which are conducted by another interviewer shortly after the actual survey interview. Respondents are interviewed on how they answered selected questions in the first interview. They are encouraged to remember how they interpreted and understood the questions, both overall and in terms of particular words and phrases. They may also be asked about how they arrived at their answers, how accurately they felt the answer given actually reflected their views and how much importance they attached to arriving at an accurate answer. Follow-up interviews tend to be lengthy and therefore have to be limited to selected questions.

In *experiments* different variants of a questionnaire (or data collection methods) are tested in a controlled way in order to determine measurement effects. Experiments may be conducted as part of a field or pilot test, or they may be embedded in ongoing surveys. Experiments can be performed on single questions, a set of questions or an entire questionnaire. The sample size in an experiment must be large enough so that significant differences can be measured. In addition, it is imperative that the respondents are assigned to the different variants in a randomised way so that differences can be attributed to the question or questionnaire and not to the effects of incomparable samples.

■ Post-evaluation methods

Post-evaluation methods represent a set of analyses of the data commonly performed by the survey practitioners, indirectly aimed at evaluating the quality of specific questions.

- *Analysis of item nonresponse rates*. The questions that present the highest nonresponse rates are commonly investigated to search for the possible causes. More detailed analysis can take into consideration the nonresponse in combinations of variables, and study if some nonresponse patterns are associated to certain respondents' profiles.
- *Analysis of response distributions*. The analysis of the response distributions is particularly useful in order to compare different questionnaire versions or for validation with external data, and usually is performed in conjunction with other methods, such as the respondent debriefing.
- *Analysis of the editing and imputation phase*. The amount of edit failures can suggest possible problems with given questions. Therefore the evaluation of this phase should be exploited for preventing error in next editions of the survey.
- *Reinterview studies*. The results of reinterview studies, in general estimating simple response variance or response bias, can be used to explore hypotheses on problematic questions.

5. Towards a Strategy for Questionnaire Design and Testing

From the specification of the concept all the way to the use of the questionnaire for the actual data collection, questionnaire design and testing should be guided by a consistent strategy. The strategy should be adapted to the conditions in a particular survey and allow a maximum number of problems to be detected and eliminated by the selected testing methods. For example, questionnaires for new surveys need different testing approaches than those for surveys which have already been running for several years. For the latter, the effects of – even apparently minor – changes in the questionnaire on the continuity of the data have to be observed carefully, necessitating an experimental field test before the actual data collection. Establishment surveys have other requirements than household surveys, given the sometimes fundamental differences in the response process. In many establishment surveys (compared to many household surveys), the focus is much more on the process of information retrieval. Cognitive interviews, for this reason, will often have to be carried out on-site at the enterprise. And finally, the testing strategy has to take into account the possibilities and limitations of the data collection mode. For technical reasons, CATI and also CAPI surveys are for example particularly suited for behaviour coding and response latency based approaches.

When developing a strategy, the entire cycle of questionnaire design and testing has to be covered. Five main steps have to be distinguished (see figure 1). All five steps must be covered by the strategy.

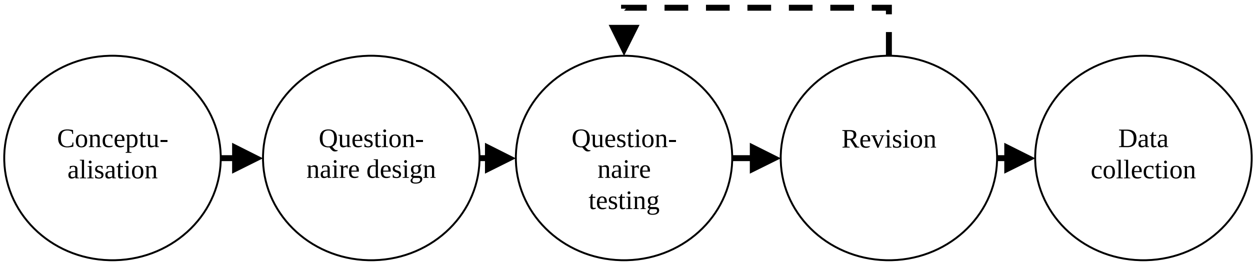
1) *Conceptualisation:* In contrast to what one might expect, questionnaire design does not start with the questionnaire. Before one can even start to think about the wording of the questions, the conceptual basis of the questionnaire has to be specified and operationalised. The complexity of the theoretical concepts – even in apparently simple cases – requires a strict selection of empirical traits (often referred to as indicators) which can be observed in a survey. These indicators are deemed to be a suitable (though never direct or perfect) representation of the concept. Although this translation is at least in part based on conventions within the scientific community, some methods like facet design, symbolic interactionism, or semantic analysis facilitate this crucial step.

   During this phase there are mainly two perspectives to cover: (a) What are the definitions the users and subject matter experts are interested in? (b) And how do those definitions match with the conceptual models of the respondents? Expert groups, focus groups and in-depth interviews are important instruments in this process. The work on the conceptual frame is of course more important for entirely new surveys, whereas in existing surveys the concepts may already be well established. In this case, less attention is required at this stage. However, the (further) development of the conceptual basis must be an integral part of almost every change in a questionnaire. The main output of this stage is the list of target variables (besides basic decisions on the target population, data collection mode etc.).

2) *Questionnaire design:* After the conceptual basis has been specified and a suitable data collection mode defined, a first draft questionnaire can be designed. Questionnaire design starts with the structure of the questionnaire. Based on the contents of the questionnaire and the requirements of the data collection method, the sequence of the thematic sections in the questionnaire is decided. The wording of the individual questions translates the target variables into concrete questions. Apart from the basic rules described in section 3.2 and 3.3 of the handbook, no particular methods and tools can be recommended for this stage. It has to be noted that the way a questionnaire is worded and structured is heavily dependent on the data collection mode chosen. For example, a computer-assisted questionnaire has different requirements and possibilities than a PAPI instrument. In CATI surveys, questions have to be short and the number of response options strictly limited. Slightly different rules apply to self-administered questionnaires, as the respondents do not have the opportunity to ask an interviewer for further instructions.

   Once a draft wording is available, the questionnaire has to be implemented technically. During the implementation, visual design elements must be considered. For computer-assisted questionnaires, an effective usability and functionality testing should take place. PAPI questionnaires should be designed using a professional desktop publishing software package. Standard text processing applications usually do not fulfil the requirements of a professional layout.

The five stages of questionnaire design and testing



3) *Questionnaire testing:* A minimum recommendation is to test the questionnaire at least once by contacting potential respondents (not only experts). Questionnaire testing should not start unless a fully implemented draft questionnaire is available. At least a version should be used that is very similar to the one which will be used for the actual data collection. A wide range of pre-field and

field methods can be used to test the questionnaire. The suitability of using one or more of these methods to test a particular questionnaire and the intensity of testing depend on various factors and circumstances. These include the type and size of the survey, the characteristics of the target population, the survey's content, the utilisation of previously tested standard questions, and the method of data collection. Furthermore, the testing approach depends on whether it is an ongoing survey or not, and last but not least on the project schedule and the available budget.

Entirely new surveys require the most intensive testing. In these cases, it is recommended to use at least one of the pre-field testing methods (e.g. cognitive interviews) and (after a first revision) a larger scale pilot study under field conditions. Furthermore, in new surveys, the entire questionnaire should be tested. In ongoing surveys, the testing can sometimes be reduced to questions which have proven to be problematic (e.g. post survey evaluation like analysis of item nonresponse rates, editing and imputation rates, inconsistent results, or external validation). In these cases questionnaire testing is indispensable also in ongoing surveys, but might be restricted to some questions or modules of the questionnaire. Nevertheless, questionnaire testing is always fruitful, even if no problems are obvious. A further requirement of ongoing surveys is the observation of possible effects of questionnaire changes on the time series. In this case an experimental design might be an appropriate solution. As regards revisions to questionnaires, the entire questionnaire or at least the revised portion of it should be tested.

Finally, in selecting respondents for the purpose of testing questionnaires, care should be taken to ensure an appropriate representation of the respondent population.

4) *Revision:* As a general rule, two or more phases of questionnaire testing are recommended. If a questionnaire has undergone changes following the results of the testing, a new round of testing is normally indispensable. This involves testing the questionnaire at an early stage of its development, making revisions to the questionnaire based on the test findings, and then testing the revised questionnaire. This process may be repeated through two, three or even more phases of testing. Different methods of testing the questionnaire may be used during each phase of testing. In ongoing surveys, the evaluation of former survey waves can provide important input for questionnaire revision.

It is vital that the testing strategy covers all stages described above. Defining an appropriate and efficient testing strategy requires some experience. It is therefore recommended to have a specialised unit inside the statistical office which can provide the survey managers with practical hints and support them in the implementation of the strategy. Some of the testing methods, like cognitive interviews, almost necessarily require some central facilities. In addition, even though it is widely accepted that testing is important to undertake, it is often neglected that these procedures also takes time and resources. Potential users and clients as well as researchers should take this fact into account when planning new surveys or revisions of ongoing surveys.

Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System

List of the acronyms

CAI: Computer Assisted Interviewing
CAPI: Computer Assisted Personal Interviewing
CASI: Computer Assisted Self-Interviewing
CATI: Computer Assisted Telephone Interviewing
CAWI: Computer Assisted Web Interviewing
CBM: Current Best Methods
CoP: Code of Practice
DBM: Disk By Mail
EQ: Electronic questionnaire
ESS: European Statistical System
LEG: Leadership Group
NSI: National Statistical Institute
PAPI: Paper and Pencil Interviewing
QDET: Questionnaire Development and Testing
RP: Recommended Practice
SPC: Statistical Programme Committee
TDE: Touch-tone Data Entry
WBS: Web Based Survey

Contents of the handbook

Chapter 1. Introduction

1.1. Background

As stated in the Quality Declaration of the European Statistical System (ESS), one of the major objectives of the ESS is the continuous improvement and harmonization of European statistics. Systematic questionnaire design and testing is essential to achieve this goal. The European Statistics Code of Practice (CoP), endorsed by the Statistical Programme Committee (SPC) in February 2005 (Commission of the European Community, 2005), further specifies this general objective. Its intention is "to reinforce the quality of the statistics produced and disseminated by the statistical authorities, by promoting the coherent application of best international statistical principles, methods and practices …". In the context of the ESS, Recommended Practices have been recognised as a highly suitable approach to identify a set of standard methods to be used under different conditions. Recommended Practices leave enough flexibility to the Member States, thus enabling further harmonisation.

The Code, based on 15 principles, whose implementation is reviewed periodically according to *Indicators of Good Practice*, implicitly and explicitly deals with the issue of questionnaire design and testing. It underlines, indeed, the importance of processes to guide in the planning of existing and emerging surveys (principle 4), the need for sound methodology supported by adequate tools, procedures and expertise (principle 7) and finally the adoption of appropriate statistical procedures from data collection to validation (principle 8). More specifically, it requires that, in case of statistical surveys, all questionnaires be systematically tested prior to data collection.

These objectives, together with the growing attention to data quality, have led to consider the development of standardising tools as an approach towards harmonization of statistical production. The decision to face the issue of questionnaire development and testing methods derived from many considerations. First of all, questionnaire quality is a keystone since it affects quality of data. From a European perspective, on the one hand, it is important to establish common methods for questionnaire design, pursuing ex-ante harmonisation. On the other hand, it is relevant to provide a systematic review of methods used in the testing phase which allow to reach acceptable quality levels. Although there is awareness of the importance of questionnaire and its testing prior to data production, there is lack of guidance about testing methods, and survey quality reports usually do not say whether questionnaires were tested and which were the results (Presser *et al.*, 2004).

This handbook was developed in the framework of the European Project "Recommended Practices for Questionnaire Development and Testing (RPs QDET)", carried out in 2005/06 with the financial support of European Commission. It meets Eurostat requirements to support the European Member States in the accomplishment of LEG on Quality Implementation projects. The LEG on Quality recommendation no. 11 states that "A set of recommended practices for statistics production should be developed. The work should start by developing recommended practices for a few areas followed by a test of their feasibility in the ESS" (LEG on Quality, 2001). The project was coordinated by the Italian National Institute of Statistics (Istat) and involved the participation, as partners, of the Federal Statistical Institute of Germany (FSO), the National Statistical Institute of Portugal (INE) and the German Centre for Survey Research and Methodology (ZUMA). The project could count on the support of experts from other European National Statistical Institutes (NSIs), and in particular Statistics Austria, Statistics Finland, Statistics Netherlands, Statistics Norway. The handbook has been discussed in the ESS Network of Quality Managers and revised after a feasibility study.

1.2. Objectives

The main objective of this handbook is to improve harmonisation and quality both at a national and European level. The questionnaire is the means by which data are collected and a communication tool, as well. This means that a questionnaire should collect what it is intended to collect in a correct way, facilitating the work of the interviewer with a limited burden for the respondent. Therefore, the first issue to face is the validity and reliability of the measurement instrument. The cognitive processes required from the respondents are far from simple, so that even apparently straightforward questions might be difficult to be submitted. Some concepts require a set of questions instead of a single one. Sometimes, even if a single question is able to identify a trait, its different formulation may affect the validity of resulting distribution of responses. In addition, the formulation of questions interacts with the respondent profile, i.e. some terms could be correctly understood by a given population subset rather than by others. Such a complexity leads to the need of a recurrent process of questionnaires design, testing, re-design and re-testing that requires an overall strategy integrating designing tools, laboratory cognitive testing methods and field testing, in order to refine questions before the questionnaire is used in the data collection phase. Furthermore, the presence of the interviewer adds a further element that should be considered when planning a survey and relevant activities for preventing and limiting errors. This manual reports a wide set of methods that can be used in this recursive process.

The underlying idea is that preventing errors by designing, testing and re-testing questionnaires is better practice than proceeding to a correction phase during the data treatment. Sometimes, evaluating quality during the survey process or at the end of it allows to identify problems in the questionnaire and to improve its quality for next edition of the survey. Such an approach is particularly relevant for repeated surveys.

At a European level, this handbook defines a set of suitable methods for questionnaire design and testing in the ESS. Depending on their national context, NSIs are free to choose those methods which suit their needs best. This handbook, thus, contributes to a further harmonisation of statistics production while providing NSIs with the possibility to select those methods which are most appropriate in their national context.

At a national level, this handbook aims at reducing unwilling variation and increasing the quality of the data collection phase, by improving and standardising the procedures adopted. At the level of individual surveys, the objectives of this handbook are to support survey managers in the adoption of sound methodology for questionnaire design and testing and to promote a systematic activity of testing new questionnaires or questionnaires already in use, but not appropriately tested before. Researchers and survey managers facing the objectives of planning a new questionnaire, changing or adding some questions within a questionnaire already in use, evaluating the quality of an adopted questionnaire, can find in the handbook a guide for their tasks. It can also be used as a checklist for evaluating the performed activities in the field of the development and testing of the questionnaire. In addition, the handbook can fruitfully be used for training new personnel or to study how to build up an optimal testing strategy tailored for different types of surveys.

1.3. Recommended Practices

This handbook has been developed as a recommended practice manual. Recommended Practices (RPs) are defined as a handbook that describes a collection of proven good methods for performing different statistical operations and their respective attribute (Bergdahl *et al.*, 2001). NSIs or survey managers can then choose among recommended practices those most suitable to guarantee the highest level of quality. Contrary to Quality Guidelines, that report “what to do” but not “how to do” it, RPs represent an operational guidance for survey managers. In addition, NSIs will find RPs more flexible compared to other standardisation tools like Current Best Methods (CBM) and Minimum Standards. Indeed, it is rather difficult to define minimum standards or best methods valid for a large group of countries, with different levels of complexity and different cultural and organisational contexts, whereas it is more useful to provide a set of “good” practices. RPs’ strong point is their capacity to motivate users to strive for excellence, induce improvements and reduce unjustified variation. They appear to be feasible to construct and accessible in the sense that they are easy to communicate and understand. They have been, therefore, identified as the best tool at a European level and the LEG on quality results suggested to invest more resources on the development of Recommended Practices rather than on Minimum Standards, whereas, translation in English, if not yet available, was recommended for existing CBMs.

1.4. Current Situation in the ESS

The development of this handbook started with an analysis of the current situation in the ESS, based on:

- Literature review
- Review on the existence and degree of application of recommended practices, quality guidelines, current best methods, minimum standards and manual in general on questionnaire design and testing, by means of a survey

The literature review highlighted a wide availability of papers and monographs on testing methods, but a lack of systematic review handbooks on good practices in official statistics. Such a result was confirmed by a review performed within the project by means of a questionnaire, referred to as state-of-the-art survey, sent to all European NSIs and some relevant overseas statistical agencies.

Main objectives of the review questionnaire were to gather information on:

- Existing tools (guidelines, manuals, committees, etc.) for questionnaire development and testing and the extent to which they were used in general and in the Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) surveys;
- Practices in use for questionnaire development and testing in general and for CATI and CAPI in households and establishment surveys.

Other issues were also explored, such as the spread of the CATI, CAPI and Computer Assisted Self-Interviewing (CASI) surveys, the existence of recommended practices or similar tools in other production process areas, information on management and training of interviewers and information related to web surveys. Some results of the survey are briefly summarised hereunder.

First of all, specific references in the recommended practices handbook to CATI surveys seem appropriate since, from the state-of-the-art survey, it came out that CATI surveys are expected to increase in the future, especially in the area of social surveys. Other results concern the great attention towards CASI surveys, not only in the establishment area, as expected, but also in households/individual surveys. Such results may be used to orient future work. As for the existence of manuals for QDET or other standardising tools, the results of the survey have shown a general lack of such instruments, with the exception of few Institutes (like the US Census Bureau, Statistics Canada, Statistics Netherlands or Statistics Finland). On the contrary, general Quality Guidelines are widely available. More than 50% of the Institutes have a unit supporting in the development of the electronic questionnaire for CATI and/or CAPI surveys, however about the same amount of Institutes do not regularly perform a functionality test on the implemented software. Regarding the attitude towards the testing of the questionnaire, the results of the survey showed that, apart from few leader statistics Institutes, most of the other Institutes do not have a strategy for testing new questionnaires or periodically revising those in use, and that there is only little systematic questionnaire testing activity, especially in economic surveys. These findings further underline the importance of implementing recommended practices for questionnaire development and testing. Together with the review questionnaire, the participating Institutes provided all the available material on the adopted standards. This material was taken into consideration when developing recommended practices.

Based on all collected material, a draft of recommended practices was developed and, as indicated by the LEG on Quality recommendations, a feasibility testing in the ESS was carried out, before the handbook were finalised.

1.5. Organisation of the volume

This handbook provides a methodological and practical overview of the standard methods to be applied in questionnaire design and testing in the ESS. It reports descriptions of methods and recommendations on their use, independently from the way data are collected. However, particular emphasis is put on personal and telephone computer-assisted interviewing (CAPI and CATI, respectively), whereas specific issues exclusively applicable to CASI (web or Disk by Mail) surveys or mail surveys are dealt with. When describing methods, if specific modifications are required to better suit the situations of household or establishment surveys, reference to such a customisation of the methods was made.

In Chapter 2, the cognitive model of response process is described as a background. The analysis of response process in a cognitive perspective supports survey researchers in the task of understanding possible faults in the questions and can lead to improvements. This model was initially developed for social surveys conducted on households and individuals, however the approach has meanwhile been customised to economic surveys. The issues faced in this chapter constitute the basis for some recommendations that will be drawn in the next chapters, and in particular in the design of questions.

Chapter 3 reports the methodology for questionnaire design. This activity can be split in a conceptual phase and in a more operative one, consisting in the implementation of the draft questionnaire. In the first phase, some steps such as literature review, definition of survey objectives and concepts, development of conceptualisation schemes, exploration of the best way to observe the phenomenon of interest are carried out. Testing methods, such as focus group and in-depth interviewing, are described in this chapter, since they are performed in a very early stage of the design and testing process, when a draft version of the questionnaire is not yet available. In the second stage, methodologists have to face the task of wording each question and taking care of the order of questions, the number and order of answer options, the length of the entire questionnaire as well as the design of non-verbal elements in the questionnaire.

Especially in the European context, the issue of harmonisation of questionnaires developed in different countries and different languages has become increasingly relevant. Suitable approaches towards translation of questionnaires as well as accepted instruments for the harmonisation of demographic and socio-economic variables are presented in Chapter 4.

Chapter 5 reports structural and graphical issues to be considered when adopting a computer-assisted data collection mode therefore requiring the implementation of the electronic questionnaire. The chapter is expressly tailored for a CATI or a CAPI survey, however several considerations are valid also in the CASI applications and this will be underlined. The additional issues deserving special attention in CASI technique will be only mentioned. It has to be considered that, apart from validly and reliably measuring the concepts of interest, the electronic questionnaire should ease the interviewer's job, and take advantage of the automated means for introducing the right amount of on-line checks on data, thus improving the quality while maintaining the fluency of the interview. Soon after the electronic questionnaire is implemented, or when finalising it, functionality and usability testing are performed. The principles of these testing methods apply to any kind of electronic questionnaire. Moreover, they have many aspects in common with methods such as expert reviews.

Chapter 6 is devoted to the description of the different testing methods. First of all, in this handbook the term testing has been chosen as the set of evaluation methods carried out before the questionnaire is actually used for the survey data collection phase or during it. Very often, the term pre-testing is used with the same meaning.

In order to better orient the reader, testing methods have been split into two groups: pre-field and field techniques. Such a division reflects some principle and operational differences that, indeed, are present in the two sets. Pre-field methods are normally (although not necessarily) applied under laboratory conditions. They are generally used in a preliminary stage of the questionnaire testing process and mostly qualitative in nature. The pre-field methods include: informal testing, expert groups, cognitive interviewing and observational interviewing.

In the field testing methods, testing is carried out under field conditions, i.e., most conditions mirror the real survey situations or the test is conducted during the data collection phase. Field testing often includes larger sample sizes and allows quantitative analyses. Field testing includes: behaviour coding, interviewer debriefing, respondent debriefing, follow-up interviews, and experiments. Other evaluation techniques, particularly suitable for on-going surveys, based on the analysis of item non-response, imputation rates, edit failure or response distributions, are also described (post-evaluation methods).

It has to be noticed that some methods, such as behaviour coding, can be performed both under laboratory and field conditions. For these methods, it has been decided to include them where they appeared to be more frequently used.

Field testing often takes place in the context of a more comprehensive pilot study. Typically, pilot studies are conducted in the field. Generally, they aim at testing the whole set of theoretical and operational conditions of the survey, from data collection to tabulation and estimation process. Very often, the testing of the questionnaire is nested in the pilot survey. In such cases, it is not always possible to disentangle what is attributable to the questionnaire and what is due to other factors.

Finally, in Chapter 7 all the methods included in this handbook are summarised and their main objectives, conditions of applicability, phase of testing, advantages and disadvantages are presented in a table. From a practical point a view, survey methodologists will have to select a strategy for testing questionnaires, based on his/her hypotheses, context conditions and time and resources available. This chapter aims at supporting survey managers in this task, by providing an overall vision of testing methods.

All chapters, excluding Chapter 2, have a common structure. First of all, an overview of the method is provided, with the specification of definitions and aims of the method. Subsequently, the method and the tools are described reporting most significant methodological developments and studies, with specific paragraphs dedicated to CATI and CAPI techniques and to the customization of the method and tools for household/individual or establishment/institution surveys. Practical experiences available from the NSIs which came out from the state-of-the-art survey or from the literature are then reported. Furthermore, a list of recommendations is provided. They suggest, given the conditions, how the method should be applied and which tool should be picked up. A checklist is also provided. The checklist is a list, following an operational order, of what should be done. It has to be noticed that sometimes recommendations and checklists may overlap. However, in general, they integrate each other, resulting a helpful instrument for survey managers.
Finally, a complete list of references is provided at the end of the handbook.

This handbook has a modular structure. Each chapter is self-contained and can be read separately. Within some chapters, too, some specific methods can be read autonomously. This allows readers to select the topic they are more interested in and easily access to it, without having to go through the entire volume, if not desired.

Chapter 2. The cognitive model in answering questions

Cognitive sciences concern the study of understanding language, remembering and forgetting, judgement, etc. All these topics are relevant in the process of survey interviewing. In order to understand what can reasonably be expected from a respondent, every survey manager designing a questionnaire should be aware of some basic findings from cognitive science. The application of cognitive research to statistical surveys[1] provides insight into the cognitive dimensions of the response process and is thus vital for solving various data collection problems. In particular, analysing the respondent's task in answering a survey question by means of a cognitive perspective can help identifying sources of error in questions thus providing ways to improve questionnaires. Furthermore, the cognitive approach and the related methods are frequently being used in questionnaire testing. Cognitive testing methods will be presented in section 6.1.4.

In the following section the general cognitive model of the response process will be described, as well as the most relevant cognitive theories related to the different stages of the respondent's task. The model has been initially formulated in Tourangeau (1984) and then further developed by several authors (e.g.: Eisenhower *et al.*, 1991). As we will see, the general original model was, actually, best suited for surveys gathering information from individuals and households, consequently, it has then been extended to embrace establishment surveys as well (Edwards and Cantor, 1991; Biemer and Fecso, 1995; Willimack and Nichols, 2001). Section 2.2 presents the main differences between the general model and the one specific for establishment surveys. While the cognitive theory interprets the response process mainly as an individual task performed by the respondent, other approaches highlight the similarities among survey interviews and conversations ("communication theory"; Sudman *et al.*, 1996) and emphasize the role of the interviewer-respondent interaction ("symbolic interactionism"; Foddy, 1993). Some insights on this issues will be provided in the concluding section 2.3.

2.1. The general model

The cognitive model of the response process can be formulated in different, although similar, ways. Here we assume a model made up of the following stages (Tourangeau, 1984; Eisenhower *et al.*, 1991; Biemer and Lyberg, 2003; Groves *et al.*, 2004):

0) Encoding: the process of forming memories from experiences.
1) Comprehension: the process of interpreting the question, trying to identify its meaning.
2) Retrieval: the process of recalling information relevant for answering the question from memory.
3) Judgment: the process of combining or supplementing what has been retrieved.
4) Reporting: the process of selecting and communicating an answer.

It should be noted that the model is a simplification of the real response process and that some stages of the model can actually be skipped or may overlap (Groves *et al.*, 2004, p.203).

The Encoding stage can take place even long time before the actual survey, and usually respondents do not know that they will be asked about it in future. During the Encoding stage the respondents register or store in memory the event the question asks about. Consequently, not

[1] A strict cooperation between survey methodologists interested in QDET and cognitive theory researchers has started in early 1980s. This cooperation amongst others led to the creation of an important movement named CASM (Cognitive aspect of survey methodology). A brief history of CASM can be found in Snijkers (2002).

so much can be done by survey designer to limit the effect of encoding errors. Cognitive studies analysed how memories are formed. Tulving (1972) makes the distinction between episodic and semantic memory. The former is concerned with personal experiences and with spatially and temporally well defined events. The latter contains more general knowledge, the meaning of words and concepts. Specific memories reside in episodic memory, while the semantic memory is used for interpreting. Tulving's work enlightened the importance of context in encoding and recall. It is simpler for the respondent to retrieve a particular event if the context of retrieval is similar to the one of encoding.

Other problems related to encoding are incompleteness, distortion and inaccuracy (Eisenhower *et al.*, 1991). Biemer and Lyberg (2003) report the example of a survey on farm operators asking farmers to estimate the value of the land they own. Some of them probably had no idea of the value requested (have not encoded the information), however they provided an estimate, that would probably be distort and inaccurate.

Another problem related to encoding could arise in the case of proxy respondents (Eisenhower *et al.*, 1991; Biemer and Lyberg, 2003). A proxy respondent may have not encoded the personal experience of someone else whom he/she is answering on behalf of. Obviously, questions about attitudes should be avoided when proxy responses are allowed, but also questions about frequency of behaviours can cause inaccurate answers, as "memories concerning oneself may be organised differently from memories concerning others" (Eisenhower *et al.*, 1991).

The Comprehension stage is the first step of respondents' task during the questionnaire administration. The respondent reads or listens to the question and has to understand its meaning. Cognitive research reported in Tourangeau (1984) shows that the process of comprehension can follow two different approaches: the top-down and the bottom-up processing. According to the first, a text is understood by imposing a general conceptual structure on it. The bottom-up approach is more data-driven and consists in the understanding of a text building up its meaning piece by piece with the help of prior knowledge. Both approaches stress the importance of context (the general conceptual structure or prior knowledge) in comprehension. In survey methodology, context effects are meant as the influence of other questions and other information (instructions, section headings…) in the interpretation given to a question by the respondent. Context effects can be either a source of interpretation errors affecting the quality of the results, or, if taken into account during questionnaire design, can be useful to improve questionnaire comprehension. More detailed information on this topic can be found in section 3.2.1.

In questionnaire development, comprehension is also connected to question wording and to the use of difficult (e.g. technical) or ambiguous terms. Accurate wording of a question should be one of the priorities of questionnaire designers. Efforts should be made for searching the simplest and least ambiguous way to formulate questions. Validation techniques like expert review, focus groups, informal testing and experiments can be used to choose the best wording for the most relevant survey questions. Use of difficult, technical terms should be limited or avoided. If used, it is advisable to provide also a definition of the term. Ambiguous terms are also a major problem, in particular considering that some studies (Belson, 1981; Belson, 1986; Groves *et al.*, 1991) have proved that also ordinary terms are often interpreted in different ways by different respondents. A good strategy to limit the effect of this kind of problem is to exploit the response categories to clarify the meaning of the terms used in the question.

At the Retrieval stage, respondents have to recall the information asked by the survey question. Despite the event of interest has been encoded, the ability of the respondent to recall it accurately is not granted. The process of retrieving information from memory has been studied by psychologists for a long time. Here we try to resume only the most relevant issues related to questionnaire design and testing.

One of the most common problems related to retrieval is forgetting. Tourangeau (1984) tried to summarise the following different reasons for forgetting:

- The relevant information may not have been transferred into long-term memory (encoding error).
- The information cannot be simply retrieved due to the length of time between the occurrence of the event and its recall. It has been proved by several studies (reported in Eisenhower et al., 1991) that the greater the length of recall, the greater the forgetting problem.
- Given the importance of context in encoding and recalling information, failure of recalling can be due to the fact that the original context of encoding is no longer present and also memory clues and aids can not be useful to retrieve information as they can not be accurately interpreted any longer.
- Over time, it can be difficult to distinguish between similar events. Eisenhower et al. (1991) referred to this problem as intervening events. This leads to omissions and loss of accuracy in reporting events. Indeed, for similar events, inference tends to be used to add not encoded details when recalling.

Forgetting is a particularly serious problem because it leads to a systematic error of underestimation in survey results. As already mentioned, one of the main factors that affects forgetting is the length of the reference period. Consequently, a way to reduce errors due to forgetting is reducing the length of the reference period, considering, if possible, the impact on survey budget (Biemer and Lyberg, 2003). Also the use of memory aids and retrieval cues included in the question text can help respondents' memory. As an example, consider the question "On average, during the last 6 months […] how often have gone shopping? For example, at drug, clothing, grocery, hardware and convenience stores" (U.S. National Crime Victimization Survey, reported in Groves *et al.*, 2004, p. 229): listing explicitly several kinds of stores helps the respondents in the retrieval process. As already mentioned, retrieval cues should be oriented to recreate the original context of the information encoding.

Another problem related to retrieval is telescoping: it consists in moving an event over time, in general from out of (before) the reference period to inside it, thus leading to overestimation. This sort of telescoping is referred to as "forward" telescoping. Groves *et al.* (2004, p. 218) report that recent studies suggest that also "backward" telescoping, that is moving an event back in time, is common. Telescoping can be also related to particularly salient and emotional events. Those kinds of events are easily remembered and their importance for the respondent often makes him/her report them in the reference period even if they happened out of it.

Besides the salience of the event to be remembered, another factor that can affect information retrieval is proximity to temporal boundaries; events near significant temporal boundaries, indeed, are easier to recall. Consequently, it can be helpful to use significant life events (e.g. births, anniversaries, birthdays…) as a reference during a survey, in order to promote the recall of events that happened near them.

The Judgment phase of the model of the response process is also referred to as estimation (e.g. when the question involves counting the frequency of a behaviour) or response formatting (e.g. in a closed-ended question, when formulating a response according to the requested format). Referring to cognitive theories, Kahneman and Tversky (1971) presented three general rules that are used to formulate judgments: "availability", "representativeness", and "anchoring and adjustment" heuristic. These rules can lead to imprecise judgements. The availability heuristic concerns the fact that respondents judge the frequency of an event on the basis of the availability and speed with which they can recall an occurrence of the event. In the representativeness heuristic events that are considered more representative are also more likely to be judged, so inference and generalisation on nonstatistical basis lead to inaccurate estimations. Anchoring and adjustment heuristic consists in determining an initial answer (anchor), for example starting with what is considered the average value, and then adjusting it referring to personal experience.

From the survey methodology perspective, the judgement stage is mostly connected with the decisions taken on response categories. Thus, the researcher should be aware of such a connection when designing the questionnaire. First of all, the choice between open or closed-ended questions is important. Open-ended questions increase the response burden, can be difficult to answer for persons with low verbal skills, difficult to collect for the interviewer and imply a resource consuming ex-post coding phase. With closed-ended questions, risks include forgetting some important answer categories and formulating overlapping response categories when a single response is asked. Furthermore, in closed-ended questions the tendency of the respondents to choose the first or the last categories in a long list has been observed. Another typical error, whose origin is in the judgment stage, consists in the respondent positioning in a scale. It is well-known that respondents tend to choose the intermediate positions, maybe considering them the “typical” behaviour or attitude (this is connected with social desirability).

The last stage of the response process is Reporting. The respondent has already retrieved and formatted a response and now has to edit and communicate it to the interviewer. Here respondents have to make a choice, to make a decision. Following cognitive theories, they can proceed in two different ways: i) assigning a value to each response category: the probability of providing a particular answer corresponds to the proportion between its value and the total value assigned to all options (Luce, 1959), ii) eliminating the response categories that result less desirable (Tversky, 1972).

Whatever the approach followed by the respondent be, from the survey methodology point of view, problems connected with social desirability, acquiescence, fear of disclosure and sensitivity of the questions come out in the reporting stage. For various reasons, the respondent provides voluntarily an inaccurate response or refuses to answer. The interviewer behaviour plays a relevant role: he/she should appear neutral and, at the same time, should make the respondent feel at ease. When dealing with sensitive topics, a further evaluation is needed on the usefulness of the presence of the interviewer. In any case, a good questionnaire design can help the interviewer in his/her task. Including a clear statement on confidentiality assurance is a first step. Secondly, it is advisable to reach the survey topic gradually, with some easy preparatory questions introducing the main theme softly. With regard to question wording, being as neutral as possible and to introducing the topic in such a way as to make the respondent feel at ease is essential.

2.2. The model for establishment surveys

The model of response process introduced in the previous section is quite general, but it is certainly best suited for household surveys. More specific models tailored on establishment surveys have been developed (Edwards and Cantor, 1991; Sudman *et al.*, 2000; Willimack and Nichols, 2001), that basically integrate the stages of the general model with peculiarities of the establishments’ response process. In this section the differences between an establishment survey specific model and the general one will be described.

First of all, the encoding phase can be revisited considering also the Record Formation. Usually, in establishment surveys, information needed to answer official statistics surveys must have been previously recorded and stored in a database. If the specific information has not been recorded it will not be possible to obtain accurate responses, no matter how well the question has been formulated. It could be very helpful to contact business or trade associations before adding an item in a questionnaire in order to know if the information is generally recorded or can be accurately estimated with available data.

However, Record formation does not completely substitute personal encoding in memory, they are rather complementary. Indeed, personal knowledge of respondent can be useful both to answer directly survey questions, and to know where to search for information that can be stored in several data sources across the company.

After the Encoding/Record Formation stage, the model of response process for establishment surveys introduces two additional phases: the respondent selection and the assessment of priorities. As already mentioned, given that the information requested by the survey question has been stored, the respondent should be able to search the information, he/she should know where to search, what to search and should be willing to do the search, otherwise he/she can provide inaccurate information or nonresponse at all. Willimack and Nichols (2001) enhanced some general features that the selected respondents should own: the respondent should have a wider knowledge of the existence of a variety of types of requested data, and should also have the needed authority to gather the requested data from multiple sources and to release it. Sudman *et al.* (2000), within their study on the response process in a group of 30 multi-unit companies, observed the problems related to respondent selection. They noticed the importance of clearly addressing the questionnaire to a particular department or specific respondent in order to avoid the questionnaire "floating" around in the company for a long time looking for the right respondent.

The assessment of priorities and, in particular, the importance recognised by the establishment to answering the survey, affects survey quality as it influences the respondent motivation, and, consequently, his/her responses accuracy.

Also the information retrieval process assumes quite a different meaning in establishment surveys. It consists in the retrieval of relevant information not only from memory but also from existing company records. Research performed on the response process in big companies (Willimack, 1999; Sudman *et al.*, 2000; Willimack and Nichols, 2001) discovered the complexities and the different strategies adopted by large establishments concerning the retrieval stage. Once a responsible for the data request has been identified at a central level, this respondent can either assume the mere function of coordinator or personally fill-in all questionnaire forms. In the first case, the respondent asks each local unit or other department to fill-in the questionnaire forms of their competence. Then he/she collects the complete questionnaire, reviews it globally and fills-in the corporate-level information. In the second scenario, the respondent contacts the units with direct access to the data in order to gather the requested information, and takes care personally of compiling all the questionnaire. In both cases more than one data provider is involved and the research authors suggest that probably additional nested rounds of response process should be considered in such a situation.

A final step that is added to the traditional ones of judgment and reporting in the model for establishment surveys is the Release of data. It is connected to the problem of the authority the selected respondent should have to provide company data to an external institution.

2.3. Further considerations

The cognitive theory explains the answering process as a series of cognitive tasks performed by the respondent and the cognitive model reveals its practical advantages. Indeed, following the model stages, it results quite simple to identify the measurement errors that can occur during the questionnaire administration, and recognise possible solutions or ways to reduce their effect. Nevertheless, different approaches to the analysis and interpretation of the survey interview process have also been outlined in the literature and should be mentioned.

On the one hand, a survey interview represents a social encounter (Sudman *at al.*, 1996). It recalls a conversation with a stranger and should follow the tacit assumption of that kind of exchange. Grice (1975, reported in Sudman *at al.*, 1996) defined four principles that "govern" conversation: i) the maxim of quality, for which speakers avoid to say something that they consider false; ii) the maxim of relation, for which speakers contribute to the conversation only with what they consider "relevant"; iii) the maxim of quantity, for which speakers are "as informative as required but not more informative than required"; and iv) the maxim of manner, that invites to be clear in the exposure. Accordingly, speakers in a conversation are supposed to be truthful, relevant, informative, and clear. Thus, respondents try to be relevant and informative in their answers and expect researchers to be clear in the question formulation. This implies, for example, that if respondents find a question whose answer can be redundant with respect to a previous answer provided, they will ask themselves if they have really understood the question and will try to provide the less redundant plausible answer, in order to be informative and relevant. For more information, see Sudman *at al.* (1996, pp.63-64).

On the other hand, a survey interview represents an interaction between the researcher/interviewer and the respondent. The application of symbolic interactionist theory (introduced by Herbert Blumer, reported and adapted to the question-answer process by Foddy, 1993) to survey interview implies that the respondent is constantly looking for a shared definition of the situation with the researcher/interviewer. As the researcher/interviewer tries to formulate the questions and to interpret the answers provided by the respondent taking into account the respondent's knowledge and characteristics, the respondent tries to interpret the questions and then to provide the answers bearing in mind what he/she knows about the researcher/interviewer.

Both approaches imply similar consequences and mainly affect the comprehension stage of the response process: when respondents find something not clear in the questionnaire, in the attempt to understand the purposes of the researcher, they look for the plausible meaning throughout the various clues that can be embedded in the questionnaire (response options, related questions…) and try to provide a meaningful answer. Once more, the context in which questions are asked stresses the importance of preventing measurement errors by improving questionnaire design.

In conclusion, the communication and interaction theories should not be considered in contrast with the cognitive model perspective, but rather complementary in order to better understand not only the respondent task in answering a question, but also the whole survey interview "exchange" process. The most fruitful approach to follow is reasonably represented by the integration of the different theories following the purpose to achieve better results in improving questionnaire development and testing by preventing measurement errors.

Chapter 3. Questionnaire design

Questionnaire design consists of three successive steps, namely the development of the conceptual frame of the questionnaire, writing and sequencing the questions and establishing the visual design elements of the questionnaire. Whereas the art of writing questions has been extensively dealt with in numerous textbooks, the conceptualisation and visual design of the questionnaire are often not covered to the extent required. However, also these steps have to be prepared thoughtfully and seriously.

Therefore, section 3.1 introduces the tasks to be taken into consideration in the context of the conceptual frame and outlines the most important steps and methods. Section 3.2 summarises the basic rules for writing questions and gives an overview on types of questions and answering categories. It also contains recommendations on the sequencing of questions in the questionnaire. The last section is devoted to visual design elements, which is more than simply the layout. It refers to the fact that not only wording and sequencing has an impact on how a questionnaire is being answered, but also the visual features of a questionnaire with regard to emotional, functional and reflective levels have to be addressed when designing a questionnaire.

3.1. Conceptual frame

Whereas survey methodologists have paid much attention to the problem of wording questions and defining the answering categories, much less effort has been spent on the process of specifying the objectives, concepts, and definitions of the survey (Hox, 1997). Before questions can be formulated, it is strongly advocated to start with a review of relevant literature as well as analyses of available data from other surveys and studies on the topic. In the next step, the objectives and concepts of the research domain need specification (conceptualisation). Subsequently, these concepts are translated into suitable indicators and measured by observable variables (operationalisation) of relevance. Even though methodologists have come to the conclusion that there is no direct representation of “reality” in a statistical measurement, the ultimate goal is to be as close as possible to what we wish to investigate. At the same time, this process inevitably leads to information losses (as information only becomes measurable after a considerable reduction of complexity) and has to be operated carefully and in close co-operation with the users. Thus, it is a sensitive translation of theoretical concepts into meaningful, operational statistical concepts and variables. As the final aim, the transfer and definition of variables and values should enable: i) the questionnaire designer to implement the questionnaire; ii) the respondents to understand and answer the questions properly; iii) the subject matter specialists and users to carry out the necessary analyses.

The process of conceptualisation and operationalisation requires intensive co-operation of the users, the subject matter specialists, the questionnaire designer and the respondents. This iterative process can be supported by different strategies and methods: dimension/indicators approaches, semantic analysis, facet design techniques, content sampling, symbolic interactionism as well as concept mapping can be applied. Such strategies and methods should however be supplemented by direct discussions with experts, key users and respondents. For this purpose informal methods (like discussions with colleagues and experts) as well as formal methods and tools (like focus groups, expert groups, in-depth-interviews) help to explore whether the concepts and definitions are accessible to the respondents and meet the user’s requirements.

With regard to the process of conceptualisation it should be noted that in official statistics there is in many cases already a legal base in place which defines the survey variables in some detail.

In these cases the degrees of freedom for conceptualisation and operationalisation might be rather limited. However, users and survey managers in any case should have a clear picture of the objectives and concepts behind the variables.

As a precondition for the concept formation a clear idea of the survey design should be available (e.g. target population, sampling, probable data collection mode, main objectives etc.). Although many features of survey design have some impact on questionnaire design, this chapter focuses on the stepwise transfer of the general survey objectives into concrete variables which could be implemented in the questionnaire. Consequently, issues regarding the survey design in general are only outlined briefly. However, it has to be noted that many features of the survey design have to be considered before questionnaire design can start.

Accordingly, the output of the conceptual frame is a precise description of the survey concepts (conceptualisation), a list of indicators and variables, a questionnaire scheme and a preliminary order of variables replacing in future the questions (operationalisation). These steps are presented in section 3.1. The concrete wording and sequencing of questions is outlined in section 3.2.

3.1.1. Literature search

As soon as the users or other bodies have announced their demand of data, it is essential to verify if and which information on the topic is already available[2]. In this starting phase it is a basic standard to collect literature on the issue, to run a search on published statistics and contact potential experts/ministries and other institutions which may have data (possibly unpublished). Many questions dealing with demographic variables are common to many surveys. Some surveys are available on the web, including copies of original survey instruments, testing and quality reports, so that a set of material will often be already available. Thus, taking advantage of the findings from the work of other researchers is highly advisable. Nevertheless, all questions need to be verified, since findings from other surveys can not guarantee that a question is appropriate in the new survey context (Fowler, 2002).

Special emphasis must be put on the need to consult reports and tests from similar surveys, since in the last decade testing of survey instruments has increasingly become a standard procedure in official as well as in academic statistics. To look through documents and pre-test papers of other surveys on comparable topics and conditions is thus advisable. This helps to find general problems and methods to detect them (SCB, 2004; ABS, 2001). Questions you may consider are e.g.:

- What almost equal surveys have been developed?
- What kind of testing was conducted and which were the results?
- Which recommendations on the design are presented?

This first review is not so costly and time intensive, but helps to get into the topic, identifies basic problems and helps the researcher to structure his/her work. It is a method which should be applied at the very beginning of the questionnaire development before any draft questionnaire is being prepared.

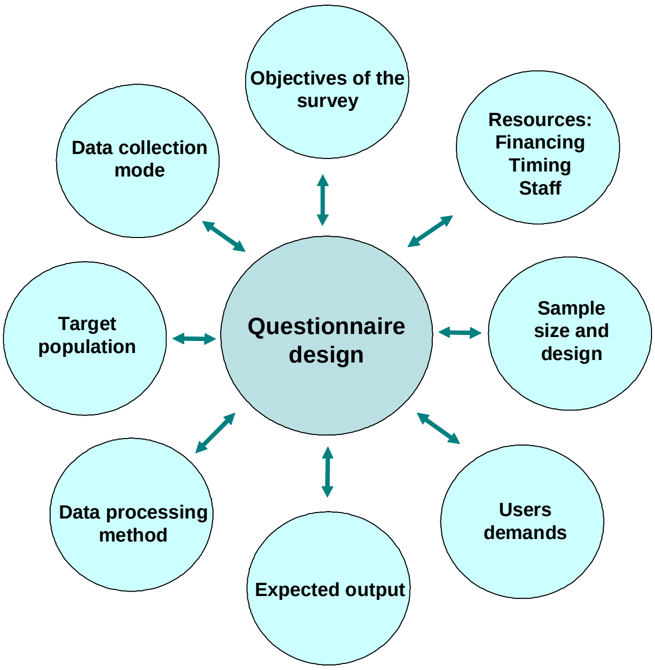
3.1.2. Survey objectives

“A prerequisite to design a good survey instrument is deciding what is to be measured” (Fowler, 2002).

[2] Even though it has been decided to launch a new survey - as data requested were not available – similar statistics may have already been presented and tested which may give advice on what to do and what to avoid.

After literature research the second step is to recapitulate the general survey design and objectives. Basic concepts such as the target population or the sampling design, as well as the available resources and the preferable data collection mode, thus the entire, planned survey design must be clarified and defined. All these factors constitute an important background when designing the questionnaire (see Figure 3.1.). For example, when there is no information on the availability of PC and web connection among the target population, one may not select Computer Assisted Web Interviewing (CAWI) as the only technique to be used; or when data are required shortly after the data collection and the estimated length is not too long, one may choose a CATI.

Figure 3.1. Interplay of survey design and questionnaire design (adapted from Statistics New Zealand, 1995)



When the basic elements of survey design are defined it is highly recommended to transfer the general survey objectives into concrete research questions of the survey. This is when the conceptualisation phase starts. Often this important exercise is neglected as time constraints press researchers to draft a questionnaire before thinking in detail about the concept. However, to ignore such basics may result in unreliable and non-valid data, respondents' and interviewers' dissatisfaction and high nonresponse rates as well as increased time for checking data (SCB, 2004). In respect of the conceptualisation two main perspectives are of relevance: a) users' needs, their objectives and perspectives and b) the theoretical concepts and the feasibility of operationalisation.

With regard to users' needs it is frequently observed that survey objectives and concepts - even if they are being regarded as established and well defined - need further inquiries and consultation of users. Thus re-contacting of users to assess their data needs is advisable. In defining the aims there is one basic rule of questionnaire design:

"Ask what you want to know, not something else" (Bradburn, 2004)

This simple rule has numerous implications making it quite a difficult task to implement. The first implication is that we have to be sure about what we want to know. In official statistics, for almost each survey, there is a multiplicity of users with their own specific requirements and statistical institutes are often acting as an agency for data production. Thus, as a first step the user requirements have to be documented and reconciled (in case of trade-offs). In many cases, there will be a small number of key-users with whom the main objectives should be defined. So, in communication with key-users the definition of objectives should become more precise. It is an iterative process between users and survey manager to define the aims. The iterative process can be conducted by different communication systems (mail, phone, meetings etc.).

For this task different methods are available:

- Informal communication (particularly useful in the very early stages)
- Meetings with users, subject matter experts and researchers (see expert groups, section 6.1)
- Contacts with well-established bodies in order to reflect on their objectives

With regard to the conceptualisation and operationalisation a brief overview on strategies is presented below.

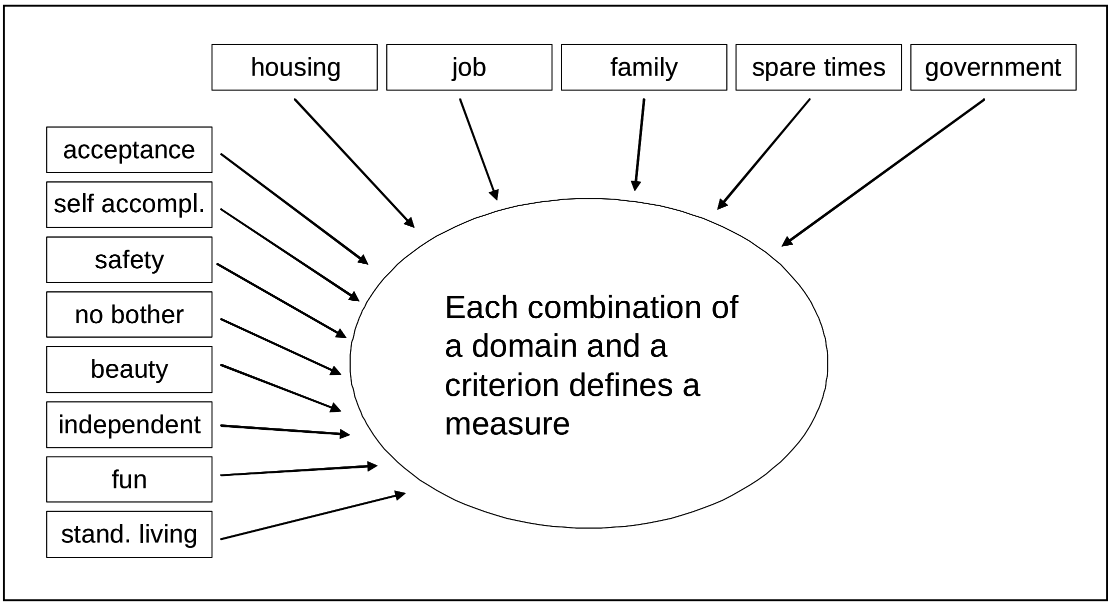
3.1.3. Research strategies and methods for conceptualisation and operationalisation

At the same time, when being in communication with the potential users, the theoretical concepts (conceptualisation) and its translation into measurable variables (operationalisation) need specification. Whereas conceptualisation means the formation of concepts and the definition of sub-domains, operationalisation refers to the translation of the theoretical concepts into observable variables by specifying empirical indicators for the concept and its sub-domains (Hox, 1997). However, methodologists and researchers came to the conclusion that a direct observation of the concepts is impossible in a statistical measurement. In the process of conceptualisation and operationalisation survey designers can only try to limit the losses which inevitably result from the reduction of complexity of “reality”. Referring to Hox and De Jong-Gierveld (1990), there are two perspectives to bridge the gap between theory (concept) and observable variables (operationalisation or measurement): i) the theory driven approaches (as dimension/indicator analysis, semantic analysis and facet design methods) and ii) the empirical driven approaches (as content sampling, symbolic interactionism and concept mapping). The mentioned methods are presented as an overview, based on Hox (1997).

From conceptualisation to operationalisation: theory driven approaches

Theory driven approaches define the theoretical concepts by splitting up the topic into sub-domains as separate components or dimensions (see Fiske, 1971; Lazarfeld, 1958; Lazarfeld, 1972). The sub-dimensions form a further, more detailed specification of the meaning of the theoretical construct. As a good example Hox (1997) refers to Andrews and Whitney (1976), who specified and conceptualised subjective well-being by visualisation as presented in Figure 3.2., where the domains are represented on the top and the criteria are reported on the left.

Figure 3.2. Conceptualisation of well-being (Andrews and Whitney, 1976)



Taking this theory driven approach further, the facet design approach presented by Guttman (1954) helps to identify and illustrate the different perspectives or facets relevant for the topic. Applying this method, the universe of observations is classified by three kinds of criteria: a) the population (source of information), b) the content facets or variables of interest and c) the response categories. To clarify the issue further on, the facet structure is worded by mapping a sentence (see Figure 3.3., example by Gough, 1985; see also Borg and Shye, 1995).

As a further tool, semantic analysis focuses on the linguistic system of the topic and helps to reflect the vocabulary as a global perception of the world (Satori, 1984). The subject of verification is the terms and constructs in use by connotation and denotation. Semantic analysis can be almost regarded as qualitative research and does not lead directly to survey questions, but rather checks the ambiguity of terms and concepts.

Figure 3.3. Example of Mapping Sentence (Gough, 1985)

Question: "To what extent does person (X) feel that

Source		**Reason**
(her own experience) (her husband) (her doctor) (the media)	led her to believe that she would	(feel healthier) (feel fitter) (be more physically attractive) (have fewer clothing problems) (suffer less social stigma) (be less anxious in social situations) (feel less depressed)
	Response	
if she lost weight, as rated	(not really at all) (not very much) (to a slight degree) (to a fair degree) (quite a lot) (very much) (very much indeed)	

where (X) are married women attending slimming groups"

From operationalisation to conceptualisation: data driven, empirical research

In opposite to the theory driven approach the empirical approaches use either data to develop concepts and theories or involve the group of observation as the starting point. When applying content sampling methods, a set of questions relevant for the topic are collected and tested by means of factor analysis. Consequently, starting from data, theoretical constructs are inferred and may be altered in empirical research. However, this approach implies that relevant data as well as a profound knowledge on the topic are available. There is a certain risk when selecting items of relevance to miss an important one. So at the beginning of research potential sets of questions are selected by exploratory factor analysis or other similar means. Then these factors are analysed separately to test whether and which single factors might be of use.

A similar method using qualitative methods has been developed in the context of symbolic interactionism and the "grounded theory" (Fiske, 1971; Kerlinger, 1986). Here, the aim is to get an understanding of how the theoretical concepts are used and understood in everyday life. Concept formation here starts with an observation of how common language concepts are used by the target population. This technique starts by discussing the issue with potential respondents to determine concepts which are accessible. Possible methods to be applied include focus groups and in-depth interviews (see below).

Whereas the symbolic interactionism is rather explorative, concept mapping as a technique observes and discusses the issue with potential respondents by conducting focus groups, but using the technique of mapping. In summary, there are two stages for the discussion: the generation stage is explorative and aims to produce a comprehensive list of statements, whereas the structuring stage seeks to reduce the statement list to a smaller list of key concepts.

3.1.4. Exploring concepts: focus groups and in-depth-interviews

Either in the phase of the development of concepts (see above), or when there is a clear picture of the survey concepts and objectives, it is advisable to explore whether the concepts the users are after are compatible with those the respondents have on their mind. In other words, are people able to answer what we like to ask them and do they understand the concepts in the same way as the survey designers do? There are mainly two ways or procedures to verify concepts and definitions released by researchers against the view of possible members of the target population: as a form of group discussion one may conduct focus groups, or via personal interviews it is advisable to conduct in-depth/qualitative interviews. Both methods are presented below.

A) Focus groups

"A focus group is a discussion among a small number of target population members guided by a moderator" (Krueger and Casey, 2000).

The first step in the organisation of focus groups is the definition of the critical aspects and topics to be treated in the discussion. It is recommended to invite 6 to 10 participants per focus group and to conduct several focus groups to ensure better representativeness. Participants to each session should not be too heterogeneous, as otherwise discussions may become quite unstructured and confusing. Nevertheless, people working in the same group (e.g. supervisors and employees, teachers and students) should not be involved in the same focus group. There should be a moderator structuring the discussion. The moderator shall create a relaxed, permissive atmosphere. At the beginning, the moderator may outline purposes and basic rules of the discussion and reassure the participants on confidentiality. Then he/she should lead the discussion and control that all participants provide their opinions. Each focus groups session should preferably last 1-1 ½ hours. The discussion can be tape-recorded, video-recorded or observed by one-way mirror and then it would be useful to have a full transcription of it.

The general aims

Focus groups are used to reflect the perspective of the target population regarding the topics of a questionnaire. It is mainly applied at an early stage of questionnaire development in order to test acceptance and comprehension of the topic to be focused on. The researcher can learn about the nomenclature of concepts, how terms are used, what common perspectives are taken by the population (Groves *et al.*, 2004). “Focus groups are a good way to get fresh ideas from the world outside statistical production” (SCB, 2004). Focus groups provide a social context for research, and thus an opportunity to explore how people think and talk about a topic; they are ideal for creative thinking and checking the concepts and terminology (Ritchie, 2003). These group discussions can provide a wealth of information not just about what potential respondents think, but why they think the way they do (ASA, 1997).

Specific aims (ABS, 2001; Groves *et al.*, 2004):

- To detect sensitive topics and domains to check
- To identify what potential respondents know about the subject, how they structure the subject and what they think about it
- To find out if people understand the terms and how they define them
- To learn from the respondents what issues of the topic are of relevance for the respondents’ point of view in order to get in good communication with the respondents in a future survey
- To determine the feasibility of conducting the survey
- To develop survey objectives or data requirements
- To determine data availability and record keeping practices
- To explore and define concepts
- To clarify reference periods
- To evaluate respondent understanding of terminology
- To evaluate alternative question wording and formats and to understand respondent burden.

Establishment surveys

To conduct focus groups when planning establishment surveys is highly recommended. Terms to be used can be checked against how these are applied in reality. While conducting focus groups it is also possible to check to whom it is advisable to address the questionnaire, what people may give the information and what processes take place when filling in the questionnaire (SCB, 2004).

Experiences

From the survey conducted in the framework of this handbook, only some NSIs use frequently focus groups as a method to check their concepts against the viewpoints of their potential respondents (3% conduct focus groups for all surveys; about 60% for some surveys). This might somewhat be related to the fact that many statistics are ongoing and thus might not be considered worthwhile to be tested further.

Some institutions use the method to check and develop a self-administered questionnaire, thus at a later stage of questionnaire development. The Centre of Survey Research and Methodology (ZUMA) in Germany uses the method as follows (Prüfer and Rexrodt, 1996):

- First step: Participants shall fill in the questionnaire without having the opportunity to clarify misunderstandings (To be noted: duration of completion of the questionnaire for each participant);
- Second step: Discussion on the general content, duration to fill in, degree of difficulty;
- Third step: Discussion on each question: comments, understanding, answering categories etc. Besides, questions prepared by the researcher can be discussed.

The Australian Bureau of Statistics assesses the method as particularly useful because it allows to test with a specific group of population (ABS, 2001). They experienced medium costs (participants must be paid, but number of participants and duration is limited) and limited time for preparation, conducting and analysing the results (around three weeks).

B) In-depth or qualitative interviews

In order to explore the perspective of potential respondents, personal qualitative interviews can be applied too. Basically, in-depth or unstructured interviews are one of the main methods used in qualitative research (Legard *et al.*, 2003). In-depth interviews are often described as a form of conversation (Burgess, 1982; Lofland and Lofland, 1995) with a purpose (Webb and Webb, 1932). Similar to focus groups the purpose of such interviews is to learn how potential respondents use terms, how they understand general concepts or specific terminology, and how they perceive the questions in terms of sensitivity or difficulty. Due to their explorative character, they are less structured than focus groups.

However, conducting interviews is by far not without any structure (see Legard *et al.*, 2003). Thus key features are:

1) They intend to combine structure with flexibility. Even in the most unstructured interviews the researcher will have some sense of the themes he/she wishes to explore, and interviews will generally be based on a topic guide.
2) A second key feature is their interactive nature. The researcher will have an initial question to encourage the conversation. The next intervention of the interviewer will be the reaction to the interviewee’s answers and so on.
3) When the interview continues, probes which are techniques from cognitive interviews (see section 6.1.), might be applied like follow-up questions.
4) Fourthly, the interviews are generative in the sense that new knowledge and perspectives of relevance are arising and can be discussed.

Additionally, with regard to the method’s purpose to explore a new topic, one may distinguish between two types of questions: a) content mapping questions, and b) content mining questions, which are in use in qualitative research. Whereas content mapping questions are designed to open up the research territory and to identify the dimensions or issues which are of interest for the potential respondents, the mining questions are to explore details and in the case of questionnaire development could be used to check terms. With regard of conducting the interview, principles of good wording and addressing respondents presented in this volume, see section 3.2. on writing questions, apply too.

In-depth interviews demand flexible, experienced interviewers and researchers with professional skills and broad knowledge on the subject. It is highly recommended to tape-record the interview. Depending on the potential respondent to talk with, an interview should not take longer than 1 ½ hours. Venues can be quite different, either at home, at a laboratory or, for establishment surveys, in the business environment. Even though at the beginning of conducting interviews researchers may feel better to join each other (thus e.g. two interviewers), respondents prefer only one interviewer at a time, as it is already an unusual experience to participate in such a test. Due to its explorative character, qualitative interviews are mainly implemented at an early stage, as for the development of a questionnaire (SCB, 2004).

Nowadays qualitative interviews for pre-testing purposes are rather seldom conducted. Focus groups are the main channel to get an idea of the perspective and understanding of the topic by potential respondents in the preparatory phase. Personal interviews are usually postponed to later testing, when either a draft questionnaire or specific questions need to be tested. This task is mainly executed by conducting cognitive interviews (see pre-field testing).

3.1.5. Questionnaire schemes

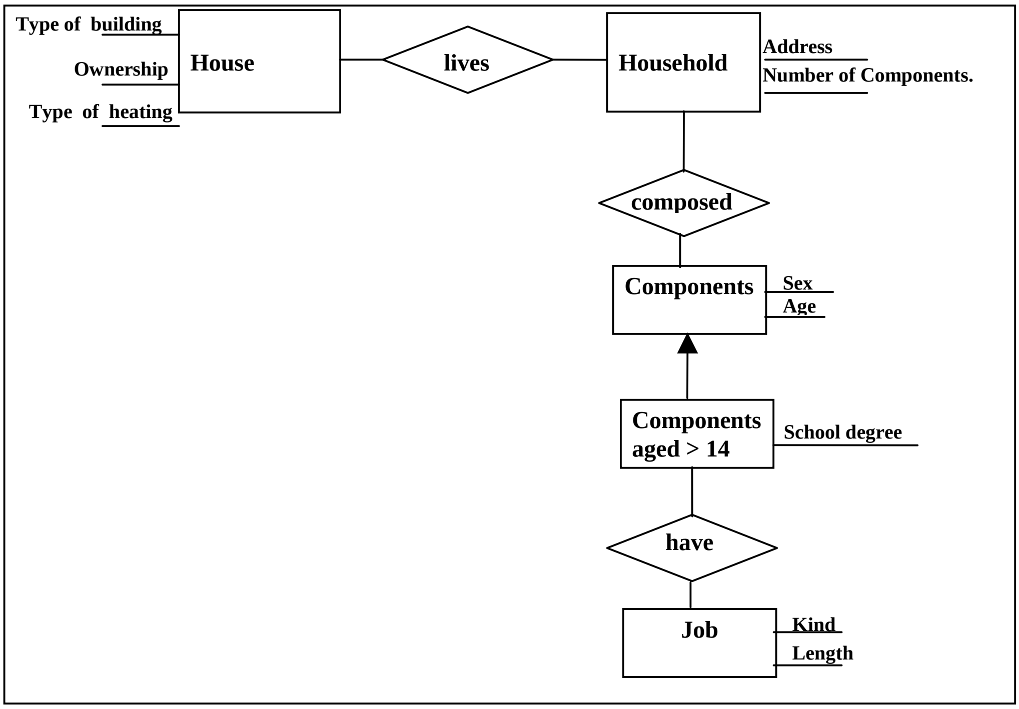
Whereas in section 3.1.1. to 3.1.4. it has been described how to transfer reality into observable statistical concepts, in this section this activity is presented through the use of graphical tools, which allow to detail these concepts and their characteristics.

To this purpose, it can be useful to adopt questionnaire schemes in order to illustrate the contents and scope of the questionnaire in a systematic way. One suitable approach is based on “entity/relationship schemes” (or “conceptual schemes”) (Chen, 1976). These schemes are based on methods stemming from information systems design methodology, which can also be applied to questionnaire schemes. Entity/relationship schemes provide an overview on units, items and relationships covered by the questionnaire. The basic structure of entity/relationship schemes consists of entities, the logical links between the entities (relationships) as well as the entities' attributes.

The entities are the concepts of interest for the survey (for instance: house, household, components of household, job, etc.); they are represented in the scheme by rectangles. The relationships are the logical links between entities (for instance: a household *lives* in a house, a household component *has* a job; "*lives*" and "*has*" are two relationships in our scheme); they are represented in the scheme by rhombi. The entity's attributes are the characteristics to be known of each entity, so they will constitute the questions of our questionnaire (for instance: the attributes of the concept *house* could be the dimension, the type of heating, etc.; the attributes of the concept *household component* could be age and sex, etc.). They are written in the scheme above the lines connected to the rectangles. It is also possible to define attributes of the relationships. In addition, subsets of the already defined entities can be identified (they are represented by rectangles connected by arrows to the entity they belong to). These are particular elements of the entity, with certain characteristics, which the survey wants to study in depth.

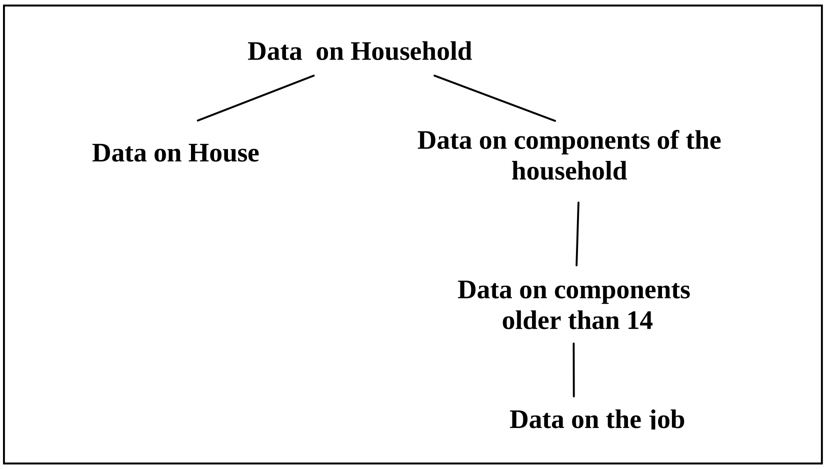
As shown in figure 3.4, entity/relationship schemes give a useful visual support to analyse the questionnaire structure and detect further requirements, redundancies or conceptual errors. Drawing and revising the scheme during the conceptual framework is an iterative process and should be finalised together with the definition of the variables and the drafting of the tabulation plan. In addition the final scheme of the questionnaire can be of use when programming databases for processing and analysis. It helps to structure and define files, cases and number of variables foreseen.

Figure 3.4. Conceptual scheme: entity/relationship scheme



After this analysis of the elementary concepts of the questionnaire, it can be advisable to carry out a macro analysis with the purpose to group topics and to check main skips, as sets of questions do not apply to all respondents. Following our example, five sections have been identified and are consistent with the entity/relationship scheme (see an example of area tree in Figure 3.5.). Some variables are addressed to sub-populations (like the sections regarding data on components older than 14) to be identified with filter questions. On the other hand, at this stage, it has still to be decided which branch of the tree to set first in our questionnaire: the one regarding data on the house, or the other one which constitutes the greater part of the questionnaire. However, by drawing the area tree by topics to cover, the issue to be decided on is how question can be grouped into sections (see section 3.2.) and, in addition, some hypotheses can be made regarding the questionnaire flow.

Figure 3.5. Example of area tree


Data on Household
Data on House
Data on components of the household
Data on components older than 14
Data on the job

3.1.6. Define variables and draft a tabulation plan

When producing statistics the reality of daily life is transferred into figures. The mode to do this transfer is the questionnaire. The data required are to be collected via questions and answers, operationalised and digitised into variables and values.

Thus, as soon as the objectives are rather detailed and have been approved by responsible bodies and persons, it is of very useful to create a tabulation plan and to state the hypotheses to be analysed. At this stage the research becomes very concrete and helps researchers to achieve the objectives of the questionnaire. When the expected output in form of planned tables and analyses has been clarified, the variables and values have to be defined. The variables and values list is to be seen as a simple list of names and defined values. With regard to the variables list, ideally retrieved from a standard variables database, it is recommended to structure it by background variables (e.g. demographic variables, some of which are clearly defined today and are advisable to use) and variables by domains to be covered in the survey (SCB, 2004). Applied common variables and their definitions should also be checked against definitions of previous or very similar surveys and data bases, as a matter of comparability and coherence of data.

In addition, it is worthwhile to check which variables are to be calculated or created via other variables and which are direct measures. The process of preparing a tabulation plan and defining variables is somewhat iterative and can not be worked on separately. However, at the end of this task there should be lists of variables and values demanded and a draft tabulation plan.

3.1.7. Data collection mode and impact on the questionnaire

Over the last decades data collection modes have changed a lot: whereas at the beginning of survey research face-to-face interviews or mail surveys were the standard, nowadays a variety of modes is in use (Groves, 2004). With the almost complete equipment of households with telephones, telephone interviewing became most popular in the seventies, and in some countries even predominant (de Leeuw, 2005). With the rapid growth of computer technology the world of data processing and interviewing changed once more with positive effects on data quality.

Based on any kind of computer applications, e.g. CAPI or CATI questionnaires, a lot of surveys use the computer already for data collection, not only for data processing. Nevertheless, also mail surveys continue to be one of the major data collection modes, as the method has its own quality and the competence in terms of design has improved a lot (Dillman, 2000).

At the same time the decision on the data collection mode has an impact on how questions, response categories and questionnaires have to be designed (SCB, 2004). With regard to the former short review on the development and application of data collection modes, there is one observable tendency to look at: modern technology on one hand and budget constraints on the other hand push the data collection modes towards less interviewer-driven and more self-administered questionnaires. This can either be implemented by CASI (CAWI, DBM etc., see below) or still by a self-administered Paper And Pencil Interviewing (PAPI) survey. Consequently, the question of whether an interviewer is involved might be an important criterion today to distinguish between questionnaire designs. Thus, without being totally exhaustive, one may distinguish between two main dimensions to describe the features of mode: a) technology and b) involvement of an interviewer.

Table 3.1. Data collection mode

Technology	**Type of administration**	
	Interviewer administration	**Self-administration**
Computer-assisted	CAPI, CATI	CASI: WBS (or CAWI), DBM, EMS and TDE
Paper and Pencil (PAPI)	PAPI face-to-face interview	PAPI (mail surveys)

CAPI: Computer-Assisted Personal Interviewing, CATI: Computer-Assisted Telephone Interviewing, CASI: Computer-Assisted Self-Interviewing, WBS: Web Based Survey, EMS: E-Mail Survey, TDE: Touch-tone Data Entry, PAPI: Paper-and-Pencil Interviewing, DBM: Disk by Mail, CAWI: Computer-Assisted Web-Interviewing.

To summarise, the following aspects should be taken into account when having to decide on the data collection mode:

- the subject of the survey: if the subject treated is very sensitive, a self-interviewing mode is advisable;
- the complexity of questionnaire: if the questionnaire is very complex, in terms of skipping rules and necessity to include consistency controls between variables, a computer assisted interview technique with interviewer is advisable;
- the estimated interview length: if the interview is thought to be very long, CATI and CASI techniques are not advisable;
- the characteristics of the target population: the available information regarding the technical equipments of the target population (telephone, computer and web connections) may lead to the exclusion of some data capturing techniques;
- the budget at disposal for the survey.

Therefore each data collection mode has its strengths and weaknesses, which must be balanced against each other, also taking account of their impact on the questionnaire (SCB, 2004). It follows that the decision on the "optimal" data collection strategy is a demanding task, which must be carried out before structuring the questionnaire, writing questions and creating visual design elements can start.

The analysis of all the characteristics of the survey and of the features of the data capturing techniques often outlines the opportunity to define a mixed mode strategy, which means to implement one questionnaire to be submitted by different modes (see later on in this section).

Whereas general features of the data collection mode have been presented above, the following pages summarise advantages and disadvantages of the different data collection modes, with particular focus on the impact on questionnaire design.

A) PAPI - Mail surveys

In mail surveys paper questionnaires are sent to the respondents by mail, which should be filled in by self-completion and sent back to the statistical office. These surveys are cheaper, but in general associated to lower response rates and time-consuming in terms of data collection and processing.

Advantages

- Respondent has time to give the answers, to reflect on them and, if necessary, to verify data regarding some particular questions (i.e. precise dates, amount of expenses, etc.)
- Detailed instructions are possible
- Sensitive questions can be asked better than in personal interview (credibility is more likely (ASA, 1996)
- Long lists of answering categories are possible
- Aids like maps, pictures and symbols are possible

Disadvantages

- Particular care must be taken when designing the questionnaire in order to make it as easy as possible to be read and understood by the respondent (wording must be very simple, short and precise, instructions must be very clear and complete, etc.)
- The same care must be taken when designing the questionnaire layout, in order to make it attractive and as easy as possible to be filled in by the respondent (the page must not be too dense, skipping rules must be clearly displayed, etc.)
- The length of the questionnaire needs to be limited (it is suggested around 12-16 pages, so that it may take 30-45 minutes to fill in)
- The questionnaire can not be too complex in terms of skipping rules, because this could increase the risk of incorrectly filled in answers
- Complex matrix tables to be filled in are not advisable, since respondents are often mistaken in columns and rows
- Good reading and writing skills of the respondent are needed
- There is no way to control the quality of filling in, like completeness, meeting question objectives, etc.
- The only way to influence positively the response rates is in the questionnaire design

B) PAPI – Face-to-face personal interview

This technique implies the paper questionnaire and the presence of the interviewer. It follows that, if the paper questionnaire by itself can not guarantee a good response rate and the avoidance of data capturing errors, the interviewer has positive effects on these aspects. On the other hand, interviewers can introduce biases and surely lead to an increase of the costs.

Advantages

- Long interviews are possible (they may last until one hour), and may even include additional probes for clarification
- More complicated questions can be asked
- Additional material can be used (maps, answering sheets etc.)

Disadvantages

- The questionnaire can not be too complex in terms of skipping rules, because this could increase the risk of incorrectly filled in answers
- It is necessary to provide the interviewers with a deep and homogeneous training on the survey issues and on the questionnaire, because they are the only means to avoid data capturing errors
- Generates social desirability bias associated with sensitive questions

C) Computer Assisted Personal Interviewing (CAPI)

This technique implies the presence of the interviewer and the use of a computer (laptop) with an electronic questionnaire to carry out the interview. This means that all the effects connected with the presence of the interviewer are possible, but the use of the electronic questionnaire allows the management of more complex interviews, helps to prevent data capturing errors and makes data available in a short time (the data entry phase is not necessary); on the other hand, this is surely the most costly data capturing technique.

Advantages

- Long interviews are possible (they may last until one hour), which may even include additional probes for clarification
- Complex interviews are possible because skipping rules are managed directly by the software
- More complicated questions can be asked
- A complex set of rules to avoid errors (range and consistency errors) can be included in the electronic questionnaire, thus enhancing the quality of collected data
- Controls and reconciliations with data already available are possible (with data from a previous surveys or from administrative archives)
- Interviews can be very smooth and pleasant because the entire questionnaire and the questions' wording can be customised
- Assisted coding of textual variables can be performed during the interview
- Long instructions and definitions can be given to the interviewers through online helps
- Additional material can be used (maps, answering sheets etc.)

Disadvantages

- Particular care must be taken when designing and developing the electronic questionnaire in order to minimise the "segmentation effect", to provide the interviewer with the maximum help and assistance (he/she is alone while interviewing) and to guarantee the efficiency of the software procedure
- First programming and testing phases can be complex, time consuming and costly
- Less table styled questions are possible
- Generates social desirability bias associated with sensitive questions

D) Computer Assisted Telephone interviewing (CATI)

With this technique the interviewer carries out the interview by phone, using a computer with an electronic questionnaire, generally working in a call centre where other interviewers are present and at least one supervisor is at disposal to give help, if necessary. This means that the considerations made for CAPI are valid also for this technique, with the exception that the interviewer is not "alone", as he/she can be supported by the survey experts. On the other hand, the rhythm of the interview is more pressing because the conversation is held by phone. In addition, the interviews are not so costly and the sample can be spread all over the territory, but coverage problems could arise for the population without phone.

Advantages

- Complex interviews are possible because skipping rules are managed directly by the software
- A complex set of rules to avoid errors (range and consistency errors) can be included in the electronic questionnaire, thus enhancing the quality of collected data
- Controls and reconciliations with data already available are possible (with data from a previous surveys or from administrative archives)
- Interviews can be very smooth and pleasant because the entire questionnaires and the questions' wording can be customised
- Assisted coding of textual variables can be performed during the interview

- The interviewing activity can be monitored directly as the interviewer is present in the call centre or through the daily analysis of a set of indicators, which allows to intervene immediately to optimise the questionnaire or to further train the interviewers (reducing the interviewer bias)
- The attribution of telephone numbers to be called by the interviewer is randomised by the software (scheduler), so that the possible interviewer bias is not concentrated on certain units
- Sensitive issues can be managed more easily than with face-to-face interviewing

Disadvantages

- Very long interviews are not advisable (not longer than 30 minutes)
- Questions and the lists of response items must be short
- Questions should not require long thinking, as it interrupts interview flow
- Care must be taken when designing and developing the electronic questionnaire in order to minimise the "segmentation effect", to optimise the navigation inside the questionnaire and to provide the interviewer with a layout which makes it very easy to find what he needs in each interview moment
- First programming and testing phases can be complex, time consuming and costly
- Less table styled questions are possible
- Not too many textual responses are advisable
- Generates social desirability bias associated with sensitive questions
- No visual aids for respondents are possible

E) Computer Assisted Self-Interviewing (CASI): WBS (or CAWI), DBM, EMS and TDE

With these techniques, the interview is always self-administrated with the use of an electronic questionnaire. Although the various types are not very similar, they are treated here together because, as already said, this manual is not intended to analyse them in depth.
Just to give a short definition:

- Web Based Surveys (WBS): the data collection is managed online. The respondent accesses the questionnaire via a web connection and fills it in.
- Disk by Mail (DBM): the respondent receives by mail a floppy disk or a CD with the electronic questionnaire to be filled in. He/she is expected to install it on his/her PC, to fill it in and to send it back.
- E-Mail Survey (EMS): it has the same characteristics as DBM, with the only difference that the questionnaire is sent by e-mail.
- Touch-tone Data Entry (TDE): the respondent dials a number (free of charge), which is connected with a computer. He/she answers the questions the computer asks, using the telephone keyboard. This technique can be used when requested data are numerical or can be directly associated to numeric codes.

With all of them, it is possible to include in the questionnaire the skipping rules and a simple set of rules to avoid errors.

They are all quite cheap, but the questionnaire must be very short and simple and they present limitations related to non-coverage errors and nonresponse errors (de Leeuw, 2005). For further detailed discussion see Couper (2000) and Dillman (2000). Due to this last aspect, they are preferably used in a mixed data collection strategy.

Advantages

- Skipping rules can be automated in the electronic questionnaire
- A simple set of rules to avoid range and consistency errors can be included in the electronic questionnaire

- The entire questionnaires and the questions' wording can be customised
- Assisted coding of very simple textual variables can be done during the interview

Disadvantages

- Only simple and short questionnaires can be managed
- Questions and the lists of response items must be short
- Particular care must be taken when designing and developing the electronic questionnaire in order to make its navigation and filling in as easy and pleasant as possible
- Less table styled questions are possible
- Filling in should not take longer than ½ hour

F) Mixed modes

Nowadays a lot of surveys are conducted by mixed mode, which means that one questionnaire is implemented by different modes. According to Biemer and Lyberg (2003), they are the norm these days. This is mainly due to two objectives: firstly, to improve the response rate by using another possibility to contact a respondent and, secondly, to use the appropriate mode for different groups (for example, in order to reduce coverage errors in a web survey due to different access of the population to the internet).

On the other hand, it must be taken into consideration that the use of multiple techniques may result in possible measurement errors even if the same questionnaire is used. The most important issue when applying mixed methods is to recognise from the start that mode effects are existing and the transfer from one mode to another has many implications. For example, the changeover from CATI to CAPI is less problematic (since telephone interviews are less complex than face-to-face interviews), whereas the transfer from a mail survey (with the possibility to put longer response categories into the questionnaire) to a telephone interview creates problems: a paper questionnaire is not always suitable to be "translated" into an electronic questionnaire. Accordingly, the use of the computer has a great impact on the quality of data, so data collected with the two techniques might present comparability problems.

There are different strategies to cope with the mixed-mode approach:

- One strategy is to focus on the main mode and optimise as far as possible the design, the structure, the wording and the layout rules. It is the most practical approach but is a strategy of compromise, because it might not be covering demands of the second mode.
- However, developing a questionnaire with regard to different modes in use leads also to compromise: it is not possible to exploit to a maximum all the potentialities of each mode (de Leeuw, 2005).

Thus, it is a matter of balancing between increasing the response rate by using a second approach to reach the respondent by a different mode, and possibly introducing measurement errors due to the change of data collection mode. For further reflection on the issue see de Leeuw (2005) and Dillmann (2000).

3.1.8. Draft a first questionnaire

The schematisation of the questionnaire, the identification of the lists of variables, the development of the tabulation plan, and the selection of the data collection mode are the starting point to draft a first version of the questionnaire. Nevertheless, many activities are still necessary to obtain the final version. These topics will be dealt with in section 3.2.

3.1.9. Recommendations

A thorough plan on survey concepts is a precondition to deliver relevant and high quality data to meet user's demands. Designing a questionnaire before defining the concrete aims and setting up the total survey design is to be avoided: it will result in high response burden for respondents and interviews, long data checking and less valid and reliable data. Good planning is the base for the development of an appropriate questionnaire which needs to be tested.

Survey objectives and concepts

- Before facing the task of questionnaire development and testing, a search on if and how the topic of interest has been dealt with should be performed
- The general objective of the survey, as well as the main design characteristics, i.e. the target population, the sampling design, the preferable data collection mode, should be clearly stated.
- Close contacts (informal and formal) with the key users are indispensable. Informal contacts and tests as well as expert meetings are of help.
- The general survey objectives should be transferred into concrete research questions of the survey.
- Topics that are not relevant for the survey objectives should not be investigated.
- During conceptualisation the ultimate goal of operationalisation (transfer into indicators and observable variables) should be considered.

Focus group and in-depth interviews

- As regards the communication of concepts and topics to potential respondents, focus groups and in-depth interviews are the main methods of testing.
- If the aim is to get a general overview of the topic from the respondents' perspective, comprehend the social context of the research, and understand how people think and talk about a topic, focus groups should be organised.
- In-depth interviews provide an opportunity for detailed investigation on people's personal perspectives.
- In organising a focus group is recommended to invite 6 to 10 participants.
- Focus groups participants should not be too heterogeneous, as otherwise the discussion may become quite unstructured and confusing. Thus inviting sub-populations and organising several focus groups with the aim of exploring specific issues is helpful.
- People working in the same group (e.g. supervisors and employees, teachers and students) should not be involved in the same focus group.
- There should be a moderator structuring the discussion.
- Length of discussion in focus groups should be preferably within 1-1 ½ hours

Questionnaire schemes

- Develop the entity/relationship scheme in order to have an overview of the units, items and relationships among them.
- The area tree helps to structure the envisaged questionnaire by topics and to make hypotheses on the questionnaire flow.
- The primary structuring of the questionnaire by means of entity/relationship scheme and area tree is the precondition for writing questions.

Definition of variables, values and a preliminary tabulation plan

- A tabulation plan including the list of variables and possible values they can assume should clearly be defined.
- For common variables the use of harmonised definitions is recommended.
- For common variables the reference to existing questionnaire databases or to the same or similar variables used in previous editions of the survey on in other surveys should be made.

3.1.10. Checklist

Survey objectives, conceptualisation and operationalisation

- Review the information already available in the survey area. Review the relevant methodological research papers (literature search).
- Check the survey design and conditions (objectives, resources, sampling, preferable data collection mode, statistical concepts to be estimated).
- Define the user's information requirements and the proposed questionnaire content. If necessary, carry out a meeting with key users in order to make sure that the topics of interest are well covered (see expert groups, section 6.1.).
- Specify the survey objectives (conceptualisation).
- Check the comprehensibility of survey concepts and terms by conducting focus groups and in-depth interviews.
- Transfer the concept into observable variables (operationalisation)

Focus groups

- Define the aim of the group discussion and which topics are the main focus.
- Define critical aspects and approaches to get along.
- Program several focus groups.
- Check date, room and equipment.
- Select and train the moderator.
- Invite participants.
- Check time schedule for e.g. the recruitment of participants.
- Consider to reimburse participants for their time and travel expenses.
- Identify collaborators and methods to work on the results (e.g. tape-recording and transfer to a written format).
- Tape-record and video-record the focus group or observe the discussion by one-way mirror and have full transcription of it.

Questionnaire schemes

- Identify concepts and relationships in the entity/relationship scheme.
- Structure the variables by topics.
- Present the topics by an area tree.

Tabulation and definition of variables and values

- Define variables and values
- Check the availability of harmonised definitions of variables
- Develop a preliminary tabulation plan

3.2. Writing questions

The formulation of a question - any set of words which ask the respondent to give information - and its place in the questionnaire determines how the respondent will interpret it and answer it. The ways in which questions are asked have a major impact on respondent behaviour and on interviewer performance, and consequently on data quality. Moreover, the efficiency of the data collection process, the simplification of coding and the reduction of the amount of editing that is required is also highly dependent on the overall quality of the questionnaire.

Therefore, the major concern with writing questions for a questionnaire should be to provide a set of questions that globally contribute to minimise errors due to the questionnaire, the respondent and the interviewer. Accordingly, good writing – as part of good questionnaire design – should contribute to attain a questionnaire that is both respondent and interviewer-friendly and may also reduce response burden.

In order to minimise the errors arising from writing questions, a set of principles should be taken into account concerning the relevance of the questions, the type of questions to be used, the logical sequence and wording of questions.

As stated in section 3.1, a comprehensive list of relevant variables, ideally retrieved from a standard database of variables, should be available for the questionnaire designer when formulating questions, instructions and response categories. It should cover the items to be investigated and the appropriate definitions to support the questionnaire designer when drafting the questions. The questions should be checked against the variable list, in order to assess whether all topics were covered. For many items, it is not possible to translate a variable directly into one single question. It might be necessary to provide a definition of the variable's meaning, clarifying which kind of information is required and/or to develop a group of questions which together measure the phenomenon.

It is of special importance to test the questions once there is draft version and before finalising the questionnaire with respect to the respondents' capability to understand the questions' wording and meaning and to provide the information requested. An overview of the main testing methods is described in chapter 6.

3.2.1. Principles of writing questions

There is a general agreement on some fundamental principles in questions' wording. The respondent should:

- clearly understand what he or she is being asked,
- in principle be able to answer to the question, and
- understand how the answer has to be given

From the start, one should aim at writing the questionnaire in a way that allows gaining optimal co-operation from all units of the target population and not only from the "optimal respondents".

The questionnaire designer needs to understand the level of the respondents' education, their knowledge of the language, if they are used to express themselves in written form, etc. and should take this into consideration when wording the questions. The extent to which definitions, explanations, or instructions have to be provided must be adapted to the sample persons' situation.

It is important to make sure that the actual respondents have the knowledge and necessary technical skill to answer the questions, i.e. questions should be written in a way that presupposes they are relevant to the respondents. 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. The aim of the question, the type and format of the data requested, the definitions and concepts behind the questions as well as the logical order of the questions should be made clear to the respondents. Conducting focus groups or intensive interviews (see section 3.1.4.) is one way of getting insight into what can be expected from the respondents before starting with the actual wording.

In household surveys the respondents are usually private persons. For this target population, it is necessary to balance the need for information against what the respondent can manage to answer with adequate certainty. In many cases, it is assumed that the respondents have the answers (the information) "in their head", for example life history, consumption, memory of an actual occurrence, knowledge, opinions and attitudes. If the respondents are asked to get certain information from their personal records, it should be clear that they can easily retrieve the required information. When the questions concern the entire household, it is assumed that more than one person in the household is able to give the answer. If a particular member of the household is thought to be the right person to answer a questionnaire, it is very important to provide clear instructions in the questionnaire on who that person is.

In surveys where the sample units are enterprises, municipalities, organisations, etc. it is often a problem to figure out to which person or bureau 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 and to leave the decision internally to the enterprise or organisation who should answer the questionnaire. However in that case, one should not assume that the respondent is automatically a person with good knowledge of the subject matter or that the questionnaire will internally be forwarded to a person having the relevant knowledge. In continuous surveys, efforts are made to establish contact with those persons who can best provide the required information. This is particularly important for large establishments, whose answers have a large impact on the results.

During the writing process, the questionnaire designer should bear in mind some major effects derived from respondent behaviour that can introduce error in writing questions. Some of the most common effects are listed below.

A) Context effects

Context effects comprise all sorts of influences that other questions or information (instructions, section headings, etc.) might have on the respondent’s interpretation of a question (Biemer and Lyberg, 2003). Context effects arise in the comprehension and retrieval phases of the response mechanism and, according to Smith (1991), are “more likely to occur with questions that require wide-ranging memory searchers, access memories that have not been previously organised into a summary evaluation or utilise ambiguous terms and/or have certain intent”.

Context effects are better grasped if we consider the survey interview as a communication process. During this process one is trying to create a dialogue, or conversation, between the questionnaire and the respondent. The respondent needs to be able to easily understand and take part in this dialogue. However, in a survey interview and especially when filling in a questionnaire the respondents do not have the opportunity to verify their conclusions as they would have in the case of an ordinary conversation between two people (Ahola and Lehtinen, 2002). Therefore, the influences of context effects will higher the lesser is the clarity of the respondent position in relation to what him/she is being asked to answer. Accordingly, survey respondents are more dependent on the context of a question than they would be in an ordinary conversation in which meanings can be discussed.

This might lead to a diversity of interpretation on the purpose of a question, because the respondents might have interpreted the concepts and expressions in different ways. In this context, psychological, social and cultural factors play a very important role in the respondent orientation to a given question (or to a given set of questions), thus, affecting the results of a survey.

In the framework of questionnaire design, understanding a question and the reason why it is asked reflects two things: semantic comprehension and its practical and pragmatic understanding Sudman *et al.*, 1996). A respondent chooses an answer according to the purpose the he/she thinks his or her answer will be used for (Ahola and Lehtinen, 2002). This is particularly important for socially sensitive questions, where the respondents are likely to consider their answers more closely and seek cues from the interview situation on the purposes for which their answers would be used.

Testing methods as described in chapter 6 should be used to study how respondents understand the meaning of questions therefore reducing context effects from the question-answer process. However, if used in the right way, the context effects can become a powerful tool to improve comprehension. For instance, it is recommended to specify the perspective all respondents should adopt in answering (e.g. whether a respondent's answer should be oriented towards the future or the past). Grouping all questions on the same topic in a section of the questionnaire permits to introduce the topic once for all, ensures that the context is clear until the beginning of the next section, and reduces the response burden via shortening the questionnaire. The questionnaire may provide a brief description on the use of the statistics collected. Apart from any legal obligations, respondents may require an indication of *why* they are required to complete the questionnaire, and what NSIs will do with the information collected.

B) Recall or memory effect

As already described in Chapter 2, some of the most common problems related to retrieval stage are telescoping and forgetting. One should be aware of the possible bias, when requiring information from the respondents' long-term memory: it is the short-term memory (temporary memory) which is known to be used extensively when completing questionnaires (ABS, 2004a).

Telescoping arises if respondents report events as occurring either earlier or later than they actually occur, incorrectly bringing events into the reference period. This effect can be somewhat prevented by being very specific about when the reference period begins and ends, for example using expressions such as “the week ending Saturday 1st September” rather than “last week” or by anchoring the beginning of the reference period to a relevant life-event (temporary boundaries).

A significant degree of error can be introduced in questions which require respondents to recall events, expenditure etc., particularly if details are requested to be remembered for a long period. The quality of the data collected from recall questions is influenced by the importance of the event for the respondent and the length of time since the event took place. This effect is enhanced when questions require information that is too specific. For example, questions about the number of times respondents did behave themselves in a specific way during a certain time span may produce very inaccurate data because respondents simply may not remember. Similarly, requests to rank large numbers of behaviours, characteristics, etc. may be too demanding and generate unreliable, invalid results. Subjects which are of greater importance or interest to respondents, or events which happen infrequently, will be remembered over longer periods and more accurately. Where possible (e.g. with financial information), questions should be framed so that respondents can refer to their own records which would enhance accurate reporting. This leads to the question of defining the reference period. “The choice of reference period is usually determined by the periodicity of the target events, how memorable or patterned the events are likely to be, and the analytic goals of the survey” (Schaeffer and Presser, 2003).

A study, conducted by U.S. Census Bureau on National Crime Victimization Survey in early 1980s (Bushery, 1981) and reported in Biemer and Lyberg (2003) has proved that the effect of forgetting can be limited by reducing the length of the reference period. Unfortunately, this can have an impact on the survey budget (to obtain the estimates for a longer period more than one survey edition is needed).

In general, what is of utmost importance is that the time context information should not be ambiguous to the respondent. In this perspective, the reference period should be fully specified at the beginning of a question and, if it stays the same, it is advisable to be given in abbreviated form, and in a parallel location, in subsequent questions, thus conserving the cognitive processing (Schaeffer and Presser, 2003).

Another kind of action that can be taken to reduce the error due to forgetting is the use of memory aids and retrieval cues that can be included in the question wording and can help the respondent memory in taking the right direction.

In business surveys, the information retrieval process assumes a quite different meaning (see chapter 2). It consists in the retrieval of relevant information from memory and/or existing company records. It is very important to give clear definition of the economic concept whose values are usually requested in business surveys, so that the respondent could search for the right quantity.

In CATI business surveys a recommended practice is also to send by mail the paper version of the questionnaire together with the survey presentation letter before the data collection period in order to allow the respondents to retrieve the needed information in advance.

C) Sensitivity effects

Questions on topics which respondents may see as embarrassing or highly sensitive can produce inaccurate answers. Their content is often perceived by respondents as an invasion of their privacy. Accordingly the respondents, especially in face-to-face interviews, may provide the interviewer with responses they believe are more “acceptable” because they anticipate the risk of disclosure of the “true” answers which are perceived as socially undesirable to third parties (Tourangeau *et al.,* 2000).

Ways of overcoming difficulties associated with sensitive questions may include reassuring respondents that the information they provide is confidential, and not requiring respondents to write their name anywhere on the survey form. Also, a self-administered questionnaire may produce more reliable responses than an interview, and it may also reduce nonresponse, although this could not be valid for certain sub-populations because of the social context of the respondent (e.g. females in certain immigrant families because of controlling husband, teenagers because of parents, etc.).

However, in an interviewer administered interview, the interviewer behaviour plays a very important role in reducing this effect. Interviewers who appear neutral, and at the same time make the respondent feel at ease should obtain more reliable and valid answers. A good questionnaire design can help the interviewer with this task.

D) Social desirability

As already mentioned (see Chapter 2), it is during the process of reporting an answer to the interviewer that the respondents have to integrate the information retrieved in their memories into an appropriate format of communication. During this process the respondents may disclose only the impression they want to give of themselves. It is in this phase that problems connected with social desirability come out, a source of nonresponse and/or data distortion. A common way of reducing the social desirability effect without harming the respondent cooperative disposition is telling him/her in a direct way, in the instructions at the beginning of the questionnaire, that all the answers are equally good and acceptable, or that there is no “good” or “bad” answers and that people have different opinions. Again, these problems can be due to a combination of factors, such as the personality, the education level and the social position of the respondent as well as conditions of the interview or design of a self-administered questionnaire.

E) Fatigue point

Finally it is important to be aware that poor use of any questionnaire design element, be it language, question sequencing, length, layout etc., creates an obstacle for the respondent. Each obstacle may be only minor, but they all accumulate in the person’s mind until a point is reached when it becomes too much and the person no longer cares about what goes on in the questionnaire. This point is known as the *fatigue point,* and its presence can introduce serious error into the data (ABS, 2004a), or can compromise the completion of the interview.

In the following points it is presented a possible depiction of the major concerns and recommendations about writing questions. Its contents are valid for all data collection modes unless otherwise stated.

3.2.2. Types of questions

Questions can be classified in several ways, using different criteria. With regard to the information or data that can be obtained, there are four main types of survey questions: factual, behavioural, opinion and hypothetical questions.

A) Factual Questions

In these questions, fact based information is required from the respondent rather than an opinion. Respondents could be asked about possession of items (e.g. "Do you have a driver's licence?") or characteristics of the business (e.g. "How many employees does this business have?").

Two specific factual questions are the following:

- **Classification or Demographic Questions:** these are used to distinguish the main groups of respondents in a survey for later analysis (e.g. age, sex, industry).
- **Knowledge Questions:** these questions test the respondent's knowledge about current issues etc. For example, "Who is the Prime Minister?", "Are you aware of these industry support groups?"

B) Behavioural Questions

These questions require information about the activity of the respondent or business (e.g. "How many times did you go to the theatre in the last 12 months?"). Behavioural questions need to be used with care because they often require difficult recall tasks from the respondent. They should be restricted to topics respondents remember easily or are likely to have records for, and cover a reasonable and specific time frame.

Schaeffer and Presser (2003) assert that "the first consideration in asking about an event (or characteristics of events) or behaviour is whether members of the target population are likely to have encoded the information. […]. For events that respondents do encode, two major types of errors affect self-reports:

- Omissions result when individual events are forgotten because of dating errors (e.g. when events are "telescoped" backward in time and so are incorrectly excluded from the reference period), because similar events become conflated in a "generic memory", or, because the wording of a question leads the respondent to search some areas of memory while neglecting others.
- Intrusions result when events are telescoped forward in time (and are incorrectly included in the reference period) or when memories are altered by scripts, schemata, or embellishments from retelling over time."

C) Opinion Questions

These questions seek to measure subjective opinions ("Are you in favour of …?") rather than facts. There are many problems associated with opinion questions. For instance:

- A person's attitude may not be fully developed or respondents may not have given it much thought;
- Opinion questions are very sensitive to changes in wording;
- It is impossible to check the validity of responses to opinion questions.

Opinion questions have two basic components: an *object* and an *evaluative dimension*. Most common dimensions are about agreement (approval or disapproval), truthfulness (true or false), assessment (good or bad), importance (important or not important), and intensity (minimum, maximum).

“The first decision in writing a subjective question is selecting names or labels for the object and the evaluate dimension. The goal is to select names that are easy to understand and that will be understood similarly by all respondents. In addition, the evaluative dimension must seem appropriate for the object. Focus groups are often used during questionnaire development to identify appropriate names and relevant dimensions” (Schaeffer and Presser, 2003).

One of the frequent response attitudes in opinion questions is acquiescence. Passive assent or agreement without protest can be found when respondents have a general tendency to agree rather than disagree with anything. This process occurs when respondents are asked whether they agree or disagree with a statement, especially when the supplied statements are presented as plausible generalities. It can also appear for questions requiring a yes or no response. Respondents tend to agree when the question is *ambiguous* or otherwise difficult to answer. The effect may be exaggerated when the respondent is fatigued or has to answer a long string of questions with the same response categories. A related effect is satisfying, where respondents select the first reasonable answer rather than make the effort to find or remember the best answer. To offset the effects of acquiescence, the literature recommends balancing the direction of agree-disagree items (by posing the same number of statements on each side of an issue) or use forced-choice questions.

There are several types of opinion questions. Two of the most common are addressed in 3.2.4.: rating scales and rankings.

D) Hypothetical Questions

These are the “What would you do if ... ?” type of questions. The problem with hypothetical questions is that one can never be certain how valid any answer to a hypothetical question is nor can you 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 a statement than to go against it. This applies especially if the statement is more socially acceptable to agree.

Hypothetical questions should be avoided because:

- Most people do not predict their behaviour very well
- Many people respond to hypothetical questions based on their perceptions of the probability that events will occur.

Therefore, hypothetical questions should be avoided or used only when referring to a hypothetical occurrence of a situation a respondent will be familiar with.

3.2.3. Questions formats

Questions can be also classified into different categories based on their answer formats. Generally they are classified as one of two types - open or closed - depending on the amount of freedom allowed in answering the question.

A) Open Questions

Open questions allow the respondents to answer the question in their own words, rather than having to select from options.

An example is “What is your occupation?” The advantages of this type of questions are that they allow many possible answers and they can collect exact values from a wide range of possible values. However, they are more demanding than closed questions, both to answer - because the respondent or the interviewer have to write out an answer - and to process because one needs to create a coding frame to classify the variety of responses, which will differ from each other in detail and accuracy. Open questions can also cause problems because of poor handwriting. Furthermore, they present the risk of orthographic errors, which could be solved automatically, when data are recorded, only through orthographic correctors implemented according to the subject treated. Finally, costs and time necessary for the data entry activity of textual variables increase.

A good approach for the use of open questions is to give some examples or directions on how to answer. It should be made clear what the respondent should do if a nonresponse, not applicable, or zero answer applies and it should be made clear what the unit of measurement is. The answer box or area should allow sufficient space for a high percentage of the likely answers.

Open questions are often used in pilot tests to determine the range of likely responses.

- **Numeric Open-End:** *What is your age in years?* ________________________
- **Text Open-End:** *What was your main occupation?* ______________________

B) Closed Questions

Closed questions provide respondents with a range of the most likely, or, ideally, with all possible answers to choose from. The respondent only needs to choose the most appropriate answer.

All possible alternatives in a closed question should be provided, including choices for a nonresponse, zero, or not applicable answer. No alternatives should be left out. The alternatives should be self-explanatory and mutually exclusive. One should take care to provide answer categories that reflect the respondents' characteristics or experience. As far as possible, answers should encompass the full range of variation across the survey sample or population, while keeping the list of options to a manageable length. This is particularly delicate for CATI surveys, where:

- if the list of items has to be read loudly by the interviewer, this causes two disadvantages regarding the increase of the interview length and the so called "memory effect" which make the respondent remember only the first or the last items of the list;
- if the list has not to be read loudly, it anyway requests time for the interviewer to find out the right item to be associated to the response and it implies a high risk of error.

The memory effect could be reduced by the rotation of items or by splitting the question in a set of hierarchical questions, while time necessary to the interviewer to find out the item corresponding to the response can be reduced by listing the category items according to their probable frequencies.

Closed questions are cheaper and easier to process and to respond to (assuming that all the answer choices are applicable to the respondents). Processing time of closed responses is significantly reduced compared to open ended responses. They produce uniform responses and so make summarisation easier. Closed questions are advantageous when:

- All (or most of) the responses can be anticipated.
- An exact value is not needed.

However, they require more effort than open questions in the development and testing stages. The respondent is not given the opportunity to compensate for a poorly framed closed question, i.e., when the choices presented are not exhaustive.
Some examples of types of closed questions are given below.

Limited choice. Limited choice questions are those which require a respondent to choose one of two mutually exclusive answers. A typical example is "yes/no" answers. However, one should be careful in using "yes/no" answers when referring to attitudes, opinions values due to the acquiescence effect.

Multiple choice. These are questions where the respondent is required to choose from a number of response categories provided, from which only one should be ticked, that applies to the respondents.

Checklist (or check-all questions) and Forced choice. In checklist (or check-all) questions more than one answer can be chosen, so allowing a respondent to choose all response categories that apply to him or her. Many questionnaire design experts recommend the use of forced choice instead of checklist questions. Similar to the checklist format, a selection of response categories is provided. The difference is that the respondent is "forced" by a yes/no answer for every category to provide an answer for every response category individually. The slight increase in response burden is justified by the fact that response categories (especially in long batteries of response categories) are less likely to be overlooked (Smyth *et al.*, 2006).

Partially closed. These questions provide a set of responses where the last alternative is "Other, please specify", followed by an appropriately sized answer box for respondents to specify the nature of their "other". Partially closed questions are useful when it is difficult or impractical to list all possible choices or where most responses can be anticipated but the remainder can not. Partially closed items reduce the probability that respondents decide not to complete a questionnaire because they can not find categories that reflect their personal characteristics, beliefs, attitudes, or opinions. However, it has to be noted that the pre-selection of response categories is an useful guide for the respondents. For this reason they should be selected carefully in order to avoid biased results.

C) Choosing between questions formats

In choosing between these two alternatives (open or closed questions), consideration should be given to factors such as the data requirement, the kind of information required, the level of accuracy needed, processing facilities, e.g. resources for doing the coding, the position of the questions on the form and the sensitivity of the question.

Open and closed questions make different demands upon cognitive capacity of the respondents. In the case of open questions respondents have to identify the background and meaning of a question by themselves. When using closed questions, in addition to the question itself the answer categories or scales are presented to the respondents. So, in the case of closed questions although it can be assumed that respondents get a neutral frame which is equal and obligatory for all, research in cognitive psychology proved that the presence of response categories can influence the answers: they help the respondents to clarify questions' meaning, and to build a frame for an adequate response.

In general, closed questions are easier to handle for both the interviewer and the coder. Also, a closed question generally reduces the response burden of a question. For example, asking for income by using income classes is likely to be less sensitive than asking for an exact value.

Open-ended questions should only be used when the researcher does not know or can not predict beforehand all the variety of possible responses and, in addition, when the respondent's answers are considered to add value to the survey objectives. On the other hand, they increase the response burden, can be difficult to answer for persons with low verbal skills and to collect for the interviewers. Finally, they imply a heavy ex-post coding phase. Thomas and Purdon (1994) suggested that answers to open questions tend to be shorter over the telephone as the whole interview procedure tends to proceed more briskly than in the case of face-to-face.

The risks of closed-ended questions are to forget some important categories and to formulate overlapping response categories when a single response is asked. In general terms, the recommended compromise is to use closed-ended questions with the open last category “other, please specify”, at least until an extensive pilot survey has been conducted.

The choice between open or closed-ended question depends also on the level of knowledge the survey designers have on the survey subject. A good level of knowledge is needed to be sure to be exhaustive in the formulation of response categories for closed-ended questions. This is one of the reasons why cooperation between survey methodologists and subject matter experts in questionnaire development is extremely important.

3.2.4. Response categories

The most generally accepted principle about response categories states that it is very important to ensure that they are adequate, exhaustive and disjoint. Response categories also need to be worded carefully, as respondents will use them to clarify or extend the meaning of the question.

In the case of sensitive questions, with several options for the respondent to select from, the less desirable options should be presented first in the list. This indicates to respondents that it is acceptable for them to choose those options.

A) Number of responses options

“The choice of the number of categories represents a compromise between the increasing discrimination potentially available with more categories and the limited capacity of respondents to make finer distinctions reliably and in similar ways” (Schaeffer and Presser, 2003).

The number of response categories can influence the quality of the data as both too few and too many categories can cause errors. Too many can cause respondent fatigue and inattention, resulting in ill-considered answers. Also, questions with a high number of response categories may place greater demand on the storing of information in short-term memory, making the branching instructions more difficult to remember, even if they have been read. If there are too few categories respondents may have difficulty in finding one which accurately describes their situation.

“Based largely on psychophysical studies, the standard advice has been to use five to nine categories, although even that number of categories can be difficult to administer in telephone interviewers” (Schaeffer and Presser, 2003). It should also be considered that choosing from several alternative replies (more than 6) requires visualisation, which is not possible in CATI.

B) Order of response options

The order of response options can also introduce bias. The options presented first may be selected because they make an initial impact on respondents, or because respondents lose concentration and do not hear or read the remaining options. Equally, the last options may be chosen because they are more easily recalled, particularly if respondents are given a long list of options. Thus, the order of response options has a greater effect on data quality when a question includes a large number of response options.

If possible, options should be presented in a meaningful order: whenever there is an inherent order in the list of options, this should be used. If some options are more socially desirable than others these should go last to reduce bias. For example, an education question should present the qualifications in order from lowest to highest. For some self-completed lists, alphabetical order is the most appropriate to help the respondent find which option he/she wants to choose.

Unordered answers are those where there is no graduation of answers. When using unordered answers you should try to control bias in the pattern of answers. Possible solutions are to list the most common alternative first (which also reduces the reading that respondents have to do), or to list choices randomly.

The longer the list of options, the more pronounced the effects of order become when perceived with certain responses styles, namely, the respondents' tendency to choose the first alternatives in self-administered surveys or the last categories when presented in an interview.

To avoid respondents choosing from only the first few options of a long list so they do not have to read the others, options should be categorised into groups or sub-questions. Normally unordered lists should not be longer than about ten items. Questionnaire designers should also phrase the alternatives similarly. The alternatives should be of similar length and be similar in other respects. The use of questions in forced choice format is another possibility. Similar considerations are valid for CATI surveys, although this mode offers the possibility to rotate the response items.

Questionnaire designers should also be aware that the respondent may interpret the provided options as having an order, when, indeed, it was not intended. This perceived order could lead the respondent to misinterpret the question or give inconsistent answers.

When the survey is face-to-face, the response options can be presented either verbally or on a show card. A show card is a list of possible responses to a question which are shown by the interviewer to assist the respondents. This helps to limit errors resulting from respondents being unable to remember all the options read out to them. However, respondents with poor eyesight, migrants with limited native language skills, or adults with literacy problems will experience difficulties in answering accurately.

C) Special Cases of Response Categories

Tables and matrices. Several factual or behaviour questions can sometimes be put together into one table or matrix and give the interviewer – and respondents, in self-administered surveys – a better overview of the questions, although they may affect the respondent preference for sequential order. They are also used in mail surveys due to space constraints. However, experience has shown that, particularly in self-administered questionnaires, extensive or complicated tables with several dimensions (e.g. when it is expected to respond using information from both rows and columns) are rarely completely and correctly filled in, even though they appear practical and logical to the questionnaire designer.

If tables are used, the statements included in tables need to be very carefully drafted to get unambiguous answers in all response alternatives.

Use of standard classification systems. In factual questions some considerations should be made on the use of standard classifications systems. They represent a formal attempt to describe in structured categories something which is almost infinitely varied. In adopting a formal classification system, two main problems may occur:

- Ensuring a consistent understanding of the classification categories, in particular the residual categories of the classification (i.e. "other" categories), as classification systems may not identify the full range of things these categories include.
- Adapting the formal classification if it does not correspond to the informal classification system used by respondents.

When developing questions one will sometimes have to make compromises with the formal classification used in order to get useful answers from respondents. This should always involve consultation with the appropriate standards' experts.

Any classification scheme has to be understood by the respondent before a correct answer can be given. Therefore, all classifications intended to be used should be tested with potential respondents to find out:

- How well the categories are understood by respondents.
- How well they apply to the respondent.

When the respondents show difficulties in using the classification, it is recommended to investigate alternatives in the following order:

1) To maintain the conceptual basis of the classification categories but change the examples of the categories to suit the population of the survey.
2) To use specifically designed response categories and map these to the output classification after the data has been collected.
3) To re-word the question in order to enable the respondent to choose among the categories or to provide an accurate textual response.
4) Include a partially closed question in the classification to give the respondents the opportunity to create different responses.
5) To use open-ended questions only and interpret the responses according to the formal classification system.

Rating scales. There is a type of ordered closed question that is commonly used when the researcher seeks to locate a respondent's opinion - the favourability of an item, the frequency of behaviour etc. - on a rating scale with a limited number of points. There is a wide variety of response scales and they could be characterised by **type of labelling** used (verbal or numeric), **number of scale points** (even or odd), **dimensionality** (bipolar or unipolar) and **direction** (ascending or descending). Some examples of scales are reported below.

Verbal scales:

Strongly Disagree	Disagree	Agree	Strongly Agree
□	□	□	□

Numeric (endpoint-labelled) scales:

Strongly disagree □ □ □ □ □ □ □ □ Strongly agree

Even Scales:

Unimportant □ □ □ □ □ □ Very important

Odd Scales:

Unimportant □ □ □ □ □ □ □ Very important

Ideally, a good response scale should: be easy to interpret by respondents, have a discrimination that fits the respondents' perceptions, and cause minimal response bias. In practice, determining a scale with a "certified" minimal respondent bias is a difficult task because each of those above-mentioned varieties produces specific effects on the responses. Therefore it is important to state that scales – even neutral looking ones, like the numerical scales – are not at all "neutral". Research findings proved that scales which are formerly equivalent but differ in, for example, their numeric labelling, are often answered very differently.

In general, from the scale respondents get "information" about:

- The distribution of "real" situations, or behaviours and,
- Their own position in this distribution.

Thus, as already mentioned, respondent knowledge of the subject matter is important to consider. It is worth considering how familiar the potential respondents are with the issues that are addressed in the survey. As it will be shown later (Chapter 4), also the social and cultural background have an impact.

Although much research has been devoted to scales, its results show a great deal of disagreements, so it would be unwise to give hard-and-fast recommendations on how to define scales. There is, however a set of important issues that should be taken into consideration when constructing one:

1) Category Labels.
   The words used as labels in a rating scale require some caution. For creating interval scales (scales in which the respondent perceive equal-sized gradations between the points on the scale) category descriptors that are truly equal-interval should be carefully chosen. For verbal scales, it is helpful to look for possibly existing lists providing the scale value means of selected adjectives and descriptors that might be used to create rating scales.

   If the rating scales contain numbers, then these numeric values can change the meanings of the scale descriptors. For example, research has shown some evidence that respondents perceive the negative-evaluation side of the scale as being more negative when there are negative numbers on that side rather then positive numbers. Therefore, it is advisable that the points of scale should not be associated with negative digits.

2) Balanced/unbalanced rating scales.
   Rating scales can be structured either balanced, with an equal number of favourable and unfavourable response choices, or unbalanced. Generally, rating scales should be balanced, with an equal number of favourable and unfavourable response choices. This means that there should be an equal number of positive and negative options to choose from. Using more opinions for one side (usually positive) biases the responses towards that side for several reasons. One of them is social desirability, where the lack of sufficient negative options leads the respondent to believe a negative response is undesirable (Tourangeou and Smith, 1996). Using the categories as information the respondent might incorrectly assume the range represents the true distribution of the population (Schwarz and Hippler, 1991).
   Friedman and Amoot (1999) say that "the only justification for using an unbalanced rating scale is in a situation where it is known *a priori* that virtually all respondents are leaning in one direction (if you know that one side of the scale will not really be used.)".
   As part of balancing the options it is important to test whether the descriptors chosen for negative and positive options, e.g. "Difficult" and "Easy", are actually considered to be opposite by respondents. The negative and positive options should also be equivalent in intensity, e.g. "Very difficult" versus "Very easy" rather than "Extremely difficult" versus "Fairly easy".

   To prevent that the question itself does not bias the respondent's answer, it is important to word the rating question in a balanced way, e.g. "Do you favour or oppose …" to imply that responses in either direction are acceptable (Schuman and Presser, 1981).

3) Neutral option.
   When a rating scale has odd numbers of points, the scale has a midpoint at which there is a transition. This midpoint can indicate either indifference (e.g. neither unimportant nor important) or ambivalence (e.g. unimportant in some ways and important in others), so that the definition of the midpoint potentially affects the meaning of other points as well.

   Research has shown that when a middle category is offered responses tend to concentrate in this category. Arguments exist for including and not including a "neutral" point: if not included, people place more frequently the responses in the positive side of the scale, due to the general willingness to be "nice" rather critical. Therefore, it is generally preferable that rating scales include a "neutral" option such as "Neither satisfactory or unsatisfactory", or "Partly satisfactory" in the middle of the scale or "Not applicable" or "Don't know" at either ends. This could be important because, again, respondents should have a formal way to indicate when (they think) a question does not apply to them. If forced, there might be a risk that respondents will produce an opinion on the spot that is neither accurate nor stable (Flynn, 1996). However, from the point of view of item nonresponse, the general recommendation is to use at least the "don't know" category as seldom as possible (de Leeuw *et al.,* 2003).

4) Order effects in rating scales.
Friedman and Amoot (1999) cite several studies that show evidence of a bias towards the left side of the scale. This seems valid either placing the negative or positive end of a rating scale first. Although they conclude for the non-existence of a way of determining which scale is more valid, they nevertheless warn for the ethical issue of placing deliberately the desired response on the left side of the scale.

5) Number of scales points.
"Ideally, a rating scale should consist of enough points to extract the necessary information" (Friedman and Amoot, 1999). When the concern is on the maximisation of the scale's reliability and the ability to discriminate between degrees of the respondent's perceptions of an item, there is no single number of points in literature review for a rating scale that is appropriate for all situations. In general, it is suggested the use of five-to-nine categories. For five categories, these would consist of one low and one high intensity option for each direction and one neutral. If required, seven or nine categories may be used but any more than that generally makes too fine a distinction between ratings that are of necessity vague (ABS, 2004a).

   It is reported (Thomas and Purdon, 1994) that in questions involving response scales, respondents on the telephone have been found to be slightly more likely to choose on of the extreme categories. This should probably be taken into account when designing CATI surveys. The frequency of the choice of extreme points on scales has furthermore been found to be quite heavily dependent on the cultural background of the respondents. For this case international comparisons based on questions with rating scales require special attention.

6) Consistency.
When selected a scale format, this should be used throughout the whole questionnaire.

Rank order scales. Ranking questions are those opinion questions which ask the respondent to number the different options in a question in order of importance. In general ranking questions should be avoided for two reasons:

- They are quite complicated to explain and respondents often have difficulty completing them correctly.
- The output from ranking questions is quite difficult to deal with unless you are looking for a "winner" alone as in voting procedures. In particular, the problem is with ranking only some of the items. When a respondent ranks every item, it is easy to assign a value to each item that can be assembled in some way, but when a respondent only ranks some items it is very difficult to decide how to value the items left blank. The interpretation of those items is unclear – are they equally unimportant, are they not applicable, and are they equivalent to those ranked last by other respondents or not?

The recommended way of measuring items that require a ranking is to ask respondents to rate each item individually, using a verbal scale rather than a numeric one similar to other rating questions.

"Don't Know", "Don't Remember", "Not Applicable" categories. "Don't know" and "Not applicable" are response categories directly related to the relevance issue: sometimes it is appropriate to include those categories, for example, when the researcher has previous sufficient awareness that a fairly good amount of respondents might have “no opinion” or when he/she knows that a particular question does not apply to a subset of the target population.

For factual questions, these categories are kinds of item nonresponse. For opinion questions “Don’t know” or “Don’t remember” might be used by the respondent as a neutral answer, if this is not present among the responses’ options. The decision about whether to include or exclude these categories depends to a large extent on the subject matter. Excluding these options may not be a good idea, as respondents may be forced to give an answer when, for example, they really do not know what their attitude is to a particular subject, or they do not know the answer to a factual question that has been asked. When respondents are forced to develop an attitude on the spot, e.g. through a forced-choice rating scale, this attitude might be highly unreliable. This approach may be reasonable only when the researcher has good reason to believe that virtually all subjects have an opinion. In general terms, if the question is crucial for the survey these options should not be allowed. If the question is sensitive and not essential, the possibility of nonresponse might be considered.

In CATI/CAPI surveys it is recommended to include these options among the response categories, but to instruct the interviewer not to read them aloud.

3.2.5. Question phrasing

A) Language

As it was mentioned earlier, a basic rule in good questionnaire design is that only through appropriate language the respondent can fully understand what is asked, and provide the appropriate information. Nonetheless, considering that we are dealing with people and the meaning of words, the potential for misunderstanding is enormous.
At least in general population surveys, respondents are people of all social strata and ability levels, with no special awareness of statistical language. While the ability of respondents to understand questions and explanations varies, it is very easy to overestimate this level of ability.

It was suggested that questionnaire designers should work with the following minimal expectations of respondents (ABS, 2004a):

- They have a limited vocabulary. They can understand short sentences with simple punctuation, but are likely to be confused by long and elaborate explanations;
- They can understand positive instructions more easily than negative ones.
- They know nothing about National Statistics Institutes procedures or structure.
- They might not understand why the information asked for is needed. If they can not understand what a questionnaire is for, they are less likely to fill it in promptly and accurately (see *context effects*). The same applies to individual questions.

Another factor one must be aware of is the tendency to believe, particularly in NSIs, that respondents understand the author’s definitions and jargon, and that explanations are well understood. Unfortunately, that might not be generally the case. Actually, despite the fact that the use of words with precise and standard definitions are commonly used in NSIs questionnaires, unless evidence from tests and evaluations shows that the respondents understand the meaning, we must make the meaning clear in appropriate language. If, through testing, you know your group of respondents understands, e.g. a jargon word, then you can include the word in the questionnaire with confidence. In essence, the way to make the respondents understand is to “talk” to them in the language they understand. The principles of appropriate language are described in the next paragraphs.

B) Words

The most common principle in wording states that simple and unambiguous terms, that all respondents might understand in the same way, should be used. Therefore, the use of vague, difficult, technical or foreigner terms should be limited or avoided if possible. Exception to this recommendation should be considered if respondents are subject matter experts and will answer easier to specific technical questions than to more common terms. If it is necessary to use them, a definition of the terms should be provided. In CATI and CAPI surveys, definitions should be implemented through the use of on-line helps (not within the question wording in order to not increase response burden). Finally, also contractions or abbreviations should be avoided.

However, questionnaire designers should be aware of the limits of word simplification: on one hand, simplification can not lead to, for example, the use of childish language; on the other hand, when asking factual questions to establishments, it might be more important to use a correct terminology than trying to phrase questions as simple as possible.

Directly linked to the adopted language, is the ambiguity factor i.e. the unclearness of terms which have more than one meaning. If ambiguous words or phrases are included in a question, the meaning may be interpreted differently by different people. This will introduce errors in the data since different respondents will be virtually answering different questions. The more ambiguous the semantic or pragmatic meaning of a question is, the more the respondents seek clues from the context of the questionnaire or from the data collection situation.

Questions items should therefore be as precise as possible – words such as *typically* and *usually* should be used cautiously, if at all, because respondents will define them differently.

A question may also seem straightforward, but allow for a variety of different kinds of answers. Other important issue is that questions should be always defined in time and space. Also, the measurement unit required should be included wherever it applies, e.g. euros, days, litres.

As it was mentioned earlier, the use of words in a particular context may change the meaning. For example, when used alone, a word may be quite familiar to people, but when used in a specialised context it may lead to confusion. In many cases the reader will have a clearer concept of what is to be filled in if the reasons for requiring the information are known. Apart from any legal obligations, respondents may require an indication of why they are required to complete the questionnaire, and what the producers will do with the information collected. Therefore, questionnaires should have a brief description on the use of the statistical information collected.

In writing questions it also should be kept in mind that, especially in the context of statistical systems, it is crucial that the data produced allow comparability over time, across geographic areas and among sectors of activities. Likewise, the data produced should as well allow for integration with data from other organisations, national and international. Thus, to facilitate such data integration, the questionnaire designers should use standard definitions for the thematic concepts, variables and classifications, as well as the populations and statistical units to which they apply.

C) Sentences length

Lengthy or complex questions can exceed the respondent’s capacity to process them. If questions with higher number of words are more difficult to understand, then it may be that respondents need to concentrate on understanding the question, at the expense of the navigational features of the question. Therefore, short sentences are easier to understand than long ones. Although it is unrealistic to specify a maximum length, each sentence should convey a single item of information.

D) Tone

Minor changes in wording can also have a significant effect on responses. The characteristic style of the phrasing of a question is an indicator of the attitude or view of who produced it. A change in wording can result in a change in responses. For example, different responses may be obtained through using the following two questions:

"Do you think that gun ownership should be forbidden?" , or

"Do you think that gun ownership should not be allowed?"

One should therefore be careful when looking at alternative wordings. The use of *negative words* like "*not*" should be avoided in questions as they are easily missed by respondents. Research has shown that people who are asked negative questions are much slower in responding. The slower response is caused by people having to think about the answers rather than giving them automatically (ABS, 2004a). Although generally it is favourable that respondents think about questions in this case it could mislead to give answers that differ from their actual attitude. Negative questions cause additional effort for respondents, for no good reason. Other reasons for avoiding negatively stated questions, specifically negative verbs (e.g. *failed* or *prohibit*) are that it may have emotional connotations. Respondents may:

- Distort their responses to such statements, intentionally or unintentionally;
- Provide neutral responses (if available) or not respond to the statement question;
- Refuse to complete the entire questionnaire.

In addition, using "*not*" in a scale such as "Satisfied", "Neither satisfied nor unsatisfied" and "Not satisfied" does not provide a true opposite. "Dissatisfied" would be a better alternative however "Unsatisfied" could also be used and would mean something slightly different to respondents.

Some data collection modes, e.g. CATI, allow to personalise the interview by means of the customisation of the wording, based on prior information on the respondent (name, names of the children, etc.), thus making the tone of the interview more conversational (see Chapter 5).

D) Order of clauses

Sentences should have clauses in chronological order to aid comprehension. In the example below, the first action is to read the instructions, but this does not become clear until the end of the sentence:

"Before you fill in the questionnaire, read the instructions"

A better structure of the sentence is:

"Read the instructions then fill in the questionnaire"

E) Active versus passive voice

Most people find sentences in the active voice easier to understand then the passive voice. Passive voice is when the subject is acted upon (example: "This questionnaire is to be completed by the operator whether operating for the full year or only part of the year"). Active voice is when the subject performs the action (example: "The operator is to complete this questionnaire, whether operating for the full year or only part of the year").

F) Punctuation

Punctuation aids comprehension if used correctly. However not all punctuations improve clarity because only commas, full stops and question marks are readily understood by most people. Additionally, the elongation of sentence using too many clauses should be avoided. Instead of using e.g. a semicolon or a long succession of commas a new sentence is the favourable alternative.

G) Instructions

There are two types of instructions used in questionnaires. The first type deals with *what to do* with the questionnaire, where to get help and so on. The second type deals with *how to* answer questions.

Instructions are frequently ignored on questionnaires. Instructions placed at the start of questionnaires, if read at all, are often forgotten by the time the person has answered one or two questions. On the other hand, it is too late to give instructions or explanations after the respondent has jumped to the new question. Also, one should have in mind that instructions might only be read if respondents get stuck, i.e. respondents only read what seems to be necessary to achieve the filling task. Therefore, the instructions need to be placed as closely as possible to the place where they apply, preferably between the question and the answer space. This often means repeating the same instruction each time it is to be acted upon. This is particularly the case with instructions on *how to answer* questions.

Instructions related to *what to do* with the questionnaire in general are best grouped together, either at the very beginning or the very end of the questionnaire. As it was stated earlier, it is very important to provide clear instruction in the questionnaire on who is the right person to answer (especially in establishment surveys). Some characteristics the selected respondent should own have been enlightened by Willimack and Nichols (2001): he/she should be the person who has broader knowledge of the existence of a variety of types of requested data, and that has also the needed authority to gather the requested data from multiple sources and to release it.

For mail surveys to establishments, it is recommended that instructions should show clearly which position or competence the person answering the questions should have. It can not be assumed that the recipient 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.

H) Use of Qualifiers

Another aspect to keep in mind concerns the use of qualifiers in a question or questions with a lot of informative words. Qualifiers may impose or add an unfamiliar concept to what was a familiar one. Too many information carrying words may interfere with the respondent's grasp of the main element of a question. Short qualifiers can be presented at the end of a question in brackets. For example, the question:

"Have you vacated any land including change in tenancies?", could be improved as follows:

"Have you vacated any land?"

- Including:
- Change in tenancies

I) Special cases of critical questions

Double-barrelled questions

Apparently single questions which actually incorporate two different questions should be avoided altogether. For example: " Do you intend to leave work and return to full-time study this year?" A person may be intending to leave work, but not to return to study, and vice versa. When different parts of the question may have different answers, or parts of the question are not relevant, respondents may not be able to answer the intended way. When attempting to interpret answers to such questions, it is unclear to which part of the question the answer corresponds.

A double-barrelled question, i.e. a question open to two possible interpretations, is a source of ambiguity and therefore a potential source of nonresponse and increases the likelihood of errors.

An indicator of such questions is the use of conjunctions "and", and "or". Thus, double-barrelled questions should be avoided: one question at a time should be asked.

Leading questions, Unbalanced questions
Error will be introduced if questions lead respondents towards a particular response. For example, the question "How many days did you work last week?", if asked without first determining whether respondents did in fact work in the previous week, is a leading question. It implies that the person would have or should have been at work. Respondents may answer incorrectly to avoid telling the interviewer that they were not working. 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*?"

Another form of leading question is unbalanced questions. It occurs when the 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 proportional if it is written: "Are you positive or negative towards...?".

3.2.6. Defining the questionnaire flow

Another important phase in the questionnaire development is the definition of its structure that can be performed with the support of the analysis made in the operalisation stage by means of the area tree.

For the overall readability of a questionnaire, a smooth progression through the questions is particularly important to minimise the nonresponse and measurement errors. Here the structure and layout elements of designing questionnaires play the fundamental role.

Good practice (ABS, 2004a) suggests that one should design the sequence of questions to:

- Initially provide and later maintain respondents' motivation to complete the questionnaire. Beginning with simple questions is a good way to achieve this; also, the opening questions should establish that the respondent is a member of the survey population;
- Aid respondents' recall;
- Direct respondents to the information source or sources they should use;
- Be relevant to respondents' own records;
- Appear sensible to respondents.

As a general recommendation, and regardless of the method used to administer the questionnaire, the questions on a questionnaire should follow a sequence that is logical to the respondents, i.e. the sequence in which material is presented on the questionnaire should be the same as the sequence in which you expect respondents to work through the questionnaire. Also, the conventions adopted should be consistent throughout the questionnaire. Two specific aspects related to sequencing are outlined below: the order of the questions and the case of filter questions.

A) Order and grouping of questions

The order and grouping of the questions should be carefully considered because a preceding question can influence the attitude toward a following one.

In general, it is recommended to keep similar topics together and place items in logical order when it exists. For example, place items in an order that reflects the sequences in which events occurred. In addition, items should be grouped by content and type within content areas. This enhances the comprehension of the respondent since when similar items are located in close proximity they are seen as a whole. Arranging the elements carefully so that the related parts are seen as a group is therefore essential: all the elements within a question (text, notes, response options, answer boxes), should all be close to each other. When a questionnaire contains groups of questions topics, it can aid the respondent to give these groups distinct part or section headings, such as "Income" and "Expenses". This is especially useful for surveys where the respondents do not need to complete all the groups of questions. Sections or parts should be based on fairly broad topics and the questions within the section or part should always be more related to each other than they are to the questions in the surroundings groups. Also, section headings can be useful where respondent is directed to skip over sections of a questionnaire. In this situation section headings can help in the quick identification of the relevant sections of a questionnaire.

However, overuse of sections or parts, like any other element on the page, makes the questionnaire too visually complicated. Questionnaire designers should always try to avoid having a section or part that contains only one question and in general the headings should not replicate question text as this increases clutter without aiding comprehension. Sections or parts should not be used instead of questions – they are headings only.

The order of the sections is another important issue to consider. Here the general rule is to proceed from less to more complex topics so to have the time to build up a confidentiality climate with the respondent, and also to ensure a substantial amount of information to be collected before an eventual interruption.

With respect to the level of abstraction, two opposite approaches can be used in ordering the questions: from general to particular or vice versa. Which approach is better depends mostly on survey objectives. If the "from particular to the general" progression is used, the respondents would probably reach the more general questions more conscious of the particular aspects the researcher considers as part of the topic of interest. If the "from general to the particular" approach is followed, the respondents would probably be less aware and more superficial in their answer to the general questions. The first alternative could seem better than the second one, but it presents its own risks. For example the respondent can provide an overall opinion different from the real one only to be coherent with the answers already given. Furthermore there could be important aspects of the topic that the producers have not considered and that could come out with the second approach, but not with the first one.

The location of sensitive questions is also to be considered. The negative effect of sensitive questions may be aggravated if they are placed at the beginning of the questionnaire and can therefore contribute to nonresponse if respondents are unwilling to continue with the remaining questions. Therefore, questions which may be sensitive to respondents should generally not be placed at the beginning of a questionnaire. Rather, they should be placed in a section of the form where they are most meaningful to the context of other questions. In this way the format of the questionnaire can act as a buffer to help the respondent feel more comfortable with sensitive questions after establishing rapport.

B) Filter Questions

Some questions, usually referred to as filter questions, ask respondents to make a choice, where one of the choices leads them to the next question, while the other leads to a different question or topic. A filter question is used to exclude respondents from subsequent questions if they do not apply. Therefore, filter questions help the respondent to understand the sequence of questions.

Filter questions are also used in interviewer based surveys to direct interviewers to follow a series of questions according to answers given by respondents. Filter questions need to be used with care, as respondents (and interviewers) need to have sufficient information about the branching condition to judge whether to skip. To this aim, an instruction box with definitions or a qualifier will often be necessary. Filters should also generally be avoided for sensitive topics as respondents might give the answer that avoids answering the sensitive questions.

If the instructions are not clear and straightforward, interviewers or respondents can follow an incorrect sequence or miss questions. In general, only one or two conditions should be placed in each sequence guide. Large numbers of skip patterns may indicate that the target population has been inadequately defined or that much irrelevant information is being requested.

In general, it is recommended to place the sequenced response first in filter questions because that response takes the respondent to a subsequent question and then they do not have to read the other option: respondents may become annoyed if they are required to read numbers of items that do not apply to them. Also, questionnaire designers should avoid questions that allow respondents to correctly leave the answer space blank when they have been routed to that questions, as it is then unclear whether the respondent should have recorded an answer or not.

Electronic questionnaires can make complex sequencing much easier (see Chapter 5). One of their primary advantages is the ability to program skips so that branching errors are eliminated. Errors of this type exist only insofar as the programming has not been able to anticipate every possible combination of answers and allowed for a proper track through the questionnaire (Bradburn *et al.*, 1991). Similarly if a set of questions has to be asked a number of times (for example, for everyone in a household), the computer will automatically repeat the questions (go round the “loop”) the correct number of times and then move on (Sainsbury *et al.*, 1993).

3.2.7. Length

The length of a questionnaire can be described in different ways. On section 3.1 it was already mentioned the importance of avoiding questions that are not strictly relevant with respect to survey objectives.

Survey designers sometimes worry about the number of pages (especially on mail surveys), whereas the number of questions (especially mandatory ones) and the time taken to complete are usually more important. The questionnaire length should be balanced considering the response burden, the mode of data collection and the fulfilment of survey goals. Towards the end of a long questionnaire, respondents may give less thought to their answers and concentrate less on the instructions and questions, thereby decreasing the accuracy of the information they provide (fatigue point effect). On the other hand, there is little use in cramping too many questions on the pages just to save few pages of the questionnaire. Moreover, it is known that the respondents’ motivation has a significant influence on their willingness to fill in the whole questionnaire, so the higher their motivation the less important is the length issue. Thus, the general recommendation is not only to try to “save pages”, but also to work on respondents’ motivation and on questionnaire fluency. Well structured and designed questionnaires might, therefore, contribute to minimise problems caused by questionnaire length, especially in surveys with mandatory reporting, where the burden on the respondent can be significantly larger than in surveys with voluntary participation.

As mentioned earlier, the greater the number of words in an item, the greater the possibility of introducing bias, and it should be remembered that efforts to enhance clarity at the expense of brevity might be wasted if respondents do not read the entire item. Again, this calls for finding the right balance in question phrasing, taking into account that short items are generally best for written questionnaires, but longer items may produce more complete responses in interviews.

The recommended length of interview is strongly linked to the data collection mode applied as has been stated in section 3.1. Whereas a face-to-face interview should not exceed one hour, the duration of telephone interviews and the completion of electronic questionnaires should only last about 30 minutes, and 45 minutes for mail surveys. Of course, these limits can be exceeded when respondents are easy to motivate.
For CATI and CAPI, it is important to note that the actual length of the interview depends on three factors: the length of the questionnaire itself, the efficiency of the CAI instrument and the experience of the interviewers. In CATI and CAPI it is also a matter of consideration the additional time necessary for the interviewer to do on-line editing and consistency checking, which may increase the time an interviewer spends, on average, with a respondent. A critical issue on CAWI is due to the possible dropouts in the web connection and to the costs for the respondents.

Finally, and although this handbook does not address specific respondent burden issues, it can not be ignored that the quality of the answers is also dependent on the time the respondent must devote to produce the correct information. In surveys to establishments, this corresponds to 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.

Moreover, the same person can be called upon to answer several different questionnaires sent out by National Statistical Institutes 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. Thus, the use of standardised questions reduces the burden on the respondent and increases the accuracy of the answer. Also, good coordination between statistical agencies and the business units is essential in reducing burden. Therefore it is recommended that NSIs design strategies that address the specific issue of the relationships with establishments.

3.2.8. Recommendations

- *Context and sensitivity effects*: the questionnaire should be supported by a presentation on survey objectives and a clear confidentiality assurance. It is advisable that, before the data collection, an advance letter (and the paper questionnaire to business units interviewed by phone) is sent by mail.
- *Memory effect*: memory aids, retrieval cues and appropriate reference periods should be used.
- *Hypothetical questions*: this type of questions should be used with caution, particularly when concerning opinions and attitudes.
- *Response categories*: there should be no overlapping among the response categories and they should cover all possible answers; long lists of response categories, particularly in CATI surveys, should be avoided; in numeric scales, it is advisable that the points of scale should be not be associates with negative digits; in CATI/CAPI surveys, in general terms, the “Don’t know” option should be included among the response categories, advising the interviewer not to read it aloud.
- *Order of response options*: if possible, options should be presented in a meaningful order: whenever there is an inherent order in the list options, this should be used; if some options are more socially desirable than others these should go last; options should be categorised into groups and sub-groups; list the most common alternatives first when using unordered options and they should never be longer than ten items. Especially for long batteries of items, the use of forced choice questions is recommended.

- *Use of standard classification systems*: when available, standard definitions for concepts, variables and classifications should be applied; validating techniques should be used to look for possible difficulties in using classifications.
- *Language*: simple language for questions and instructions should be used; technical words should not be used; long sentences should be avoided; negative words should also be avoided; sentences should have clauses in chronological order; active voice is preferable to passive voice; ambiguous expressions, if necessary to be used, should be defined; general, abstract and deductive questions should be avoided; validation techniques should be used to choose the best wording for the most relevant survey questions; if possible, instructions should be located as closely as possible to the place where they apply; for business surveys, clear instructions on who is the right person to answer together with the definitions of the economic concepts and the required values should be provided in the questionnaire.
- *Double-barrelled questions:* they should not be used: not only they are confusing for the respondent, but also create problems in the analyses: which part of the question have people in fact answered? Instead, a single information should be asked about at a time; also, questions with too much information should be avoided: when complicated questions are asked, containing several clauses and determinations, the respondents may give answers to questions which they have simplified.
- *Leading questions and unbalanced questions*: questions may easily become leading or appealing or may contain persuasive definition; therefore wording should be designed with caution not to construct leading questions. Attitude questions should be balanced: the question should reflect both sides of an opinion.
- *Sequencing*: the questionnaire flow should follow a logical stream; questions should be arranged into logical groupings: let the subject decide the grouping of the questions; the use of checks should be carefully evaluated against the increase of difficulties in controlling the interview; filters should be avoided for sensitive questions; in general, place the sequenced response first in filter questions; only one or two conditions should be placed in each sequence.
- *Order of questions:* the questions must be scrutinised with respect to their context. The respondent's interpretation of a certain question may be influenced by earlier questions; questions on the same topic should be grouped in the same section; with regard to sections' ordering, proceed from less to more complex topics; with sensitive questions, the topic should be reached gradually, in a way that makes the respondent feels at ease and the wording should be as neutral as possible. In general, sensitive questions should not be placed in the beginning of the questionnaire.
- *Length*: the questionnaire length should be balanced considering the response burden, the mode of data collection and the fulfilment of survey goals; questions that are not strictly relevant with respect to survey objectives should be avoided.

3.2.9. Checklist

- Think about the language: keep it simple.
- Ask short and concrete questions.
- If used, define ambiguous expressions.
- Ask about one thing at the time.
- Define the questions in time and space.
- Give instructions in connection with the questions: examples are good but must be simple, common, not leading.
- Find out which national and international standard classification systems need to be followed to meet requirements for comparability with other statistics.
- Be careful with hypothetical questions.
- Avoid leading questions.
- Always balance attitude questions.

- The answer alternatives must cover all possible answers.
- Make the alternatives mutually exclusive.
- Remember that the answer alternatives should show the real variation as far as possible.
- Use open ended question only when absolutely necessary.
- Do not use too many categories – think about the limitations of the short-term memory.
- When you choose the alternatives:
  - Consider that the respondent can feel there are too few alternatives.
  - Do not force the respondent into stating an opinion she or he never thought of. Include the middle or neutral, “do not know”.
- Test your questionnaire: make a pretest; compare the questionnaire with the tabulation plan.
- Place the questions in logical groups.
- Think about the order of the questions.
- Ensure there are good linking questions between different subjects.
- Limit the skips in the questionnaire; if used keep skips’ instructions simple.
- Consider the response burden.

3.3. Visual Design Elements

From a cognitive point of view, questions in questionnaires present standardised stimuli to the respondents in order to generate measurements. As a precondition for valid and reliable results, questionnaire design has to make sure that each respondent receives the same stimulus in the same way. Verbal stimuli like the terms used for question wording, the order of the words in a question, and the sequence of the questions have received much attention in the literature. However, the importance of visual stimuli should not be underestimated.

Visual design elements comprise all design elements except the wording of the questions and the sequence of the questions (section 3.2). They include the layout of questionnaires (either on paper or as computer screens), the fonts, the structure of the tasks required in the questions and also the “look and feel” of the questionnaire. The technical development of electronic questionnaires is a special case of non-verbal design and will be discussed in detail in chapter 5.

Visual design elements guide the respondent as well as the interviewer when filling in the questionnaire. They include the questionnaire layout as well as the use of symbols and other graphical elements, but also the setting in which an interview is taking place. The importance of visual design elements is obvious in self-administered questionnaires (either as PAPI or as CASI instruments). In these cases the visual design should directly support the respondent in filling in the questionnaire, for instance, by transmitting an instruction on where to start or on what to fill in without any verbal instructions, etc. However, also interviewer-administered questionnaires consist of more than words: in CATI and CAPI questionnaires, the design of the screen determines how the interviewer will read the question to the respondent. In many CAPI questionnaires, the respondent will also be shown screens in the course of the interview which contain numerous visual design elements. Visual design must facilitate the interviewer’s work and make it effective.

3.3.1. Visual design elements in questionnaires

Cognitive science usually distinguishes three levels of visual design which questionnaire designers have to take into account (Norman, 2004; Jenkins and Dillman, 1997). At the *emotional level* (determined by the connection of the brain with the visceral nervous system), the respondent and the interviewer produce positive or negative reactions (emotions) without reasoning. In order to be understood correctly, it is important that questionnaires should be visually pleasing and look easy to complete. The *functional level* determines the usability of the questionnaire, i.e. whether the information is cognitively processed by the respondent as intended by the survey designer. The structure of tasks as well as the cognitive processes required of the respondent and the interviewer should be designed in a way as to enable them to provide the answer correctly. Therefore, questionnaire designers have to be aware of how visual information is processed. Finally, the respondents take a conscious and reflected decision when participating in a survey or answering survey questions. This level is referred to as the *reflective level*. At the reflective level, respondents and interviewers attribute meaning also to non-verbal features of the questionnaire and to their activity of providing answers (Norman, 2004, pp. 63-98; Dillman *et al.*, 2005). Those involved in questionnaire design should be aware that this can influence the readiness to participate in a survey as well as the responses to the questions. It has to be noted that the distinction of these levels is an analytical one. In the empirical reality the three levels are in complex interaction (Norman, 2004, pp. 21-24).

After the structure and wording of the questions have been completed, the next step therefore is to implement them into a questionnaire which sets the right visual stimuli. The implications of these three levels of visual design elements will be discussed in the following. As many questionnaires are used in National Statistical Institutes, it is advisable to standardise the visual questionnaire design. If carefully tested prior to its implementation, a standardised approach can considerably reduce the effort needed to develop the questionnaire. At the same time, it can promote a corporate design for all questionnaires of a statistical office.

The visual design of a questionnaire has to be developed with respect to the main users of the questionnaire and the data collection mode. For example, a self-administered questionnaire should focus mainly on the requirements and capabilities of the respondents, whereas interviewer-administered instruments will also have to take into account the requirements of the interviewer.

3.3.2. The emotional level of visual design

Questionnaires, like other everyday objects, evoke responses connected to the visceral nervous system, i.e. immediate automatic responses which are genetically determined. These responses are responsible for registering a questionnaire as pleasing, looking easy to complete, attractive etc. As Norman (2004) argues, responses at the emotional level are quasi universal as they are genetically driven. Abstracting from the numerous interactions with cultural factors, respondents produce the same positive or negative reaction to a given questionnaire without reasoning.

These immediate responses are of major importance for the measurement process as objects that are perceived as visually pleasing or beautiful also tend to function better. Questionnaires should therefore respect a small number of principles in order to look visually pleasing and be easy to complete. An immediate positive reaction towards the questionnaire will at the same time improve the usability of the questionnaire.

Unfortunately, until today little information has been available on universal principles for achieving visually pleasing questionnaires. Building on a list proposed by Norman (2004, p. 29), questionnaires around the world (as well as other objects) are being associated with positive feelings if they are

- harmoniously designed,
- rhythmically constructed,
- contain smooth or rounded shapes, and are
- composed of symmetrical shapes.

In order to prevent immediate negative reactions they should avoid

- extremely bright lights and dark colours,
- sharp or discordant shapes,

- crowded as well as empty pages, and
- sudden changes in the questionnaire flow.

Although these general principles should be taken into consideration when constructing a questionnaire, one should not forget that there is an interaction between the emotional level and the functional and reflective levels. The final questionnaire has to fulfil the requirements at all three levels: it should inspire immediate positive effects without showing deficiencies pertaining to the usability and the cultural context.

3.3.3. The functional level of visual design

At the functional level, questionnaire design focuses on the function and performance of the questionnaire (without taking its appearance into consideration): does the design succeed in communicating the concept of the survey designer to the respondent's mind? Since survey methodologists started making use of the methods and findings from cognitive science for questionnaire design, the cognitive aspects of the question-answer process have constantly been subject to scientific research, which has resulted in a quite elaborate set of cognitive methods for questionnaire testing.

It is important to note that not only the wording of the questions, but also the visual features of questionnaires have a strong impact on their usability (see section 5.7.). As far as the respondent and the interviewer are concerned, information is communicated through the use of verbal *and* non-verbal stimuli, which are processed by the respondent and interviewer and given meaning (Jenkins and Dillman, 1997). There are a number of general rules which should be taken into account when constructing a questionnaire, described below.

A) Reducing the complexity of the structure of tasks

The questionnaire designer should be cautious not to cognitively overburden all or some of the respondents. A valid and reliable measurement is only possible if the entire target population has the capability to cognitively process the questions. Survey designers normally have a profound knowledge of their subject matter area and are used to handle questionnaires in their everyday work. For this reason, there is a tendency to underestimate the complexity of the cognitive tasks required of the respondents.

The structure of tasks can be simplified by the application of a number of simple rules.

1) One question should be asked at a time and no more. In order to reduce the number of pages in a questionnaire, survey designers sometimes ask several questions at a time, which would better be asked separately using a simple skip instruction (like "In 2004, did you receive any self-employment income?" and "How much did you receive?"). Integrating several questions can easily confuse the respondent so that the questionnaire designer may partly lose control of the answering process. Putting two or even three questions into one can also result in parts of the question (like the reference period or other important details) being forgotten while such "multiple" questions are being processed. Furthermore, there is some empirical evidence that double or even triple question formats increase item nonresponse, whereas there is hardly any evidence that an increase in the length of the questionnaire (resulting from asking every question separately) leads to an increase in unit nonresponse.
2) Similarly, the use of matrices should be minimised as far as possible. Especially in business surveys, matrices can sometimes not be avoided. However, at least parts of the population have serious problems in filling in matrices correctly, so that any alternative approach should be taken into consideration.
3) In closed questions, the number of items per question should not be exaggerated. Short-term memory is normally limited to no more than five separate items at a time. If a questionnaire designer goes beyond this limit there is the risk that certain items will be overlooked or that the results will be biased depending on the order of items proposed.

B) Using “natural” mappings

The respondent and the interviewer should navigate through the questionnaire with natural ease. It should be self-evident from the visual design which task is required. For instance, the visual design should clearly show the order in which the tasks are to be done and where responses are required. Respondents and interviewers should see at a glance what they are supposed to do. The best instructions in the questionnaire are those which do not need to be written down because it is self-evident from other design features what the respondent and interviewer are supposed to do. With verbal instructions there is always the risk that they will be skipped or misunderstood.

A number of simple rules can help to improve the construction of the questionnaire. In the case of self-administered questionnaires, these rules mostly concern the layout of the questionnaire or the screen design. In the case of interviewer-administered questionnaires, the visual design must take into account the requirements of the interviewers (except for design elements directly addressed to the respondents, like graphical or interactive elements in CAPI surveys). The following rules are also important for providing interviewers with appropriate training (e.g. specific telephone and communication skills for CATI and CAPI surveys, respectively).

1) The questions should start where it is expected by the respondent and the interviewer, i.e. in the upper left corner of the page or the screen. Placing information which is not directed at the respondent in the upper left corner will almost necessarily create confusion and lead to item nonresponse. If necessary, additional graphical features should be used to make clear where the respondent and interviewer are supposed to start their task.
2) Elements that belong together in the questionnaire (like the question, the respective answer space and response options, as well as the instructions on that question) should be grouped together. For the same reason, it should be made clear to the respondent and the interviewer which questions form a section in the questionnaire and where a new section begins. This can be done for instance by separating the individual questions or sections graphically.
3) The visual design should also always clearly show where the respondents and interviewers are expected to enter their responses. A technique often recommended is to use the principle of figure-ground contrast, i.e. to keep all answer spaces and fields that are supposed to be filled out white, with a lightly coloured background providing a contrast to all other spaces. Cognitively, the white boxes will be perceived as figures against the coloured background and automatically attract the attention of the respondent and the interviewer. Additionally, on each new page the respondent will see at a glance which parts have to be filled in when completing the questionnaire. For the same reason, information directed at the statistical institute should always be printed in a way less visible to the respondent. Furthermore, it is recommended to organise questions and response alternatives in a column so that the fields the respondent is supposed to fill in all appear on one side of the page. This can further facilitate navigation in the questionnaire.
4) In case of paper questionnaires, strong visual symbols (like arrows, bold printing, larger font) should be used in order to make sure that skip instructions do not get overlooked. In addition, one should place skip instructions near the response option to which they apply. When reading a text attentively, the human eye captures no more that eight characters at a time (“foveal view”). With a skip instruction placed outside this very limited space there is the risk that it will not guide the respondent as intended. Symbols should always be used consistently throughout the questionnaire.

C) Standardising question patterns

The use of natural mappings unfortunately is not a cure-all. In some cases additional instructions are necessary because some things simply do not have a “natural” order. If the questionnaire can not make reference to a pre-established conceptual model (a so-called mapping) in the mind of the respondent or interviewer, one should use a similar visual design for similar questions throughout the questionnaire. For example, in the case that matrices can not be avoided in a questionnaire, they should (apart from being kept as simple as possible) at least be designed in a consistent and regular form across the questionnaire. Similarly, the use of symbols and graphical arrangements should be standardised in the entire questionnaire. Some experts recommend using *italics* for instructions, CAPITAL letters for response alternatives not to be read out in CATI interviews or reverse print for navigational information like section headings or question numbers (SCB, 2004). The use of reverse print is supposed to provide respondents or interviewers with orientation as soon as they start looking at an entire new page or screen without yet concentrating on any particular item (a process often referred to as “preattentive processing”; see Dillman, 2000).

3.3.4. The reflective level of visual design

The cognitive aspects of questionnaire design in many ways depend on the cultural background of the respondents. For example, different cultures attach different meanings to symbols like arrows, crosses, etc. At the same time, it is important to note that many respondents make a conscious and reflected decision to participate in a survey. Thus they do not necessarily do what the survey designer expects them to do, i.e. to answer the questions accurately and truthfully, but may provide erroneous answers on purpose. Similarly, for individual questions, respondents often give a specific incorrect answer on purpose (e.g. for reasons of social desirability). Such effects must be examined carefully in the conceptual stage as well as during the test of the questionnaire.

Regarding the visual design elements, the most important issue may be the decision to participate in a survey or not. Respondents might decide not to participate if the visual design gives the impression that the data are not safe, that the data collection is carried out for purely commercial reasons, or that the results produced are not relevant to them or society as a whole. Thus, a questionnaire does not only have to be attractive, look easy to complete and be functional, but at the same time must transmit an intended image of the NSI. Research has shown, for instance, that questionnaires of official statistics should make a serious, “official”, and trustworthy impression in order to achieve higher response rates (Dillman, 2005; Budowski and Scherpenzeel, 2005). In this context, the information provided together with the questionnaire (like the advance letter or information leaflets) has to be considered as well. Such documents should be tested together with the questionnaire, because they are of vital importance for the attitude of the respondents towards the survey. Survey research has shown that apparently minor details (like the use of real stamps on the envelope of the advance letter, a personal signature on the advance letter, and a personalisation of the correspondence) may lead to significant differences in response rates.

3.3.5. Experiences

In the field of official statistics, non-verbal design elements have been extensively studied by a number of North American statistical agencies (e.g. U.S. Census Bureau, U.S. Bureau of Labour Statistics, U.S. National Agricultural Statistics Service). Dillman and colleagues also carried out a number of experiments on how to optimise questionnaire design at the emotional and functional levels. The focus of these studies is on self-administered questionnaires, both as PAPI or CASI instruments (Dillman, 2000; Dillman *et al.*, 2005).

Some NSIs have developed standards for questionnaire design. For example, the Federal Statistical Office (FSO) Germany uses a set of modules for questionnaire construction in order to make sure that every questionnaire used in surveys of German official statistics fulfils the basic requirements for questionnaire design and at the same time is clearly recognised as a product of official statistics. In the development of the modules, communications design specialists were consulted, thus guaranteeing coherence between the questionnaire design and the general corporate design of German official statistics (Schwamb and Wein, 2004).

The standardisation of questionnaires’ visual design in NSIs is better achieved with the support of a centralised unit, where the specific expertise can be available. In the survey conducted for the purpose of this handbook, it was observed that, within the ESS, such a centralised unit or committee currently exists in seven NSIs, whereas no information on the degree of standardisation of the appearance of the questionnaire was collected.

3.3.6. Recommendations

- Survey designers should be aware of cultural, cognitive, and emotional aspects of questionnaire design. Ideally, a behavioural scientist should be a member of the questionnaire development team and a communications design specialist should be consulted whenever significant decisions regarding visual design elements are taken.
- Questionnaires should be tested for usability and their emotional level responses.
- The visual design should be developed involving professional expertise. Being this activity quite resource consuming, the standardisation may be a solution in this context. If developed and tested carefully, standardised questionnaire modules could be used.
- In order to make the relevant expertise available inside the statistical offices, at least a centralised unit for questionnaire design and testing should be set up, which reviews all questionnaires before they are used in the field.

3.3.7. Checklist

Design for a good "look and feel"

- Take into account that the design of a questionnaire leads to immediate emotional responses on the part of respondents and interviewers.
- Review the standard used for questionnaire design with the help of behavioural scientists and communications design specialists.

Design for cognitive functionality

- Reducing the complexity of the task required of the respondent:
  - Ask one question at a time.
  - Reduce the number of items or response options in such a way that every respondent is able to process them.
  - Use as few matrices as possible. If using matrices, reduce complexity, build them consistently and regularly.
- Using natural mappings:
  - Start the questions in the upper left quadrant, where interviewers and respondent expect them.
  - Make information that is used exclusively for internal purposes of the statistical office (like coding instructions) less visible.
  - Help the respondent and interviewer to identify thematic sections by separating them visually.
- Standardising question patterns
  - Establish consistency in the use of symbols and graphical arrangements across the questionnaire.
  - Highlight the answer space by providing a figure/ground composition.
  - Use font size, brightness and colour to attract attention, if needed.
  - Place instructions where the information is needed. If the respondent is supposed to enter a code, place it as near as possible to the answer space.
  - Provide strong visual guides for changes in the pattern of questions and skip instructions.

Take into account that the respondent will fill in the questionnaire on the basis of a reflective decision:

- Transmit the image of the statistical office via the questionnaire and make it look reputable to the respondent.

- Use non-verbal design elements in the questionnaire and additional information material to convince the respondent that his/her participation is necessary and rewarding.

Chapter 4. Cross-National Comparability of Questions

"Direct measurement is based on definitions by fiat. … Direct measurement requires that the language of measurement be common to all observations, reflect relationships among the phenomena observed, and be consistently applied." (Przeworski and Teune, 1970, pp. 96-97). These are the problems of cross-national research to this day. With direct measurement we even have problems in national survey research, because national societies are different in class affiliations, in education and in cultural history. The common answer to the problem: the researcher often ignores problems of question comprehension.

"Cross-system comparisons of single variables will be dependent upon the units and the scale of measurement within each social system." (Przeworski and Teune, 1970, p. 42). Therefore, also historically, the *first step* towards comparability in cross-national survey or statistical research is to overcome language barriers. Researchers learned about functional equivalence, which points out the importance of transferring meaning over literal translation and highlights that a process of repeated translations enhances the face validity in intercultural use of measurements. Face validity is achieved when a test appears valid to examinees who take it, personnel who administer it, and other untrained observers (Duquesne University, 2005). Face validity requires that the measure appear relevant to the construct to an innocent bystander, or more specifically, to those that should be measured in the population of interest (Rymarchyk, 2005).

In the 1970s, cross-national projects began to establish a process of forward and blind backward translation to increase face validity. In view of functional equivalence, Przeworski and Teune (1970) demanded that comparative indicators as well as national indicators should be applied in intercultural research. Today, the international data collecting projects are aware of methodological problems and routinely use different established translation techniques.

The *second step* to comparability in cross-national survey or statistical research is to harmonise demographic and socio-economic variables. Demographic and socio-economic variables describe the context in which a person is acting. Context variables or background variables are variables that "contain information necessary to define homogeneous subgroups, to establish causal relations between attitudes and societal facts, and to define differences between scores on scales." (Braun and Mohler, 2002, p. 112). In cross-national comparable research standardised instruments or indices exist only for a very small group of variables as "occupation", "education", and "status". For the variables "income", "family", and "ethnicity", there is preparatory work in progress. Aside from these instruments there are rules for developing further measurement instruments for measuring socio-demographic variables in cross-national research (Hoffmeyer-Zlotnik and Wolf, 2003a).

The *third step* of comparability in cross-national survey or statistical research is scaling of attitude measurement. Scaling is very problematic, even in national research, because nobody really knows the effects of changing scale endpoints and of omitting scale midpoints. The next problem is that of the distance between two verbalised scaling points. Smith (2004, p. 437) is discussing the calibration of response scales. Research on changing scale endpoints is carried out by Krebs and Langfeldt (2005) and on middle alternatives of rating scales by Krebs (2001) for the case of Germany and by O'Muircheartaigh *et al.* (2000) for the case of the United States. However, similar results are reported so that we can hope that the findings of these research activities are also usable in cross-national research. Distances between verbalised scaling points were analysed in national research decades ago, and as a two countries comparison (Germany and U.S.) by Mohler *et al.* (1998), pointing out that verbalised scale points and their distances are culture dependent. But the project, being more a starting point than a completion of this research, was a singular one.

4.1. Conceptual framework

Before describing techniques of translation and harmonisation and presenting instruments for measuring socio-demographic or socio-economic variables a short introduction is necessary to explain key concepts for translation and harmonisation.

Functional equivalence refers to the role or function that behaviour plays in different cultures. One can not assume that behaviours play the same role or function across cultures; therefore, assumptions about the function of behaviour in a cultural group must be verified. Conceptual equivalence refers to the similarity in meaning attached to behaviour or concepts. Certain behaviours and concepts may have different meanings across cultures (see: Chang, 2005, p. 4) [3].

In the framework of systemic functional grammar Matthiessen (1999, p. 27) discusses translation equivalence in the environments of translation, and identifies the environments relevant to translation in different dimensions of contextualisation. He says, "the wider the context, the more information is available to guide the translation" and "the wider the environment, the more congruent languages are likely to be; the narrower the environment, the more incongruent languages are likely to be". The concept of equivalence has been one of the key words in translation studies. In earlier work on translation equivalence, Catford (1965, p. 20) defines translation as "the replacement of textual material in one language by equivalent textual material in another language ".

"Translation is a transfer process which aims at the transformation of a written source language text into an optimally equivalent target language text, and which requires the syntactic, the semantic and the pragmatic understanding and analytical processing of the source language text" (Wilss, 1982, p. 3).

Scales become a very special problem when deciding in which cultural context they will be applied, because enough information is available to know that they differ in their "effect" on the answer. Different comparative studies evidence different levels of rationality for different cultures, where it is perfectly reasonable to suspect that different rationales require different "functional equivalence" in the scales (Worcester *et al*., 2000, pp. 9-10). But we know too little about scaling effects to correct these problems.

"Sociodemographics" often is a black box because in cross-national research comparative demographic and socio-economic indicators, with the exception of sex, gender, and questions with national categories or a crude inquiry (with the categories: "low", "middle", "high") for education, normally are non-existing. However, there are techniques for harmonisation. The goal of data harmonisation is to create data that measure the same conceptual variable and that are measured in the same units or categories. Data can be harmonised prior or after the data gathering stage. International data collection programs use different techniques of harmonisation but all share a high-level methodological consciousness. Generally, one distinguishes two main approaches towards harmonisation: input harmonisation (measurements based on harmonised survey questions and methods) and output harmonisation (measurements based on harmonised target variables that allow national variation on the level of the questions). Output harmonisation can be done as ex-ante output harmonisation or as ex-post output harmonisation.

[3] For example, in the field of mental illness there are a lot of idioms which are largely specific to one target population and quite inappropriate for other cultural groups. No literal translation from one language to another is possible for 'feeling blue' or 'butterflies in the stomach'. Such idioms require alternative conceptually equivalent terms in another language.

Input harmonisation takes internationally agreed standards (such as definitions, concepts, aggregations, classifications) as a starting point and then uses harmonised survey methods for implementing these standards. “All survey countries use precisely the same survey procedures in an ideal case. Country-specific particularities are only permissible where they are indispensable” (Information Society Technologies, 1999, p. 1). Input harmonisation is always ex-ante harmonisation. For input harmonisation, a project needs a methodology group constructing a set of key indicators for some socio-demographic/socio-economic core variables (as was done by the “European Community Household Panel” and by the “European Social Survey”). If the (final) international categories were defined before data collection is started, then the national measurement instrument can be adjusted to the final categories.

Output harmonisation normally is ex-post harmonisation, retroactively done. “The decisive characteristic of the ex-post strategy [...] is that existing national statistics are subsequently adapted by means of conversation procedure in such a way that comparable statistics can be created” (Information Society Technologies, 1999, p. 2). Output harmonisation starts from a common, internationally agreed definition for a variable representing a common indicator. The goal or the target value to be surveyed is determined. The selection of suitable survey methods is left to the participating researchers and is accomplished by a national measurement instrument using national categories. Here the national researchers should aim at the best operationalisation of the common indicator. But they have to take care of the national concepts and structures behind those socio-demographic variables they want to harmonise. Different cultural concepts or different national structures result in incompatible data. If the measurement procedure is valid for the national as well as for the international concept, then the approach is called ex-ante output harmonisation. “With the ex-ante strategy, the harmonisation process is already a part of planning the survey, which means that comparable structures are created in the survey design from the outset” (Information Society Technologies, 1999, p. 2). This ideal case is rare because normally national indicators are not culture-free. If a transfer from national to international categories is necessary, then documentation is essential because often classification of categories in different categorical schemes depends on interpretation and can not be reconstructed without documentation[4].

A major problem of harmonisation is a lack of accepted general guidelines and of central coordination.

Often the techniques for harmonisation and translation are geared to specific U.S. American research questions. However, it should be noted that the situation in the U.S. is completely different from Europe. In the U.S. context, translation means the transfer from one language to another, from English to Spanish. All respondents live in the same country with the same institutions and the same norms. Only values differ from subculture to subculture. In Europe, translation alone will not do for harmonisation. Even in the European Union there is a mixture of different languages, different institutions, different laws, different norms and different values – largely different from country to country. Only in a small group of countries there are the same problems as in the U.S. – e.g. in Switzerland, in Luxembourg, in Spain or in some regions of Italy.

In the following, first the process of translation is shown, outlining translation strategies and procedures disposed for the European Social Survey and by the U.S. Bureau of the Census. Second, the rules for harmonisation are explained and existing measurement instruments for cross-national comparison of socio-economic variables are listed.

[4] As an example for individual classification, in the definition of the correspondence between the national German categories for the variable education and the international ISCED 1997-system, two different approaches would lead to different results. Eurydice (European Commission, 2002) would not codify any cases in the category “post secondary, non-tertiary”, the German coder would classify to this category those persons who finished education with a degree of “master craftsmen” and those having done obligatory practical courses and internships after having reached the university entrance diploma.

4.2. Translation strategies and procedures

De la Puente *et al.* (2000) performed a search in the World Wide Web and found that no more than a handful of key international statistical agencies such as Statistics Canada, Statistics New Zealand, the Australian Bureau of Statistics, and the Centre for Survey Research and Methodology (ZUMA) in Mannheim, Germany, provide some guidance for translating questionnaires. Later on, WHO (2002), the European Social Survey (2002), and U.S. Bureau of the Census (2004) followed with guidelines for questionnaire translation.

Behling and Law (2000, p. 15) state the following regarding the importance of equivalence in translated questionnaires: “Demonstrating that the translated questionnaire possesses the basic characteristics required of all measurement instruments is not enough. In addition, the researcher must show that it exhibits appropriate levels of semantic and conceptual equivalence relative to the source language measure and that it and the procedures through which it is administered minimise any problems created by lack of normative equivalence.”

The World Mental Health Survey used a complex procedure of translation or a controlled procedure to increase face validity (WHO, 2002):

> Forward translation followed by an expert panel. Then blind back-translation (by different interpreters); discussion about social and cognitive structures, cognitive interviewing, focus groups, pretest.

Finally, functional equivalence can be controlled in the process of data analysis. But the practice is not so complex:

- Normally, cross-national research starts with the agreement that one language (mostly the English language) is the reference language.
- In the next step a drafting group is established to formulate the questions of the questionnaire. Native speakers of the reference language (English) in this drafting group will be not only experts for the language but also experts for the cultural background of the drafting questionnaire. If two source questionnaires are existing – the Eurobarometer has an English and a French questionnaire – then translation, back-translation and pretesting by bicultural experts should be done for a “calibration” of both source questionnaires because both questionnaires have a different cultural background. Normally this procedure is too much time- and money-consuming.
- The third step is the translation process. Normally a bilingual but “unicultural” member of the national project team translates the questionnaire. Problems are often not detected before analysing the data. Cultural differences are ignored if an item is only translated from one language to another without analysing the cultural background of the national question wording or the item in the (English) master copy. On the other hand the term “functional equivalence” is sometimes interpreted in such a way, that the national translator will transfer a question or an item into his mother tongue by changing the sense of the item (Braun, 2003).

The problem, however named, is a problem of lack of knowledge, sometimes of ignorance concerning cultural differences. It is problematic if our translators are bilingual, but not bicultural. Therefore a team should do the translating and discuss the right formulation of a question.

For the European Social Survey the process of translation is documented in great detail (European Social Survey, 2004, pp. 15-16):

Each country translates its own version(s) of the source questionnaire. The procedure regarding translation is as follows:

- National co-ordinators are required to find suitable individuals to fulfil the three key roles in the approach: translators, reviewer, and adjudicator.
- Translators have to translate from reference language (English) into their mother tongue.
- Two translators are required per translated draft questionnaire.
- Countries should budget for two independent draft translations, although countries sharing languages can use a "split" translation approach, which reduces costs.

The expertise involved:

- Translators should be skilled practitioners who have received training on translating questionnaires.
- Reviewers need to have translation skills at least as good as those of the translators, but should also be familiar with questionnaire design principles, as well as the study design and topic. One reviewing person with linguistic expertise, experience in translating, and survey knowledge is sufficient. If a person with these skills can not be found, two can cover the different aspects.
- The adjudicator is responsible for the final decisions about which translation options to adopt, preferably in co-operation with reviewer and translators, but at least after discussion with a reviewer. Adjudicators must a) understand the research subject, b) know about the survey design, and c) be proficient in the languages involved.

The U.S. Census Bureau is also working with translators, translation reviewers and translation adjudicators. The translation guidelines of the U.S. Census Bureau (2004; see also: Pan and de la Puente, 2005) document five steps:

1) Prepare: The up-front preparation for the conduct of the translation is specified;
2) Translate: The process of translation is described. There is only a translation in one direction without back-translation;
3) Pretest: The translation is followed by a pretest as an integral and necessary part of the translation process. "Translated questionnaires should be tested as thoroughly as questionnaires designed for one context, and most of the techniques used for testing monolingual questionnaires are equally relevant for testing translated questionnaires. Incidentally, assessment should include everything translated for a study, including hidden CAPI instructions to interviewers and any support materials, such as show cards, diagrams, etc. Attention should also be paid to any culturally anchored visual components" (Harkness, 2003, p. 41);
4) Revise: "translation team members reconvene after results from the pretest are available to discuss revision to both the source language and target language document based on pretest results" (U.S. Census Bureau, 2004);
5) Document: Documentation begins with step 1 and the written specifications and should be carried through all other steps.

Methodological consciousness is high. There are well-developed instruments, specifications in deep detail, and help from specialists is available. It is necessary that the whole team have enough cultural consciousness. Attitudes and behaviour depend on a cultural background. In attitude measurement, the cultural background should be taken into account by the adjudicator. This person has to be integrated in two cultures, that of his or her own country and that of the country represented by the master copy. The specific accentuation of the pretest as an integral and necessary part of the translation process is as important as bicultural specialists for the revision of the translation. Only by pretesting the researcher can control what stimulus will reach the respondent.

Often "with one measure it is impossible to know if observed differences (or nondifferences) are societal or merely linguistic. With two measures a consistent pattern on both items establishes a clear finding" (Smith, 2004, p. 434). Therefore sometimes in attitude measurement but often in the measurement of socio-demographic variables a second measure is necessary.

In transferring socio-demographic variables from the drafting group's version into the specific country's version, the problems are more evident than in attitude measurement.

4.3. Harmonisation of demographic and socio-economic variables – accepted instruments

Only a small number of tested and accepted measurement instruments exists. The most established instruments are developed for measuring "occupation":

- The International Labour Organisation (ILO), a specialised agency of the United Nations, first started in 1958 with an "International Standard Classification of Occupations, ISCO". The actual instrument is ISCO-88 (ILO, 1990). The revision is announced for 2008 (ILO, 2005). The "International Standard Classification of Occupations, ISCO" was developed for comparative UN statistics.
- ISCO-88 COM (Elias and Birch, 1994), the European Union variant of ISCO-88 has small modifications.
- Sociologists very soon started to use the ISCO classification scheme as a starting point for developing a prestige-score: "Standard International Occupational Prestige, SIOPS" (Treiman, 1977; Ganzeboom and Treiman, 1996), a socio-economic index of occupational status: "International Socio-Economic Index of Occupational Status, ISEI" (Ganzeboom et al., 1992; Ganzeboom and Treiman, 1996), and nominal class categories: "EGP Class Categories" (Erikson et al., 1979; Ganzeboom et al., 1992). These three indices, real sociological instruments for comparative research, actually based on ISCO-88, are documented for comparable research by Ganzeboom and Treiman (2003).

Other instruments by the International Labour Organisation (ILO) can also be found for official statistics and survey research, such as the "International Classification of Status in Employment, ICSE-93" (Hoffmann, 2003), and amongst other guidelines, the "guideline concerning the implications of employment promotion schemes on the measurement of employment and unemployment" (ILO, 1987) and the "guideline concerning treatment in employment and unemployment statistics of persons on extended absences from work" (ILO, 1998).

The Institute for Statistics (UIS) of the United Nations Educational, Scientific and Cultural Organization (UNESCO) developed an instrument for comparing education by school leaving certificates (general and vocational): the "International Standard Classification of Education" from 1997, ISCED-1997 (UNESCO, 1997), which contains a minimal consensus on the definition of education.

Therefore, for these variables, one can find a lot of competitive instruments for international comparative research based on combinations of different variables:

a) an index of general and vocational education: the CASMIN (Comparative Analysis of Social Mobility in Industrial Nations) Educational Classification (Brauns *et al.*, 2003),
b) an index of general and vocational education combined with a mean of occupational prestige one can reach by specific education: the "Hoffmeyer-Zlotnik Educational Index" (Hoffmeyer-Zlotnik, 2003) and the "Hoffmeyer-Zlotnik/Warner Matrix of Education" (Hoffmeyer-Zlotnik and Warner, 2005),
c) school leaving certificates combined with earned income, or
d) years of schooling, asking for grades, asking for years in the educational system, or asking for life time learning.

RAMON (2005), Eurostat's classifications server, concentrates on international classifications, restricted to statistical classifications. Altogether 60 different classifications (current and earlier forms of classifications, not only those developed by Eurostat or other EU organisations) are documented. The ultimate and very ambitious goal of the RAMON project is to present all available information on statistical international classifications. This means that RAMON is making available all necessary material, for each classification identified in the database, to know everything about that particular classification; this covers information about:

- the methodological principles applied when building the classification;
- the structure of the classification;
- its explanatory notes, if any;
- its links with other international classifications; and
- other relevant information (for instance: case law).

The two best known classifications in the field of economics are:

- NACE, "Nomenclature statistique des Activités économiques dans la Communité Européenne", that is the classification of economic activities in the European Community. The 2002 version of NACE (RAMON, 2005, classification: p. 3) has 514 classes on the 4th level and is in agreement (in principle) at two-digit level with the "International Standard Industrial Classification of all economic activities, 3rd revision, ISIC Rev.3.1", constructed by United Nations Statistics Division. Next revision of NACE is announced for 2007.
- CPA, the "statistical Classification of Products by Activity in the European Economic Community", is the European version of the CPC, "Central Product Classification" by United Nations Statistics Division. CPA provides a common EU framework for the comparison of statistical data on products, i.e. goods and services, and is more detailed in order to meet the specific needs of the EU. The 2002 version (RAMON, 2005, classification: p. 1) has 2608 sub-categories on the 6th level; next revision of CPA is announced for 2007.

Preparatory work for the measurement of income is widely done by an UN specialised group: the "Canberra Group. Expert Group on Household Income Statistics" (2001). The "Canberra Group" was formed with the aim of improving national statistics on household income distribution and inequality and with a desire to improve the quality of international comparisons in this area. A first meeting of experts to further the development of statistics on household income in particular and, more generally, on household economic well-being was organised in 1996 by the Australian Bureau of Statistics. "The primary objective of the Canberra Group was to enhance national household income statistics by developing standards on conceptual and practical issues related to the production of income distribution statistics. Its work was in support of a revision of international guidelines ... " (Canberra Group, 2001, p. xi). Their income concept "seeks to establish conceptual groundrules for defining and measuring household income", the way from concept to practice "provides an overview of the practical considerations which will determine the parameters for the production of a set of income distribution statistics" (Canberra Group, 2001, pp. xiii-xiv).

For all other socio-demographic variables there are no commonly accepted instruments available. There are initiatives by the European Society for Opinion and Marketing Research (ESOMAR) for developing a Standard Demographic Classification (1997), by Eurostat, with rules for harmonising socio-economic variables in EU statistics (Mejer, 2003), and in social survey research some big international comparative survey programs established their own methodology groups for controlling the quality of functional equivalence and harmonisation (e.g. "European Social Survey" and "International Social Survey Programme"). All other measurement instruments for cross-national comparison should be developed by the researcher her- or himself using harmonisation.

4.4. Rules for harmonisation of demographic and socio-economic variables

For harmonising socio-demographic or socio-economic variables the researcher should follow different steps from “concept” to “measuring instrument”.

The *first step* is the “concept”. The term “concept” is used here in two different ways. On the one hand the concept is the idea of what the researchers in a comparative survey want to measure. The term “concept” is used here to denote ideas concerning a specific domain of interest. In this sense the researchers need one common definition of their variable: what should be measured? what should be explained by the measured variable? On the other hand “concepts” are the result of a process of cultural development that took place within the historical context of a country. Since these concepts are formed by history and cultural experience of a people, they are always culture or nation specific. Often they are so specific that it is not possible to compare different concepts across nations.

For international comparison it follows that the researchers:

a) need an internationally accepted common definition of the background variable they want to measure;

b) have to clarify the national or cultural background of this variable.
The following influences can be named for the six main socio-demographic variables:
   - “household” and “religion” are influenced by culture with a historical regional background;
   - “education” and “race and ethnicity” are influenced by national political concepts and national experiences partly in history;
   - “occupation” and “income” are influenced by actual national laws and by the reaction of the parties involved in politics, economy, and finance;

c) have to find out about the national conceptualisations behind the variable.

Every society has developed a value system which is transformed into national legislation by the state. National structures arise from concepts comprising (historical and current) political ideas, cultural values and national norms that are often unique to a given society.

In a *second step* these different national structures that are embedded in national concepts have to be elaborated. The measurement of background variables has to be based on profound knowledge of the national concepts and the related national structures that are relevant for the variables of interest.

The *third step* is to find or to construct a measurement instrument – consisting of questions and response-categories in a questionnaire – which is valid in national surveys and useful in international comparisons. In national surveys, measurement instruments are constructed on the basis of concepts and structures. Therefore, for cross-national research, measurement instruments are useful if they are based on related concepts and structures.

The *fourth step* normally is the harmonisation of the measurement instrument if output harmonisation is chosen. In the case of input harmonisation, this step precedes the choice of the measurement instrument.

The end product of output harmonisation is a common scale or a common system of categories recoding the national measurement instruments.

4.5. Experiences

The European Community Household Panel (ECHP) is one of the best datasets for methodological research in cross-national comparison. Conceptually, the ECHP is an input harmonised instrument. For later waves of the panel, however, three nations used output harmonisation by harmonising the data from national surveys instead of collecting data with the ECHP questionnaire. By content analysis of the questionnaire we can learn about the problems with input harmonisation and by comparing input and output harmonisation for the early waves (e.g. of the German data) we can learn about the problems with output harmonisation. The end of ECHP after 8 waves shows a dilemma of input harmonisation: the complexity is very high and the product is inevitably based on compromises.

The following instrument, the "European Community Statistics on Income and Living Conditions, EU-SILC" (European Commission, 2003) is no longer an input harmonised instrument. Rather, only target primary variables are defined, and each country is constructing its own questionnaire. In this instrument, there is a lot of cultural bias of the participating countries. A task force consisting of methodological specialists from different countries with different cultural background seems necessary (like the controlling of the European Social Survey as done by intercultural groups of methodological specialists on different topics, such as translation, sampling, interviewing, and so on).

The "Eurobarometer" is a negative example with respect to the translation process, because two source questionnaires with two different cultural backgrounds are existing. Both source questionnaires are corresponding with each other. But it is not tested by whether both source questionnaires are measuring the same.

4.6. Recommendations

- *Translation*: translation of questions about attitudes and behaviour should be integrated in a process of forward-translation, pretesting, revision, and a final decision by bi-lingual and bi-cultural experts in coordination with the whole team of translators and researchers. The translation-team should act as agent between the culture underlying the master copy and that underlying the target respondents of their translation. In addition, it is very important to pretest the translation with cognitive and quantitative methods.
- *Harmonisation*: for harmonisation of socio-demographic variables there is only a small number of real harmonised measurement instruments. The most important are:
  - a Standard Classification of Occupations by the International Labour Office which can be combined with indices measuring prestige, or socio-economic status, or class affiliations;
  - a Classification of Status in Employment by the International Labour Office;
  - a classification of industries by the European Union;
  - statistical Classification of Products by the European Union;
  - a Standard Classification of Educations by UNESCO;
  - the "Hoffmeyer-Zlotnik/Warner Matrix of Education" by sociologists.

  Beside these instruments, there are conceptual groundrules for defining and measuring some variables, particularly "household income".

4.7. Checklist

Translation

- Organise forward-translation from blueprint to target language by (at least two) different translators.

- Test the translation by cognitive methods if the cultural background between source questionnaire and target culture is too different.
- Analyse by means of statistical procedures the pretest-data in order to ascertain whether you can find misclassification of items or variables. Pre-requisite: comparative data from another country.
- Organise the revision of translation by bi-lingual and bi-cultural reviewer.
- Last decisions have to be done in a process of discussion by the whole team (or after discussion with the team by a special expert). Important are also persons who are experts of the topic of the research and the research process.
- Documentation is necessary during the whole process of translation.

Harmonisation of socio-demographic and socio-economic variables (Hoffmeyer-Zlotnik and Wolf, 2003b, p. 405)

- Search for a common definition of what should be measured.
- Guarantee that this common definition works in each country which is involved.
- Identify the similarities of the national concepts and structures underlying the variables.
- Locate a valid indicator or a set of indicators (depending on the variable of interest as well as on the national specifics).
- Decide if the variable of interest should be measured by the same measurement instrument in every country or culture (input harmonisation), or if the variable of interest should be measured by nation or culture specific instruments, to yield data which are harmonised after data collection (output harmonisation).
- Test whether the chosen instrument reflects the empirical structures found in the different countries or cultures and if it is logically related to the common definition.
- Take care that your measurement instrument is understood and can be answered sensibly by all respondents, i.e. people who live in different national and cultural contexts.

Chapter 5. CATI and CAPI Electronic questionnaire design and implementation

This chapter deals with the issues related to the design and implementation of the Electronic Questionnaire (EQ), with particular focus on CATI and CAPI surveys. Some of the conclusions that will be drawn apply also to the CASI mode, and these will be referred to in the text. However there are many additional considerations that are necessary for an adequate treatment of this topic and that will only be mentioned here. Therefore the development of recommended practices for CASI surveys is deferred to an eventual next project, when in our opinion more technical and methodological knowledge is available

The use of computer to support telephone surveys started in the early 1970s originating CATI surveys. Later on, with the advent of the laptop computers, it has been explored its use to support field surveys involving face-to-face interviews and that produced the acronym CAPI (the first large scale application of CAPI has been developed in 1987 for the Labour Force Survey in The Netherlands) (Weeks, 1992). The rapidly expanding of the computer technology, especially for CATI, had a great impact on the basic survey tasks of organisations using these interviewing techniques. One of these tasks was the questionnaire design: the questionnaire designer had different responsibilities than before that concern not only the survey’s objectives but also the new way of administering questions using a new and powerful tool (House, 1985).

This powerfulness allows for an easier realisation of surveys based on very complex questionnaires. The management of the different branching, of the skipping between questions and the customisation of texts’ wording (fills) based on prior information are all set up by the software thus simplifying the interviewer’s job. The powerfulness of the EQ relays also on the management of errors: range errors or inconsistencies among questions can be solved during the interview allowing higher quality of the final data and, therefore, speeding up the editing and imputation phase which is performed after the data collection.

Due to this complexity, in terms of both survey contents and technical requirements, the design of an EQ requires, first of all, detailed and accurate survey specifications. This mean**s** that the paper questionnaire must contain not only the questions and the answer categories but also the rules for administering questions (skips and branching), the rules for customising the texts of questions and errors’ messages, the texts for the instructions for the interviewers, the set of inconsistencies to be detected and the way they have to be solved, the rules for the several attempts to contact the sample units and the conditions to make the substitution among them. In addition, the nature of the EQ makes that the best results are achieved either from the cooperation between the survey manager and the expert of the CAI (Computer Assisted Interviewing) questionnaire design or from a designer skilled on both disciplines (House, 1985).

Powerfulness of the instrument, anyway, must not be misunderstood with the possibility of increasing the questionnaire complexity as much as one could wish. To exceed in complexity means: to increase the risk of making errors, to make the interview difficult to manage for the interviewer and burdensome for the respondent, to make the implementation process, as well as the testing phase, heavy to cope with. For example, it is not necessary to activate, in the data collection phase, the entire set of the inconsistencies rules that will be implemented later in the editing and imputation phase; the EQ has to prevent from the major and most frequent inconsistencies and from those regarding the most important survey variables: to exceed in controls will not only be time consuming but will also tire interviewers and respondents since, for any detected inconsistency, an error message is displayed and the confirmation of keyed data is asked.

In designing an EQ it is extremely important and useful to adopt standards covering various aspects of the implementation, that is: standards of texts’ wording (questions, instructions, errors messages), standards for the management of the different types of errors and, finally, standards for the screen layout.

5.1. Aims

The EQ is the Information Technology (IT) translation of the paper questionnaire which has to be already planned taking the interviewing technique into consideration: what makes a "paper questionnaire" a "CATI/CAPI questionnaire" is not its appearance on a PC video but its design that has to be thought with a continuous focus on the chosen mode.

In general, the tools and principles for the EQ development for CATI and CAPI surveys are the same. When different tools are in use, this will be underlined.

The main objectives to take into consideration when designing an EQ are the following ones:

- the collection of information: the EQ must be able to collect information according to the survey informative needs;
- to ease the interviewer's job: the design must take into account the setting of the screen layout, the management of the various events of errors and the customisation of texts' wording. The screen layout setting has to make the interviewer immediately able to understand what has to be read and where to find it; the error management has to make the interviewer quickly understand what kind of error happened and which questions where involved in it; the customisation of texts has to help the operator in reminding the information previously collected;
- the agreeability of the interview: customisation of questions wording must be taken into account to make the interview more pleasant for the respondent;
- to detect and reconcile inconsistencies: an EQ must be designed in such a way to solve the greatest number of inconsistencies (major or more frequent inconsistencies, etc), paying attention, at the same time, to the fluency of the interview;
- to limit the segmentation effect : the segmentation effect is typical of EQs and consists in the display of one question per screen thus restricting the view of the questionnaire; the design of an EQ must limit this effect not only considering how many questions to display per screen, but also planning how to manage the backing up of the interviewer with the help of messages on the screen;
- to ease the navigation of the questionnaire: the EQ should allow the possibility of modifying some responses previously given, without missing any relevant information (i.e. to allow the change of route when backing up) and, if the questionnaire contains *ad hoc* modules and some of them are "standalone", it should allow the possibility of completing them in any order;
- to avoid sample distortion: for CATI surveys, the call scheduler must be planned in such a way to give the sample units the same probability to be contacted, that is without altering their inclusion probability.

From a technical point of view, the main objectives of the design of the EQ can be borrowed from the IT science (Saris, 1991). An EQ must be:

- effective, measure what is intended to measure;
- flexible, i.e. easy to modify when changes occur and easy to correct in case of errors;
- structured in modules, in order to be able to use these modules for other studies or survey's waves. (A module represents a set of instructions that can be implemented and tested independently from the other parts of the questionnaire);
- portable, i.e. easy adaptable to different hardware or software platforms;
- efficient, in terms of response time for screen replacement.

5.2. Transition from paper to CAI questionnaire

As previously said, an EQ is not a mere translation of a paper questionnaire in a set of code lines since the computer technology affects the design of the data collection tool. Anyway, in some surveys it could happen a change of the interviewing technique from paper-and-pencil to a CAI mode. In this case the questionnaire design could be incorporated in the EQ implementation phase. An analogue circumstance can apply to mixed mode surveys where the same questionnaire has to be administered with different techniques involving paper and electronic questionnaires.

In all these examples the transposition of a paper questionnaire into an electronic one, can be considered as a translation. The result of this translation has to be an EQ whose aims and design characteristics (flow, structure, questions) are those described previously in this handbooks (above paragraph and chapter 3). Anyway, it is important to underline that the translation must consider also those changes relative to the presence/absence of the interviewer. The main elements to be considered when transforming a paper questionnaire into an EQ can be summarised in the following list:

- the question wording has to be more colloquial; example: in a self administered paper questionnaire the question “Respondent’s marital status” will become that for a CAPI questionnaire: “Which is your marital status? (*Read out*)”;
- the questions’ texts should be shortened especially for a CATI surveys;
- the answer categories should be modified to always allow for an answer, in other words empty values should not be accepted;
- for questions with long list of answer categories that can not be shortened for comparability reasons, the item rotation is extremely useful (especially for CATI surveys);
- the instructions relative to skips and branching will be implemented in the software, while those relative to the interviewer\respondent will be reported on the video screen, without crowding it to much, using, when necessary, the help system;
- control rules that were performed at the end of the data entry phase for the paper questionnaire, can be implemented in the EQ, but making them manageable during the interview through the use of dialog windows containing error messages;
- longitudinal controls using administrative archives or data from previous waves of the survey can be carried out during the data collection phase and, therefore, the relative control rules can be implemented in the EQ.
- the graphical layout has to be adapted to the video screen.

5.3. Electronic questionnaire structure

In developing the EQ it is important to take into consideration some important factors in order to achieve the main aims described in the previous paragraph.

First of all the EQ should be designed to collect all the information requested by the survey questionnaire in all eventual situations: therefore all possible branching paths should be considered in the design, also those relative to rare but possible situations.

Concerning the management of errors two aspects have to be always considered: a) the typology of controls and b) the content of the error messages that are shown to the interviewer.

a) An EQ can manage two types of errors: range errors and inconsistencies errors. To detect them it is possible to use two kinds of controls: hard controls and soft controls. In case of an hard control, the keyed wrong data must be corrected to let the interview continues, while in case of soft control the keyed data can be modified or left as it is. Which type of controls is to be used should be decided according to the type of error, the variable’s nature and the interviewing technique:
    - hard controls can be used to detect range errors or inconsistencies for objective variables (like the year of birth);
    - soft controls can be used to highlight eventual inconsistencies for attitudinal or knowledge variables (for example: if the answer to the question “How many hours do you usually work per day” is 10 then a soft control, asking for the confirmation of the given answer, should be used);
    - for CASI interviews the majority of controls should be soft (the respondent is alone and can not know how to solve the inconsistencies), whereas for CATI and CAPI interviews the presence of the interviewer makes the management of hard controls easier;

b) The content of the error messages should ease the navigation of the questionnaire: it should always indicate, to the interviewer, how to skip backward to the variables involved in the errors and how to go forward to the last answered question.

Another important aspect in developing the EQ is where to place the instructions for the interviewer: instructions can be put before or after the question text according to the moment the interviewer has to carry them out. For example, the instruction "*Read out*" should be displayed after the question text while the instruction "*Precise that the answer is not compulsory*" have to be placed before. In the first case, the interviewer immediately knows that the next action after reading the question text is reading (aloud) the questions items; in the second case, he soon understands that he has to read this sentence before the question text to make the respondent more comfortable with the question and, therefore, more prone to provide an answer (this instruction is, indeed, particularly useful for sensitive questions).

The last factor to consider in developing the EQ is the customisation of texts (fills). As already mentioned, the use of fills simplifies the interviewer's job and makes the interview agreeable for the respondent. Therefore their use is highly recommended; anyway, it is necessary not to exceed with them since too many rules for fills would make the implementation and the test phase too heavy. Example of fills that can be omitted are: fills for verb tenses ("How many days per week *did/do* you work?", the interviewer has to read the right tense according to the situation) or fills for the feminine/masculine of words (for languages it applies).

5.4. Graphical Layout

The graphical layout of the EQ plays an important role in easing the interviewer's job and preventing data capturing errors (Couper *et al*., 2000).

The screen must be drawn in a way that the interviewer can immediately find what he needs. The most important elements, obviously, are the question texts, the response items (whenever present) and the specifications related to texts not to be read aloud.

Additional information can be shown in the screen, like:

- instructions regarding particular actions the interviewer has to perform;
- helps for the interviewer regarding specific questions, more general specifications or procedural aspects concerning the EQ management;
- error messages following entered values;
- names of sections in order to reduce the so called "segmentation effect".

The graphical layout must be planned so as to constitute a "guide" for the interviewer, who must never feel to be in doubt on how to go on with the interview. This aspect is very important for both CATI and CAPI techniques, even if for different reasons:

- in CATI surveys the rhythm of the interview is very intense (time is a very important resource)
- in CAPI surveys the interviewer is alone and can not rely on anybody to ask for help.

All these features apply also for CASI surveys substituting the role played by the interviewer with that of the respondent; peculiarities of CASI techniques can be drawn from the "Visual design elements" section (Chapter 3) where the basic principles of the graphical layout are described.

Despites many information can be shown on the screen, the first aim to keep in mind, in designing the graphical layout, is that the screen must not be too dense. To this aim, the use of on-line help is strongly recommended: it can contain definitions of key concepts of the questions – e.g. the meaning of part-time job - or instructions about the alternative way of asking a question in case of probing or procedural help on the EQ management. The presence of helps linked to specific questions must be shown in the screen so that the interviewer is informed, while more general or procedural help should always be available for the interviewer through the same actions.

The use of on line helps is particular important for CASI surveys where the respondent is not trained as interviewers are. In this case specific on-line helps have to be used together with interfaces objects (list box, option box, text box, buttons, etc) that will help the respondent in answering and navigating the questionnaire.

To implement an optimal screen layout, other elements have been studied, like:

- the number of questions to be shown in each screen;
- the identification of different parts of the screen to be used for different purposes;
- the length of displayed texts;
- the use of colours;
- the use of graphical solutions (CAPITAL, **bold,** underline, *italics*);
- the use of icons;
- etc.

There is not a common solution in literature for each of the above mentioned elements, even if the shared recommendation is to apply the same set of rules to each element in a consistent way throughout all the screens of the questionnaire.

The design of the graphical layout must be planned in order to reduce the segmentation effect typical of CAI questionnaires: more questions concerning the same topic should be displayed on each screen with the constrain of not exceeding in the number of questions per screen and, when possible, the title of the section should be displayed each time it changes. The recommendation is therefore to establish a good balance between the content of the questions to be asked and the necessity of not to crowd the screen too much.
Examples on how some of these elements can be implemented with the software Blaise con be found in the appendix (Appendix 5.1).

5.5. Software system components

Apart from the administration of the survey questions, an EQ can perform other functions that are implemented in additional modules. These modules are briefly described in the following table.

Module	**Description and Objectives**
Call scheduler	It is implemented only for CATI surveys with the aim of processing all the sample units, taking into account the results of previous telephone contacts and the sample units' characteristics with the strong constraint of not altering the inclusion probability of each unit.
Interviewer agenda	This module is used only for CAPI surveys and it allows the interviewers to set appointments with respondents respecting their needs and their availability. This can positively influence the response rate.
Quota system	When the sample is structured per quota, this module allows not to contact or not to ask some questions when quota are reached.
Substitution system	Once a sample unit can not be contacted anymore because it refuses or is not eligible or because reaches the maximum number of possible calls, etc. this module has to automatically change this sample unit with its respective substitute-unit according to some defined criteria. Despite it is quite useful for the survey management, this module is rarely supported by the various CATI/CAPI software and substitutions are made with the help of external software.
Assisted coding	If a questionnaire contains open-ended questions like Occupation, Nationality, Place of birth, etc., this module can provide the coding of the free text answers during the interview itself (on-line coding or assisted coding). When properly used, it can provide high quality coded data and can reduce the time for the coding process to be performed at the end of the fieldwork by limiting it to those cases where the assisted coding failed.
Interview monitoring	This module is aimed at producing sets of quality indicators to monitor the interviewers' job. The periodical (daily or weekly) analysis of these indicators permits to detect those interviewers with poor quality indicators and who require a further training. Since this training is performed while the interviewing is in progress a high level of data quality is potentially assured.
Output files system	This module transforms the final database, proper of each CATI/CAPI system, into data files of different formats, which can thus be elaborated by any kind of software.

Technically speaking, a good EQ must work accordingly to the underlying CATI/CAPI software system; in order to make an EQ performing what the designer intends to do it is necessary that the designer knows how the system works and which are its potentialities. One of the outcomes of the European project DATAMED

(Data Capturing and Interchange in Mediterranean Countries) (Datamed, 1998) was the evaluation of the main software available on the market in terms of the implementation of different functions like: implementation of the electronic questionnaire, definition and implementation of substitution criteria, automatic or assisted coding, execution of ex-post control procedures on compiled forms, etc. One of the results of this analysis, that compared software like Quantime, Survey Craft, Bellview, MaCati, Blaise, NIPO and VisualQ, is the listing of the main functions or components a CATI/CAPI software should support. These functions/components are described in the following table.

Main Functions of CATI/CAPI software
Creation of survey unit list
Definition and implementation of the electronic questionnaire
Definition and implementation of control procedure (ex-post)
Definition and implementation of substitution criteria
Definition and implementation of reminder criteria
Survey presentation
Transmission of forms/records
Automatic or assisted coding
Returns management and monitoring
Reminders management
Execution of ex-post control procedures on compiled forms
Re-interviews management
Research, consultation and extraction of metadata from non-electronic archives
Extraction of data from archives, on the basis of selection criteria
Management of data bases

These main functions can also be applied for CASI surveys adding to them the “Authentication of the respondent” function, typical of this technique, which can be implemented with different methods each having a proper impact on the EQ design. In general, the authenticity of the respondent as well as the assurance of the confidentiality of the information are guaranteed by a well implemented web environment where the data exchange is made through certified websites and secure protocols (HTTPS) and respondents are provided with their own credentials (user-id and password).

5.6. Functionality testing

The testing of the EQ is aimed at evaluating if the EQ performs according to the given specifications and if the system is robust with respect to unexpected events (Tarnai and Moore, 2004). Such a test represents a first step, concerning the functionality of the system, of the more complete approach to testing that is described throughout this handbook.

The functionality test should be performed during or at the end of the EQ implementation step, therefore before it is released for the data collection on the field.

In the literature, it is recognised that to completely identify the errors is virtually impossible; therefore, the goal should be to design a testing procedure capable to catch as many errors as possible. Due to the nature of the EQ, the testing is also based on criteria used in software evaluation. Tarnai and Moore (2004) reported a taxonomy of main errors in CAI questionnaires and proposed an approach for a structured EQ testing. In detail, for each questionnaire feature, the possible CAI errors and recommended methods for detecting them are presented in table 5.1., adapted from Tarnai and Moore work.

Some of these features can also be detected by means of the pre-testing phase with survey respondents (see chapter 6).

Very shortly, by *question-by-question testing* (Q-by-Q) is meant the analysis of single questions or screens. *Testing by task* refers to the assignment of specific tasks to different testers, so that each of them focuses on given issues. *Scenario testing* consists in hypothesising some real situations, entering them in the questionnaire and checking the performance and the results, whether in the *simulation* the software generates random responses, which results are then evaluated by the testers. Finally *data testing* refers to the analysis of the data output from the CAI program.

Table 5.1. Taxonomy of CAI questionnaire errors and methods for detecting them

Questionnaire Feature	***Possible errors***	***Detection methods***
Screen appearance	Poor visual design, inconsistent formatting	Q-by-Q testing
Preloaded data and sample administration	Incorrect data formats, incorrect data order or appearance	Q-by-Q testing Testing by task
Question wording	Missed words, spelling errors	Q-by-Q testing Testing by task
Response ranges and formats	Formats do not match specifications, missing response options, inappropriate formats	Q-by-Q testing Testing by task Data testing
Missing values	Unavailability or inconsistent use of: refusals, don’t know, not applicable or other response options	Q-by-Q testing Testing by task Data testing
Skip patterns	Not all response options branch correctly, skips to wrong question	Testing by task Data testing Scenario testing Simulation
Calculations and fills	Division by zero, missing values, incorrect formulas, insufficient space reserved for fill variables	Q-by-Q testing Testing by task
Randomisation	Biased processes	Testing by task Data testing Simulation
Function keys and instructions for the interviewer	Not accessible, inaccurately worded	Testing by task
Rosters	Incorrect branching, insufficient calls to a roster	Q-by-Q testing Testing by task Scenario testing
Attempt tracking	Insufficient variables to track call attempts, inappropriate call slots	Testing by task
Screening questions	Inaccuracies in determining eligibility	Q-by-Q testing Scenario testing Data testing
Termination questions	Insufficient termination codes	Q-by-Q testing
System issues	Abnormal terminations, corrupt output files	Scenario testing Testing by task

When the CAI questionnaire concerns a CASI survey, i.e. a web, Disk by Mail (DBM) or an e-mail survey, the list of the features to consider in the previous table should be adjusted with respect to the additional modules that are relative to such modes (e.g. authentication step, speed of page delivery, appearance under different browsers operating systems and screen resolution, pick loads in the hosting platform, etc.). In addition, it has to be considered that some features will be common to more than a questionnaire (e.g. robustness of the hosting platform), whereas others will be related to the specific questionnaire.

The EQ testing phase should involve different people and integrate the methods in order to be able to identify the different types of errors. Obviously, the complete test will be achieved with the successive testing phase involving the respondent and the interviewer.

Based on the software testing literature and on a review of the current practices gathered by means of a survey on CAI testing, Tarnai and Moore (2004) developed a model for testing CAI applications. First of all, a person responsible for the testing phase and in charge of coordinating a team of testers should be identified. The team should be formed by personnel with different background (programmers, subject-matter experts, researchers), since different typologies of errors are more easily identified by different experts. The team defines the EQ testing procedure, which usually starts with the control from the programmers. This includes also the assessment of the loading of the instrument in PCs and that the EQ can be correctly navigated. The team analyses the specifications, the problematic branching paths and questions that should specifically undergo the testing, establishes the methods to use, and then the tasks are assigned. For relatively simple surveys, the analysis could start with the written version of the questionnaire and the programming instructions. The encountered errors are reported in a document or registered in a database. The method usually includes the data testing and/or the simulation of respondent data. The corrections of the errors are made and attention is posed in order not to overwrite changes made by other programmers (version controlling). This EQ testing is an iterative process consisting in testing, making changes and corrections and re-testing, until the version is as much as possible error-free.

In the literature (Levinsohn and Rodriguez, 2001), there is evidence of attempts of automating at least some functions of an EQ testing process. A prototype system, *RoboCAI*, has been developed, that creates test scripts or questionnaire scenarios, which are integrated in a Blaise program to test the EQ. However, such a kind of systems, although very useful for large size and complex surveys, are only aimed at identifying given kinds of errors, while do not detect others.

5.7. Usability testing

As the design and layout of paper questionnaires can affect the data collection process, similarly the interface and the system in Computer Assisted Interviewing (CAI) may introduce additional aspects which should be evaluated. The initial research on CAI has focused on attitude of interviewers towards CAI and data quality, with positive results in terms of increased sense of professionalism and reduction of post-interview editing (Hansen and Couper, 2004). Later on, the awareness of the existence of an interaction among the interviewer, the respondent and the CAI instrument, also affected by the survey setting (including the administration mode) led to the development of models for the Human-Computer Interaction (HCI), which permit the analysis of sequences of actions in the computer assisted interviews. For example, in CATI interviews, a typical sequence is: a) the computer displays the question, b) the interviewer reads the question, c) the respondent provides an answer, d) the interviewer enters the response, presses [*enter*] and moves to the next question, mostly on a new screen. In such a setting, the computer interacts with the interviewer, while it is less relevant for the respondent. The HCI sequences involve cognitive tasks in question interpretation and response formulation. For this reason, the approach used in the usability testing methods has many overlaps with some cognitive methods, for example the behaviour coding (section 6.2.1.).

What reported in this section applies to any kind of CAI questionnaires, however the methods need to be customised whether the data collection mode is self-administered or with interviewer. In addition, it has to be considered that in web surveys, the HCI has an unpredictable element inherent to the respondent environment.

Besides, for this technique, the usability test should include both the usability of the website and of the EQ. The first has to be aimed at testing the functional completeness and the easiness of navigation while the second has to be made paying particular attention to the direct interaction between the respondent and the electronic application.

More specifically, the usability test addresses issues on how efficiently and effectively the users (interviewers and respondents, the last in self-administered questionnaires) make use of the system. In usability testing the focus shifts from system functionality to the evaluation of instruments from the user’s perspective. In a well performed usability test the entire package, including software, hardware, manuals and training are evaluated. Here, issues like operating the CAI program and the computer, appropriate screen layout in relation to the data collection mode, readability of the questions, duration of an interview, clarity of instructions (in the manual, visualised by the computer or given during training) are tested.

The idea that the screen layout may affect the interview is the basis of the previous sections, which provide indications on standards for the screen design for CATI and CAPI surveys and, in general, of section on “Visual design elements” .

In the literature, usability evaluation methods are divided into two main sets: i) usability inspection methods and ii) end-users evaluation methods. In usability inspection methods, also known in the literature as heuristic evaluations, a group of experts evaluate the instrument on the basis of a set of heuristics or evaluation criteria. Examples of these criteria are: navigation easiness, consistency, minimization of user memory load, and so on. The method can be considered as part of the functionality test and is equivalent to the expert review in the evaluation of surveys questionnaire. It is reported that, in general, three to five experts are required for this task and that it can be completed within few days. The end-user evaluation method involves the ultimate users of the systems. This test can be carried out in the laboratory or in the field, on experimental or nonexperimental basis. In the laboratory, the method can take advantage of the equipment usually available in such an environment (video-recording, possibility to observe scan-converted images of the interviewer's screen, …). In the field, the advantage is to test the questions into the survey environment away from the artificial setting of question development.

Through field usability testing other issues can be considered, such as the interviewer's ability to physically hold the computer and enter the responses in a poorly lit doorway in CAPI mode.

The usability evaluation may use a wide range of methods: utilisation of performance data, observational methods, users' debriefing, behaviour coding, and in general cognitive methods. A method characterised for being costless to collect and permitting the analysis at a question level, is represented by trace files or keystroke files, which are automatic by-products of the CAI systems (Couper *et al.*, 1997). However, in general they can reveal what happened at a particular moment, not why it happened, resulting difficult to be interpreted. For this reason it is suggested to use them in combination with other methods. As an example of experimental usability testing, the method can be performed with the aim of evaluating two different screen configurations.

Hansen and Couper (2004) suggest a procedure to conduct usability testing. First of all, it should be based on the definition of a plan stating: objectives of the test, user profile, test design issues, participant tasks, test environment and equipment, test monitor task, evaluation measures and report content and presentation. Based on this report containing the findings of the usability test, appropriate solutions are identified for improvement. Sometimes also to do mock interviews can be useful: in this case, for example, interviewers can act as respondents.
It is reported that, depending on the length of the CAI questionnaire, two to five tests a day are feasible. It has to be considered that, a considerable amount of time is required for the analysis of the data, the preparation of the report and the identification of the solutions.

By usability testing, increased confidence can be gained that the delivery and reception of the questions are in accordance to the original aims of the questions. In addition, it can help identifying problems of different nature such as question wording and criticism in interviewers' training. Although there is not a given phase in which to perform the usability testing, it is advisable to have it at an early stage of the testing process, in order to be able to correct the questionnaire on the basis of the test evidence.

5.8. Documentation

The documentation of an EQ is extremely important to verify the adherence of the EQ to the specifications, for the testing phase, for the debugging of the code and to assist in making changes (House, 1985). The documentation can be organised in different ways:

- specifications (rules, variables format, admissible values, types of error, etc) can be written on the paper questionnaire itself or on a separate document: this represents the main body of the documentation;
- a flow chart can be used to better understand the questionnaire flow and the branching paths;
- other specific documents or tools can be used. For example, a document or a database specific for the errors can be created, containing the rules under which errors are automatically detected; an univocal number, associated to each error, would make it easy to correct or modify the rules when necessary.

The importance of documentation is demonstrated by the growing interest in the development of software that automatically produces documentation having as input the EQ itself, e.g. Delta (Gatward, 2004).

5.9. Experiences

Generalised procedures and tools for the design of the EQ seem not to be so spread among the European NSIs. From the survey conducted for the purpose of this handbook it results that even if most of the Institutes have Quality guidelines and Manuals for questionnaire design, however, only a very limited number of them has specific handbooks for CATI/CAPI.
More than a half of the Institutes have a unit of experts supporting the development of the electronic questionnaire, moreover, about half of the Institutes relay on the outsourcing for the computer assisted data collection phase. Finally the majority of the Institutes use the software Blaise (see later in this section).
The functionality test for CATI/CAPI electronic questionnaire is rather applied, but very few Institutes adopt software packages for the interviewers' simulation.

Statistic Netherlands produces the software Blaise, which is widely used in the world, that allows the implementation of most of the above mentioned standards; sometimes its default settings can be considered themselves as standards to be respected in designing an EQ. For CATI surveys the software manages also the call scheduler whose architecture has been planned to avoid sample distortion. The manual "Blaise User Guide" can be considered as one tool specific for the design of an EQ.

Extremely important for CAI surveys is the role played by the users of the EQ that could be the interviewer or the respondent. Their role is important whatever technique is used, but for these techniques the presence of the computer has made it necessary to pay attention on how to develop user interfaces (Wensing *et al*., 2003). Many NSIs have developed screen layout standards that cover different aspects of the EQ: from the colours of texts according to their function, to the way response categories have to be displayed on the screen (one or more columns according to their number); from the need of displaying instructions for probing according to the type of questions (fact or opinion questions, attitude or knowledge questions) (Kuusela, 2003), to the use of capital letters only for specific type of instructions, e.g. for probing instructions; and, finally, from the extent to which customisation of texts is to be used (Gatward, 2003), to the use of icons or symbols instead of texts, to provide instructions or information to the EQ user. Some examples on how to develop standards for users interface can be found in the experiences of Statistic Finland, ONS and Australian Bureau of Statistics described in the proceedings of the "8Th International Blaise Users Conference – IBUC 2003" held in Copenhagen on May 2003.

The Italian National Statistical Institute is moving towards standardisation of data collection procedures based on CAI techniques. This standardisation basically consists in designing and developing "in-house" all the survey aspects and giving in outsourcing only the hiring of call centres and the selection of interviewers. The new aspect of this internalisation regards the software procedure: EQ questionnaires are developed in-house with positive results in terms of quality of data. This new strategy is mainly adopted for CATI surveys but it is starting to be applied also for the other CAI techniques.

5.10. Recommendations

Electronic questionnaire structure recommendations

- The EQ should be designed in order to collect all the information requested by the questionnaire planned for the survey and to collect them for all eventual situations.
- The EQ should follow the purposes of the paper questionnaire adding those elements, like automatic skips, branching paths or fills, allowed by the computer.
- For CATI surveys the setting of the call scheduler should avoid sample distortion.
- If some response items are presented more then once in the EQ, it would be better to associate them always to the same codes (e.g.: Yes =1, No = 2).
- Automatic controls to detect errors should be implemented in the EQ balancing the need of high quality data and the necessity of a fluent interview.
- Questions' names should be short and meaningful so as to facilitate the programmer job and, if displayed in the screen, to be clear for the interviewer.

- The types of controls, hard and soft, should be decided according to the nature of variable, the type of error and the interviewing technique.
- Instructions can be displayed on the screen before or after the question text according to the moment the user carries them out.

Design recommendations

- The design should reduce the segmentation effect typical of CAI questionnaires: more questions concerning the same topic have to be displayed per screen and, when possible, the title of section displayed each time it changes.
- The screen should not be crowded with non-useful information. The question to be read and, eventually, the response items should be located on the screen in order to be quickly identifiable by the interviewer.
- If a question has a long list of response items (which anyway should not be used in a CATI and/or CAPI questionnaire), they should all be displayed in the screen avoiding the scrolling.
- The description of the keyed response code should always be shown clearly on the screen to the interviewer.
- Particularly when CATI is used, the question text should not be too long and, anyway, the lines in the screen which contain it should not be too long.
- When in CAPI a “show card” is to be shown to the respondent, this action should be emphasised in the screen (always in the same way).
- Text’s functions should be differentiated by means of different colours: it is important to use always the same colour for a specific aim and to optimise the contrast between the selected colours.
- It can be useful to give emphasis to some particular words or instructions to catch the interviewer’s attention, displaying them in different ways, e.g. using CAPITAL, bold, underline or italics. It is recommended to adopt always the same solution and not to exceed with its use.
- On-line helps have to be used for those instructions that would fill too much the screen if displayed on it.
- The navigation of the questionnaire should be as easy as possible. The design should contain error messages indicating how to skip backward to the variables involved in the error and how to go forward to the last answered question.
- Customisation of texts with prior information should be used in order to simplify the interviewer’s job and to make the interview agreeable for the respondent.

Testing recommendations

- A testing team should be formed, including survey responsible, programmers and interviewers; a person responsible for the testing phase should be identified.
- The objectives of the test should be defined and a testing procedure should be outlined, integrating different methods such as question-by-question testing, testing by task, data testing, testing by scenario and simulation.
- To ensure the version controlling, especially when many programmers work on different parts of the EQ, only one master copy of each source file should exist.
- Depending on the complexity of the survey questionnaire, a testing log, i.e. either a database or paper documentation, that keeps a log of all the problems detected during the EQ testing phase and their resolution should be implemented.
- The human-computer interaction and users’ friendliness with the CAI instrument should be evaluated by means of a small-scale usability test.
- A criterion to end the usability testing phase should be established.
- A complete and organised documentation of the development of the EQ should be produced.

5.11. Checklist

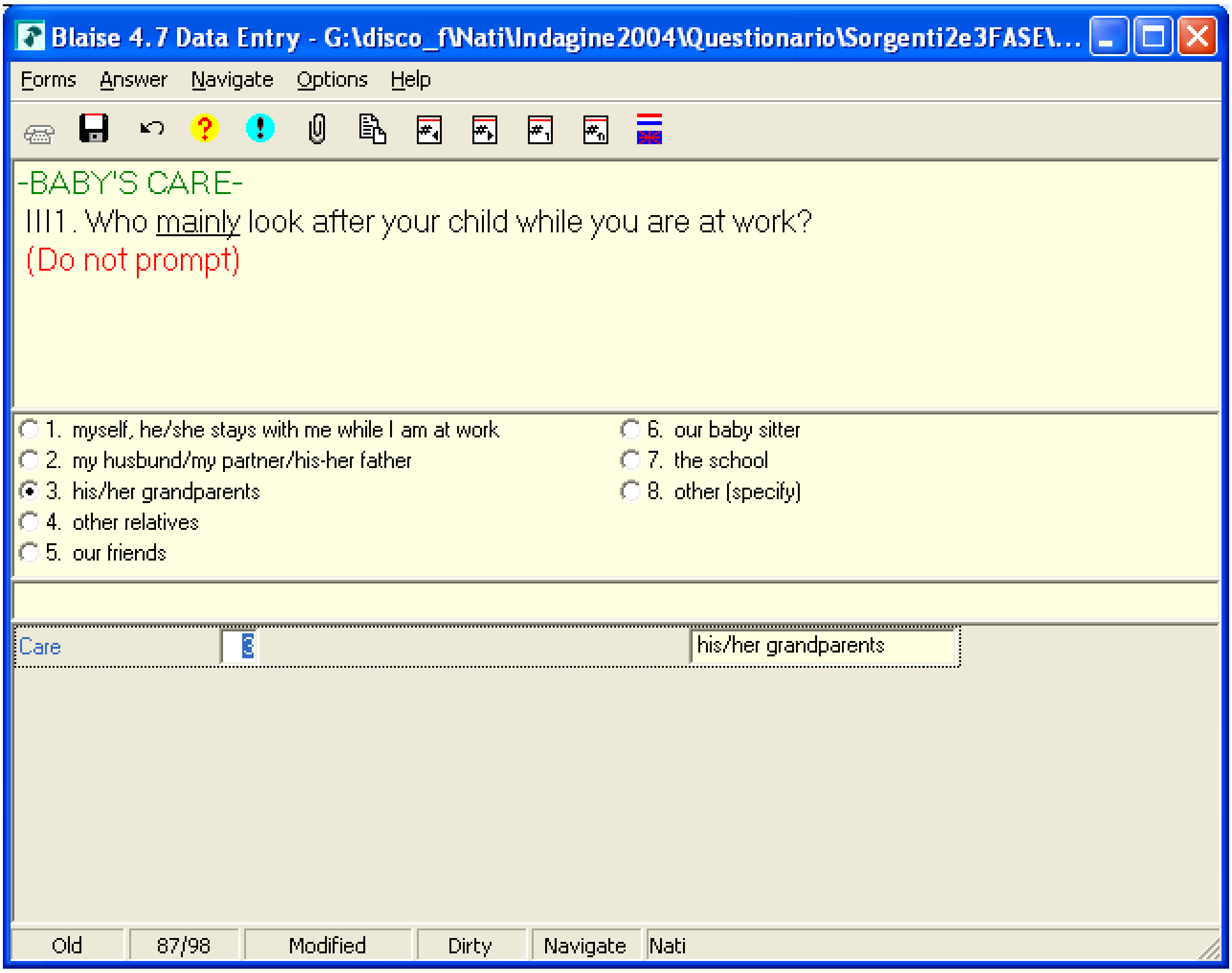
- Define all the skips and branching paths among variables.
- Define the range of variables.

- Define which type of control (soft, hard) has to be activated and when.
- Define the messages to be shown in relation to control activation.
- Specify the rules for text fills.
- Define the standards for the text to be read by interviewer, the answers appearance, the probing actions, the help instructions.
- Define the questions per screen and eventual standalone modules.
- Define appropriate splits for the code, helpful in the simplification of the construction of the EQ and its functionality testing.
- Establish the background colour and the colours to use for the text according to each specific function;
- If show cards are in use, define the icon or symbol to be associated.
- Check the adherence of the EQ to the specifications on paper.
- For CATI surveys, set the scheduler parameters according to the sample units, the survey period and length, the amount of available phone numbers, the type of phone number archive.
- For CAPI surveys, set the agenda parameters.
- Identify a responsible and a team in charge of performing the functionality and usability test of the EQ.
- Document the specifications, questionnaire flows and branching, EQ software, functionality and usability testing results.

Appendix 5.1. Examples of the adoption of recommended practices for design and implementation of an EQ for CATI using Blaise

The following examples are taken from the "Birth Sample Survey" carried out by Istat in 2004. The examples show video screens of the electronic questionnaire developed with the software Blaise.

Example 1: The use of different colours and styles for texts


Blaise 4.7 Data Entry - G:\disco_f\Nati\Indagine2004\Questionario\Sorgenti2e3FASE\...
Forms
Answer
Navigate
Options
Help
Old
87/98
Modified
Dirty
Navigate
Nati

-BABY'S CARE-

III1. Who <u>mainly</u> look after your child while you are at work?

(Do not prompt)

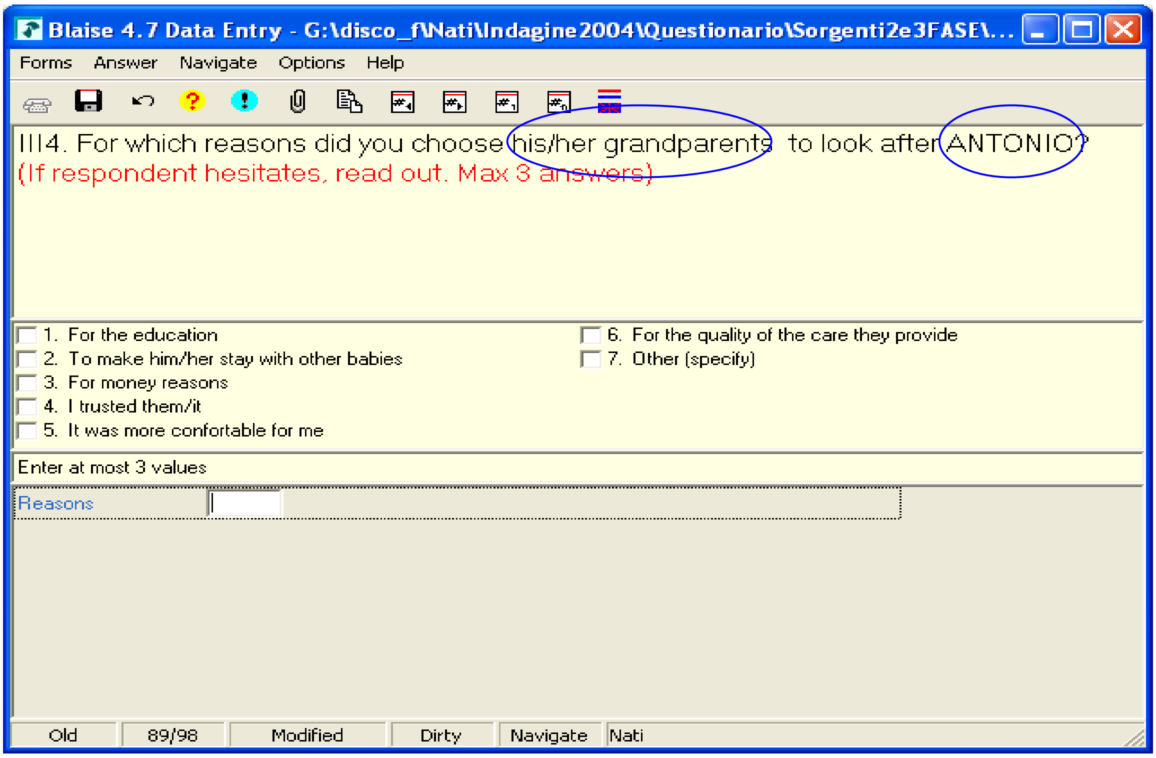
○ 1. myself, he/she stays with me while I am at work
○ 2. my husbund/my partner/his-her father
● 3. his/her grandparents
○ 4. other relatives
○ 5. our friends
○ 6. our baby sitter
○ 7. the school
○ 8. other (specify)

Care 3 his/her grandparents

The following colours and styles were associated with the same function through the entire questionnaire:

- **Green**: for section's heading;
- **Black**: for texts of questions or of error messages to be read aloud to the respondent;
- **Red**: for instructions to the interviewer not to be read to the respondent;
- <u>Underlined</u>: to stress particular words. In this example the underlining of word "<u>mainly</u>" was aimed at collecting only one answer, the most important one.

Example 2: The use of fills


Blaise 4.7 Data Entry - G:\disco_f\Nati\Indagine2004\Questionario\Sorgenti2e3FASE\...
Forms
Answer
Navigate
Options
Help
Old
89/98
Modified
Dirty
Navigate
Nati

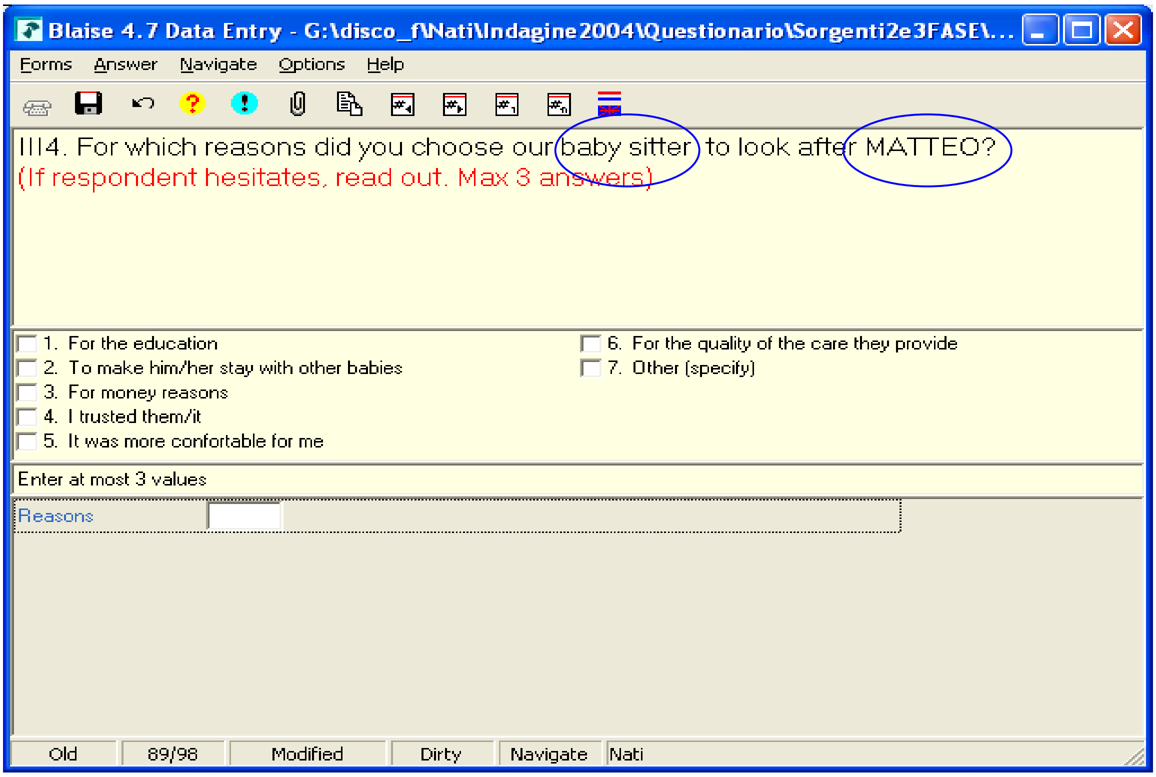
III4. For which reasons did you choose his/her grandparents to look after ANTONIO?
(If respondent hesitates, read out. Max 3 answers)

□ 1. For the education
□ 2. To make him/her stay with other babies
□ 3. For money reasons
□ 4. I trusted them/it
□ 5. It was more confortable for me
□ 6. For the quality of the care they provide
□ 7. Other (specify)

Enter at most 3 values

Reasons

The above example shows two fills that use different source of information: the fill “his/her grandparents” is an answer to a previous question (question III3 – see example 1), while the baby’s name “Antonio” is stored in the input file. A different answer to question III3 and a different sample unit would change the values of the fills as follows:


Blaise 4.7 Data Entry - G:\disco_f\Nati\Indagine2004\Questionario\Sorgenti2e3FASE\...
Forms
Answer
Navigate
Options
Help
Old
89/98
Modified
Dirty
Navigate
Nati

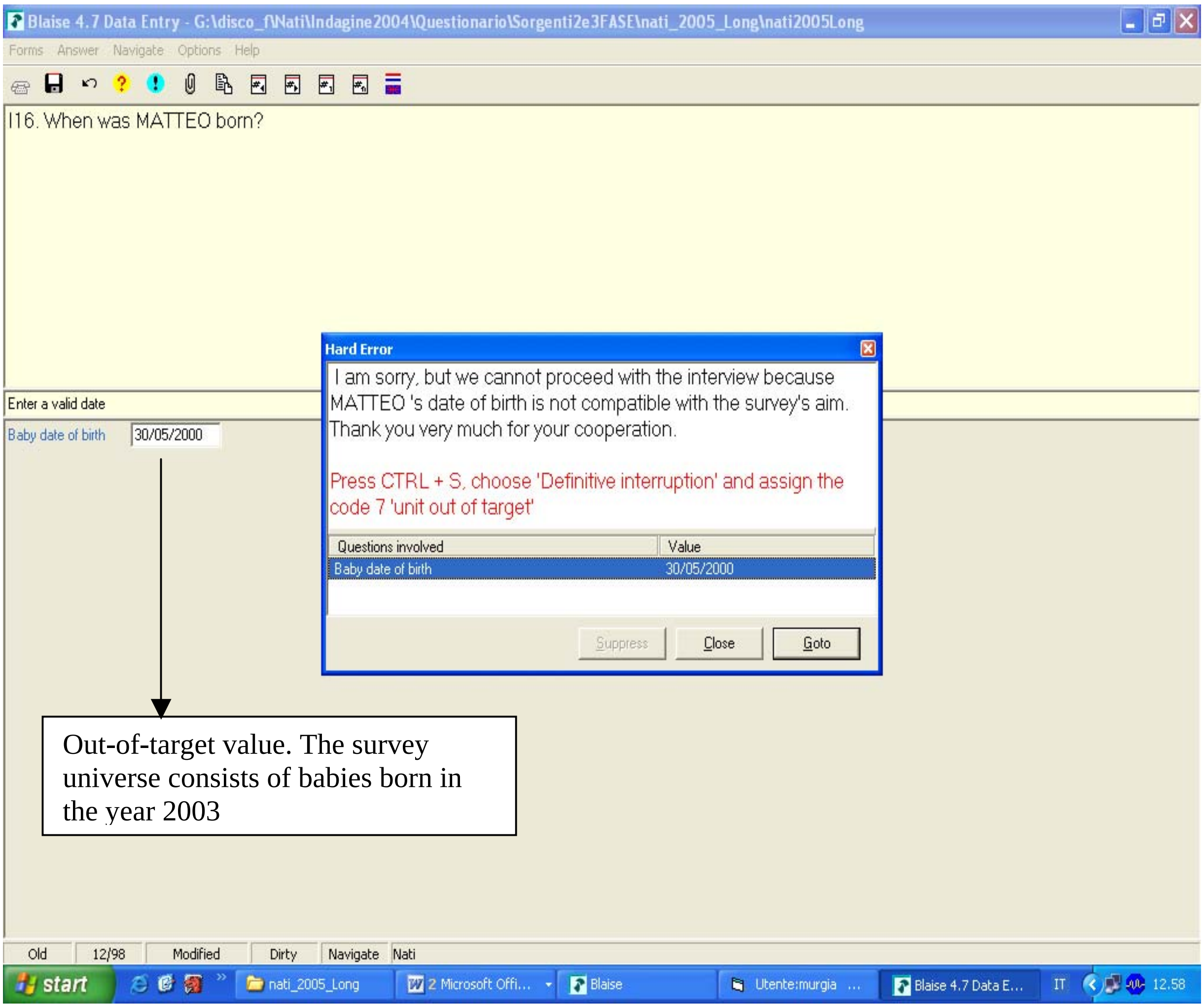
III4. For which reasons did you choose our baby sitter to look after MATTEO?
(If respondent hesitates, read out. Max 3 answers)

□ 1. For the education
□ 2. To make him/her stay with other babies
□ 3. For money reasons
□ 4. I trusted them/it
□ 5. It was more confortable for me
□ 6. For the quality of the care they provide
□ 7. Other (specify)

Enter at most 3 values

Reasons

Example 3: Hard error and instructions to interviewer


Blaise 4.7 Data Entry - G:\disco_f\Nati\Indagine2004\Questionario\Sorgenti2e3FASE\nati_2005_Long\nati2005Long
Forms Answer Navigate Options Help
Hard Error
Suppress
Close
Goto
Old 12/98 Modified Dirty Navigate Nati
start
nati_2005_Long
2 Microsoft Offi...
Blaise
Utente:murgia ...
Blaise 4.7 Data E...
IT
12.58

I16. When was MATTEO born?

Enter a valid date

Baby date of birth 30/05/2000

I am sorry, but we cannot proceed with the interview because MATTEO 's date of birth is not compatible with the survey's aim. Thank you very much for your cooperation.

Press CTRL + S, choose 'Definitive interruption' and assign the code 7 'unit out of target'

Questions involved	Value
Baby date of birth	30/05/2000

Out-of-target value. The survey universe consists of babies born in the year 2003

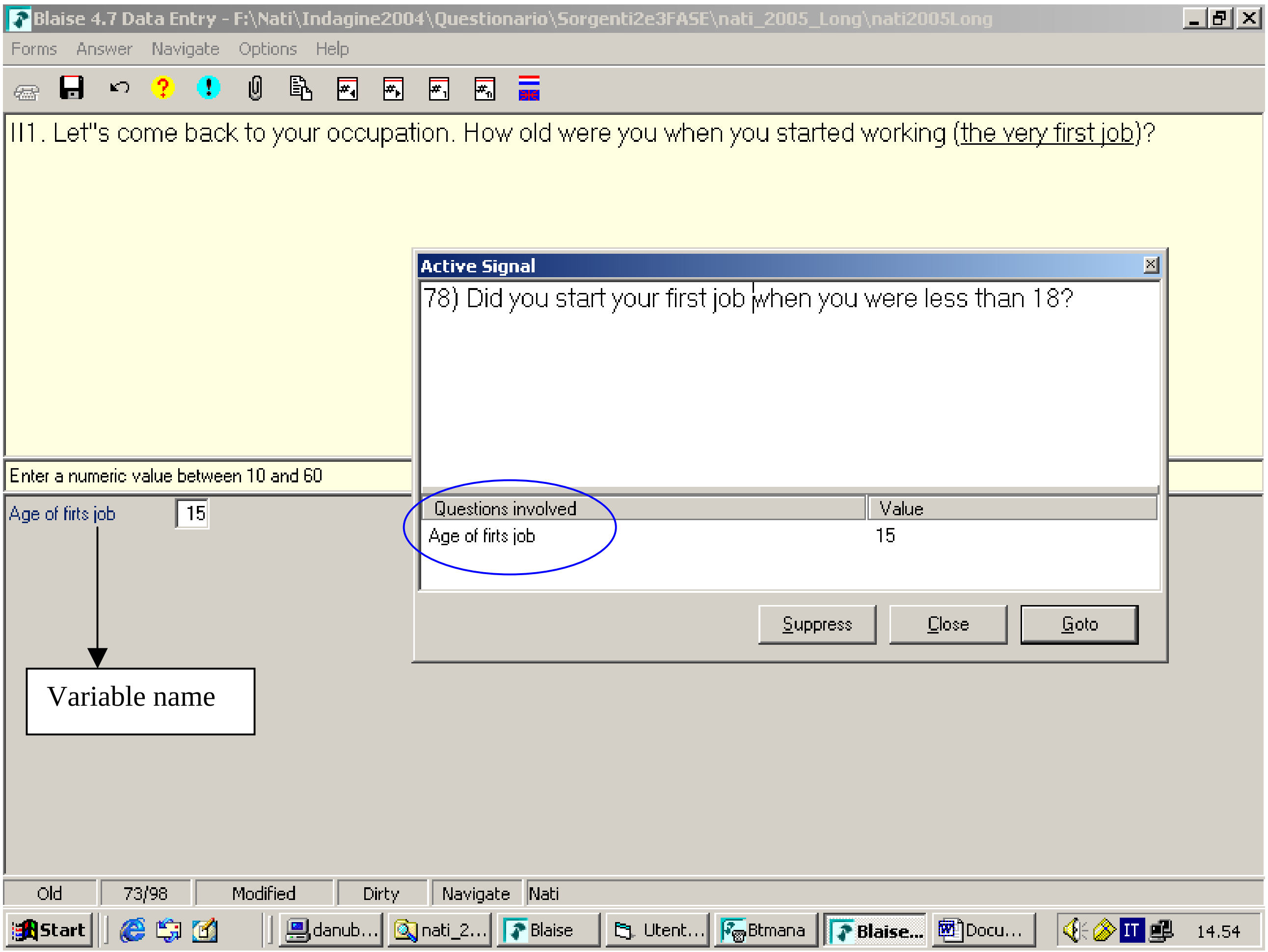
The control underneath the “date of birth” variable is a <u>hard</u> type: this is because the variable is both an objective one and it is very crucial for the survey, since the sample unit’s eligibility depends on its value.

In case of a non eligible unit, an error message is shown to the interviewer. It contains:

- a black colour text, to be read to the respondent, to explain the reason why the interview can not proceed;
- a red colour text to provide instructions to the interviewer (and not to be read to the respondent) on how to code the results for the non eligible unit;
- a fill, the baby’s name, to create a customised message.

This example also shows how messages for respondents and instructions to interviewers can be provided without crowding the video screen: they are indeed included in the error message window so as to appear only when necessary.

Example 4: Soft error and variable name for interviewer



II1. Let''s come back to your occupation. How old were you when you started working (the very first job)?

Enter a numeric value between 10 and 60

Age of firts job 15

78) Did you start your first job when you were less than 18?

Questions involved	Value
Age of firts job	15

The control chosen is a soft one because there are not objective rules to establish the right age to start a job.

This example also shows how important is to give to each variable a meaningful name: in case of error, the list of variables involved in it (in this case only one variable is involved) is shown to the interviewer who can easily understand which of them have to be corrected. This would have been more difficult if instead of a meaningful name the technical one was shown.

Chapter 6. Testing methods

For several decades researchers and statisticians have been working on the development of testing methods in order to reduce non-sampling errors and to improve the quality of data. Until the mid-1980s, methods to evaluate questionnaire design and data collection modes focused on launching small sample studies with about 20 to 200 respondents. Feedback on the design was mainly linked to interviewer debriefings, based on a debriefing just after conducting a small scale survey. Methods were standardised only to some extent and were generally fairly unsystematic.

This type of questionnaire evaluation has often been, and sometimes still is referred to as standard pretest. The approaches changed in the second half of the 1980s when the concepts and findings of cognition psychology and behavioural science gained more attention with respect to the response process. The psychological background of the cognitive processes led to different methods to scrutinise the response process (see Chapter 2). Cognitive laboratory methods, based on cognitive-psychological techniques were used more and more in order to assess the question answering process and to evaluate questions and questionnaires (Prüfer and Rexrodt, 1996).

During the 1990s, behaviour coding was introduced as a further instrument (see section 6.2.). Interviewer or/and respondent behaviour with regard to specific questions was classified in a coding scheme. The method was implemented either in the field or in the lab with the sample sizes considerably larger than in cognitive interviews (enabling researchers to make use of quantitative analysis methods).

Furthermore, the enormous development in the field of information technology also enhanced the possibilities for questionnaire testing. Developments like screen-captured devices, audit trails, eye-tracking etc. opened possibilities to analyse non-verbal reactions as well as the interviewer-respondent interaction via the screen (when using computer assisted questionnaires) additionally to the verbal reactions (see Beukenhorst, 2004; Redline and Lankford, 2001).

Given this historical background, the further development and variety of methods, it is rather difficult to establish an overarching definition of questionnaire testing methods. Different authors classify the techniques by different perspectives. In this handbook we distinguish pre-field and field testing methods.

Pre-field testing methods are (although not necessarily) applied under laboratory conditions. This means that the interviews are not carried out in exactly the same way as later on in the field. The term "laboratory conditions" refers to an observational environment which may differ totally or partially from the actual field conditions. For example, the respondents may not be interviewed at home, but in a testing environment facilitating the use of specialised testing methods (e.g. a cognitive laboratory). Only small parts of the questionnaire might be included. Additional questions can be added regarding how the respondents perceive the questions. Consequently, the interview flow might deviate quite substantially from the later fieldwork. Pre-field methods are particularly suitable to collect information on how respondents proceed when answering the questions during the preliminary stages of questionnaire development. They include mainly informal tests, expert groups, cognitive interviews and observational interviews. Focus groups and in-depth interviews are sometimes also classified as pre-field methods, while we treat these methods under questionnaire design as they are in most cases related to the specification of the concepts to be measured.

Field testing methods are those used to evaluate questionnaires tested under field conditions. This means that the interview is carried out in a very similar way as later on in the fieldwork (regarding setting, lengths, choice and order of questions etc). Field methods include behaviour coding, interviewer debriefings, respondent debriefings, follow-up interviews, analysts' feedback, experiments.

These testing methods are generally conducted to verify the questionnaire concepts and the data collection mode. In contrast, *pilot surveys* are conducted on a bigger scale and – in addition to questionnaire design related issues – also aim at an evaluation of the entire survey design, including sample design, recruitment, phase of data collection, fieldwork administration and data processing (Biemer and Lyberg, 2003).

Finally, some analyses, aimed at exploring possible questionnaire's shortcomings, can also be performed on data coming from testing methods as well as on data deriving from the actual data collection phase (analysis of item nonresponse rates, imputation rates, edit failures, or responses' distributions). When applied to the real data collection phase, such methods result particularly useful for ongoing surveys. Sometimes, also the reinterview studies provides some clues on possible questionnaire's shortcomings.

It has to be noted that questionnaire testing is an iterative process which requires the application of different subsets of these methods (Groves, 2004). Only with the choice of different testing methods, it is possible to detect different kinds of problems. An adequate questionnaire evaluation process therefore usually requires a well-balanced mix of various methods which have to be applied at different stages of the development process. The result of each method will improve specific aspects of the questionnaire.

Finally, there is a considerable difference in the selection of suitable testing methods for either individual/household surveys or establishment surveys because in these surveys the respondents differ, as do the mode of contact and technologies used. However, the research on testing methods specifically targeting the aspect of establishment surveys has started only fairly recently, as reflected by a quite limited number of publications in this field. Consequently, only some information on the test of establishment surveys can be provided, whereas the recommendations are mainly focused on individual/household surveys.

The aim of the recommendations presented here is to give an introduction to widely accepted methods, to explain their objectives and specific qualities, to discuss their advantages and drawbacks and to provide information during which stage of the testing they are optimally applied.

6.1. Pre-field testing

All testing methods which are conducted more or less under laboratory conditions (not applied in the field, but in the institutes/agencies) are classified as pre-field methods. The most important pre-field testing methods are: informal tests, expert groups, cognitive laboratory methods and observational interviews. Focus groups and in-depth/qualitative tests sometimes are also referred to as pre-field testing methods. However, since these methods are mainly implemented during the stage of questionnaire development, they are presented in section 3.1. in this handbook. Sometimes respondent and interviewer debriefings are also applied before a field test has been conducted, however this is rather seldom and, accordingly, they are considered as field testing methods in this chapter.

In addition, new technologies, like screen-capture devices, audit trails, eye-tracking devices (Beukenhorst, 2004) are under development. Until now, these new approaches have not been applied as daily routine testing methods. Consequently, the recommendations do not cover these methods in detail, but present a list of methods and references.

The general aim of pre-field testing is to detect problems of a questionnaire such as wording, terminology, skip instructions, answering categories and visual design before the survey is conducted in the field to prevent respondent, interviewer and processing errors. It is important to notice, however, that pre-field testing aims at qualitative testing of the questionnaire, followed by more quantitative testing in the field, but not in checking the shortcomings of the entire survey design (e.g. sampling).

6.1.1. Informal tests

"A questionnaire designer, working alone, can easily become blind to defects in his/her own work" (SCB, 2004). Therefore, to prevent potential "blindness", researchers distribute a first draft of the questionnaire to their colleagues and acquaintances to get their attitude, a method which is called informal test or desktop pretest (Biemer and Lyberg, 2003). With this technique the researcher gets a first feedback on the questionnaire design, commonly used to detect mistakes in design, wording and skip instructions. Informal tests with colleagues can help to avoid “unnecessary” errors (SCB, 2004) and lead to a questionnaire which can be tested by a broader audience.

Informal tests are typically unstructured reviews, either with individuals or in a group-setting. It can be done with questionnaire design experts or other colleagues (depending on the perspective of the test).

Informal tests can already provide very useful information on shortcomings of the instrument as questions that might be misunderstood, confusing layout, misleading instructions, typographical problems and so on can be detected during a very early stage of the questionnaire. However, it is a first, unstructured feedback, which may not identify all important problems. In fact it should be considered as the starting point for any further tests.

It is advisable for an informal test to gather colleagues from different departments who will get in contact with the questionnaire during any stage of the survey process. The informal test should ideally involve fieldwork specialists, methodologists and the people involved in data entry and editing. Any potential “user” should be confronted with the draft questionnaire in order to test if the instrument conforms to his of her requirement. At this early phase of testing it is vital to check the general design and layout of the questionnaire e.g. with the data entry and processing team in order to avoid basic mistakes of the design.

6.1.2. Expert reviews (Expert groups)

In contrast to focus groups and in-depth interviews which assess the viewpoint of potential respondents, expert reviews (sometimes known as expert groups) allow feedback from potential users of the data: subject experts and questionnaire design experts. Recently, also data processing experts are asked to give a feedback. However, in contrast to focus groups, the discussion is based on a first draft of the questionnaire allowing more concrete feedback on the survey instrument.

Since expert reviews are conducted with a group of people, the discussion is chaired by a moderator (ABS, 2001). There are two ways of moderating: Either by a structured discussion on each question with a standardised coding scheme, to be filled in before the meeting or a (less) structured discussion on each question, but without a standardised coding list. The subject experts and questionnaire design experts involved in expert reviews aim (a) to ensure that the questionnaire collects the information needed to meet the analytic objectives of the survey and (b) to check the design and the questions themselves. Therefore they focus specifically on:

- Terms and wording of questions;
- Structure of questions;
- The response alternatives;
- Order of questions;
- Navigational rules of the questionnaire;
- Instructions to interviewers on questionnaire administration;
- Confusing layout;
- Typographical errors (Groves, 2004, p. 242; Biemer and Lyberg, 2003, p. 262).

Additionally, data processing experts might be asked to give feedback on the design of the questionnaire with respect to aspects of processing. Sometimes, expert reviews are conducted in concert with other testing methods, for example as part of a usability evaluation (see Chapter 5).

Two experts of each area of expertise might be asked to review the questionnaire. As the reviews can be either unstructured or structured they require different resources concerning facilities, time and equipment. However, it is becoming more and more common to implement structured reviews in order to collect more comprehensive data (Biemer and Lyberg, 2003).

A) Structured expert reviews

For structured expert reviews a set of criteria by which each question is to be examined are implemented (Lessler and Forsyth, 1996). These coding schemes are based on the cognitive response process, they are used as a guide for the systematic appraisal of surveys and help the reviewer to structure his/her feedback. The coding schemes are very concrete because they are linked to each question to be tested. A proposal for a coding scheme (Biemer and Lyberg, 2003; Lessler and Forsyth, 1996) is presented below which is structured into seven basic items: (1) Problems with reading, (2) problems with instructions, (3) problems with item clarity, (4) problems with assumptions, (5) problems with knowledge/memory, (6) problems with sensitivity/bias and (7) problems with response categories.

The coding scheme needs to be filled in for each question and consequently the draft questionnaire needs to be completed by the standardised coding categories after each question. This questionnaire is to be sent to the experts, who should check and fill in their standardised feedback.

The experts may need two to five days for their review (depending on the complexity of the questionnaire) and it is reasonable to plan another week for one person to check and summarise the results (ABS, 2004b). After reviewing the standardised feedbacks an expert group meeting might be conducted to discuss the findings of the review.

Figure 6.1. Coding categories for a structured expert review system (Biemer and Lyberg 2003, originally by Lessler and Forsyth, 1996)

1. **PROBLEMS WITH READING:** Determine if it is difficult for the interviewers to read the question uniformly to all respondents.
   1a – What to read: Interviewers may have difficulty determining what parts of the question are to be read.
   1b – Missing information: Information the interviewer needs to administer the question is not contained in the question.
   1c – How to read: Question is not fully scripted and therefore difficult to read.

2. **PROBLEMS WITH INSTRUCTIONS:** Look for problems with any introductions, instructions, or explanations from the respondent's point of view.
   2a – Conflicting or inaccurate instructions, introductions, or explanations.
   2b – Complicated instructions, introductions, or explanations.

3. **PROBLEMS WITH ITEM CLARITY:** Identify problems related to communicating the intent or meaning of the question to the respondent
   3a – Wording: The question is lengthy, awkward, ungrammatical, or contains complicated syntax.
   3b – Technical terms are undefined, unclear or complex.
   3c – Vague: The question is vague because there are multiple ways in which to interpret it or to determine what is to be included and excluded.
   3d – Reference periods are missing, not well specified, or are in conflict.

4. **PROBLEMS WITH ASSUMPTIONS:** Determine if there are problems with assumptions made or the underlying logic.
   4a – Inappropriate assumptions are made about the respondent or his/her living situation.
   4b – Assumes constant behaviour: The question inappropriately assumes a constant pattern of behaviour or experience for situations that in fact vary.
   4c – Double-barrelled question that contains multiple implicit questions.

5. **PROBLEMS WITH KNOWLEDGE/MEMORY:** Check whether respondents are likely to not know or have trouble remembering information.
   5a – Knowledge: The respondent is unlikely to know the answer.
   5b – An **attitude** that is asked about may not exist.
   5c – **Recall** failure.
   5d – **Computation** or calculation problem.

6. **PROBLEMS WITH SENSITIVITY/BIAS:** Assess questions for sensitive nature or wording, and for bias.
   6a – Sensitive content: The question is on a topic that people will generally be uncomfortable talking about.
   6b – A socially acceptable response is implied.

7. **PROBLEMS WITH RESPONSE CATEGORIES:** Asses the adequacy of the range of responses to be recorded.
   7a – Open-ended question that is inappropriate or difficult.
   7b – Mismatch between question and answer categories.
   7c – Technical terms are undefined, unclear, or complex.
   7d – Vague response categories.
   7e – Overlapping response categories.
   7f – Missing response categories.
   7g –Illogical order of response categories.

B) Unstructured expert reviews

Unstructured expert reviews are carried out with three to six experts, they should be taped or video-recorded and the discussion normally takes about 2-3 hours. Even though there is no checking list to be filled in by the reviewer, a moderator should structure the discussion by items like wording, terms, skip instructions, instructions, layout, comprehension etc. After the discussion a report on results should be written and even be checked once more by tape or video (Biemer and Lyberg, 2003)

C) Experiences and recommendations

The survey conducted for the purpose of this handbook revealed that almost half of the European NSIs and other international statistical institutions are conducting expert reviews. It seems that it is the second most applied testing method in the NSIs questioned (after interviewer debriefings). Presser and Blair (1994) launched a study on the productivity and usefulness of different testing methods in the early 90ths. They compared four methods: conventional testing[5], behaviour coding, cognitive interviews, and expert reviews, by the criteria: productivity, costs, amount and types of problems detected. In this study it could be shown that:

- Expert reviews identified most of the problems, followed by behaviour coding, conventional tests and cognitive interviews.
- Behaviour coding and expert reviews proved to be the most reliable methods.
- Expert reviews were the cheapest method, followed by cognitive interviews, whereas conventional testing and behaviour coding were the more expensive (at almost a similar level).
- Although expert reviews suggested some reasons for problems, they failed to provide any idea of the perspective of interviewers or respondents, which is essential at a certain stage of testing.
- Experts flagged most of problems, but it is not clear whether they were all important problems.

Summarising these experiences it can be recommended that expert reviews should be conducted during the initial phase of the questionnaire pre-field testing process. Referring to Presser and Blair (1994) they are less expensive than other methods and quite easy to implement (ABS, 2004b). However, whereas structured expert reviews are very concrete with regard to each question, they do not provide evidence on the interplay of questions, on skip instructions and on problems associated with overlaps and/or relationships regarding the entire questionnaire. The feedback is focussing on isolated questions, not on the overall flow of the questionnaire.

It is highly recommended to use structured expert reviews for getting structured feedback. However, even then one has to be aware of the fact that the appraisal is subject to the experts' interpretation and judgement. Consequently the results of expert reviews should be regarded as guidance toward potential problems in the questions which in turn need further testing (Groves, 2004). In addition there is no reflection on the perspective of either the interviewer or the respondents (Prüfer and Rexrodt, 1996).

6.1.3. Observational interviews

An observational interview is a method to test and evaluate self-completion forms where a trained observer watches the survey process (e.g. form completion or response within an interview) to better understand the respondent's thought processes during questionnaire completion.

"Observational interviews aim to identify problems in wording, problems in question order, presentation or layout, and to estimate the time taken to complete the questionnaire or parts of the questionnaire. Survey designers look for behaviours that result from an error of the instrument, including the participant's behaviour (e.g. reading all the questions before responding), non-verbal cues, reactions and observed cognitive processes (e.g. counting on their fingers or writing calculations on the page). This technique can also use follow-up probes to elicit information about why the respondent behaved as he or she did" (ABS, 2001).

"Observation" actually is the striking word of new technologies and becomes more and more frequently used. It shows that not only verbal communication is of interest, but that video recording and observing computer screens are regarded as useful means to understand the question answering process as well. A brief overview of these new technologies is presented in section 6.1.5.

[5] Conventional testing is essentially a dress rehearsal in which interviewers receive training like that for the main survey and administer a questionnaire as they would do during the real survey (Presser *et al.*, 2004)

6.1.4. Cognitive interviewing

Cognitive interviewing is a collection of different techniques for studying the comprehension stage, thought and answering processes of respondents during the interview using expanded or intensive interviewing approaches (Biemer and Lyberg, 2003). It is based on the assumption that verbal reports from the respondents are a direct representation of their specific cognitive processes elicited by the questions (Ericsson and Simon, 1993).

The cognitive interview can be conceptualised as a modification and expansion of the usual survey interviewing process. However, the one-on-one in-depth interview is conducted by a specially trained cognitive interviewer rather than by a survey field interviewer, and this interviewer administers questions to a test person instead of the usual survey respondent (Willis, 2004). Test persons are recruited especially for the cognitive interview, they have not seen the questionnaire before and have similar characteristics to the actual respondents later in the field. Due to the qualitative nature of cognitive interviews only a small number of test persons are involved. Usually a round of cognitive interviews is conducted with a maximum number of 15 test persons (Willis, 2005). Cognitive interviews are usually carried out in somewhat controlled rooms like laboratories or other suitable rooms and the interviewer, as well as the test person take part actively.

The stages of the cognitive model (encoding, comprehension, retrieval, judgement and reporting) drive the respondents' survey response (chapter 2) and they are potential error sources of specification, measurement (especially respondent), and probably nonresponse error as well as mode effects (compare e.g. DeMaio and Rothgeb, 1996; Biemer and Lyberg, 2003). Cognitive interviewing provides a qualitative method to detect sources of these errors during the testing stage by elucidating how respondents understand or interpret a question and how they reach an answer in accordance to the stages of the response model. The idea is that an understanding of cognitive processes as described in the cognitive model supports the development of design rules that govern choice of response categories, question ordering, and so on to construct, formulate and ask better survey questions. The general objective of cognitive interviewing is therefore to reduce sources of response error via providing detailed information about these errors and some information about mode effects during the testing stage.

Cognitive interviewing actually comprises a suite of techniques that are suitable for investigating if survey respondents understand questions in the way it is intended by survey designers and if it is feasible for the respondents to recall and provide the necessary information. It is used as a testing method to obtain qualitative information on how a draft questionnaire is understood and answered by actual respondents, to revise the questionnaire and to eliminate deficiencies. In other words it aims to assure survey developers that respondents finally are answering the question survey developers think they are asking and to provide information on diverse respondent reactions to sensitive or difficult questions.

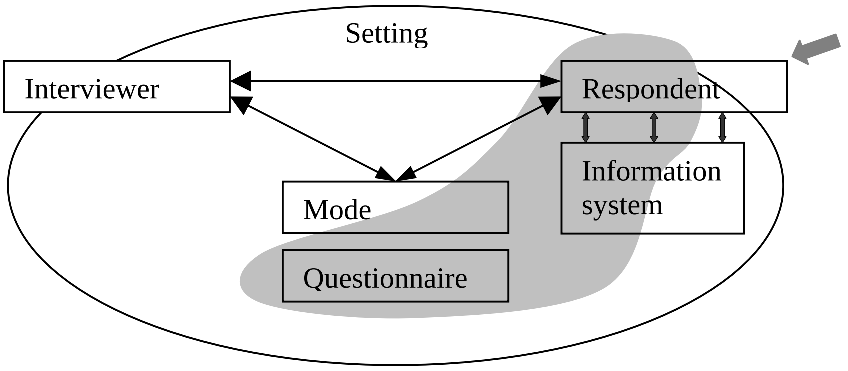
However, cognitive interviewing is not carried out primarily for the purpose of developing general principles of questionnaire design, but rather to evaluate targeted survey questions, with the aim of modifying these questions when indicated. The strength of cognitive methods is that they provide information about the existence of a problem in a question, its possible source(s) as well as information toward the problem's solution. They can determine whether a specific question wording communicates the objective of the question and quickly identify problems such as redundancy, missing skip instructions and awkward wording with only a few interviews. Additionally, they can provide information on sources of response error that are usually unseen by the interviewer and the survey practitioner (DeMaio and Rothgeb, 1996).

When considering cognitive interviewing and the response process, two aspects are important to ensure appropriate application: cognitive interviewing is qualitative in nature and the cognitive process model concentrates on the respondents' site of measurement error.

The qualitative manner of cognitive interviewing has implications on its timing of use during testing. Given the variety and flexibility of cognitive interviewing methods it is ideal to investigate in depth the respondent-question interaction e.g. how the respondent understands and interprets a question. This investigation can be very detailed and very informative but it is restricted to a limited number of respondents and interviews.

Therefore cognitive interviewing should be applied when a first profound draft questionnaire has been developed. However it should never replace any quantitative testing of the questions which should be scheduled after cognitive interviewing.

Figure 6.2. Potential measurement error sources and the focus of cognitive interviewing (after Biemer and Lyberg, 2003)



Area where cognitive methods can support the reduction of measurement error

The concentration of cognitive interviewing on the respondents' site of measurement errors can best be visualised with the help of a figure from Biemer and Lyberg (2003) which illustrates the concept of potential sources of measurement error (e.g. interviewer, respondent, questionnaire…) (copied and slightly changed in figure 6.2). In addition to the original figure, in figure 6.2 a grey region is added which visualises where cognitive interviewing supports the reduction of measurement error. It is important to note that there are white areas left in the figure for which measurement error needs to be reduced with the help of other testing techniques.

In the literature, cognitive interviewing is discussed with different focuses. Sometimes only techniques requiring the test persons to speak out loudly what they are thinking when answering a question, and techniques where interviewers probe for information to elicit the respondents' understanding and thought process are comprised under the term cognitive interviewing. Nevertheless, methods like the test persons' judgement of their own answers, the repetition of questions in the test persons' own words, the arrangement of a set of particular cards or objects, the judgement of hypothetical situations as well as measuring the time it takes a respondent to answer a question are also introduced in the following subsection to present an overview of existing methods (figure 6.3).

Figure 6.3. Systematic list of cognitive interviewing methods discussed in this chapter:

A) Think aloud
    Concurrent think aloud
    Retrospective think aloud
B) Probing
    Follow-up probing or concurrent scripted probes
    Post-interview probing or retrospective probing
    Comprehension probing
    Information retrieval probing
C) Confidence Ratings
    Paraphrasing
    Sorting
        Free Sort
        Dimensional Sort
D) Vignette (classifications)
E) Response Latency

Cognitive interviewing has the advantage that different cognitive techniques can be combined in one interview and that the choice of techniques is a very flexible, iterative process with the option to compile different (rounds of) interviews (e.g.: Willis, 2005).

However, conducting a cognitive interview or some cognitive techniques does not represent a panacea, because no universally accepted standard for cognitive interviews exists (Beatty, 2004). There are trade-offs to balance with respect to the prevailing testing aim, e.g. which aspect of a question or a suite of questions should be clarified. Further more, it has to be decided if it is important to obtain a codable answer or if a response that consists of the subject's own words is needed. Similarly, a decision has to be made if probes on special parts of a question are needed, e.g. to test different potential problems or hypotheses, or whether the test persons should be asked to elaborate on a given answer (Willis, 2004). Accordingly, due to this variety of approaches it is difficult to decide on how to conduct cognitive interviews.

The overall aim of this chapter on cognitive interviewing is to point out that probably the best way of conducting cognitive interviews is to adopt a strategy that is closely oriented towards specific questions to be investigated in detail. Once that choice of questions has been made a continuous flexible decision process on the application of different techniques or their combination should be adopted.

Accordingly, first of all cognitive interviewing techniques are introduced in an overview indicating prevailing aims, potential strengths and weaknesses. Afterwards it is tried to discuss how they might be combined and what can be achieved with a flexible selection strategy.

A word with respect to further reading: Willis (2005) introduces the various techniques and beneficially discusses all aspects of cognitive interviews in his book "Cognitive Interviewing". Anyone who is further interested in cognitive interviewing is referred to this volume. After introducing the method it provides a valuable chapter discussing the think aloud technique and devotes a special part to verbal probing. Years of experiences result in compendious practical information on how to plan and realise cognitive interview as well as on how to analyse their results.

A) Think aloud

In a think aloud interview, the test person is asked to formulate loudly all thoughts which lead to an answer. The aim is to get information about how the test persons understand a question or term. Think aloud sessions are used to identify difficulties in question comprehension, misperceptions of the response task, types of recall strategies used and reactions to sensitive questions.

The think aloud technique was originally used by psychologists to elucidate cognitive processes for problem solving tasks. Its effectiveness in testing is, however, discussed controversially because it strongly depends on the ability and willingness of the test persons to articulate their thoughts. Although an aimed support and guidance of the test persons is inevitable, still most test persons are not capable of speaking out their thoughts loudly (Prüfer and Rexrodt, 2005) and alternative methods need to be chosen in many cases.

> **Concurrent think aloud.** In the concurrent think aloud technique the test person is asked to speak out the thoughts immediately after the question is posed until the answer is given. DeMaio and Rothgeb (1996) have found the concurrent think aloud technique to be very useful in discovering which parts of the question test persons read and how they move around from question to question in addition to helping the questionnaire designer learn how they comprehend and answer the questions.

> *Example (Biemer and Lyberg, 2003):*
> *Question: "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?"*
> *Question is immediately followed by the instruction: "Now, tell me exactly what you are thinking as you try to answer that question. What is the first thing you are trying to think of?"…*

Pronounced disadvantages of the concurrent think aloud technique are its high respondent burden, its disruption of the normal flow of the interview and the fact that speaking out the thoughts loudly might alter the cognitive process to be studied.

The interview's flow is not as much disrupted in the **Retrospective think aloud** technique. With this technique the respondent explains either after answering the question or even after completing an interview, under conditions that are similar to the actual survey, how the answer was developed. However, detailed information about the stages of the response process might be lost applying retrospective think aloud because of the time gap between the response process and the verbal report.

The use of the following described methods of cognitive interviewing – probing, confidence ratings and paraphrasing – focuses on comparatively particular aspects of the response formation process in order to gain a more complete understanding of how test persons accomplish their task, e.g. if the information given is incomplete or when it became obvious that a question was not understood in the right way (DeMaio and Rothgeb, 1996).

B) Probing

Probing in a way examines an answer with the help of additional questions (probes) asked by the interviewer about it with the purpose of achieving additional information. For the formulation of probes, it is important to have already an idea or hypothesis about what the cognitive difficulty of the question might be. Accordingly, probing questions are used, e.g. when the information provided by the test person during the think aloud is incomplete or revealed any potential problem with a specific question and the researcher wants to find out how respondents chose among response choices or how they interpreted reference periods or a particular term.

Probes can be categorised in different ways. Either they are grouped according to the time they are asked (follow-up or retrospective) or they are grouped according to their aim (to get information about how the question was understood versus to get information on how the test person came to a particular answer). The following examples present some typologies of probes (follow-up, retrospective, comprehension and information retrieval). However, there is no sharp distinction between the different specific objectives of each of them. For example, retrospective probes can as well have an information retrieval aim.

Follow-up probing or **concurrent scripted probing** refers to probing immediately after an answer is given.

> *Example (Biemer and Lyberg, 2003):*
> *Question: "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?"*
> *Answer: "About two days"*
> *Series of questions about the respondents cognitive process resulting in this answer: "How did you decide on that number of days? - Describe for me the illnesses and injuries that you included in your answer. – Did you have any difficulty deciding whether days were 'good' or 'not good'?"*

Similarly to concurrent think alouds, the follow-up probing disrupts the normal flow of the interview and might itself have effects on the cognitive process.

In **post-interview probing** or **retrospective probing** the questions are asked after the interview is completed.

Example (Biemer and Lyberg, 2003):
Question: "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?"
Probes – after the interview: "What were you thinking when you responded to that question? – How did you arrive at four days as your answer? – What did the term 'no good' mean to you? – Did you have difficulty recalling whether you were ill in the last 30 days? – How did you go about remembering you illness?"

Comprehension probing refers to questions asked especially about the original question's understanding.

Example (after Prüfer and Rexrodt, 2005):
Question: "To what extent do you agree to the following statement: 'My health in mainly a result of good disposition and luck?"
Probe after the test person has given his or her answer: "How do you understand the term 'luck'?"

Information retrieval probing aims to reveal certain aspects concerning the way how the information given in an answer was obtained.

Example (Prüfer and Rexrodt, 1996):
Question: "When have you been to the dentist the last time?"
Probe: "How difficult did you find it to answer this question?" or "How did you remember your last visit to the dentist?"

C) Confidence ratings

With confidence ratings the test person has to assign the degree of reliability to her or his answers, usually with the help of a given scale. Confidence ratings attempt to identify questions that test persons find difficult to answer by having them rate their level of confidence in the answer they have provided. The theory is that low confidence ratings are often the result of a lack of knowledge or a difficult recall task.

Example (Prüfer and Rexrodt, 1996):
Question: "For how long altogether did you watch TV during the last seven days?"
Confidence rating: "What would you say is your answer 'very precise' – 'rather precise' – 'rather inaccurate' – or 'a rough estimate'?"

D) Paraphrasing

Paraphrasing means that after answering, the test persons have to repeat the questions in their own words. This permits to examine whether the test person understands the question and interprets it in the intended manner. Paraphrasing may also reveal better wordings for questions, for example, if different test persons consistently use the same terminology (DeMaio and Rothgeb, 1996). There are two main test person strategies to paraphrasing: either it is tried to remember the question word by word or the question's context is repeated in own words. Usually it is easier to assess whether the test person understood a question if the repletion is given in own words.

Paraphrasing is especially useful to reveal complex and/or confusing questions. Additionally problems can be revealed when respondents can not remember all important details e.g. time periods the question refers to (Ehling, 1997).

> *Example (Prüfer and Rexrodt, 1996):*
> *Question: "In comparison with other people living here in Germany: do you think you receive a fair share, more than a fair share or less or substantially less?"*
> *Request after answering the question: "Now please repeat the question I just read out for you again in your own words. What was the question?"*

E) Sorting

This method aims at providing information about how test persons categorise terms or understand the underlying concepts. The test persons are given a set of objects that are arranged in no particular order and are asked to sort them according to their own or given criteria. According to Brewer and Lui (1995) sorting procedures provide an efficient method for assessing perceived similarities between a large number of stimulus objects within a particular domain (for example visual patterns, social institutions, political policies, consumer products and so on).

The analysis of the resulting subjective groups reveals information on how test persons cognitively represent and organise their knowledge in that domain but it may also provide a means for assessing differences between respondent populations in such cognitive representations.

The results of sorting can orient decisions about the structure and ordering of questions to facilitate efficient and reliable retrieval of relevant information. Perhaps even more important is the potential applicability of comparative sorting analysis to understand differences in the interpretation of questions between different respondent populations or between different survey questionnaires.

Sorting can be used with any form of stimulus objects – from verbal labels or statements (printed on index cards) to visual representations and photographs as well as physical objects. Thus, this technique is not limited to verbal representations or assessments of semantic knowledge. Judgements obtained from sorting do not rely on test persons' ability to verbalise the bases for their assessments of similarity or to represent their decisions verbally.

Free Sort. With this technique test persons group a given set of discrete randomly presented objects according to their own criteria into any number or type of groups without constraints. The free sort methodology has the advantage that it can be used even when the stimulus set contains a large number of objects.

Dimensional Sort. It differs from the free sort in that grouping occurs according to previously defined criteria.

Example 1 (Biemer and Lyberg, 2003):
Respondents might be asked to sort certain given topics (e. g. health, income, activities, family structure) to evaluate which topics are considered sensitive by the respondents.
Free sort: if respondents can order the topics in a way they want
Dimensional sort: if respondents are asked to order the topics according to the criteria 'most desirable' to 'least desirable'.

Example 2 (Brewer and Lui, 1995):
To develop an efficient and reliable probing strategy to elicit information about chronic illnesses with the aim to reduce a long list of about hundred (chronic) conditions a sample of respondents who had personal experience with at least one chronic condition were asked to sort the chronic illnesses in the list into categories.

F) Vignette classifications

Vignette classifications could be regarded as a form of dimensional sort but also as a technique of its own. Test persons are given certain descriptions as vignettes to determine by themselves whether or not these have to be included into the answering process, or to decide if they relate to a particular survey question or concept. A hypothetical vignette may describe a certain behaviour or activity of a fictitious person and the test person is asked: "How should the person described in this scenario respond to the following question?".

The main objectives of vignette classifications are to test consistency of respondents' interpretation and to evaluate the effects of different alternative questions on the interpretations of survey concepts. Additionally, by determining the wordings that resulted in the most correct classifications of the vignettes by the test persons one can determine the best question wording. However, the disadvantage of the vignette approach is the artificiality of the situation imposed on the test person.

Example (after Biemer and Lyberg, 2003):
This example's purpose was to compare alternative wordings of a question about drug abuse.
Question: Have you ever, even once, used a drug that is available only by a doctor's prescription and that was (1) either not prescribed for you OR (2) taken for the experience or feeling that it causes?
Vignette1: J.G. has had some teeth pulled and the dentist gives him a prescription for Tylenol with codeine. He is supposed to take two pills every four hours, but because his mouth hurts so badly, he takes three pills every four hours. For J.G. what would be the correct answer to this question?
Vignette2: G.W. has become addicted to cough syrup. He goes to different doctors all over town with a fake cough, and gets a number of prescriptions for cough syrup. For G.W. what would be the correct answer to this question?
…
The vignettes are followed by the instruction to make a cross in the 'YES' box if the respective persons used the drug without a doctor's prescription and make a cross in the 'NO' box if the respective persons did not use the drug without a prescription.

The use of vignette classifications is considered here as a pre-field testing method, due to its cognitive nature. However, it has to be noted that applications exist as a field testing technique as well. The main advantages of conducting vignette classifications in the field are that i) the context validity is saved (i.e. to grasp respondents’ interpretations in the real survey setting rather than in the laboratory environment) and ii) probability samples can be used instead of convenience samples.

The methodological advantages of vignette classifications in the field are illustrated by an experience conducted for the U.S. Current Population Survey (CPS) estimating unemployment in the U.S. (Campanelli *et al.*, (1989) reported in Martin, 2004). Vignette classifications were used to figure out the test persons’ interpretations of specific questions and to evaluate alternative questionnaires based on these interpretations. First of all, a set of vignette classifications portraying irregular employment situations were administered to about 2300 test persons in a computer assisted telephone interview, just after the final CPS interview. The vignette classifications were also submitted to the interviewers. The findings highlighted that key questions did not adequately communicate the intended meaning of important concepts. When the CPS questionnaire was reviewed, vignette classifications were again used in a split-sample comparison between the old and the new questionnaire for the purpose of evaluating the new instrument.

G) Response Latency

Response latency is the time measured between question and answer either by a stop watch or qualitative estimation (named qualitative timing). Response latency methods have been used to identify problems in the interpretation of questions, memorial retrieval and response selection.

However, it has to be taken into account that the process of answering a question, namely understanding the question, cognitively processing the question, assessing possible answering alternatives until the answer is given, is individually different. Accordingly the effectiveness of the response latency method depends, to a large extent, on the techniques that are used to analyse the response-time data. To reduce this problem one can produce a baseline measure by using the time test persons take to respond to straightforward questions.

> *Example (Biemer and Lyberg, 2003,* citing Bassili, 1996*):*
> *In a study it became obvious that respondents needed very much time to answer a question with a ‘don’t know’. The interpretation of this observation is that a longer response latency followed by a ‘don’t know’ may be indicative of a true ‘don’t know’ rather than a ‘don’t know’ that is given in lieu of refusing to provide the requested information (i. e. a hidden refusal). This type of reasoning can be used to assess the strength of attitudes or opinions.*

In conclusion the results of response latency studies can be quite useful for identifying potential problems in the questionnaire. However, extensive follow-up of the results may be required to verify the problems suggested by the latency results and to identify an appropriate remedy. Accordingly, this technique is probably working best when combined with one or more other cognitive interviewing techniques.

H) Potentials and weaknesses of cognitive techniques

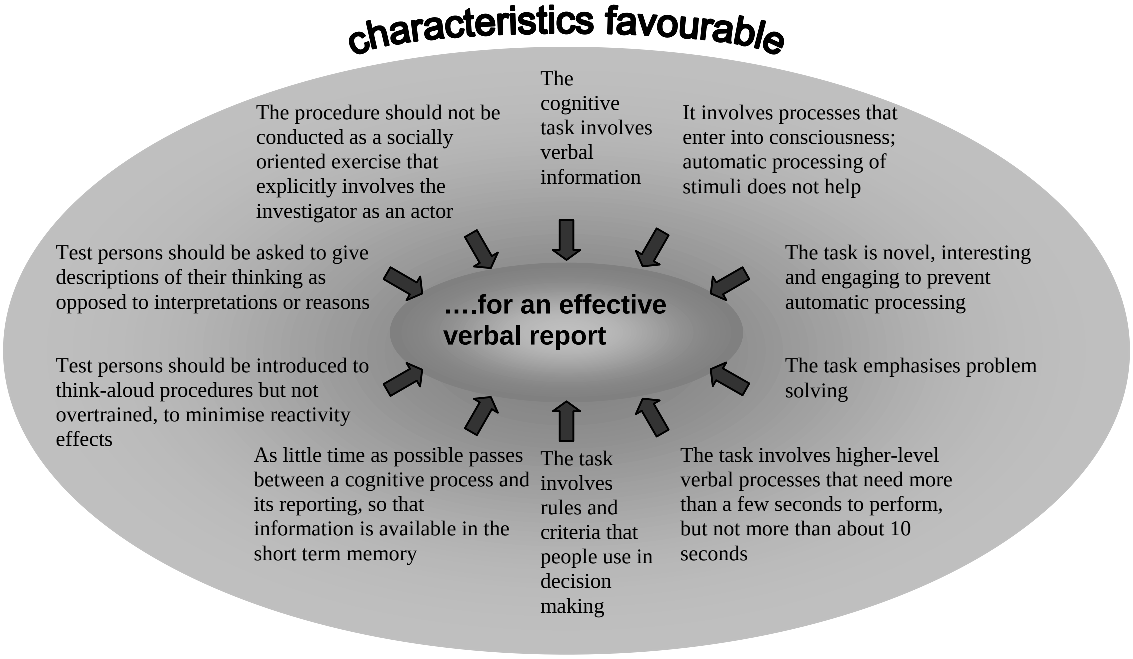
The theoretical background of cognitive processes originates in psychology where think aloud techniques were used to elucidate cognitive processes for problem solving tasks. The very strength of cognitive techniques is that they provide information about the existence of a problem in a question, its possible source(s) as well as information toward the problem’s solution.

However, the transferability of some techniques to the survey testing environment might be problematical, for example when respondents are not capable to fulfil the think aloud “task”. There are very different experiences concerning verbal reports using the think aloud technique in different countries. While this technique seems to be widely and successfully employed in North America it yet did not seem to work well in Germany, for example (Prüfer and Rexrodt, 2005). This suggests that the capability of test persons might indeed have a cultural component as well.

Acknowledging that more effort has to be paid to the basic assumptions of cognitive interviews, some recent studies are now discussing the theoretical background and fundamental prerequisites of cognitive interviews. These and future works have great potential with respect to the standardisation and tailoring of cognitive interviews for specific tasks.

One approach is to recognise the importance of the test person’s verbal reports (as the basic results of cognitive interviews) about their cognitive processes. A comparison of the theoretical basis of the original application of verbal reports in experimental psychology to their application for survey testing reveals the need for more research to provide a thorough theory as a fundament for the application of verbal report methods for cognitive interviewing. Without such a basis it needs to be carefully considered if certain preconditions which apply to verbal reports (as summarised in figure 6.4) are satisfied in cognitive interviews (Willis, 2004).

Figure 6.4. Preconditions of cognitive tasks and verbal report procedures (simplified from Willis, 2004)


characteristics favourable
The procedure should not be conducted as a socially oriented exercise that explicitly involves the investigator as an actor
The cognitive task involves verbal information
It involves processes that enter into consciousness; automatic processing of stimuli does not help
Test persons should be asked to give descriptions of their thinking as opposed to interpretations or reasons
….for an effective verbal report
The task is novel, interesting and engaging to prevent automatic processing
Test persons should be introduced to think-aloud procedures but not overtrained, to minimise reactivity effects
The task emphasises problem solving
As little time as possible passes between a cognitive process and its reporting, so that information is available in the short term memory
The task involves rules and criteria that people use in decision making
The task involves higher-level verbal processes that need more than a few seconds to perform, but not more than about 10 seconds

For example, one precondition is that the verbalisation of thoughts does not fundamentally alter the thought processes. However, due to the response burden of think alouds in a cognitive interview the articulation of thoughts might have substantial effects on the thought process of the test person. An approach addressing this problem would be to investigate if and in which testing situations probes may trigger a more valuable verbal protocol and meet the above mentioned criteria closer than think alouds due to reduced respondent burden (for discussion on the techniques, see Willis, 2005).

On the other hand it has been shown that certain probes or ways of probing may “generate” some imprecision in the answers of a cognitive interview. Accordingly, a distinction between discussion oriented (elaborating) probes and probes that “push” test persons to provide codable responses in a manner similar to standard survey interviewing’ is sensible in certain test situations (Beatty, 2004). In the field, the uncertainty whether extensive field probing introduces measurement error is known. However, in the context of cognitive interviews any relationship between what interviewers do in the laboratory and what test persons report should be precluded as far as possible. Accordingly, an evaluation of the test person’s ability to answer questions and the quality of their responses should for example aim at getting the test persons to answer survey questions in the desired response format. Elaborating probes in this situation may make it very difficult to judge whether participants could not answer questions, or whether they simply did not (Beatty, 2004). An obvious conclusion is that cognitive interviews are substantially improved by training interviewers to recognise important distinctions among cognitive techniques, their potential impacts on the interview and the situations in which each ought to be employed.

However, not only the training of the interviewers with respect to the actual test but also with respect to producing valuable standardised reports is vital. Conrad and Blair (2004) suggest that the assessment of interviewer reports provides potential for improving the quality of the results of cognitive interviews. Due to the effect different techniques of cognitive interviews have on the validity and probably the reliability of the interpretation of the test persons' verbal reports, it is recommended that the quality of the interviewer reports (which in a way is the transcription of the test persons' verbal report) should be assessed in terms of two essential objectives: problem detection and problem repair. While problem detection addresses the effectiveness of a cognitive method in identifying problems, problem repair issues the assessment whether or not a problem reoccurs in subsequent testing. Accordingly improving the quality of verbal reports in future is an important step also towards improving the quality of data measured (Presser *et al.*, 2004).

In summary it can be concluded that a great deal for the further development of cognitive interviews lies in the initialisation of more theoretical background for cognitive interviewing which would also provide guidance on which methods from the "cognitive toolbox" may be more or less useful under specifiable conditions.

Most examples and discussions in the literature on cognitive interviews refer to household surveys. In establishment surveys the focus of measurement error might be different from the household survey situation. The challenge for establishment surveys is to find appropriate persons within an enterprise to respond to the respective questionnaire and to investigate how difficult it is for that person to provide the information asked for (e.g. Dillman, 2000).

Supposing it is possible to identify "the" right person in the enterprises, cognitive interviews may provide valuable insight if this person can provide the information requested, which questions are found difficult to answer, the reasons for that and how the persons understand and interpret a question. The results from cognitive interviews might be of special value for establishment surveys in identifying sensitive questions (like financial information) that could lead to nonresponse. The results of cognitive interviews may suggest if eliminating or changing these questions (e.g. categorising financial groups instead of asking precise numbers) would be an alternative.

Cognitive interviews could also indicate if a person, identified as the right person to fill in the questionnaire of the establishment survey, still needs support from other people in an enterprise to complete the questionnaire. If this is the case it might be considered e.g. to change questions to a format that enables the person to complete the questionnaire on his or her own.

With respect to organising the cognitive interviews for establishment surveys one also needs to take into account that the test person most likely will not come to a lab but that the interview has to take place in the enterprise. Further information about cognitive interviews in establishment surveys can be found in Ware-Martin (1999).

6.1.5. Experiences

The short survey conducted for the purpose of this handbook revealed the extent to which NSIs in Europe, North America and Australia have experience with cognitive interviewing and that these experience vary significantly. About 10% of the institutions who took part in the survey conduct cognitive interviews for most or all of their surveys. Forty-five percent of the institutions test some of their surveys with cognitive interviews. The other institutions (about 45%) do not conduct any cognitive interviews at all.

Two NSIs, the Australian Bureau of Statistics (ABS) (ABS, 2001) and Statistics Sweden (SCB) (SCB, 2004), have reported their experiences with cognitive interviews in a quite elaborate way. First of all, these experiences are summarised with some additional experiences in ZUMA (Prüfer and Rexrodt, 2005) and Statistics Netherlands (Snijkers, 2002). From these experiences and from the chapter above which discussed the potentials and weaknesses of cognitive techniques practical recommendations are drawn which aspects of cognitive interviews one should make allowance for when planning to conduct cognitive interviews.

In the institutions mentioned above cognitive interviews are commonly used as a testing method and the interviews are applied in a middle stage of the survey (questionnaire) development process when a profound draft of questionnaire has been developed. Ideally, before that stage, informal tests, focus groups and expert reviews should have influenced the questionnaire already, to achieve a certain standard. Additionally, with the help of the experience from these previous test phases some of the questions might already be identified as being critical. To test them, these cognitive interviews are particularly suitable.

Although cognitive interviews are considered to be costly and resource intensive in comparison with most other testing methods, they are also considered advantageous with respect to their capability of producing results very effectively. Some problem analysis may even occur during the interviews themselves (Presser and Blair, 1994; ABS, 2001).

Ideally, cognitive interviews should take place in a cognitive laboratory where they can be videotaped. Video tapes can be analysed acoustically and visual behaviour patterns can be taken into account in the interpretation of results. The questions to be tested should be presented to the test person in the same mode as the survey is conducted (Prüfer and Rexrodt, 2005).

It is very important to note that different or several rounds of interviews are regarded to be beneficial as an iterative process allows for immediate testing of solutions to problems. If there is only one round of cognitive interviews and a question is changed according to a problem detected there needs to be a validation if the question changed works better than the previous one. Such validations can only be done if the testing plan provides for subsequent cognitive interviews.

The maximum time limit of an interview is reported to be one or one and a half hours and about 15 interviews per interviewing round are reported to be sufficient. Due to this time limit the concentration on key items or questions is strongly recommended.

With respect to the recruitment of test persons a focus should be on the choice of “representative” groups that are similar to the actual respondents or respondent groups in the survey.

It is vital that the cognitive interviewers obtain special training at minimum. Ideally they need to be very experienced in interviewing and should possess the capability of appraising the effect of their questions/ interference, their social interaction with the test person and – in most cases – they need to provide a detailed valuable report of any interview. With respect to the quality of the interviews’ reports it is very beneficial if the interviewer also has knowledge about the concepts of the survey in question as well as the capability to process the qualitative data he or she just collected via the interview.

Any form of probing seems to be used extensively. However, according to the specific questions, different methods of cognitive techniques are appropriate. Figure 6.5 summarises the different cognitive techniques discussed in the previous chapters to assist the choice of methods. However, when deciding on cognitive interviewing techniques, it has to be taken into account that they are employed in an artificial environment and/or situation (laboratory like room). Consequently, probes, queries and the style of interviewing used may elicit cognitive processes that are not part of regular interviewing and might therefore not be replicated in a cognitive interview (Biemer and Lyberg, 2003).

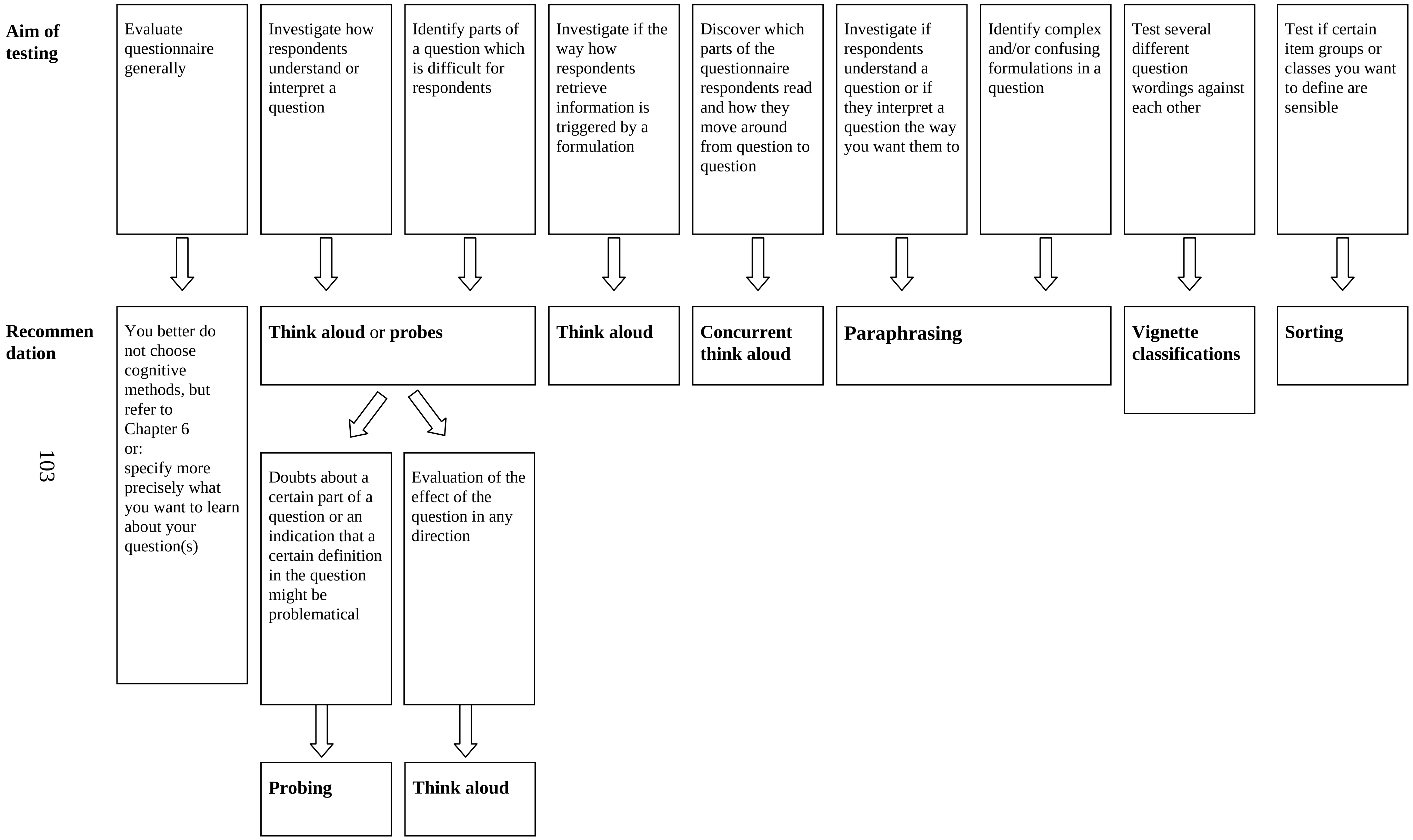
There should be a protocol for each set of cognitive interviews which outlines the specific probing and other supplemental questions. The protocol should be developed by means of an expert appraisal conducted by the professional research staff assigned to the project. The staff member’s expertise is needed to suggest potential questionnaire problems and develop a protocol that elicits information and allows evaluation of the extent to which these or other problems are encountered by test persons. It is recommended that the protocol should be a common denominator of issues to explore and as strict as necessary but it needs to leave the interviewers enough freedom to react appropriately to any situation during the interview that could not be foreseen beforehand e.g. when the test person encounters any problems other than the ones suggested before (e.g. Snijkers, 2002). Accordingly, the balanced application of scripted protocol questions which ensure some standardisation and spontaneous questions which allow some flexibility to investigate problems or situations which were not foreseen should be considered as crucial.

It is advisable to leave space in the protocols so that the interviewers can note any verbal report from the test person to the appropriate instruction. In this way the protocol changes into a standardised report which can be further processed. As the reports are crucial for any analysis and the conclusions drawn from the cognitive interviews it is vital to determine their form and level of detail in advance and communicate these decisions to the interviewer. It is very important to note that complete, somehow standardised reports are indispensable because they are the only means to differentiate between actual results and their interpretation.

To achieve a certain level of standardisation it is useful to establish a routine to enter the results of the cognitive interview, documented in the protocol, directly after the interview into a standardised database. Even if there is only a limited number of cognitive interviews, the results need to be documented and stored in an appropriate, sufficient way, e.g. by building up a database on results.

Whenever cognitive interviewing is implemented, it should not be used as a stand alone technique but its indications about problems or solutions should be cross-checked via other supplementary field testing techniques following the cognitive interviews, like debriefings.

Figure 6.5. An approach of searching for the appropriate cognitive method


Aim of testing
Evaluate questionnaire generally
Investigate how respondents understand or interpret a question
Identify parts of a question which is difficult for respondents
Investigate if the way how respondents retrieve information is triggered by a formulation
Discover which parts of the questionnaire respondents read and how they move around from question to question
Investigate if respondents understand a question or if they interpret a question the way you want them to
Identify complex and/or confusing formulations in a question
Test several different question wordings against each other
Test if certain item groups or classes you want to define are sensible
Recommen dation
You better do not choose cognitive methods, but refer to Chapter 6 or: specify more precisely what you want to learn about your question(s)
Think aloud or probes
Think aloud
Concurrent think aloud
Paraphrasing
Vignette classifications
Sorting
Doubts about a certain part of a question or an indication that a certain definition in the question might be problematical
Evaluation of the effect of the question in any direction
Probing
Think aloud

6.1.6. New technologies

With the development of new information technologies new possibilities emerged for testing questionnaires too. At the beginning e.g. it was by the implementation of CATI instruments that audio recording became less disturbing and easier to handle. It allowed to check the question answering process orally. More recently CAPI instruments have been enlarged in their functions: An interviewer conducts a CAPI interview using a laptop computer, which also serves as a digital tape-recorder in order to check actions and reaction during the interview. This technology is often used in the U.S. **(**Biemer *et al.*, 2000). In addition there are linguistic software packages available (see QAID/QUEST).

However, nowadays the major challenge is the observation of either the interviewer, the respondent or the screen. Via adequate software packages even the storage and the combined evaluation of all sources are possible. However, these methods are not yet applied as daily routines in most institutions, but a lot of explorative studies and research push the development forward, so that those methods might be the future. The technical required tools are various and are not presented here in detail, but references are given for each technique.

A) Computer based tools to check linguistic problems

QAID/QUEST is a software tool that identifies possible problems of respondents to understand a question, thus checking the questionnaire basically. QAID reviews survey questions via classifying imprecise, unfamiliar terms, vague or ambiguous noun phrases, complex syntax, or working memory overload. It is based on a cognitive computational model (ABS, 2004b; Graesser *et al.*, 2000). There is another tool, searching large bodies of text to identify the generalities of language as used by different speakers and writers. It identifies the co-location of a word with other words, the context in which a particular word is used and the grammatical frames in which the word occurs. These searches suggest sources of potential comprehension problems due to question wording (ABS, 2004b; Graesser *et al.*, 1999). At the moment these tools are seldom used internationally as they are very sensitive with regard to translation of terms and terms in use, as they are always linked to the cultural background of each country, but the language of origin of the software is American English.

B) Observing interviewers, respondents and screen

Originally observation was applied when self-completed questionnaires were the subject of testing. Trained observers watched the survey process (e.g. form completion or response within an interview) to better understand the respondent's thought processes during questionnaire completion. It was and is also to identify problems in wording, problems in question order, presentation or layout, and to estimate the time taken to complete the questionnaire or parts of the questionnaire. Survey designers look for behaviours that result from an error of the instrument, including the participant's behaviour (e.g. reading all the questions before responding), non-verbal cues, reactions and observed cognitive processes (e.g. counting on their fingers or writing calculations on the page). This technique can also use follow-up probes to elicit information about why the respondent behaved as he or she did (ABS, 2001).

Due to new technological possibilities, observation has enlarged and offers even more: besides audio-recording the behaviour of either of the two people in conversation (interviewer and respondent), the interaction of the respondent with the computer screen can be observed. As already described in chapter 5 (usability testing), it is even possible to use both audio-recording and video-recording at the same time. The idea behind is that verbal behaviour, facial expressions and similar reactions can tell a lot about comprehension problems. There are different tools to record visual reactions: a) *screen-capture devices* record e.g. the movements of the cursor, skip movements etc. for computer assisted self-completion questionnaires, b) *audit trails* record the computer screen via a fixed video camera c) the verbal conversation and the interaction between interviewer and respondent is recorded by video plus tape-recording and d) *eye-tracking devices* video-record the face of the respondent (Beukenhorst, 2004; Redline and Lankford, 2001).

C) Computerised evaluation of questionnaires

Additionally research has been started in using computerised evaluation by adding evaluation questions at the end of the questionnaire, to be implemented when CASI questionnaires are implemented (Beukenhorst *et* *al.*, 2002). The study has been conducted to verify the potentials and weaknesses of such an approach. Expected benefits were:

- Self-administered evaluation offers access to a broader and larger group of test respondents in a cost-effective way, since there is no need for face-to-face contact.
- Self-administered meta questions can be presented to the respondents in the same context as the survey questions, which is expected to make it easier for them to remember specific problems.
- Self-administered evaluation does not disrupt the interview, which represents the more realistic situation for CASI questionnaires.
- All test respondents receive the meta questions in a similar way; there are no interviewer effects.
- Especially for questionnaires on sensitive issues, self-administration is an attractive mode; respondents do not have to feel embarrassed towards an interviewer asking sensitive questions.
- Vignettes can be used as a tool for evaluating questionnaires very well in a self-administration situation. Respondents can take their time evaluating the described situations.
- Respondent behaviour can be observed (registered) automatically, using so called audit files. Examples include registering navigation through the questionnaire, response times and corrections made by respondents.

Findings on the feasibility to use such tools were positive and revealed the following:

- Respondents are willing to fill in evaluation forms.
- Self-administered evaluation forms yield useful results.
- The quantity and quality of problems found are strongly linked to the quality of meta questions asked. Especially unexpected problems are definitely difficult to detect (which is exactly the problem with cognitive methods and probing techniques).

6.1.7. Recommendations

- It is recommended to test potential questionnaire problems like visual design, wording, terms, skip instructions and answering categories by a mix of different pre-field methods before conducting a field test.
- Different pre-field methods are recommended in the sense of catching different perspectives. Thus checking the instrument with subject experts, users, respondents, data processing staff, etc.
- An informal test can already provide useful information on shortcomings of the instrument at an early stage of questionnaire design but it should not be chosen as the only pre-field test method because it might not detect all problems.
- An expert review should be conducted to ensure that the questionnaire collects the information needed to meet the analytic objectives of the survey and to check the design, the questions and potential processing problems.
- To achieve the experts' judgement and interpretation in a structured way, it is highly recommended to conduct structured expert reviews.
- Cognitive interviews should be conducted to investigate if survey respondents understand questions in the way intended by survey designers and if it is feasible for the respondents to recall and provide the necessary information.
- Cognitive interviewing comprises a suite of techniques which should be combined in a flexible, iterative process with the option to compile different rounds of interviews.
- It is vital to note that all introduced pre-field test methods are qualitative in nature. They should be conducted before any quantitative tests or field tests are done but they can never replace the latter. Accordingly, pre-field tests should be regarded as a part of the preparation for the field in a well balanced mix of different pre-field and field methods.

6.1.8. Checklist

Informal Testing

- To conduct an informal test gather colleagues from different departments who will get in contact with the questionnaire during any stage of the survey process to get a first feedback on the instrument.

Structured expert reviews

- Questionnaire design experts, subject experts (e.g. possibly clients) and probably data processing experts should be gathered.
- Standardised coding schemes should be developed that dwell on each question of the questionnaire in a concrete manner
- Decide on the time schedule for: sending out the questionnaires, feedback, preparing final results and discussion of the results
- Decide on the composition of the expert group
- Select 2-4 experts (subject-matter and design experts)
- Provide a questionnaire with the coding categories for the structured interviews
- Send the questionnaires out to the experts
- Review the filled in questionnaires
- Check for equal and different feedbacks on questions
- Summarise the findings and if possible conduct an expert meeting on the results. Check especially contradictory statements

Unstructured expert reviews

- Questionnaire design experts, subject experts (e.g. possibly clients) and probably data processing experts should be gathered.
- Define aims of the expert group discussion
- Establish time schedule and location of the reviews
- Decide on the composition of the expert group
- Develop a discussion guideline
- Check tape and video-recording facilities
- Define moderator of the discussion
- Check staff available for preparing final results of the expert reviews

Cognitive interviews

- Have a good idea about what you want to test in particular
- Be sure that cognitive interviewing is the appropriate method for that test
- Check if cognitive interviewing is applicable for the situation to be tested, e.g. whether the preconditions for verbal reports are given
- Decide which cognitive methods to choose from the cognitive tool box
- Decide on the manner of the interview (e.g. discussion oriented or elaborating probes)
- Plan the number of interviews per round, the number of test persons and the length of the interviews.
- Choose a suitable lab for the interviews
- Decide on the role of the interviewers and the degree of their flexibility allowed in the interview
- Prepare a protocol
- Prepare a database for storing the results of the protocol and the video tape
- Select and train interviewers for the specific purpose – remember to train interviewers to recognise distinctions between different kinds of probe (cognitive probes, confirmatory probes, expansive probes, functional remarks and feedback) and the situations in which they are suitable
- Decide on the form and level of detail of the report and communicate it to the interviewers

6.2. Field testing methods

Field techniques are implemented in order to test the instrument under field conditions. It corresponds to switch from the artificial setting of testing in a laboratory or office to real survey environment. This means that the questionnaire is administered in the same way or very similar way to the fieldwork (regarding setting, lengths, choice and order of questions etc).

Testing the questions in the field is a crucial stage in the development and assessment of survey questionnaire along with the development and testing of associated procedures and instructions. No matter how much work has been done, the questionnaire has to be tested under field conditions, in particular, the questions have to be tested in the mode they will be used. Field methods can be applied either in field test, in pilot surveys or in ongoing surveys.

Whereas field test can be restricted to the test of the questionnaire and the mode of data collection, pilot surveys are conducted to check all survey design characteristics, e.g. sampling methodology, target population, interviewer training, fieldwork procedures, data processing, estimate costs, etc. (see introduction on testing methods). Field tests are usually implemented by a larger number of units than pre-field tests. However, the scale of testing units is still smaller than for pilot surveys. Pilot surveys are used to conduct a final test of the questions before moving to the real survey: it is in the dress rehearsal that a final check of the procedures and instructions can be tested and evaluated.

There is a range of field testing techniques available for questionnaire designers which meet different purposes. They include, among others, behaviour coding of interviewer/respondent interaction, interviewer debriefing, respondent debriefing, follow-up interviews, and experiments.

6.2.1. Behaviour coding

Behaviour coding is a technique that consists on the systematic classification of interviewer/respondent interaction in order to evaluate the quality of the questionnaire. This method can be conducted in the field as well as in the laboratory. However, it is generally undertaken in the field and regarded as a field testing method. Interviews are either tape-recorded or live-coded. The reactions of interviewers and respondents are classified and registered systematically by the assistance of a predetermined coding system, so that frequencies on reactions to each question can be registered, highlighting the occurrence of problems with administration, wording, understanding and completion of questions. After several interviews, methodologists can identify problems related to specific questions, detect the need of revising the interview guidelines and other tasks associated with interviews. However, criticism on behaviour coding still remains as it grasp problems with questions, but does not catch anything about the causes of the problem and, thus, possible solutions. Some important problems, such as respondent misinterpretations are largely hidden because respondents and interviewers tend to be unaware of them. In addition it is a rather costly and time consuming method. Perhaps the most important reason for expanding the use of behaviour coding is that it is a more systematic, objective, quantitative, and representative means of evaluating survey questions relative to other methods such as cognitive interview and expert review (Esposito *et al.*, 1991).

Mode and coding scheme
Behaviour coding is a method used for either telephone or face-to-face interviews. Coding can be done either during the interview (using CATI, interviewer can code in live time or like “problem coding” as explained later) or afterwards by special trained coders over tape-recorded interviews.

In face-to-face interviews, additional information on the interaction process might be obtained, e.g. gestures or mimic of interviewers or respondents can be coded. However, the high costs of the procedure and the possible context effects make it rather seldom applied.

The coding scheme, presented in table 6.6. below, has been used by several researchers and only recently it has been expanded using a more sophisticated system by taking into account the model of question-answering process and final coding.

Table 6.6. Behaviour coding scheme by Oksenberg *et al.* (1991)

Abbreviation	Indicator	Comment
Interviewer question–reading codes		
E	Exact	Interviewer reads the question exactly as printed.
S	Slight change*	Interviewer reads the question changing a minor word that does not alter question meaning.
M	Major change*	Interviewer changes the question such that the question meaning is altered. Interviewer does not complete reading the question.
Respondent behaviour codes		
1	Interruption with answer*	Respondent interrupts initial question-reading with answer.
2	Clarification*	Respondent asks for repetition or clarification of question, or makes statement indicating uncertainty about question meaning.
3	Adequate answer	Respondent gives answer that meets question objective.
4	Qualified answer*	Respondent gives answer that meets question objective, but is qualified to indicate uncertainty about accuracy.
5	Inadequate answer*	Respondent gives answer that does not meet question objective.
6	Don't know*	Respondent gives a "don't know" or equivalent answer.
7	Refusal to answer*	Respondent refuses to answer the question.

* Indicates a potential problem with the question.

To apply the above scheme the coder listens to an interviewer reading the question and to the respondents' answers, then assigns one or more codes, e.g. interviewers make a major change (M) and respondent interrupts with answer (1). The coding procedure is applied in the same way for each question. Codes can be recorded on coding sheets that include the code categories along the top and the question numbers down the left margin. Coding is then accomplished simply by checking a box under the appropriate code. A less expensive and more effective method is to code directly into a computer database, which allows the results to be tabulated at any point during the operation.

Staff and training /number of interviews

Behaviour coders should go through basic interviewer training in order to understand the interviewing process itself. With adequate training, coders can code tape-recorded interviews with high degree of consistency. Short extracts of recorded interviews are played and discussed until general agreement is obtained. Two or three coders are usually in charge of the task. Following this, segments of interviews are independently coded and results are compared. Coding from live interviews is usually less reliable and less detailed, as the interviewing process can not be replicated and checked once more.

There is a discussion on the amount of behaviour coded interviews which can be of use for evaluation. It varies between 30 to 100 interviews. However, recent methodological research on the sample size of tests using behaviour coding advices a smaller size of 30 or even less. The necessary number of interviews is also a function of the number of coders (for an evaluation regarding the reasonable number of interviews, see Zukerberg *et al.*, 1995).

Example on the analysis of the results

Using codes for each questions like suggested in table 6.6., the following statistics can be analysed:

1) Percentage of correct wording of the question
2) Percentage of questions, where respondent needed clarification
3) Percentage of inadequate answers
4) Percentage of adequate answers

Table 6.7. presents an example of results from a hypothetical behaviour coding:

Table 6.7. Behaviour codes for two questions (percent)

Interviewer behaviour		
Code	**Question 1**	**Question 2**
E: Exact	82	54
S: Slight change	12	23
M: Major change	4	23
Respondent behaviour		
Code	**Question 1**	**Question 2**
1: Interruption with answer	2	17
2: Clarification	9	12
3: Adequate answer	81	38
4: Qualified answer	0	3
5: Inadequate answer	5	21
6: Don't know	2	8
7: Refusal to answer	1	1

Examining the table it appears quite clear that question 2 is more problematic than question 1 and should be revised. In particular, only 54% of interviewers read question 2 exactly and only 38% of respondents answer it adequately, whereas the correspondent percentages for question 1 are 82% and 81%. However, also question 1 should be focussed on, since for problem identification it is commonly agreed to focus on questions in which at least 15% of the responses showed problems (cut-off 15%).

Method advancements, evaluation and studies

From the literature, behaviour coding reveals a sizeable number of question problems, stores them in a systematic, standardised manner and is mainly conducted under real conditions. Its methodological limitation is related to the lack of information on the underlying causes for problems. Consequently, researchers tried to enhance the method by different tools like cognitive interviews, debriefings with either respondents, interviewers or coders, focus groups, etc. Some examples are presented below.

- Oksenberg *et al.* (1991) proposed to upgrade the method by conducting special probes with respondents after the interview and by debriefing coders and interviewers about their interview experience.
- Prüfer and Rexrodt (1996) implemented a kind of "problem coding", which is done during the interview. Live coding is performed by the interviewer, using two codes to classify respondents' reactions: adequate and inadequate. After the interview, the interviewer writes a report on questions with inadequate reactions in order to get an idea on the cause of problems. Results are presented by frequencies on the answering categories of adequacy/inadequacy and possible reasons via reports. They assess their method as direct, straight coding and helpful. However, it has to be noted that live coding is very demanding for the interviewers and requires additional training.
- Sykes and Morton-Williams (1987) conducted follow-up interviews with respondents after behaviour coding. Through a semi-structured interview, questions of special interest were scrutinised by asking questions on how respondents retrieved their answer. The method gave a better assessment on the causes of problems with specific questions. However, the researchers proposed a less costly testing by using a smaller number of codes and by limiting the interviews to about 100.
- Van der Zouwen and Smit (2004) implemented a more sophisticated coding scheme oriented by the question-answering process (15 codes). Frequencies on occurred problems could be presented. Coding was done by the researchers themselves and was assessed as less time consuming and more reliable.
- Beukenhorst (2004) proposed a coding system, using verbal and video observation and audit trails. Using the appropriate software all information can be linked and stored in one file/database.
- Burgess and Paton (1993) implemented behaviour coding using CATI, coding only respondents behaviour (see below).

Specific recommendations for CATI and CAPI surveys

There are a few studies on behaviour coding in CATI surveys. In general it seems rather logic and easy to apply behaviour coding to a CATI instrument, since tape-recording is easy to implement.

Statistics Canada (Burgess and Paton, 1993) studied the test conducted to a survey about violence against women, using CATI. In order to reduce costs and time, behaviour coding was done by the interviewers during the interview. Codes were limited to six items, four of them as introduced by Oksenberg *et al.* (1991) and the other two related to critical length of the questionnaire. It proved to be inexpensive, easy to apply and useful during the redesign. Additionally, discussions with the questionnaire development team and interviewers helped to identify problems and were essential. The results of behaviour coding could show which questions were problematic and to which extent. This was extremely useful in directing the team away from problems, whose perceived importance was not supported by the behaviour coding. The importance of the other methods of gaining insight was in trying to identify the nature of problems with the questionnaire. It was noticed, as weakness, that interviewer behaviour could not have been observed.

Experiences

The Current Best Methods on QDET of Statistics Sweden advise that, in telephone surveys, the interviewer, during the interview, should code the respondent behaviour. They report the general coding system proposed by Statistics Canada. In face-to-face interviewing, they point out that the gains of using an observer or expert personnel are relatively marginal compared to the high costs associated with their use.

From the survey conducted for the purpose of this handbook, it resulted that more than 40% of the interviewed institutes do not adopt behaviour coding at all; those adopting the method, mainly apply it only to some surveys (34.5%). Only 3.4% of the institutes adopt behaviour coding in all the surveys. However, as expected, it is never used in the field of business surveys.

6.2.2. Interviewer debriefing

Interviewer debriefing is a testing method applied to interviewers, who conducted the interview, in order to obtain their useful feedback on the performance of the questionnaire. Interviewer debriefing is a useful technique either to new surveys or ongoing surveys. It provides a better understanding of questionnaire weaknesses and captures the perspectives of interviewers and, in an indirect way, reveals some of the difficulties experienced by respondents.

Methods and Tools

There is a diversity of ways to gather information from interviewers. The most frequent method is group discussions after fieldwork has been done (focus group style) in order to obtain information about interviewers experiences in administering the questionnaire. To ensure that the discussion covers relevant questionnaire problems, an outline should be prepared to guide the discussion. Interviewers may receive the outline in advance to help them organise their thoughts and prepare appropriate comments. Digressions from the outline are to be expected, however, group atmosphere should also stimulate new ideas and topics for discussion. In this case, groups should not exceed approximately 12 persons. Individual interviewer debriefing can also be conducted but group debriefings are more common.

When interviewers are geographically scattered, debriefing could be done by telephone. In this case, using structured questionnaires, a group of interviewers participates in a telephone conference. This method keeps the benefits of group discussion (motivation and stimulation of new ideas and topics) and reduces survey costs.

Debriefings may be done daily or a single debriefing may be conducted at the end of the test. Sessions are frequently tape-recorded for later analysis, nevertheless, even when the session is tape-recorded a designated person should take notes during the debriefing.

Another debriefing method consists of requesting interviewers to fill in a standard interviewer debriefing-questionnaire for each or for all interviews, to gather information about their sensitivity on a problem, its frequency and sources, as well as the proposed solutions. This method can also be conducted to find out the extent of particular types of problems and to test the interviewers' understanding of questionnaire concepts.

When interviewers are geographically spread, interviewer debriefing can be done also by e-mail, using short semi-structured questionnaires.

Finally, rating questions is a method that consists on asking interviewers to rate each question in the test questionnaire on particular features of interest to the researchers, thus acquiring information about probable interviewer’s biases and perceptions of the respondent attitudes and behaviours. With rating forms more quantitative information is obtained.

According to U.S. Census Bureau (2003), interviewers may not always be precise reporting certain types of questionnaire problems for several reasons. When interviewers report a problem, we do not know whether it was troubling for one respondent or for many. Interviewers’ reports of questions’ problem may reflect their own preference for a question rather than respondent confusion. Also, experienced interviewers sometimes change the wording of some questions to make them work, and may not even realise they have done so.

Debriefings are classified as medium cost intensive, low staff intensive and can be conducted in about approximately two weeks. The staff involved is mainly composed by researchers, who have the task of leading the discussion. It is certainly better to structure the discussion in advance (see e.g. expert groups), defining aims of the discussion and to invite people at the beginning to give a “free comment” before going into details and structuring the discussion. The range of information obtained from such discussion is rather narrow, linked to interviewer problems and less to respondents’ problems (ABS, 2001).

Experiences
For the Australian Bureau of Statistics this method is commonly used and considered of high value. However, it is rather used after field test than at the beginning of testing (ABS, 2001), in its assessment it is regarded as a technique which “… provides detailed information about sources of interviewer error and some limited information about respondent errors. Debriefings can identify potential issues with ease and consistency of administration and sensitivity to interviewers as well as provide some information about perceived respondent sensitivity. They can also provide limited information about perceived respondent burden” (ABS, 2001).

In the Measurements Laboratory of Statistics Sweden (Henningsson, 2002), the interviewers have an internal e-mail system that enables quick and efficient communication. Experienced interviewers, scattered all over the country, are debriefed by e-mail. Whenever they find questions *difficult to understand*, *to answer* or *problematic in any other way* for the respondents, they should use 2-3 hours to:

- write their comments, starting with the more problematic questions ordered in a top down list with the most important things first.
- write their comments on survey instructions and introduction letter.
- provide their opinion on the whole survey, and suggestions for improvements.
- ask any other questions about the survey.

It is important that interviewers are trained to observe respondent’s reactions, and gain experience from the survey (doing at least 20 interviews), before they start debriefing.

From the results of the survey conducted for the purpose of this handbook, interviewer debriefing is widely adopted in the respondent institutions, in particular in social surveys as compared to business ones.

6.2.3. Respondent debriefing

The main objectives of respondent debriefing are to obtain quantitative and qualitative information about respondents’ interpretation of survey questions, to find out if concepts and questions were understood by respondents as questionnaire designers planned, and to identify the data sources that respondent consulted to answer survey questions. An additional aim could be to find out how respondents understand the scope of the survey.

In order to obtain sufficient information to evaluate the extent to which collected data are consistent with survey definition, debriefing should reveal if concepts or questions were interpreted as intended. Some of the survey questions may need to be modified or even dropped. On the other hand, it could be necessary to include further questions in the final questionnaire**.**

Methods and Tools

Respondent debriefing consists on a structured follow-up by means of an interview or self-administered questionnaire, that can be conducted directly after the interview or later on by re-contacting the respondent. Nevertheless, if carried out immediately after the interview, the respondents will still remember their answers, and re-contacting costs will be saved.

To developed good debriefing questions it is necessary that researchers and question designers be aware of the potential problems. This information is obtained after pre-field techniques conducted previous to the field test, from analysis of field technique data or through the analysis of data collected from ongoing survey. Accordingly, respondent debriefing provides a useful complement to other quantitative tests, such as behaviour coding or item nonresponse. Design and samples size for debriefing could vary according to question designers needs.

"Respondent debriefing has the potential to complement information obtained from behaviour coding (behaviour coding can demonstrate the existence of problems but does not always indicate the source of the problem). When properly designed, the results of respondent debriefing can provide information about the problem sources and may reveal problems not evident from the response behaviour" (U.S. Census Bureau, 2003).

Open-ended questions employing standardised probes can provide valuable information to indicate whether questions and concepts were well understood. Follow-up questions can be included in an ongoing survey setting on a large and representative sample of respondents and the answers can be coded and analysed like any other item of survey data. For a minimum additional cost, respondent debriefing can be used to evaluate questionnaires undergoing revision or on completed surveys to provide an additional measure of response quality.

Nevertheless, the effect that debriefing questions have on respondents needs further study, namely, regarding its potential influence on respondents' future answers.

6.2.4. Intense interviews or follow-up interviews

The follow-up interview method is a tool for investigating how respondents answered the questions, providing information about the hidden processes involved in answering the questions. Follow-up interviews can identify a question that is apparently answered in a straightforward manner but is widely misinterpreted. One of their main values is in diagnosing the nature of any problem that has been identified through behaviour coding. Follow-up interviews tend to be lengthy and do not allow for a thorough treatment of many items. They are very time and money demanding.

Methods and tools

The technique consists on: shortly after a survey interview has been conducted with the respondent, a second intensive semi-structured interview is conducted by another interviewer. Respondents are taken carefully through selected questions asked in the first interview and encouraged to remind how they interpreted and understood the questions, both overall and in terms of particular words and phrases. They may also be asked about how they arrived at their answers, how accurately they felt the answer given actually reflects their views and how important they consider to provide an accurate answer. In this way information on the frame of references and range of interpretations used by respondents can be obtained. In comparison to respondent debriefings, intensive interviews are more detailed. According to Belson (1981) follow-up interviews provide information about the covered processes involved in answering the question. It is suggested a two stage intensive interview on questions:

- Firstly, the question and the given answer is quoted by the interviewer once more. Respondent is invited to describe how the answer has been retrieved.
- Secondly, the interviewer poses some predetermined questions on aspects and concepts related to the question. The respondent is requested to give feedback.

Variations of this technique have been explored by other researcher and called (intensive) reinterviewing, or double interview.

6.2.5. Experiments

Two or more feasible solutions for questionnaires or data collection modes are compared in order to determine the best option, under the point of view of their effect on measurement quality, nonresponse rate, time and costs. To this aim, experiments can be conducted. They can be carried out before a new survey is implemented, as pre-field or field testing technique, or may be embedded within the actual data collection phase for ongoing surveys.

In general, the experimental approach, in which some of the conditions can be set by the researcher, offers a more objective and rigorous method as compared to the testing methods presented so far, thus representing a useful integration to them. However, this approach turns out to be rather resource and time-consuming, especially if all the statistical issues are taken into consideration, leading to the fact that it is seldom used on a regular basis.

Methods and tools

The planning and realisation of an experiment for the evaluation of questions' or questionnaires' alternatives should follow the good principles of any experiment. It is a good rule to write down the protocol of the experiment in which all the methodological and practical issues are faced (Pocock, 1983). The protocol should describe the general design of the experiment, the study and the confounding variables, the mode of realisation (time, target population, persons and institutions involved), the statistical aspects (hypotheses, expected variations, design and size of the sample, analyses to be carried out). Some of the methodological issues will be critically reviewed here, for a recent paper on this topic, see Tourangeau (2004).

First of all, the hypotheses to be tested should clearly be identified. To this aim, it is useful to organise a group discussion with subject matter experts and methodologists. When specifying these hypotheses, the researchers have to identify the so called "outcome", i.e. the measure that will be considered to build the acceptance and refusal areas typical of the hypothesis testing approach. In general, in questionnaire testing field, the null hypothesis of equivalence between two questions will be contrasted against the alternative stating that a question "works better" than the other. This last sentence has to be formulated in a measurable way: to "work better" could mean to correctly re-produce the underlying variable distribution in the population, or to have the smallest number of nonresponses, or to produce the highest frequency for a phenomenon (for example in sensitive questions, where the underlying assumption is a general tendency to underestimate a given behaviour). This is a very critical issue since, most of the times, no information on the true response distribution is available to draw conclusions on which of the alternative questions is more valid.

A second important step in the design of an experiment is the definition of the design and size of the sample. For inferential purposes, probability sampling should be applied. The size of the sample should be based on statistical considerations, i.e. considering the expected difference in the alternatives, the intrinsic variability of the phenomenon, and the variability related to the experimental settings, in order to have enough power to significantly detect the difference to be tested. With regard to the power, i.e. the probability to detect a real difference of a given size, the common aim is to achieve a level of 0.80 (Tourangeau, 2004). Sometimes, also resources' considerations play a role in the decision on the size and design of the sample. Indeed, in order to ensure the representativeness of the sample, it is necessary to adopt a complex sampling design involving units geographically spread, with the possibility of an increase of the costs of the experiment.

Critical issues in the definition of the sample size and design are related to the fact that often small differences are expected (thus requiring a larger number of units, or a very efficient design), and to the accuracy of the experimental work. Especially in questionnaire testing applications, the data collection environment involves many factors (i.e. respondents, interviewers, geographical areas, etc.) that can act as confounding in the experiment.

Usually, in the hypothesis-testing approach, the type I error (the probability to reject the null hypothesis of equivalence, when indeed it is true) is set to 0.05, whereas the type II error (the probability of not rejecting the null hypothesis when it is false), as already implicitly mentioned, is aimed to be 0.20. Caution should be used in defining the hypotheses and the associated errors when dealing with experiments in questionnaire testing. Indeed, the researches should be aware that such an approach puts much more importance in the type I as compared to type II error, and should evaluate if this assumption applies to the specific testing situation. Tourangeau (2004) states that, depending on the main purpose of the experiment, the importance could be placed on the practical significance rather than in the statistical one. In particular, when evaluating the practical feasibility of the alternatives, considerations on the costs, ease of administration, etc. may be more relevant.

The experiment should be designed in a way to ensure the absence of systematic error, thus increasing its precision (Cox, 1992). This means that the unit submitted to a given questionnaire/question should not systematically differ from the units submitted to the alternative, if only two versions are compared. Usually, this property is ensured by means of randomising the assignment of the units to the different versions. To this aim, randomised blocks could be used, where the units inside each block are chosen to minimise the variation (Cox, 1992). Each version of the questionnaire should be submitted to at least a unit in the block.

Finally, when the aim is comparing together the effects of several different elements in the experiments (data collection mode, questions, time for administering, sex of interviewer, etc.), each “experimental treatment” consists in the combination of the chosen elements, or factors. Obviously, factorial design experiments are more complex to organise and, in general more demanding in terms of time and resources.

The treatment of the topic of the experiments in this handbook has been necessarily short and quite general: only the issues considered more relevant have been briefly described. Many books have been written on this topic (e.g. Cox and Reid, 2000) and can be consulted for further details.

Experiences

There are many applications on the experimental comparison of questions reported in the literature (e.g. Fowler, 2004). Nevertheless, such experiences are rather specific and difficult to generalise. As an example of application in official statistics, in 1989, Statistics Sweden decided to change the data collecting method of the Labour Force Survey switching 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 (SCB, 2004).

That the experiments are rarely carried out is confirmed by the results of the survey conducted for the purpose of this handbook, that showed that 62.1% of the institutions do not perform them.

6.2.6. Recommendations

Behaviour coding

- Behaviour coding should be regarded as a preliminary activity of questionnaire evaluation and revision. Further testing and evaluation is required to ensure the efficacy of the revision.
- In live behaviour coding, the coding system should be kept as simple as possible, so as to permit the simultaneous correct classification of the observed behaviours.
- The capacity for the behaviour coding activity to provide useful information, depends on the correct classification of the behaviours. The coding system should be exhaustive and it should permit the interpretation of the results.
- It is advisable to carry on coders debriefing sessions.
- Permission for recording the interview should be asked to the interviewers and the respondents.
- In pre-field BC, the interviewers should be expert ones, so as to be able to highlight as much as possible the problems related to the questions.

- In CAPI surveys, the interview should be conducted in the respondent environment rather than in the laboratory with the presence of the coder.
- In CATI and CAPI surveys, a parallel application of the electronic questionnaire should be implemented in order to make the recording of the codes easier.
- The number of interviews to be carried out depends on the complexity of the questionnaire and on the dimension of target subpopulations to which is intended to generalise the results.
- The reliability of the coding system should be assessed.

Interviewer debriefing

- Selection of interviewers with high experience and skill levels is recommended.
- Interviewers should be well trained in order to get useful information to improve the questionnaire.
- An outline covering all the relevant questionnaire problems should be prepared to guide the discussion.
- Outline should be given in advance to interviewers to help them organise appropriate comments.
- Debriefing sessions should be tape-recorded for later analysis. Notes should be taken during the debriefing, because transcription of the entire tape is very time consuming. Tape-recording is useful for reviewing unclear notes.
- The entire process should be carefully documented throughout the testing.
- Whenever substantial changes are made in the questionnaire as the result of testing, it is essential to conduct another test to evaluate the new questionnaire.

Respondent debriefing

- Question designers and researchers must have a clear idea of potential problems.
- Debriefing should be done immediately after the interview.
- The survey mode does not have to be the same as the one used by respondent debriefing (i.e. even if survey is interviewer-administered, the mode of respondent debriefing can be self-administered).
- Open-ended questions employing standardised probes can provide valuable information to indicate whether questions and concepts are well understood.
- Respondent debriefing provides a useful supplement to other quantitative measures of quality, such as behaviour coding or item nonresponse analysis. Therefore, respondent debriefing should be conducted after behaviour coding or any test that highlighted possible problems.

Experiments

- The objectives for an experiment are best determined by a group discussion.
- The protocol of the experiment describing its design, the study and the confounding variables, the implementation settings, and the statistical aspects should be prepared before its realisation.
- The hypotheses should be expressed in measurable terms.
- The design and the size of the sample should be decided in order to ensure the necessary power and representativeness of the experiment. Sometimes, practical issues can be considered more relevant than the statistical ones.
- The type I and type II errors associated to the hypothesis testing should be evaluated in relation to the specific application.
- Randomisation of assignment should be adopted.
- Factorial designs serve when many elements need to be evaluated. However, increasing the number of factors increases the complexity and affects the feasibility of the experiment.

6.2.7. Checklist

Behaviour coding

- Identification of questions or sequences of questions to be evaluated.
- Definition of who is going to code (researcher, interviewer, …).
- Definition of the subjects to be coded (respondents, respondents and interviewers).
- Selection of the interviewers.
- Selection of the coders.

- Definition of the coding system; protocol for coding.
- Definition of the number of interviews to be carried out.
- Selection of the sample of units to interview.
- Organisation of the test: place, scheduling time.
- Training of the interviewers and/or the coders.
- Implementation of a parallel application to easily record the codes.
- Debriefing of the coders.
- Analysis of the frequencies and type of codes.
- Identification of the problematic questions/sections.
- If reasonable, evaluation of coders variability.
- Identification of further actions on the problematic questions/sections.

Interviewer debriefing

- Identify the questions or set of questions in the questionnaire that should be tested.
- Development of debriefing guidelines.
- Selection of interviewers.
- Training of interviewers.
- Conduct the field test.
- Conduct the debriefing.
- Evaluation report.
- Improve the questionnaire where deficiencies are discovered. When changes are required, test the revised questionnaire again.

Respondent debriefing

- Identify the questions, which are a potential problem, to be tested.
- Develop the follow-up questions.
- Select the respondents.
- Conduct the field test.
- Complete the follow-up questions.
- Carry out the evaluation report.
- Improve the questionnaire where deficiencies are discovered. When changes are required, test the revised questionnaire again.

Experiments

- Discuss the objectives.
- Write down the protocol.
- Determine place, time, and duration of the experiment.
- State the hypotheses and define the outcome.
- Define the design and size of the sample.
- Select and assign sample units.
- Carry out the experiment.
- Test and interpret the results.
- Make decisions.

6.3. Post-evaluation methods

Post-evaluation methods intend to indirectly analyse the quality of the questionnaire during or after the real data collection phase. However, the methods presented apply also to many testing situations. Indeed, methods like analysis of item nonresponse, analysis of response distributions, analysis of editing and imputation rates and reinterview studies can provide useful information about the performance of the questionnaire and point out questions that require further investigation and/or evaluation by other testing methods. The quality of the collected data may also be evaluated by making comparisons with data from other sources.

The results achieved by post-evaluation methods should be transmitted to questionnaire designers so that changes may be assessed. The process should be iterative, characterised by numerous feedback loops, where information obtained at any point in the process could be used to improve and update questionnaires.

6.3.1. Analysis of item nonresponse rates

Analysis of item nonresponse rates from data collection can supply helpful information about questionnaire quality. As defined by Eurostat (2003), "Item nonresponse occurs when a respondent provides some, but not all, of the requested information or of the reported information is not usable".

The analysis should begin by investigating the extent of item nonresponse. A summary table should be made including an overview of the responses for each question, and providing valuable information about the rate of missing values. The questions with the higher nonresponse rates should be investigated in detail. However, also some questions with moderate missing value percentage should be analysed, if critical for the survey.

There might be combinations of questions to which all or only certain types of respondents did not answer. The analysis of them could represent a further way to investigate possible shortcomings of the questionnaire.

The analysis of nonresponse patterns provides a clear overview of the quantity, positioning and the types of missing values in the data set. It can be helpful in identifying some respondents' profiles that encounter difficulties with some sequences or groups of questions.

Finally, the item nonresponse analysis should be performed taking into account for the nature of the questions (e.g. sensitive questions), the use of proxy respondents, and the questionnaire length.

6.3.2. Analysis of response distributions

An extensive analysis should be performed on the quality of the responses received during data collection. This can be evaluated by examining the response distribution for every question. The detection of outlying values for all questions, the detection of outlying values for combination of variables, and the comparisons with data from other sources should also be carried out.

The analysis of response distribution is essential when testing more than one version of a question, a set of questions or the entire questionnaire. It can determine if variations among the responses resulting from different question wordings or question sequencings, produce different response patterns. Possible effects on response distributions by changing from one mode of data collection to another mode should also be examined.

This analysis is useful when performed in conjunction with other methods such as respondent debriefing or interviewer debriefing. This knowledge by itself is not sufficient to evaluate modifications, hence the results are not sufficient to disclose if one version of a question produces a better understanding of what is being asked than another.

6.3.3. Analysis of editing and imputation phase

The analysis of the editing and imputation phase can be used to assess the quality of survey questions, and for the continuous improvement of the questionnaire.

"Data editing is the application of checks that identify missing, invalid or inconsistent entries or that point to data records that are potentially in error" (Eurostat, 2003).

“Imputation is the process used to resolve problems of missing, invalid or inconsistent responses identified during editing. This is done by changing some of the responses or missing values on the record being edited to ensure that a plausible, internally coherent record is created” (Eurostat, 2003).

Numerous edit failures may indicate problems with the wording of questions, the structure of the questionnaire (e.g. routing, skip patterns, etc.), the understanding of survey concepts and definitions, and the training of interviewers, among others. To carry on this kind of analysis, information, such as an identifier for imputed values or original values for imputed fields, should be captured as part of the data file.

Although plenty of work has been done examining methods to improve editing and imputation to adjust for missing data, more emphasis should be placed on using editing to learn about the data collection process, in order to concentrate on preventing errors rather than on fixing them.

It is crucial the involvement of the questionnaire designer and subject matter experts in this analysis. They are truly the ones who have the knowledge to interpret the results of this analysis and explain why a question has been asked in a particular way, or to suggest better ways to ask a question when it has been identified as a possible problem.

6.3.4. Reinterview studies

Reinterview studies are aimed at estimating the measurement error, both the systematic and the variable component. “Measurement error refers to error in survey responses arising from the method of data collection, the respondent, or the questionnaire (or other instrument). It includes the error in a survey response as a result of respondent confusion, ignorance, carelessness, or dishonesty; the error attributable to the interviewer, perhaps as a consequence of poor or inadequate training, prior expectations regarding respondents’ responses, or deliberate errors; and error attributable to the wording of the questions in the questionnaire, the order or context in which the questions are presented, and the method used to obtain the responses” (Eurostat, 2003).

In the literature, reinterview studies are usually associated with the estimation of response error, i.e. when the role of the respondent is predominant respect to the other sources, as in the self-administered questionnaires. However, results from reinterview studies can orient towards the identification of some problematic questions.

Classical methodology for the estimation of the response errors identifis two main approaches based on the reinterview studies (Hansen *et al.*, 1964; Biemer and Forsman, 1992). The first one, aimed at estimating the variable component of the response error, namely the Simple Response Variance (SRV), is based on a reinterview independent and conducted under exactly the same conditions as the original one. The hypothesis of independent and identically distributed response error permits the estimation of the variance. Recent literature refers to such studies as “test-retest reinterviews” (Biemer and Lyberg, 2003).

When the aim is to estimate the systematic component of the response error, i.e. the response bias, the reinterview survey is oriented at identifying the true value. In a sample of respondents to the original survey a more accurate reinterview is performed. Sometimes, as true value, the result of the reconciliation is taken. Indeed, during the reinterview the interviewer compares the two responses, the first provided at the original survey and the second given to the reinterview. If the two answers disagree, the interviewer proceeds to the reconciliation, ascertaining which of the two is the true response. A value different from the previous two can also be accepted as true value. These reinterview studies are sometimes referred to as “gold standard reinterview studies”. Therefore, in order to estimate both response error components, usually the reinterview survey design is based on a sample of the respondent units that is split into two subsamples: all the units are reinterviewed under the same conditions of the original surveys, concurring to the estimation of the SRV, a subsample of them is reinterviewed with reconciliation, providing the data for the response bias estimation .

Recent research developments based on the use of latent class models, overtake some of the assumptions of the classical methods. For example, by introducing opportune hypotheses, i.e. the conditional independence of the errors, it is possible to estimate both the simple response variance and the response bias without having the need of identifying the true value or without requiring a reinterview under the same conditions of the original survey (Biemer, 2001; Brancato *et al.*, 2001). These reinterviews, reported sometimes as repeated measures (Biemer and Lyberg, 2003), may be integrated with probing questions with the aim of understanding the quality of the responses to some specific questions. Biemer and Lyberg (2003) report of a reinterview conducted by the U.S. Census Bureau in relation to the Census of Retail Trade, where the interviewers asked respondents to check the establishment's files to get a presumable more accurate value to be compared with the one obtained from the original interview.

6.3.5. Recommendations

Analysis of item nonresponse rates

- The extent of item nonresponse should be investigated by means of a summary table including an overview of the responses for each question, and the rate of missing values.
- Questions with the highest nonresponse rates as well as questions relevant for the survey with moderate amount of missing values should be examined in detail.
- Nonresponse to combinations of questions and nonresponse patterns should be analysed in order to investigate possible shortcomings of the questionnaire and to identify respondents' profiles that encounter difficulties with some sequences or groups of questions.
- The item nonresponse analysis should be performed taking into account for the nature of the questions (e.g. sensitive questions), the use of proxy respondents, and the questionnaire length.

Analysis of response distributions

- The response distribution for every question should be examined.
- The detection of outlying values for all questions, the detection of outlying values for combination of variables should be carried out.
- Comparisons with data from other sources should be carried out.
- Analysis of response distributions should be carried out when testing more than one version of a question, a set of questions, or the entire questionnaire.
- Analysis of response distributions should be carried out in conjunction with other methods such as respondent debriefing or interviewer debriefing.

Analysis of editing and imputation phase

- An identifier for imputed values or original values for imputed fields should be captured as part of the data file.
- The extent of corrections applied to the survey data should be used to identify problems with the wording of questions, the structure of the questionnaire (e.g. routing, skip patterns, etc.), the understanding of survey concepts and definitions, the training of interviewers, etc.
- It is crucial to exploit the editing and imputation phase to learn about the data collection process, in order to concentrate on preventing errors rather than on fixing them, involving in this phase the subject matter experts as well as the questionnaire designer.

Reinterview studies

- Use results of reinterview studies (high simple response variance, high response bias) to explore hypotheses on problematic questions.

Chapter 7. Towards a strategy on questionnaire design and testing methods

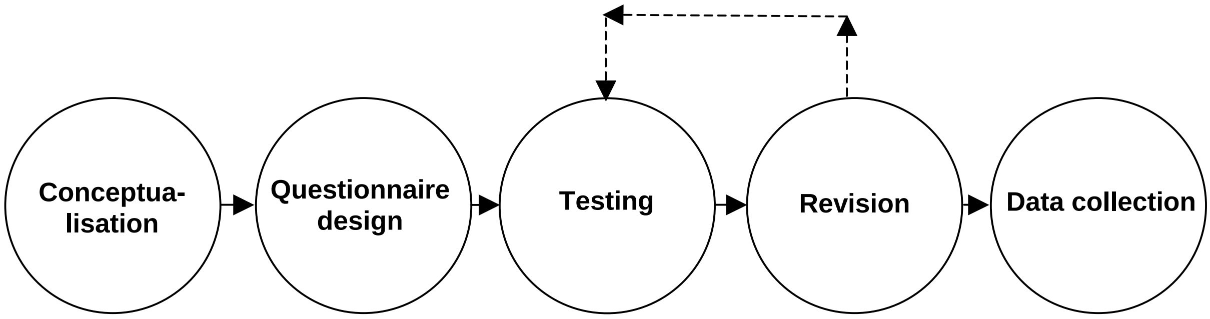
Questionnaires have a central impact on data quality as being the instrument to collect data and hence having the potential to produce measurement errors (Statistics Canada, 2002). Thus, professionally designed and systematically tested questionnaires are a precondition for valid and reliable statistical measurements. Consequently, from the conceptualisation to operationalisation, development and testing of a questionnaire a consistent strategy has to be applied. However, there is no cure-all strategy which can be employed for every survey using the same procedures and methods (Akkerboom and Dehue, 1997). Thus the strategy must be adapted to the specific conditions of each individual survey. Methods to be applied differ with regard to various factors and circumstances, e.g. if the aim is to implement a new survey or to evaluate an ongoing survey, if a social or a business survey is to be conducted, etc. In addition, the data collection mode to be used has to be taken into consideration, as well as the target population and resources such as budget, available staff, and time schedules. This makes the definition of a strategy rather complex, but exactly because of the complexity it is particularly necessary.

Accordingly, there can not be any general advice on the most appropriate method, other than to act strategically in planning and testing. The first six chapters of this volume showed details on each task and method as a kind of tool-box. The purpose of this chapter is to enable people to decide on a strategy on questionnaire development and testing, which fits the specific circumstances and demands of their survey and country. The structure of the chapter is divided into three parts: strategic planning of questionnaire development and testing methods, recommendations and a summary table on available methods and tools.

7.1. Steps of strategic planning

When developing a strategy the entire cycle of questionnaire design and testing has to be covered. Defining an appropriate and efficient testing strategy requires some experiences. For each step of the cycle several tools may be applied and the crucial point is to define the most appropriate methods and tools for the survey under consideration. Some methods might be implemented at various stages (e.g. expert groups or cognitive interviews), recommendations given as part of each planning phase in the figures 7.1 and 7.2 may help to select the most appropriate tools. Actually, the strategy consists on a sequence of tasks and methods to be implemented where various factors and circumstances are of influence. Changes in the order of tasks and applied methods depend mainly on the issue of whether a new survey is implemented or an ongoing survey is evaluated. In any survey carried out in the ESS, it is strongly recommended to systematically conduct each of the planning steps identified in figures 7.1 and 7.2. In particular each questionnaire for surveys in the ESS has to be tested using methods which involve potential respondents in a systematic way.

Figure 7.1. Strategic planning (1): new surveys



For strategic planning it is useful to distinguish the different phases of survey and questionnaire development by steps, which vary with regard to the current status of the survey (new/ongoing). In respect of new surveys there are five steps which have to be covered by the strategy:

1) *Conceptualisation.* Different from what one might expect, questionnaire design does not start with the questionnaire. Before one can even start to think about the wording of the questions, the conceptual basis of the questionnaire has to be specified. During the phase of conceptual design it is essential to start by a literature review, integrating the review of reports on comparable surveys and possible reports on testing. With regard to the objectives there are mainly two perspectives to catch and be aware of. One is the perspective of users and subject matter experts. What are their definitions, concepts and objectives? And the second perspective is how those definitions match with the conceptual models of the respondents. Expert groups, focus groups and in-depth interviews are important instruments in this process. The complexity of the theoretical concepts – even in apparently simple cases – requires a strict selection of empirical traits (often referred to as indicators) which can be observed in a survey. These indicators are deemed to be a suitable representation of the concept. Although this translation is at least in part based on conventions within the scientific community, some methods like facet design, symbolic interactionism, or semantic analysis facilitate this crucial step (see section 3.1). The work on the conceptual frame is of course more important for entirely new surveys, whereas in existing surveys concepts might be already well established. However, the (further) development of the conceptual basis must be an integral part in almost every change in a questionnaire. The main output of this stage are an entities/relationships scheme, an area tree about the structure of the targeted questionnaire, the list of target variables (direct and meta variables), and a preliminary tabulation plan (Precondition: target population, possible data collection mode, resources and time schedule have already been established).

2) Questionnaire design. After the conceptual basis has been specified a first draft of a questionnaire is to be worked on: appropriate wording, order of questions and definition of answering categories are now the ultimate tasks. The wording translates the target variables into concrete questions. Subsequently, the sequence of the questions is being defined taking care of possible context effects. There are basic rules of wording to follow described in section 3.2 of the handbook, which apply for each question regardless of the data collection mode. Supplementary rules and advice specific to each data collection mode are given below this section. For example, telephone interviewing has other requirements and possibilities than other techniques: in telephone surveys questions have to be short and the number of response options should be strictly reduced. CAI modes have the potential to use more complicated skipping rules by program. In general, as can be noticed, without the interviewer (self-administered questionnaires) the complexity of the questions and the complexity of the entire questionnaire should be limited. Thus, in a way, the striking key to consider during the development of the instrument is the availability of a person for the respondent to get help when demanded. Once a draft wording is available, the questionnaire has to be implemented visually with regard to the data collection mode. Visual design elements have to be considered (see section 3.3). PAPI questionnaires should be designed using a professional Desktop Publishing software package. Standard text processing applications do usually not fulfil the requirements of a professional layout. Specific requirements and methods when implementing CATI and CAPI instruments are vital to take into account and are presented in chapter 5. With respect to European surveys and international comparability another requirement is obvious: profound and functionally equivalent translation has to be ensured and harmonised definitions and variables should be implemented if possible (see chapter 4).

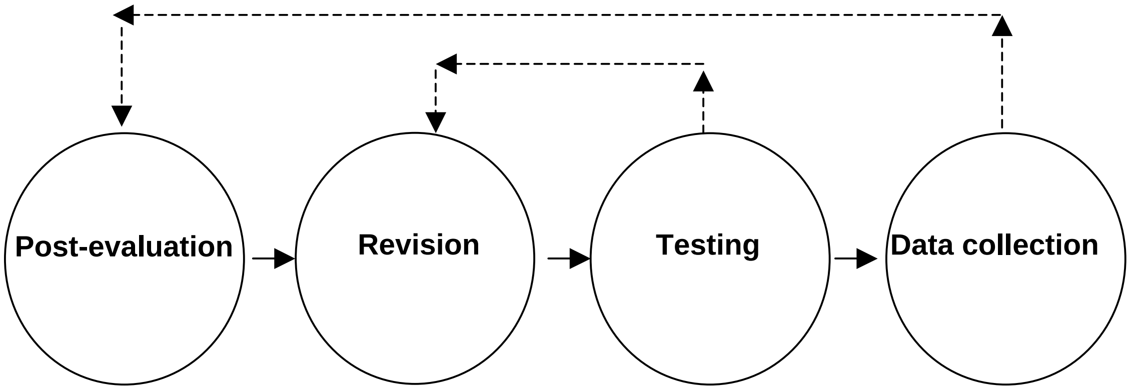
3) *Testing.* Testing a section or the entire questionnaire should not start unless design and wording reach the final version. Basically the questionnaire needs to be tested from three different viewpoints: a) wording of questions/answers, order and structure of the questionnaire; b) problems related to translation, cultural background and harmonisation aims (especially with community surveys); and c) in respect of the data collection mode and the involvement of an interviewer (technology implemented and self-administered or interviewer administered). A wide range of pre-field and field methods can be used to test the questionnaire. It is recommended to conduct pre-field and field testing methods, whereas the involvement of respondents is essential. Consequently, a combination of different methods is advisable (e.g. results of behaviour coding needs further probes by cognitive interviews) (see chapter 6). In

addition, tests for electronic questionnaires are necessary (usability tests and functionality tests, see chapter 5).

4) *Revision.* After testing, making revisions to the questionnaire based on the test findings, is necessary. Afterwards, a new round of testing is often indispensable. The revision idea involves testing the questionnaire at an early stage of its development, and then re-testing the revised questionnaire. This process may be repeated through two, three or even more phases of testing. Different methods of testing the questionnaire may be used during each phase of testing. Thus, in the context of revision the aim is to check if the changes are really resulting in a higher validity and reliability of data in relation to the specific objectives of the survey.

5) *Data collection.* With the implementation of the survey (either by a pilot study or as the real survey) the iterative process of development and revision is terminated, but the process of observation should be continued via monitoring the interviewers, interviewer debriefings and respondent debriefings (either by meetings or standard debriefing questionnaires for interviewers). Monitoring can be regarded as a continuous tool of evaluation (Akkerboom and Dehue, 1997). The monitoring of fieldwork can be essential for the further phase of post survey evaluation when conducting ongoing surveys or having implemented a new survey on full scale.

Figure 7.2. Strategic planning (2): Ongoing Surveys



When testing ongoing surveys, the steps to undertake are slightly different, offering other possibilities than post-evaluation methods, but also involving limitations since simply adapting questions possibly leads to incomparability with former surveys. The evaluation of ongoing surveys should start with the analyses of available data, thus with the data sets: analysis of item nonresponse rates, response distributions, imputation rates, edit failures, inconsistent results, or external validation are basically applied. In conjunction with the evaluation of data sets, reports and debriefings of interviewers might also help to clarify some implausible aspects. Often reports of interviewers on the fieldwork are only checked once after data collection. However, with regard to plausibility checking and post-evaluation it may help to have a look at the reports a second time. Findings from this evaluation phase can partly identify questions which have proven to be problematic. However, post-evaluation methods are not sufficient as the only method to check ongoing surveys, in fact these methods may give hints on deficiencies which need further testing on selected questions e.g. via cognitive methods, thus more qualitative testing should be performed (see section 3.2).

7.2. Recommendations

1) Minimum requirements

It is commonly accepted that pre-testing survey instruments is vital for conducting a full scale survey in order to minimise measurement errors (Sudman and Bradburn, 1982). Several recommendations in the European Statistical System stress this requirement too (see introduction, chapter 1). The European Statistics

Code of Practice requires that every questionnaire has to be systematically tested prior to being used for collecting data for European statistics. This requirement covers all existing questionnaires (given that a systematic test did not yet take place or it is evident that some questions need improvement) as well as new questionnaires.

The Code of Practice requirement is quite demanding, and much time and resources are needed to fulfil it. However, systematic testing of a questionnaire should be performed, at least if one or more of the following circumstances apply:

- legislative changes mandate a new survey,
- new questions, which were formerly not tested by the statistical institute, have to be asked,
- questions in existing surveys are being modified, even if apparently minor changes are made,
- the data collection instrument has to be changed (e.g. use of computer-assisted interviewing) or an additional data collection mode is introduced (e.g. web surveys in parallel to mail surveys), or
- poor data quality has been indicated by a review of nonresponse rates and biases, validations against other surveys or re-interview studies, deficiencies in internal consistency or other evidence.

Taking into account the methods presented, basically three minimum requirements are central:

- Only tests which are performed on the drafted questionnaire are sufficient. Thus methods such as focus groups for the general discussion on concepts and terms at the beginning phase of questionnaire design (not based on a draft questionnaire) are advisable and necessary, but not sufficient to meet minimum requirements.
- It is of basic relevance that at least one method to be applied is conducted with potential respondents. For example, informal testing of the questionnaires involving a number of colleagues is useful in early stages of questionnaire development but can not be considered a sufficient testing approach.
- It is advocated to execute both pre-field methods and field tests, as their potential use is different and the clue is to combine both information (qualitative and quantitative perspective). Moreover, experimental field tests are vital to monitor the impacts of changes in question wording on the comparability over time. They should always be carried out when questions measuring key indicators like poverty rates or unemployment rates are being revised.

2) Time schedule and documentation

However, even though broad acceptance of the necessity of systematically testing survey instruments can be noticed, it is often neglected that questionnaire testing takes time and resources. Consequently, when planning new surveys or revising ongoing surveys, pre-testing is either done within a few days, or conducted without involving additional staff resources. In addition, even though it is burdensome, a profound documentation on the testing strategy and results is essential. Establishing standards for documentation and maybe building up a standard data base for storing information on testing strategies and results can help to meet this requirement and reduce workloads for new projects.

3) New surveys versus ongoing surveys

Entirely new surveys require the most intensive testing. The whole questionnaire should be tested and it is recommended to use at least one of the pre-field testing methods (e.g. cognitive interviews) and (after a revision) a field testing method. Two or more phases of questionnaire testing are recommended. This involves testing the questionnaire at an early stage of its development, making revisions to the questionnaire based on the test findings, and then testing the revised questionnaire. Alternatively, modules or portions of the questionnaire may be tested during different phases of testing. For revisions to questionnaires, the entire questionnaire or at least the revised portion of it should be re-tested.

For ongoing surveys, evaluation should start by performing analyses of data sets as described previously. Esposito and Rothgeb (1997) described these steps as “quality assessment research” which leads to further testing and revision. However, revision of ongoing surveys is rather demanding and sensitive and needs very thoughtful strategic planning (see Akkerboom and Dehue, 1997; Esposito and Rothgeb, 1997). Consequently, a further requirement of ongoing surveys is the observation of possible effects of questionnaire changes on the time series. In this case an experimental design might be an appropriate solution (experiments, see section 6.2). For ongoing surveys, the impact on the continuity of the data of even apparently minor changes in the questionnaire, has to carefully be evaluated. When revisions to the questionnaire are made, the entire questionnaire or at least the revised part of it should be tested. The evaluation of former survey waves can provide important input for questionnaire revision. For recurrent surveys, the U.S. Census Bureau (2005) advocates cognitive interviews as they provide detailed insights into questionnaire problems whenever time permits or when a redesign is undertaken. Cognitive interviews may be more useful than focus groups with a pre-existing questionnaire because they mimic the question-response process.

4) Different methods – different perspectives

It is highly recommended to apply different methods, as each method has its own quality for development and testing (U.S. Census Bureau, 2003). Thus it is particularly desirable to apply both more objective (quantitative) and subjective (qualitative) methods – i.e. the respondent-centred and the interviewer-centred tests. This complementarity allows for both good problem identification and problem resolution and provides an evaluation of broad scope (ASA, 1997; Scheuren, 2004). In other words, checking the different perspective on the issue (users, matter experts, respondents and interviewers) is advisable and thus leads to a multiple method strategy. This strategy applies for both circumstances: new or ongoing surveys. However, due to existing experience from ongoing surveys and possible post evaluation analysis, different methods need to be applied to verify the problems. Especially errors detected via post evaluation may suggest in-depth, detailed tests on specific questions to be checked firstly via pre-field tests (mainly qualitative test as e.g. cognitive interviews). Thus a problem-oriented case to case decision on methods is of use.

5) Social surveys versus business surveys

Business surveys have other requirements than household surveys, mainly because of retrieval of information and available data. In addition, the first step to check might be to get in contact with the “right” respondent, respectively the person, who is actually responsible filling in the questionnaire, possibly not the employer him/herself. Focus groups can help to get an idea on this process (ASA, 1997). However, sometimes employers do not like to participate as they do not like to offer insides of their enterprise in front of other “colleagues” and thus personal interviews are often the “better” choice. To this aim, cognitive interviews will often have to be carried out on-site, at the enterprise. In business surveys, exploratory/feasibility studies, conducted as company or site visits, also provide information about structuring and wording the questionnaire relative to data available in business/institutional records (U.S. Census Bureau, 2003).

6) Electronic questionnaires

The electronic questionnaire (EQ) is a powerful tool that eases the realisation of surveys based on very complex questionnaires and is used more and more. While the management of the different branches is all set up by the software, the job of interviewing can be simplified and some errors can be already avoided when doing the interview. However, the transfer from a paper format to an electronic questionnaire needs further and additional testing: the mode of data collection has its peculiarities and needs to be taken into account. Thus, when planning a consistent testing strategy, it is indispensable to test the electronic version of the questionnaire by two different perspectives: a) the functionality – does the program technically proceed as it should? and b) the usability – are respondents and interviewers able to handle the electronic questionnaire?

Functionality tests should be applied when other testing methods on the content, wording, translation, layout etc. have been performed and programmers have transferred the questionnaire into the electronic version. When applying usability testing, the perspective is changing from the designers to the users: respondents and interviewers are invited to check the instrument. Usability testing is basically testing the performance of the electronic questionnaire by potential users. The test focuses on the Human-Computer Interaction (HCI) and the user friendliness of the CAI instrument. In a well performed usability test the entire package, including software, hardware, manuals and training, is evaluated. Thus, skipping usability testing is not recommendable, as it serves to test the instrument by the potential users within the scope of the data collection mode.

The documentation of testing procedures is extremely important in order to verify the EQ’s adherence to the specifications, for the testing phase, for the debugging of the code and to assist in making changes (House, 1985).

7.3. Summary table on testing methods – a tool-box

By presenting the different methods it has been distinguished between methods to be mainly used a) at the beginning of questionnaire design, b) when a first draft has been developed (pre-field testing, mainly in laboratories) and c) at the end, if testing needs a broader audience of respondents and a more natural environment, thus requiring to be implemented in the field. Additional tests and methods to be applied are related to d) the transfer of a paper questionnaire into an electronic questionnaire and e) translation and harmonisation tools to insure cross-national comparison. The distinction by these phases has been somewhat analytical and academic and researchers may distinguish differently from the presented way, due to other criteria. However, the summary table below shall offer the opportunity to get a quick overview of the features of each method, without attempting to be all-embracing nor to be totally sufficient in describing all criteria. In addition, the assessment of resources and planning has been particularly difficult in terms of standardised presentation and limited data available. Main information retrieval for the assumption on resources and planning are from the Australian Bureau (ABS, 2001), Statistics Sweden (SCB, 2004) and the U.S. Census Bureau (2003).

With regard to the complexity of selecting and defining a right strategy it is recommended to have a specialised unit inside the statistical office which can provide the survey managers with practical hints and support them in the implementation of the strategy. However, the reality might be different and survey managers are confronted to choose on their own what to do. Therefore, it is the aim of the handbook and the summary table below to assist in developing and defining a consistent strategy.

Table 7.1. Summary of testing methods

Methods and tasks	Phase of testing or implementation	Aims	Resources and planning	Strengths	Weaknesses
Focus groups (FGs) (respondent group discussion)	Early stage of questionnaire design	• To gain a reflection of the target population perspective • To check terms	• Overall duration: 3 weeks for setting several FGs • Costs: medium, related to number of FGs • Duration of FG: 1-1/2 hours per FG • Personnel: 2-3 persons (researcher and support) caring for FGs • Critical issues: time needed for recruitment of participants	• Fast resolution of issues from perspective of respondents • Useful for new surveys • Once a problems is discovered, detailed information is available	• They reflect only the perspective of respondents • Participants are not necessarily representative of the survey population • Social interaction may be misleading
Informal test (evaluation by colleagues)	Possible in each phase, preferably at the beginning	• To detect all kinds of mistakes: wording, layout, skips etc.	• Overall duration: within a few days • Costs: low, contacts "inside" the institution • Duration of each test: dependent on the length of the questionnaire • Personnel: 2-3 experts (researcher, 1-2 colleagues to contact) • Critical issues: none	• Quick to arrange • Eliminates obvious mistakes • Limited time-consuming • Basic misunderstandings and layout mistakes can be detected	• Rather subjective • Not systematic • Does not explore questionnaire in detail • Neither respondents' nor interviewers' perspective • Reflection of people used to statistics
Expert group (group discussion among design and matter experts, sometimes with users)	Initial phase of questionnaire development	• To check concepts, definitions, vocabulary against survey's objectives • To discuss data processing requirements	• Overall duration: 2 weeks • Costs: low to medium (depending on the questionnaire, number of experts and use of structured/unstructured checking) • Duration of expert group: 1/2 -1 day • Personnel: 4-6 persons (1-2 researchers, 2-4 experts) • Critical issues: expert availability in terms of timing	• Detects the highest rate of problems (however not always the most important ones) • Limited time-consuming	• Neither respondents' nor interviewers' perspective
In-depth or qualitative interviews (interviews with respondents)	Early stage of development and testing the questionnaire	• To evaluate respondents' viewpoint and understanding of the questionnaire • Explorative nature	• Overall duration: within a week • Costs: low • Duration of interview: 1-1½ hours (depending on the topic) • Personnel: 1-2 researchers • Critical: -	• Quick impression of the issue • Deep in detail • Discovers unexpected mistakes	• Limited information due to its subjective and unstructured nature
Cognitive interviews (one-to-one in-depth, structured interviews with specially trained interviewers and researchers)	Middle of the development process, once a draft questionnaire has been developed	• To gain qualitative information on how a questionnaire is understood and answered	• Overall duration: 5-6 weeks • Costs: high • Duration per interview: 1-1½ hours (depending on amount of questions, 5-15 interviews per test) • Personnel: 3-4 persons (2 researchers, 1-2 technical staff) • Critical issues: recruitment of participants from target population	• Valuable insight into the question-answering process to reduce measurement, specification and probably nonresponse error	• Qualitative nature, interpretation and generalisation of the interviews result difficult • Not representative • Time-consuming

Methods and tasks	Phase of testing or implementation	Aims	Resources and planning	Strengths	Weaknesses
Observational interviews (observation of respondents while completing a questionnaire)	Middle of development, when a tested questionnaire exists	• To check self-completed questionnaire by observing potential respondents in the lab	• Overall duration: 3-4 weeks • Costs: high • Duration of observation: ½ - 1 ½ hour • Personnel: 2-3 persons • Critical issues: recruitment of test persons from target population	• Possibility to get information on the filling in process • Intensive testing of the question	• Not representative • Time-consuming • Need of follow-up interviews to understand the reactions
Behaviour coding (coding behaviour and interaction of interviewers and respondents)	After a set of pre-field methods have been conducted	• To evaluate the question-answering process by standardised methods and coding scheme	• Overall duration: 6-8 weeks • Costs: high • Personnel: labour intensive (interviewers, coders, researcher) • Critical issues: difficulties in coding assignment	• Sheds light on the relevance of problems, since based on a high number of respondents • Quantitative feedback on problems, more systematic, less subjective (by standardised coding scheme)	• No identification of causes of problems • It does not shed light on the order of questions and context • Very time-consuming
Interviewer debriefing (either by group discussion or by questionnaire)	Possible in each phase of questionnaire development, preferably during field testing	• To obtain interviewers' useful feedback on the performance of the questionnaire	• Overall duration: 1-2 weeks • Costs: medium • Personnel: 1-2 persons (researchers or technical staff + interviewers) • Critical issues: recruitment of interviewers with different experiences	• Provides information about sources of interviewer's error, respondent's sensitivity, respondent's burden	• Not representative • Subjective information (depends on the group of interviewers)
Respondent debriefing (group discussion)	Is typically used at a later stage in the questionnaire design or evaluation process	• To determine whether concepts or questions are understood by respondents as intended by questionnaire designers	• Overall duration: variable • Costs: medium • Duration per interview: depending on the length of questionnaire • Personnel: regular interviewers • Critical issues: to get respondents to participate	• Provides information about problem sources	• Not representative • Subjective information identification • Researchers need strong hypotheses on problematic questions in advance • Discussion might be misleading
Intense or follow-up interviews (second intensive semi-structured interview, conducted by a different interviewer)	Field testing method after an interview has taken place	• Mainly limited to some questions which are problematic to implement	• Overall duration: 2-3 weeks (depending on the length of fieldwork) • Costs: medium-high (interviewers, respondents and researchers) • Duration per interview: 1-1 ½ hours • Personnel: researchers and interviewers • Critical issues: well trained interviewers and respondents willing to give feedback	• Profound discussion on problematic questions • To detect apparently right answers that indeed were non-valid replies, due to a misunderstanding of the question	• High human resources and thus costly • Very well trained interviewers
Experiments	Can be done as pre-field or field test or embedded within the actual data collection phase	• To compare two or more questions or questionnaire versions	• Overall duration: variable • Costs: high • Personnel: researchers, interviewers, design and subject matter experts • Critical issues: definition of the sample	• Reliable inferences • Quantitative information	• Very costly and time-consuming

Methods and tasks	**Phase of testing or implementation**	**Aims**	**Resources and planning**	**Strengths**	**Weaknesses**
Functionality testing (testing the performance of electronic questionnaire)	During and at the end of the software development	• To test the performance and the robustness of the CAI software	• Overall duration: variable • Costs: limited • Personnel: programmers, subject-matter experts, researchers, interviewers	• Prevents functioning errors	• Requires coordination among different expertises
Usability testing (testing the human-computer interaction)	At the end of the software development	• To evaluate the human-computer interaction and the user-friendliness of the CAI instrument by the end-users	• Overall duration: variable • Costs: variable • Personnel: interviewers, respondents, researchers • Critical issues: the interpretation of some by-product results can be very demanding	• Permits to evaluate the ability of the users with the CAI instruments	• Difficult to be performed in the real setting environment

Definition of items:

Methods and tasks:	reference to the term used in the handbook for methods and tasks
Phase of testing or implementation:	appropriate timing in implementation
Aims:	rough description of the main purposes of the method
Resources and planning:	evaluation of time, costs, and human resources needed for implementation (not including costs of training staff); critical issues for the implementation
Strengths and weaknesses:	selected, rough assessment of the method

References

ABS (2001). Pretesting in survey development. An Australian Bureau of Statistics perspective, Research Paper, ed. Australian Bureau Statistics, Canberra, Australia.

ABS (2004a). Forms Design Standards Manual. Australian Bureau of Statistics. Available on-line at http://www.sch.abs.gov.au.

ABS (2004b). Pretesting in survey development. An Australian Bureau of Statistics perspective, Research Paper, ed. Australian Bureau Statistics, Canberra, Australia.

ABS (2000). Basic Survey Design. Available on-line at http://www.sch.abs.gov.au.

Ahola, A. and Lehtinen, M. (2002). Contextuality of Survey Responses as a Challenge to the Development of Questionnaire Testing Methods. International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), November 14-17, 2002, Charleston, South Carolina.

Akkerboom, H. and Dehue, F. (1997). The Dutch Model of Data Collection Developments for Official Surveys. International Journal for Public Opinion Research, 9, 126-145.

Andrews, F.M. and Whitney, S.B. (1976). Social Indicators of Well-being, New York: Plenum

ASA (1997). How to Conduct Pretesting. American Statistical Association series: What is a Survey? Alexandria.

Bailar, B. (1968). Recent Research in Reinterview Procedures. Journal of the American Statistical Association, 63, 41-63.

Beatty, P. (2004). The dynamics of cognitive interviewing. In Methods for testing and evaluating survey questions, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Behling, O. and Law, K.S. (2000). Translating Questionnaires and Other Research Instruments: Problems and Solutions. London: Sage Publications.

Belson, W. (1981).The Design and Understanding of Survey Questions. Aldershot: Grower Publishing.

Belson, W. (1986). Validity in Survey Research. Aldershot: Grower Publishing.

Bergdahl, M., Japec, L., Magdaleno, M., Signore, S., Tzougas, I. (2001). LEG chapter on CBM and Minimum Standards. Proceedings of the International Conference on Quality in Official Statistics (Q2001) Stockholm, May 14-15.

Beukenhorst, D. (2004). Tangible Evidence: Using Modern Technology for Recording and Analysis of Interviews. In ZUMA-Nachrichten Spezial, Band 9, Questionnaire Evaluation Standards.

Beukenhorst, D., Giesen, D., de Vree, M. (2002). Computerized versus interviewer-guided evaluation of CASI questionnaires. Paper presented at QUEST workshop 2002. Washington DC.

Biemer, P., Herget D., Morton J., Willis G. (2000). The feasibility of Monitoring Field Interview Performance Using Computer Audio Recorded Interviewing (CARI). Proceedings of the Section on Survey Research Methods, American Statistical Association. Available on-line at http://www.amstat.org/sections/srms/proceedings/papers/2000_183.pdf.

Biemer, P.P. (2001). Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. Journal of Official Statistics, 17, 2, 295-320.

Biemer, P.P. and Fecso, R. (1995). Evaluating and Controlling Measurement Error in Business Surveys. In Business Survey Methods, eds. B.G. Cox, D.A. Binder, B. N. Chinnappa, Anders Christianson, M.J. Colledge, P.S. Kott, New York: Wiley.

Biemer, P.P. and Forsman, G. (1992). On the Quality of Reinterview Data with Application to the Current Population Survey. Journal of American Statistical Association, 87, 420, 915-923.

Biemer, P.P. and Lyberg, L.E. (2003). Introduction to Survey Quality. Hoboken, New Jersey: John Wiley & Sons.

Borg, I. and Shye, S. (1995). Facet Theory: Form and Content. Newbury Park, CA: Sage.

Bradburn, N.M (2004). Understanding the question-answer process. Survey Methodology, 30 (No.1), 5-15.

Bradburn, N. and Sudman, S. (1991). The Current Status of Questionnaire Design in Measurement errors in surveys. New York: John Wiley & Sons.

Bradburn, N., Frankel, M., Baker, R. and Pergamit, M. (1991). A Comparision of CAPI with PAPI Interviews in the National Longitudinal Study of Youth. 1991 AAPOR Conference.

Brancato, G., Fortini, M. and Pichiorri, T. (2001). On the use of Bayesian approach to estimate response errors in National Statistical Institutes. Proceedings of The International Conference on Quality in Official Statistics (Q2001). Stockholm, May 14-15.

Braun, M. (2003). Funktionale Äquivalenz in interkulturell vergleichenden Umfragen. Mythos und Realität. Mannheim: Habilitationsschrift.

Braun, M., and Mohler, P.Ph. (2002). Background Variables. In Cross-Cultural Survey Methods, eds. J.A. Harkness, F. J.R. van de Vijver, and P.Ph. Mohler. New York: John Wiley and Sons, 99-113.

Brauns, H., Scherer, S., and Steinmann, S. (2003). The CASMIN Educational Classification in International Comparative Research. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. J.H.P. Hoffmeyer-Zlotnik and C. Wolf. New York: Kluwer Academic / Plenum Publishers, 221-244.

Brewer, M.B. and Lui, L.J. (1995). Use of sorting tasks to assess cognitive structure. In Answering questions: Methodology for determining cognitive and communicative processes in Survey research, eds. N. Schwarz, and S. Sudman, San Francisco: Jossey-Bass.

Budwoski, M. and Scherpenzeel, A. (2005). Encouraging and Maintaining Participation in Household Surveys: The Case of the Swiss Household Panel. ZUMA-Nachrichten, 29 (no. 56), 10-36.

Burgess, R.G. (1982). The unstructured interview as a conversation. In Field research: A Source Book and Field Manual, ed. R.G. Burgess, London: Allen and Unwin.

Burgess, M.J. and Paton, D. (1993): Coding of Respondents Behaviour by interviewers to test questionnaire wording. Available on-line at http://www.census.gov/srd/papers/pdf/az9501.pdf.

Burnside, R. (2000). Towards Best Practice for Design of Electronic Data Capture Instruments. Australia: Statistics Clearing House Reference Material.

Bushery, J.M. (1981). Recall Biases for Different Reference Periods in the National Crime Survey. Proceedings of the Section on Survey Research Methods. American Statistical Association, 238-273.

Canberra Group. Expert Group on Household Income Statistics (2001). Final Report and Recommendations. Ottawa: Canberra Group.

Cannell, C., Oksenberg, L., Kalton, G., Bischoping, K., Fowler, F.J. (1989). New Techniques for Presting
Survey Questions. Final Report August, 1989. The University of Michigan: Survey Research Center.
University of Massachusetts: Center for Survey Research.

Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on interviewing techniques. In Sociological methodology, ed. S. Leinhardt, San Francisco: Jossey-Bass.

Catford, J.C. (1965). A Linguistic Theory of Translation: An Essay in Applied Linguistics. London: Oxford University Press.

Chang, C.Y. (2005). Cross-Cultural Assessment: A Call for Test Adaptation. Available on-line at aac.ncat.deu/newsnotes/y99sum1.html.

Chen, P. (1976) The Entity Relationship Model: Towards a Unified View of Data. ACM Transactions on Database Systems, 1, 1, 9-36.

Commission of the European Community (2005). Communication from the commission to the European Parliament and to the Council and Recommendation of the Commission on the independence, integrity and accountability of the national and Community statistical authorities. Brussels, 25.5.2005, COM(2005) 217 final.

Conrad, F.G. and Schober, M. (2000). Clarifying Question Meaning in a Household Telephone Survey. Public Opinion Quarterly, 64, 1-28.

Conrad, F.G. and Blair, J. (2004). Data quality in cognitive interviews: the case of verbal reports. In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Couper, M.P., Horm J., Schlegel, J. (1997). Using trace files to evaluate the national health interview survey CAPI instrument. Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria: ASA, pp. 825-829. Available on-line at www.amstat.org/sections/SRMS/Proceedings/papers/1997_038.pdf .

Couper, M.P., Beatty, P., Hansen, S., Lamias, M., and Marvin, T.(2000). CAPI Design Recommendations, report submitted to the U.S. Bureau of Labour Statistics.

Couper, M.P. (2000). Web Surveys: A Review of Issues, and Approaches. Public Opinio Quarterly, 64, 464-494.

Cox, D.R., (1992). Planning of experiments. New York: John Wiley & Sons.

Cox, D.R., Reid, N. (2000). The Theory of the Design of Experiments. Chapman & Hall/CRC

Datamed (1998). Report on methodologies and specifications second work-package deliverable 4 (D8D0101O), Commission of the European Communities - DG III - Industry de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. Journal of Official Statistics, 21, Special Anniversary Issue, 233-256.

de Leeuw, E.D., Hox, J. and Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. Journal of Official Statistics, 19, 2, 153-176.

DeMaio, T. J. and Rothgeb, J. M. (1996). Cognitive interviewing techniques: in the lab and in the field. In Answering Questions, Schwarz, N. and Sudman, San Francisco: Jossey-Bass Publishers.

U.S. General Accounting Office (1993). Developing and Using Questionnaires. Available on-line at http://www.ojp.usdoj.gov/BJA/evaluation/guide/documents/documentgg.html.

Dillman D.A. (1978). Mail and Telephone Surveys: The Total Design Method, New York: John Wiley & Sons.

Dillman, D. (2005). It takes more than words to write a question. How visual layout affects answers to internet and mail questionnaires. Short course manuscript for the Journal of Official Statistics 20th Anniversary Conference, Stockholm 24-25 August 2005.

Dillman, D., Gertseva, A. and Mahon-Haft, T. (2005). Achieving Usability in Establishment Surveys Through the Application of Visual Design Principles. Journal of Official Statistics, 21, 183-214.

Dillman, D.A. (2000). Mail and Internet Surveys. The Tailored Design Method. New York: John Wiley & Sons.

Duquesne University (2005). Validity. Available on-line at www.mathcs.duq.edu/~packer/Courses/Psy624/test.html (12-20-2005).

Edwards, W.S. and Cantor, D. (1991). Towards a Response Model in Establishment Surveys. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley & Sons.

Ehling, M. (1997). Pretest – Ein Instrument zur Überprüfung von Erhebungsunterlagen. In Wirtschaft und Statistik, Vol. 3, Statistisches Bundesamt, Wiesbaden, pp. 151-159.

Eisenhower, D., Mathiowetz, N.A. and Morganstein, D. (1991). Recall error: sources and bias reduction techniques. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley & Sons.

Elias, P. and Birch, M. (1994). Establishment of Community-Wide Occupational Statistics. ISCO 88 (COM) A Guide for Users. University of Warwick, Institute for Employment Research. Available on-line at www.warwick.ac.uk/ier/isco/isco88.html.

Ericsson, K. and Simon, H. (1993). Protocol analysis: verbal reports as data, Cambridge: MIT Press MA.

Erikson, R., Goldthorpe, J.H., and Portocarero, L. (1979). Intergenerational Class Mobility in Three Western European Societies. British Journal of Sociology, 30, 415-441.

ESOMAR (1997). The ESOMAR Standard Demographic Classification. A System of International Socio-Economic Classification of Respondents to Survey Research. Amsterdam: ESOMAR; see also: Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. J.H.P. Hoffmeyer-Zlotnik and C. Wolf. New York: Kluwer Academic / Plenum Publishers, 97-121.

Esposito, J.L, Campanelli, P.C., Rothgeb, J., Polivka, A.E. (1991). Determining which questions are best: Methodologies for evaluating survey questions. Proceedings of the American Statistical Association (Survey Research Methods Section). Alexandria, VA. American Statistical Association, 46-55.

Esposito, J.L. (2002). Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study. Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing (QDET) Methods, 14-17 November 2002, Charleston, SC.

Esposito, J.L. and Rothgeb, J.M. (1997). Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment. In Survey Measurement and Process Quality eds L. Lyberg, P.P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D.Trewin. New York: John Wiley and Sons.

European Commission (2002). Key data on education in the European Union – 2002 Luxembourg: Office for Official Publications of the European Communities. Available on-line at www.eurydice.org/Documents/cc/2002/en/CC2002_EN_home_page.pdf.

European Commission (2003). Commission Regulation (EC) No. 1982/2003 of 21 October 2003 implementing Regulation (EC) No. 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the sampling and tracing rules. Official Journal of the European Union. L 298/29.

European Community Household Panel (ECHP). Available on-line at forum.europa.eu.int/irc/dsis/echpanel/info/data/information.html.

European Social Survey (ESS) (2002). European Social Survey, Round 1. Specification for participating countries. Available on-line at www.europeansocialsurvey.org/.

European Social Survey (ESS) (2004). European Social Survey, Round 2. Specification for participating countries. Available on-line at www.europeansocialsurvey.org/.

Eurostat (2003). Glossary. Eurostat/A4/Quality/03/Glossary.

Eurydice. Available on-line at www.eurydice.org/.

Fiske, D.W. (1971). Measuring the concepts of personality, Chicago: Aldine.

Flynn, J. (1996). Constructing and Reconstructing Respondent Attitudes During a Telephone Survey. Proceedings of the section on Survey Research Methods. American Statistical Association, 5/1, 43-59

Foddy, W. (1993). Constructing questions for interviews and questionnaires: theory and practice in social research. Cambridge (UK). Cambridge University Press.

Forsman, G. and Schreiner, I. (1991). The Design and Analysis of Reinterview - an Overview. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley and Sons.

Forsyth, B.H. and Lessler, J. T. (1991). Cognitive Laboratory Methods. A Taxonomy. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley & Sons, 393-418.

Forsyth, B.H., Rothgeb, J.M., Willis, G.B. (2004). Does Pretesting Make a difference? An Experimental Test. In Methods for testing and evaluating survey questionnaires eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Fowler, F.J. (1989). Coding Behaviour in Pretests to Identify Unclear Questions. In Health Survey Methods. Proceedings of the Fifths Conference, Rockville: National Center for Health Services Research and Health Care Technology Assessment.

Fowler, F.J. (1995). Improving Survey Questions. SAGE Publications.

Fowler, F.J. (2002). Survey Research Methods, 3rd edition. Thousands Oaks (California): Sage.

Fowler, F.J. Jr. (2004). The case for more spilt-sample experiment in developing survey instruments. In Methods for testing and evaluating survey questionnaires eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Fowler, F.J. Jr. and Cannell, C.F. (1996). Using behavioural Coding to identify Cognitive Problems with Survey Questions. In Answering questions. Methodology for determining cognitive and communicative processes in survey research, eds. N. Schwarz, and S. Sudman, San Francisco (California): Jossey-Bass.

Friedman, H.H. and Amoot, T. (1999). Rating the Rating Scales. Journal of Marketing Management, Vol. 9:114-123.

Fuller, W.A. (1995). Estimation in the Presence of Measurement Error. International Statistical Review, 63, 2, 121-147.

Ganzeboom, H. B.G. and Treiman, D.J. (1996). Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. Social Science Research, 25, 201-239.

Ganzeboom, H. B.G. and Treiman, D.J. (2003). Three Internationally Standardised Measures for Comparative Research on Occupational Status. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. J.H.P. Hoffmeyer-Zlotnik and C. Wolf. New York: Kluwer Academic/Plenum Publishers, 159-193.

Ganzeboom, H. B.G., de Graaf, P.M., Treiman, D.J., and de Leeuw, J. (1992). A Standard International Socio-Economic Index of Occupational Status. Social Science Research, 21, 1-56.

Gatward, R. (2003). Developing and updating screen layout and design standards. Paper presented at the 8° International Blaise User Conference (IBUC 2003), Copenhagen, Denmark, 21-23 May.

Gatward, R. (2004). An evaluation of Delta, a documentation tool. Paper presented at the 9° International Blaise User Conference (IBUC 2004), Quebec, Canada, 22-24 September.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and non identifiable models. Biometrika 61 215-231.

Gough, G. (1985). Reason for slimming and weight loss. In Facet Analysis, New York: Springer

Graesser, A.C., Bommareddy, S., Swamer, S., and Golding, J.M. (1996). Integrating questionnaire design with a cognitive computational model of human question answering. In Answering questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, eds. N. Schwarz and S. Sudman, San Francisco, CA: Jossey-Bass, 143-175.

Graesser, A.C., Kennedy T., Wiemer-Hastings P, and Ottati V. (1999). The use of computational cognitive models to improve questions on survey and questionnaires. In Cognition and Survey Research, eds. M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, R. Tourangeau. New York: Wiley

Graesser, A.C., Wiemer-Hastings K., Kreuz R., Wiemer-Hastings P., Marquis K. (2000). “QUAID: A questionnaire evaluation aid for survey methodologist.” In Behaviour Research Methods. Instruments and Computers, 32 (2); 254-262

Grice, H.P. (1975). Logic and conversation. In Syntax and semantics: 3. Speech acts, eds. P. Cole and J.L. Morgan, New York: Academic Press, 41-58.

Groves, R. (2004). Survey errors and survey costs. New York: Wiley

Groves, R., Biemer, P.P., Lyberg, L.E., Massey J., Nicholls W., and Waksberg J., eds. (1988). Telephone Survey Methodology. New York: Wiley.

Groves, R.M., Fowler, F.J.Jr, Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). Survey Methodology. Hoboken, New Jersey: John Wiley & Sons.

Groves, R.M., Fultz, N. and Martin, E. (1991). Direct questioning About Comprehension in a survey setting. Public Opinion Quarterly, 56, 475-495.

Guttman, L. (1954). An Outline of some Methodology for Social Research. Public Opinion Quarterly, 18, 245-259

Hagenaars, J.A. (1993). Log-linear models with latent variables. London: SAGE Publications, series 07, number 094.

Hak, T., Willimack, D.K. and Anderson, A.E. (2003). Response process and burden in establishment surveys. Proceedings of the Joint Statistical Meetings, Alexandria, VA: American Statistical Association, 1724-1730

Hansen M.H., Hurwitz, W.N. and Bershad, M.A. (1961). Measurement errors in censuses and surveys. Bulletin of the International Statistical Institute, 38, 359-374.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1964). The Estimation and Interpretation of Gross Differences and the Simple Response Variance. In Contribution to Statistics, ed. C.R. Rao, Calcutta, Statistical Publishing Society, pp.111-136.

Hansen, S.E. Fuchs, M. and Couper, M.P. (1997). CAI instrument usability testing. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Norfolk, VA, May.

Hansen, S.E., and Couper, M.P. (2004). Usability Testing to Evaluate Computer-Assisted Instruments. . In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Harkness, J. (2003). Questionnaire Translation. In Cross-Cultural Survey Methods, eds. J.A. Harkness, F. J.R.van de Vijver, and P. Ph. Mohler. New York: John Wiley and Sons, 35-56.

Henningsson, B. (2002). Interviewer Debriefing by e-mail. Statistics Sweden. Available on-line at http://www.jpsm.umd.edu/qdet/final_pdf_papers/henningsson.pdf

Hoffmann, E. (2003). International Classification of Status in Employment, ICSE-93. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. Hoffmeyer-Zlotnik, J. H.P. and Wolf C. New York: Kluwer Academic/Plenum Publishers, 125-136.

Hoffmeyer-Zlotnik, J. H.P. (2003). The Classification of Education as a Sociological Background Characteristic. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. Hoffmeyer-Zlotnik J. H.P. and Wolf C. New York: Kluwer Academic/Plenum Publishers, 245-256.

Hoffmeyer-Zlotnik, J. H.P. and Warner, U. (2005). How to Measure Education in Cross-National Comparison: The Hoffmeyer-Zlotnik/Warner Matrix of Education as a New Instrument. In Methodological Aspects in Cross-National Research, eds. Hoffmeyer-Zlotnik J. H.P. and Harkness J. A., Mannheim: ZUMA-Nachrichten Spezial, 11, 223-240.

Hoffmeyer-Zlotnik, J. H.P. and Wolf, C. (2003b). Comparing Demographic and Socio-Economic Variables Across Nations. Synthesis and Recommendations. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. Hoffmeyer-Zlotnik J. H.P. and Wolf C. New York: Kluwer Academic/Plenum Publishers, 389-406.

Hoffmeyer-Zlotnik, J. H.P. and Wolf, C., eds. (2003a). Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York: Kluwer Academic/Plenum Publishers.

House, C.C. (1985). Questionnaire design with computer Assisted Telephone Interviewing. Journal of Official Statistics, 1, 2, 209-219.

House, C.C. and Nicholls, W.L. II (1988). Questionnaire design for CATI: design objectives and methods. In Telephone Survey Methodology, eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg,. New York: Wiley.

Hox, J and De Jong-Gierveld, J.J., eds. (1990). Operationalisation and Research Strategy, Lisse (NL): Swets & Zeitlinger

Hox, J.(1997). From Theoretical Concept to Survey Question. In Survey Measurement and Process Quality, eds L. Lyberg, P.P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D.Trewin. New York: John Wiley and Sons.

ILO (1987). Guidelines concerning the implications of employment promotion schemes on the measurement of employment and unemployment, 14th ICLS, 1987. Available on-line at www.ilo.org/public/english/bureau/stat/download/guidelines/emprom.pdf

ILO (1990). International Standard Classification of Occupations: ISCO-88. Geneva: ILO.

ILO (1998). Guidelines concerning treatment in employment and unemployment statistics of persons on extended absences from work, 16th ICLS, October 1998. Available on-line at www.ilo.org/public/english/bureau/stat/download/guidelines/exleave.pdf

ILO (2005). Updating the International Standard Classification of Occupations, ISCO-08. Available on-line at www.ilo.org/public/english/bureau/stat/isco/isco88/intro.htm

Information Society Technologies and CHINTEX (1999). CHINTEX Synopsis. Contract no IST-1999-11101. Available on-line at www.destatis.de/chintex/download/synopsis.pdf

International Social Survey Programme (ISSP). Available on-line at www.issp.org/homepage.htm

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

Kahneman, D. and Tversky, A. (1971). Subjective probability: a judgement of representativeness. Cognitive psicology, 3, 450-454.

Kerlinger, F.N. (1986). Foundations of Behavioral Research, New York: Holt, Rinehart & Winston

Kinsey, S.H., and Jewell, D.M. (1998). A systematic approach to instrument development in CAI. In Computer assisted survey information collection,eds. M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls, and J. O'Reilly. New York: Wiley, 105-123.

Krebs, D. (2001). How Providing Versus Not Providing a Midpoint in Response Scales Affects Validity and Reliability of Measurement. Paper presented at the AAPOR 61st annual conference, Montreal, Canada, 17-20 May.

Krebs, D. and Langfeldt, B. (2005). Reliability and Validity of Different Measures of Work Orientation: Effects of Response Scale Format on Measurement Quality. Paper presented at the conference Applied Statistics 2005, Ribno, Slovenia, 18-21 September,

Krueger, R. and Casey, M. (2000). Focus Groups. Practical Guide for Applied Research. Beverly Hills: Sage Publication.

Kuusela, V. (2003). Screen Layout Standards at Statistics Finland. Paper presented at the 8° International Blaise User Conference (IBUC 2003), Copenhagen, Denmark, 21-23 May.

Lazarfeld, P.F. (1972). Qualitative Analyis. Historical and Critical Essays, Boston: Allyn & Bacon

Lazarfeld, P.F.(1958). Evidence and Inference in Social Research, Daedalus, 87, 99-130.

LEG on Quality (2001) "Summary Report from the Leadership Group (LEG) on Quality". EUROSTAT. SPC. Available on-line at http://amrads.jrc.it/WPs%20pages/Quality/Documents/LEGsummary.pdf

Legard, R., Keegan, J., Ward, K. (2003). In-depth Interviews. In Qualitative Research Practice. A Guide for Social Science Students and Researchers, eds. J. Ritchie and J. Lewis, London: Sage.

Lessler, B.H. Forsyth, J. (1996). A Coding System for Appraising Questionnaires. In Answering questions. Methodology for determining cognitive and communicative processes in survey research, eds. N.Schwarz, S. Sudman, San Francisco (California): Jossey-Bass

Lester, A. and Wilson, I. (1995). Surveying Businesses by Telephone – a case study of methodology. Proceedings of the International Conference on Survey Measurement and Process Quality, American Statistical Association

Levinsohn J.R., Rodriguez G. (2001). Automated Testing of Blaise Questionnaires. Proceedings of the 7° International Blaise User Conference. Washington, USA. Available on-line at www.blaiseusers.org/ibucpdfs/2001/Levinsohn_Rodriguez--IBUC_paper.pdf

Lewis, J. (2003). Design Issues. In Qualitative Research Practice. A Guide for Social Science Students and Researchers, eds. J. Ritchie and J. Lewis, London: Sage.

Lofland, J. and Lofland, L.H. (1995) Analyzing Social Settings, Belmont, CA: Wadsworth

Luce, R. (1959). Individual choice behaviour. New York: John Wiley and Sons.

Martin, E. (2004). Vignettes and respondent debriefing for questionnaire design and evaluation. In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Matthiessen, C. (1999). The System of TRANSITIVITY: An Exploratory Study of Text-Based Profiles. Functions of Language, 6.1.

Mejer, L. (2003). Harmonisation of Socio-Economic Variables in EU Statistics. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. J.H.P. Hoffmeyer-Zlotnik and C. Wolf, New York: Kluwer Academic/Plenum Publishers, 67-85.

Mohler, P.Ph., Smith, T.W. and Harkness, J.A. (1998). Respondents' Ratings of Expressions from Response Scales: A Two-Country, Two-Language Investigation on Equivalence and Translation. ZUMA-Nachrichten Spezial, 3, 159-184.

Norman, D. A. (2004). Emotional Design. Why we Love (or Hate) Everyday Things: Cambridge, MA: Basic Books.

O'Muircheartaigh, C., Krosnic, J.A. and Helic, A. (2000). Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. Harris School Working Papers Series: 01.3.

Oksenberg, L., Cannell C, Kalton G. (1991). New strategies for pre-testing survey questions. In Journal of Official Statistics, 7, 3, 349-365

Pan, Y. and Puente, M. de la (2005). Census Bureau Guideline for the Translation of Data Collection Instruments and Supporting Materials: Documentation on how the Guideline was Developed. Research Report Series, Survey Methodology 2005-06. Washington: U.S. Bureau of the Census.

Petty, R. E. and Jarvis, W. B. G. (1996). An individual differences perspective on assessing cognitive processes. In (1996) Anwering Questions, Schwarz, N. and Sudman, S., San Francisco,:Jossey-Bass Publishers

Pocock, S.J. (1983). Clinical Trials: A Practical Approach. New York: John Wiley & Sons.

Presser, S., Court, M.P., Lessler, J., Martin, E., Martin, J., Rothgeb, J.M., Singer, E. (2004). Methods for testing and evaluating survey questionnaires. In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley.

Presser, S. and Blair, J. (1994): Survey pre-testing. Do different methods give different results? In Sociological Methodology, 4, 73-104

Prüfer, P. and Rexrodt, M. (1996). Verfahren zur Evaluation von Survey-Fragen. Ein Überblick. ZUMA Arbeitsbericht, 96/05, ZUMA, Mannheim.

Prüfer, P. and Rexrodt, M. (1996): Verfahren zur Evaluation von Survey-Fragen. Ein Überblick. ZUMA Arbeitsbericht, Nr.96/05, Mannheim: ZUMA.

Prüfer, P. and Rexrodt, M. (2005). Kognitive Interviews. ZUMA How-to-Reihe15, Mannheim

Przeworski, A. and Teune, H. (1970). The Logic of Comparative Social Inquiry. New York: Wiley Interscience.

Puente, de la, M., Pan, Y. and Rose, D. (2000). An Overview of A Proposed Census Bureau Guideline for the Translation of Data Collection Instruments and Supporting Materials. Available on-line at www.fcsm.gov/03papers/delaPuente_Final.pdf

RAMON (2005). Eurostat's Classification Server. Available on-line at europa.eu.int/comm/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC#top

Redline, C. D., Lankford, C.P. (2001). Eye-movement Analysis: A New Tool for Evaluating the Design of Visually Administered Instruments (Paper and Web). Paper presented at the American Association of Public Opinion Research, Montreal, Canada.

Redline, C., Lankford, C.P. (2001). Eye-Movement Analysis: A New Tool for Evaluating the Design of Visually Administered Instruments (Paper and Web). Paper presented at the American Association of Public Opinion Research, Montreal, Canada.

Ritchie, J. (2003). The Applications of Qualitative Methods to Social Research. In Qualitative Research Practice. A Guide for Social Science Students and Researchers, eds. J. Ritchie and J. Lewis, London: Sage.

Rousseau, E. and Sanders, A. (2003). Survey Research Methodology. SAEM Annual Meeting.

Rymarchyk, G.K. (2005). Validity. Available on-line at www.socialresearchmethods.net/tutorial/Rymarchk/rymar2.htm

Sainsbury, R., Ditch, J., and Hutton, S (1993). Computer Assisted Personal Interview. Social Research Update, Issue 3, UK: University of Surrey.

Salant, P. and Dillman, D.A. (1994). How to Conduct Your Own Survey. New York: John Wiley & Sons.

Saris, W.E. (1991). Computer-Assisted Interviewing. Sage University Paper Series on Qualitative application in the Social Sciences, 07-080. Newbury Park, CA, Sage.

Satori, G. (1984). Guidelines for Concept Analysis. In Social Science Concepts, Satori, G. (ed), Beverly Hills, CA:Sage.

SCB (2004). Design your questions right, Stockholm: Statistics Sweden. Available on-line at http://www.scb.se/Grupp/Metod/_Dokument/Design_your_questions_rightB.pdf

Schaeffer, N. C. and Presser, S. (2003). The Science of Asking Questions. Annual Reviews Sociol. 29:65-88

Scheuren, F. (2004). What Is a Survey? Availabe on-line at http://www.whatisasurvey.info/

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

Schwamb, H.J. and Wein, E. (2004). Development of Questionnaires - Current Practices of the Federal Statistical Office Germany (Destatis) and Perspectives. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, 24-26 May 2004.

Schwarz, N. and Hippler, H.J. (1991). Response Alternatives: the Impact of their Choice and Presentation Order. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley & Sons, 41-56.

Schwarz, N., Hippler, H.J and Noelle-Neumann, E. (1992). A Cognitive Model of Response-Order Effects in Survey Measurement. In Context Effects in Social and Psychological Research, eds. N. Schwarz and S. Sudman, New York: Springer-Verlag, 187-201

Sinclair, M.D., Gastwirth, J.L. (1996). On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data. Journal of the American Statistical Association, 91, 435, 961-969.

Smith, T.W. (1991). Context Effects in the General Social Survey. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley & Sons.

Smith, T.W. (2004). Developing and Evaluating Cross-National Survey Instruments. In Methods for Testing and Evaluating Survey Questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer. New York: John Wiley and Sons.

Smyth, J.D, Dillman, D.A, Christian L.M. and Stern L.J. (2006). Comparing Check-All and Forced-Choice Question Formats in Web Surveys. Public Opinion Quarterly, 70, 1, 66-77.

Snijkers, G.J.M.E. (2002). Cognitive Laboratory Experiences: On Pre-testing Computerised Questionnaires and Data Quality. Ph.D Thesis, Utrecht University, Utrecht, and Statistics Netherlands, Heerlen

Statistics Canada (2002). Policy on the Review and Testing of Questionnaires. Statistics Canada Methods and Standards Committee, Ottawa: Statistics Canada.

Statistics Canada (2003), Statistics Canada Quality Guidelines. Available on-line http://www.statcan.ca/bsolc/english/bsolc?catno=12-539-X&CHROPG=1

Statistics Finland (2002), Quality Guidelines for Official Statistics. Available on-line http://www.stat.fi/tk/tt/laatuatilastoissa/alku_en.html

Statistics New Zealand, ed. (1995). A guide to good survey design, Wellington: Statistics New Zealand.

Sudman, S. and Bradburn, N. (1982). Asking Questions: a Practical Guide to Questionnaire Design. San Francisco: Jossey Bass.

Sudman, S., Bradburn, N. and Schwarz, N. (1996). Thinking about answers: the application of cognitive processes to survey methodology. San Francisco: Jossey Bass.

Sudman, S., Willimack, D.K., Nichols, E. and Mesenbourg, T.L. (2000). Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies. Proceedings of the Second International Conference on Establishment Surveys. American Statistical Association, pp.327-337.

Sykes, W. and Morton-Williams. (1987). Evaluating Survey Questions. Journal of Official Statistics, 3, 2, 191-207.

Tarnai J., and Moore, D.L. (2004). Methods for testing and evaluating Computer Assisted questionnaires. In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer. New Jersey: Wiley.

Thomas, R. and Purdon, S. (1994). Telephone Methods for Social Surveys. Social Research Update, Issue 3, UK: University of Surrey.

Tortora, R. (1985). CATI in an Agricultural Statistical Agency. Journal of Official Statistics, Vol.1 N.º 3, pp. 301-314, Statistics Sweden

Tourangeau, R. (1984). Cognitive Sciences and Survey Methods. In Cognitive aspects of survey methodology: building a bridge between disciplines, eds. T.B. Jabine, M.L. Straf, J.M. Tanur and R. Tourangeau, Washinghton D.C.: National Academy of Science.

Tourangeau, R. and Smith, T. (1996). Asking Sensitive Questions: the Impact of Data Collective Mode, Question Format, and Question Context. Public Opinion Quarterly, Vol. 60: 275-304

Tourangeau, R.(2004). Experimental design considerations for testing and evaluating questionnaires. In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer. New Jersey: Wiley.

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

Treiman, D.J. (1977). Occupational Prestige in Comparative Perspective. New York: Academic Press.

Tulving, E. (1972). Episodic and semantic memory. In Organisation of memory, eds. E. Tulving and W. Donaldson, New York: Academic press.

Tversky, A. (1972). Elimination by aspects: a theory of choice. Psychological Review, 79, 281-299

U.S. Census Bureau (1985). Evaluating Censuses of Population and Housing. Statistical Training Document, ISP-TR-85, Washington, D.C.

U.S. Census Bureau (2003). Testing Questionnaires and Related Materials for Surveys and Censuses. Census Bureau Methodology and Standards Council, Washington, D.C.

U.S. Census Bureau (2004). Census Bureau Guideline: Language Translation of Data Collection Instruments and Supporting Materials. Available on-line at www.census.gov/cac/www/Paper(3)-LanguageWG_Spring2004.html

U.S. Department of Education: National Center for Educational Statistics. Response Variance in the 1994-95 Teacher Follow-up Survey. Working Paper No. 98-13, by John M. Bushery, Irwin D. Schreiner, and Amy Newman-Smith. Project Officer, Steven Kaufman. Washington, D.C. 1998.

UNESCO (1997). International Standard Classification of Education. In Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables, eds. J.H.P. Hoffmeyer-Zlotnik and C. Wolf, New York: Kluwer Academic/Plenum Publishers, 195-220. Available on-line at www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm

Van der Zouwen, J. and Smit, J.H. (2004). Evaluating Questions by Analysing Patterns of Behaviour Codes Question-Answer Sequences: A diagnostic Approach.. In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer, New Jersey: Wiley, 109-130

Ware-Martin, A. (1999). Introducing and implementing cognitive techniques at the Energy Information Administration: An establishment survey environment. Statistical policy working paper 28, Seminar on Interagency Coordination and Cooperation, Office of Management and Budget, Federal Committee on Statistical Methodology, Washington, pp. 66-82.

Watson, J.B. (1924). Behaviorism, Chicago: University of Chicago Press

Webb, B. and Webb, S. (1932). Methods of Social Study, London: Longmans Green

Weeks, M.F. (1992). Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Operations. Journal of Official Statistics, 4, 445-465.

Wensing, F. Barresi J. Finlay D. (2003) Developing an optimal screen layout for CAI. Paper presented at the 8° International Blaise User Conference (IBUC 2003), Copenhagen, Denmark, 21-23 May. Available on-line at www.blaiseusers.org/IBUCPDFS/2003/Developing_an_.pdf.

WHO (2002). World Health Survey. Translation Guidelines. Available on-line at www3.who.int/whs/P/translations.html

Willimack, D.K. (1999). Understanding the questionnaire in business surveys. Proceedings of the Survey Research Methods Section (ASA). Paper Presented at the 54th Annual Conference of the American Association for Public Opinion Research, St. Pete Beach, Florida, May 13-16.

Willimack, D.K. and Nichols, E. (2001). Building an Alternative Response Process Model for Business Surveys. Proceedings of the Annual Meeting of the American Statistical Association.

Willis, G. (2004). Cognitive interviewing revisited: a useful technique, in theory? In Methods for testing and evaluating survey questionnaires, eds. S. Presser, J.M. Rothgeb, M.P. Couper, J. Lessler, J.T. Martin, J. Martin, and E. Singer. New Jersey: Wiley.

Willis, G.B. (2005). Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage.

Wilss, W. (1982). The Science of Translation. Problems and Methods. Tübingen: Gunter Narr Verlag.

Worcester, R., Lagos, M., and Basanez, M. (2000). Problems and Progress in Cross-National Studies: Lessons learned the hard way. Paper presented at the WAPOR/AAPOR annual conference, Portland, Ore., USA, 18 May.

Zukerberg, A.L., Von Thurn DR Moore JC (1995): Practical considerations in sample selection for behaviour coding pretests. Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, pp. 1116-1121. Available on-line at http://www.census.gov/srd/papers/pdf/az9501.pdf

Index by topic