Journal of Official Statistics, Vol. 11. No. 3, 1995, pp. 309–323

Pretesting Procedures at Statistics Sweden's Measurement, Evaluation and Development Laboratory

Lars R. Bergman[1]

In 1989 the Measurement, Evaluation and Development Laboratory (MED) was established at Statistics Sweden after a period of testing and evaluating different pretesting procedures. The pilot work and MED's current procedures are described against the background of different approaches on the international scene. Much of the time, mail questionnaires are tested by personally interviewing respondents in their homes using retrospective probing on problems they had with the questionnaire. Encouraging results are reported from a survey of the laboratory's procedures and results directed to MED's clients. In the paper, the MED procedures are compared to those used for pretesting at certain laboratories in the U.S.A., and alternative directions for MED's future work are also briefly discussed. We think that this paper on the organisation we have created, as well as on the mistakes we have made, can provide some useful insights for other pretesting laboratories as well as for organisations planning such activities.

Key words: Pretesting; cognitive laboratory; questionnaire development; mail questionnaires.

1. Introduction

The Measurement, Evaluation and Development Laboratory (MED) was established at Statistics Sweden in the fall of 1989, after exploring different ways of conducting questionnaire pretests in a set of small studies. In the first four years, 27 questionnaire drafts have been tested by MED and detailed reports written. Of these 27 questionnaires, 16 were to be administered by mail, 4 as face to face interviews and 7 by telephone. A large number of less ambitious tests have also been performed, often of single questions or blocks of questions. In addition, consultation concerning questionnaire evaluation and construction have been undertaken. We think we can justly claim that the laboratory has been perceived successful and as filling a vital need. The Director General has also ordered that all new surveys must run both a qualified pretest of the questionnaire as well as a pilot study.

The present paper aims at describing MED against a background of the different approaches to questionnaire pretesting seen internationally. We believe that a description of the organisation we have created, as well as of the mistakes we have made, might provide some useful information both to similar laboratories and to agencies which plan to start such activities. In the next section the need for testing

[1] Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden.
Acknowledgements: The author is grateful to the associate editor and to two anonymous reviewers for many useful suggestions. Chris Denell has very competently struggled with my English.

questionnaires is articulated and in Section 3 an overview is given of different methods for developing and testing questionnaires. In Section 4, MED is described, first the methodological work leading to the establishment of MED and then the main methods used for field pretesting: testing mail questionnaires in the context of a personal interview and testing questionnaires for personal interviews. Our experiences in using think-aloud-protocols for testing mail questionnaires and information brochures are also discussed. In Section 5, MED procedures are compared to the pretesting procedures used by some organisations in the United States. Finally, future directions for MED are discussed.

2. The Need for Testing Questionnaires

The importance of considering nonsampling errors in surveys is now generally acknowledged (for a review, see Groves 1989). The subject matter content of the questionnaire and the formulation of its questions are often the factors that influence these errors most strongly. It is also widely recognised that more resources should be allocated to the careful development and testing of questionnaires (see for instance Hunt, Sparkman, and Wilcox 1982). Questionnaire design is often seen as the weak link in the survey measurement process (Nathan and Sirken 1986; Sirken 1986), and the pretesting that is undertaken is often primitive (Presser 1989). Jobe and Mingay (1989, p. 1053) go so far as to say that "Recent research has shown that poorly designed questions are often not revealed even by field testing. . . .this same research has shown that techniques exist to improve questionnaire design."

3. Different Methods for Developing and Testing Questionnaires

The process of developing and testing questions and questionnaires covers several activities:

1. Initial exploration and feasibility study. The aim is to determine whether the prospective respondents can answer the tentative questions in the subject matter field of interest, what kind of questions the respondents think are important in this field, and which terms they use for various items.
2. Experimenting and testing of single questions or groups of questions to determine how they are perceived by typical respondents and what cognitive processes are involved, including those activated when giving an answer. Often small and informal samples are used, but sometimes large samples are used, for instance to test the effects of variously formulated questions, using a split-ballot design.
3. Pretesting a complete preliminary questionnaire (in a laboratory setting or under field conditions), using special techniques to obtain as much information as possible. Usually small and informal samples are used.
4. A pilot study under field conditions of the not quite final version of the questionnaire. The sample and the procedures should resemble the conditions in the main survey as much as possible. Of course, this should not prevent the inclusion of embedding experiments for elucidating particular issues (Fienberg and Tanur

1989; Tanur and Fienberg 1992). Normally, in this final round the sample should comprise at least several hundred subjects. For discussions of pilot studies, see Kasprzyk (1988) and Lyberg and Dean (1989).

The aims of questionnaire pretesting and pilot surveys are only partly overlapping, and normally the one cannot replace the other. Both are needed to provide the necessary information for achieving a survey of high quality (see e.g., Bercini 1989), but sometimes elements of both can be included within the framework of a single large methods study. One can profit considerably from a simultaneous consideration of the information collected by pretesting (e.g., probing and interviewer debriefing) and the information collected in a pilot study (e.g., comparisons of response distributions and item nonresponse). A good example is given by Esposito, Campanelli, Rothgeb, and Polivka (1991) in their redesign of the U.S. Current Population Survey.

Below, we will limit the discussion to activities relating to testing questions and pretesting complete questionnaires, focusing on the latter. For a general discussion of questionnaire development, the reader is referred to DeMaio (1983) and Platek (1985) and for overviews of pretesting procedures to Oksenberg, Cannell, and Kalton (1991), Forsyth and Lessler (1991) and Nelson (1985).

3.1. Testing single questions

Single questions can be tested in many ways. One is to conduct *split-ballot experiments* which provide straightforward information about the response effects of different formulations. Using this method on a sufficiently large representative sample can considerably strengthen the external validity of the results. Schuman and Presser (1981) offer excellent examples of the results that can be obtained by this approach. However, from a pretesting perspective, split-ballot experiments are of limited value, because they usually are too expensive and time-consuming to carry out during the pretest phase, and must be conducted during the pilot phase.

Standard split-ballot experiments often do not yield much information about *why* a question is answered in a certain way. Instead, such information can be obtained by a *cognitive laboratory approach.* Here, methods from cognitive psychology are used in an experimental setting to obtain insights into the respondent's cognitive processing of the questions. Some useful tools are experimental manipulations, think-aloud protocols and special probes. In this way, information is collected which not only helps to identify potential problems with a question, but also yields relevant information on how to improve it (see e.g., Lessler, Tourangeau, and Salter 1989; Jobe and Mingay 1989; Royston, Bercini, Sirken, and Mingay 1986). Within this cognitive framework, a variety of "new" methods are coming into use, for instance, the use of paraphrases and vignette classifications. Of course, the cognitive laboratory approach is also used for testing whole questionnaires. A taxonomy and discussion of cognitive laboratory techniques is given by Forsyth and Lessler (1991).

3.2. Testing complete questionnaires

Focusing on the testing of complete interview questionnaires, Oksenberg, Cannell,

and Kalton (1991) have reported some new techniques for pretesting, tried out in a comprehensive experiment. They recommend using (a) behaviour coding (both interviewer and respondent behaviour), (b) special probes, (c) special training of pretest interviewers on question rating, and (d) a modified type of interviewer debriefing using information from (a) - (c).

Quite another approach was developed by Belson (1981), who, a few days after the survey, had the respondents re-interviewed by special interviewers, focusing on the problems the respondents had in answering certain questions in the first interview. The results of such an approach can be used for providing information about the reliability and validity of the questions in a survey as well as for questionnaire development.

In a carefully designed study, Presser and Blair (1993) compared four different strategies of pretesting an interview survey: conventional pretesting, a cognitive laboratory approach, behaviour coding, and expert panels. The four techniques were compared on costs, reliability (not validity), and productivity of identifying problem questions. Their results suggest that expert panels can be useful in pretesting, have a high productivity in identifying problems, and are less expensive than other methods. Cognitive interviewing, which is the method focused on in the present article, was found to be useful for identifying respondent problems and analysis problems, but not for identifying interviewer problems.

The above procedures have been developed in the context of interview surveys. An interesting contrast is found in a report from the United States General Accounting Office (GAO 1986) about the pretesting of self-administered questionnaires. GAO uses a special pretest administrator and, after explaining the purpose of the pretest, he or she asks the respondent to fill in the form. While this is being done, the administrator collects various kinds of information (the time each question takes, nonverbal behaviour, etc.). The information is used for different purposes, one of which is to provide input to the debriefing with the respondent that follows.

While it is not possible here to discuss all methods of pretesting, we hope the summary will suffice as background and indicate that techniques such as the ones described can be highly useful in improving questionnaires.

4. A Description of MED

4.1. Preparatory methodological work

During the last decades, an extensive testing of questions and data collection methods has been conducted at Statistics Sweden. The work has mainly consisted in testing new questions and questionnaires and data collection methods for use in studies of the general population. Many results are reviewed by Bergman and Wärneryd (1982), in Thorslund and Wärneryd (1985), and in Wärneryd (1986). Usually, various types of pilot studies were conducted, sometimes including split-ballot experiments, or re-interview studies. This methodological work highlighted the importance of developing sound procedures for pretesting.

When the MED was established, a number of small pilot studies were conducted to assess the applicability of different procedures for swift pretesting of questionnaires concentrating on questionnaires for self-administration and personal interview. The results are described below.

4.1.1. Questionnaires for self-administration

With regard to questionnaires for self-administration, two procedures were of primary interest:

1. In the first procedure the respondent is given the questionnaire as a hand-out within the context of a personal interview. His or her behaviour while filling it in is noted and coded, and afterwards he or she is asked about any problems with some or all of the questions. Including the pretest for the 1990 census, 44 interviews were conducted to test this procedure. The main findings were:
   - It was preferable to have the respondent fill in the entire questionnaire before probing rather than probing on a question-by-question basis. In the latter case, after a few questions the respondent frequently "learned" what it was all about and started to answer as a quasi-expert, giving advice instead of reacting to the questions as a respondent.
   - It was most important to follow up the questions with specific probes and not just use general probes like whether the respondent had any problems with the question. The specific probes were most frequently based on observations made by the interviewer during the time the respondent filled in the questionnaire. The most common observations related to hesitating before answering a question, changing an answer, or spontaneously commenting on a question. The probes also appeared to train the respondent as to the kind of information that was of interest.
   - It was easy to instruct and obtain cooperation from the respondents; it was much more difficult to train high-pace survey interviewers to slow down and listen for problems. Being used to interview surveys, the interviewers sometimes found it difficult to understand the different nature of the respondent's information processing in the case of a mail questionnaire. For instance, in a self-administered questionnaire the response alternatives are experienced by the respondent as parts of the question.
2. The second procedure of interest was to send out a draft questionnaire as a mail questionnaire, asking the respondent to fill it in and comment. While this is a comparatively low-quality method, it is also inexpensive and we thought it would be of interest to see whether it could be used as a complementary method. A 36-item questionnaire on living conditions was therefore sent to 70 subjects (the same questionnaire was used in some interviews in (1) above). Red and blue pens were also included, together with a small booklet of vital statistics, all to be kept by the respondent. They were asked to fill in the questionnaire with the blue pen first and then to go through it and add comments, problems, etc. with the red pen. The outcome was not encouraging:
   - Only 50% had completed the questionnaire after one reminder.

- Only 29% of the respondents had commented on their answers to the survey questions.
- Those who commented gave little useful information; usually they just elaborated on their answers.

However, some information was gained (e.g., about item nonresponse rates and problems with following routing directions), and the cost for the entire data collection equalled the cost of only three survey interviews. Consequently, it might be worthwhile to attempt a revised procedure in which the respondent is asked to comment while filling in the questionnaire. The instructions should include an example of how the respondent is to go about his or her task.

4.1.2 Questionnaires for personal interviews

With regard to questionnaires used in personal interviews, we attempted a procedure for testing the questionnaire in that context. Three interviewers interviewed 27 persons (out of 53 persons drawn from the telephone directory). The high nonresponse rate is not surprising, as the interviewers were instructed not to trace the persons they could not immediately contact and not to try to persuade reluctant subjects. In this experiment, the questionnaire about living conditions that was used in the mail questionnaire experiment described above was transformed into an interview questionnaire. Some of the results were:

- The main approach was first to conduct the entire interview according to the instructions, and afterwards ask more indepth questions about a subgroup of eleven questions that were a priori thought to be troublesome. The interviewer reread each such question and noted the respondent's original answer. Then we discussed the question following a standardised procedure. This strategy worked well, but it was evident that much more information was obtained for questions where specific, rather than general, probes were used.
- In a few cases, probes based on the respondent's answers or comments were introduced during the interview. As expected, this sometimes caused confusion of roles (is the interviewee a respondent or an expert?) and the strategy was only rarely used. (There was no difficulty in asking respondents to clarify their answers; the role problem tended to occur when they were asked how they defined various words and concepts.)
- The shift from the basic interview to the probing session was not easy. The interviewer had to change from a fast-pace survey interviewer to a slow-going listener. The respondent, on the other hand, had to understand that, while short, accurate answers were appropriate for the main interview, long reports about problems and thoughts were now encouraged. The need for thorough training of the interviewers was clear, as was the need for a very clear division of the interview in two parts, to permit the respondent to get a firm grasp of the two different roles he or she was expected to play. The respondents were annoyed only rarely by the fact that the interview continued after the survey questions. Informing them beforehand about the two parts of the interview was found to be useful.

- The results yielded by the interview procedure described above were summarised by ranking the eleven questions according to how problematic they were, based on the interviewers' reports of how many respondents had problems with each question. These results were then compared to the judgements of two experts and to the results of the LIX method for measuring the difficulty of a text (Björnsson 1968). The LIX value of the text is computed by adding the ratio between the number of words and the number of sentences to the percentage of long words (words with more than six letters). Based on the eleven questions ranked with regard to how problematic they appeared according to the different methods, the correlation coefficient (Spearman's rho) between the interview procedure results and each of the two expert judgements was .60 and .74. With the LIX method, a negative correlation was obtained (rho = −.56). It was contended that the LIX method appeared irrelevant, while the interview procedure and the expert judgements both gave relevant information that only partially overlapped. It is interesting to note, though, that although both expert judgements were substantially correlated with the results of the interview procedure, they were only moderately correlated to one another (rho = .43). There was strong agreement only on questions with pronounced problems.

The above has been intended as a short overview of the information used in planning the pretesting activities at MED. These activities are described below.

4.2. Pretesting by MED

4.2.1. Field pretesting

The basic part of MED's questionnaire testing is the collection of information from a number of genuine subjects under field conditions (i.e., collected outside the laboratory). Normally the information is collected by specially trained field representatives of the Statistics Sweden regular interviewer staff. In the case of an interview survey questionnaire, this means that the interview is *first* conducted as it would be done in the real survey, and that the probes come *afterwards*. In the case of a mail questionnaire, it is *first* filled in by the respondent without any interaction with the interviewer, and the probing follows *afterwards*. The respondent is encouraged to comment as much as he or she wants while answering the survey questions, but the interviewer only listens. It is emphasised that it is the questionnaire that is being tested, not the respondent, and that it is of great importance that we learn about any problems he or she may have in answering the questions.

Three major kinds of probes are used:

1. The interviewer asks for the respondent's general impression of (a) the questionnaire and (b) the information letter or brochure.
2. In case of an interview survey, the interviewer notes during the regular interview whether the interviewee appears to have problems with any questions (for instance, asks for clarification, or gives inappropriate, qualified or hesitant answers). Such questions are followed up after the interview, using the observations as input (for instance, 'I noticed that you hesitated when answering

Question 7 about.....Can you tell me why?'). A similar procedure is used for self-administered questionnaires, but the observations are somewhat different (e.g., the time it took to answer the question, if the answer was revised or inappropriate, body language indicating problems, comments or questions to the interviewer).

3. A specific probe is formulated beforehand and included in the instructions to the interviewers in cases when the questionnaire expert anticipates a specific problem.

MED uses nine professional interviewers at Statistics Sweden in the pretesting. These nine interviewers have been given special training and are regularly given feedback about their interview behaviour, on the basis of the written scripts they provide and on tape recordings of their interviews. The interviewers receive written instructions, reiterating the proper interviewing and probing procedures and specifying the special probes MED wants them to use. Special care is taken in formulating the instructions and training the interviewers so that the change of roles between the regular data collection and the subsequent probing phase is made clear. The interviewers report by:

1. Returning the filled-in questionnaires with his or her comments (if the questionnaire tested is self-administered, both the interviewer's and the respondent's copies are returned).
2. Filling in a special form, detailing the information collected.
3. Returning the audio tapes of the interviews.
4. Writing a brief report about what the interviewer sees as the major problems with the questionnaire and his or her suggestions for solutions.

How is a suitable sample for questionnaire testing to be designed? A strict random sample from the target population, with efficient field work tracing those not at home, etc., is usually unrealistic. Often it is not even desired, considering the small samples that have to be used. Usually it is only important to ensure that the subjects used vary strongly in their personal characteristics (e.g., educational level, sex and age). We have also found that convincing reluctant respondents to cooperate often leads to less useful interviews. They report fewer problems with the questions and volunteer less information; their answers appear to be influenced by a desire to shorten the interview. These considerations, together with the demand for swift results and a tight budget, have led us to use convenience samples. Often the telephone directory is used as the sampling frame, and the interviewers are given an extra list of subjects and instructed to obtain a certain number of interviews. They are further instructed to ensure that certain specified categories of respondents are included.

4.2.2. Pretesting in the laboratory

Two kinds of basic questionnaire pretesting are performed in the laboratory, think-aloud protocols and videotaped standard data collections. Normally MED's central staff conduct the interviews.

Think-aloud-protocols (Ericsson and Simon 1984) have proved especially useful for testing self-administered questionnaires. The protocols give information not only about problems experienced, but also about the cognitive processing of the

questionnaire by the respondent (for instance, the aspects of the content the interviewee focuses on, the order, and the aspects that are ignored). An analysis of such protocols usually provides a greatly increased understanding of how the questionnaire works from the subject's point of view, within the context provided by instructions and information material. Such an analysis can be similar to a clinical case study, where the revealed processes are thought to be generalisable to most persons. In that case, a small number of protocols can form the basis for generalised conclusions and recommendations.

It is not self-evident how to produce good think-aloud data. The procedure now used at MED for self-administered questionnaires contains four steps:

1. The instructor thinks aloud, using a test questionnaire. The entire session is videotaped or audiotaped, so that by the time the main data collection starts, the subject is familiar with the situation. If it is videotaped, zoom optics are used to provide a close picture of the subject's face, hands, pen and paper.
2. The subject is trained to think aloud on a test questionnaire.
3. The main data collection is performed. While thinking aloud, the subject fills in the questionnaire without any interruption or comments by the tester, who observes the behaviour of the subject and continuously takes notes of the time required for each question, the respondent's body language, facial expression, etc. If necessary, the notes are later checked against the tape.
4. A debriefing session is conducted, based on the subject's experiences, the tester's observations, and the preformulated probes.

4.3. The client's views of the laboratory

In May 1993 a brief questionnaire was sent to 19 clients for whom the laboratory had undertaken a complete pretest evaluation of a questionnaire, including the presentation of a full report. By October (and after several reminders) questionnaires had been filled in and returned by 17 of the clients. All the clients were survey statisticians working with different surveys at Statistics Sweden (in ten cases the survey was commissioned by an agency outside Statistics Sweden). Four of the 17 pretest evaluations concerned establishment surveys and the rest surveys of the general population. Six evaluations were for interview questionnaires and the remaining eleven for self-administered ones. Table 1 summarises the results of the survey for five closed-ended questions.

It can be seen from Table 1 that the clients were fairly satisfied with the services provided by the laboratory. However, the support was not total, in the sense that only one client was prepared to pay more for more qualified MED pretesting. Seven of the 17 clients considered the laboratory's comments only to a limited extent in the construction of the final questionnaire; the most frequent reason being the need to retain time series. Six clients reported problems with the final questionnaire, and four reported one or more problems that had not been detected by the pretest. Reviewing these problems reported by the clients, the MED assessment is that in three cases mistakes had been made by the laboratory. By more careful preparatory work before the pretesting, the mistakes could probably have been avoided. The central

Table 1. Response distributions for five questions to clients about MED. n = 17

Question	Response category	Frequency
Q1. How useful was the MED pretest in developing your questionnaire?	No use whatsoever	0
	Some use but not worth the cost	1
	Probably worth the cost	9
	It was absolutely worth the cost	6
	No answer	1
Q2. What level of ambition in the MED pretest would have been right for you?	A lower-than usual level, without a written report	3
	The usual level with a simple report based on 6-7 interviews	10
	A higher-than-usual level with a somewhat extended report based on 10 interviews	1
	Other alternative, namely.....	3
Q4b. To how large an extent did you consider our comments when constructing the final questionnaire?	To 100 percent	0
	To a large extent	10
	Partly	7
	(Almost) not at all	0
Q4c. How well do you think the final questionnaire works/has worked?	Very well	4
	Well	6
	Not so well	6
	Badly	0
	Very badly	0
	No answer	1
Q4e. Have you detected any problems with the questionnaire or the data collection which were not found by the pretesting?	No	12
	Yes, one	2
	Yes, several	2
	Nó answer	1

Note. In some cases the response alternatives have been abbreviated.

staff should have suspected the problems and transformed them into interviewer instructions for directed probes.

When asked for ways to improve the MED pretesting procedures, clients most frequently mentioned a more active MED role in the construction of the questionnaire and not, as is usually the case, contributions at a later stage, when the questionnaire construction deadline is close. Some clients would also like a stronger focus on the formulation of revised questions and corresponding less emphasis on evaluation.

5. MED Procedures and the Pretesting Procedures Used by Some U.S. Organisations

During March 1990 I had the privilege to visit experts conducting pretests at organisations in the U.S.A.: the U.S. Bureau of the Census (BC), the General Accounting Office (GAO), the Questionnaire Design Research Laboratory at the National Center for Health Statistics (NCHS), and the Center for Research in Statistics at Research Triangle Institute (RTI).

It was interesting to note the similarities as well as the differences to MED procedures. It should be stressed that many similarities exist (which is not surprising

since we extensively relied on American questionnaire research design experiences in the construction of MED). However, some differences were salient. These are discussed below, and it should be emphasised that here I merely present some impressions, covering very limited aspects of the U.S. work.

There is a major difference in how the *probing* is done. When testing whole questionnaires, MED and GAO usually probe *after* the entire questionnaire has been answered, while BC, NCHS, and RTI usually do it after each question. The question-by-question technique is more flexible and probes can be used for single questions, or blocks of questions, as well as for entire questionnaires. This procedure is common at the early stages of questionnaire development. It can also give more indepth information about the functioning of a single question. The advantage of the MED and GAO approach is that the answering process is similar to the ultimate one, in that the questions are asked and answered in the same way as in the actual survey. This gives the questions the proper context and also avoids the problem that the respondent, after being probed on a few questions, changes his or her way of reacting, and no longer behaves as an ordinary respondent.

Presumably, one reason for this difference in probing strategy is the dominance of testing interview survey questionnaires, often in the early construction stage, at BC, NCHS, and RTI, in contrast to the dominance of testing entire self-administered questionnaires at MED and GAO. After careful consideration, we have decided to continue to probe after all survey questions have been answered as the MED basic procedure, and to use question-by-question probing mainly when testing parts of a questionnaire.

BC, NCHS, and RTI all use *think-aloud protocols* to a certain extent. The respondent is encouraged to verbalise, aloud, what he or she thinks when answering a question. However, these agencies do not appear to rely on complete think-aloud protocols, such as are used at MED. The difference is natural, considering their focus on interview surveys and question-by-question probing and the MED focus on self-administered questionnaire probing after all questions have been answered. We believe that for self-administered questionnaires, the MED procedure is often preferable, as it provides insight into the complete cognitive process of responding to the questionnaire. The subject's attention focus can be traced, while he or she is generating the data which also indicates how the layout works and how he or she makes use of the information material and the instructions.

Due to its very limited resources, so far MED has focused on *testing preliminary questionnaires*, developed by the client with only a minimum of MED support and advice. This is in contrast to the laboratories listed above, where the questionnaire design expert usually is engaged from the start of the development work. This approach is most pronounced at GAO, where a new questionnaire often is developed from the very beginning by two officers: the questionnaire design expert and the client's subject matter expert. They form a team that is responsible for producing the questionnaire. The team has all the necessary knowledge and authority to make on-the-spot changes and improvements, which facilitates successive revisions of the questionnaire. The efficiency of this iterative approach is illustrated by a collection of examples in Featherston and Moy (1989).

6. Future Directions of MED

The cognitive laboratory approach for testing questionnaires tends to focus on interview surveys, also a concern of MED. However, our emphasis is on the testing of self-administered questionnaires, whether they are mail questionnaires for a sample from the general population or administrative forms. One reason for this is that high quality name and address population frames are available in Sweden, permitting mail surveys of a high quality. Consequently MED undertakes development and research work mainly on methods for developing self-administered questionnaires. In the future we plan to focus on the following activities:

1. Implementing more objective and quantitative methods in pretesting. Oksenberg, Cannell, and Kalton (1991) have pointed to the dangers inherent in relying on the kind of subjective information that is usually obtained in questionnaire pretesting. Their criticism is also relevant for the pretesting currently being carried out by MED. For example, one method we plan to explore in a laboratory setting is to connect a galvanic skin response apparatus to the respondent, synchronised to a video camera showing the respondent's hand and the paper as the questionnaire is filled in. Time lags and changes in skin conductance can then be used as supplementary information about reactions to questions (for instance, to questions suspected to be sensitive). Another method we want to explore is to vary question wording within the same instrument. In the context of a very long questionnaire administered in the laboratory, a moderate number of important questions are asked four times with a large number of other questions between each repetition (or two slightly differently formulated questions asked twice). The amount of variation in the answers to such repeated questions is then used as an indicator of problems with these questions.
2. Methods for using expert panels will be explored. Presser and Blair (1993) studied experimentally different methods for pretesting a questionnaire and found expert panels highly useful (the paper is briefly reviewed in Section 3.2). Using such a non-empirical approach can be a very attractive complement to our cognitive empirical approach to pretesting. It is true that our own results from more informal studies of expert judgements (some of which are reported in Section 4.1.2) have not been encouraging. However, the reason for this may well be our comparatively low level of ambition when acquiring the expert judgements (no qualified expert panels were used); it is probably of decisive importance how the expert panel is constructed. Following Presser and Blair, it seems reasonable to include three persons in the panel representing both expertise in questionnaire design, cognitive aspects and survey methodology. It is probably also a good idea that a majority of the panel is not heavily involved in the program responsible for the survey.
3. We plan to conduct some work on improving methods for testing and successively modifying instructions and other information material in a laboratory setting based on think-aloud protocols. Such protocols collected in the context of self-administered questionnaires have often revealed a deplorable discrepancy between, on one hand, the information provided and, on the other hand, the information that a respondent actually reads and understands.

4. We want to provide more support in the early phases of questionnaire construction, preferably following the GAO team-work model in the development of difficult questionnaires where subject matter expertise is especially important. The approach of getting involved early in the construction also receives support from many of the laboratory's clients (see Section 4.3) and is standard practice for most cognitive laboratories. However, this activity requires greater resources than does the evaluating of questionnaire drafts. We also think it is important to retain our independent status and produce our own written evaluation of the questionnaire drafts, based on empirical testing. Otherwise there is a danger that we will be absorbed into and be subservient to the client's questionnaire construction team. For instance, sometimes the client imposes various constraints on the questionnaire development that are detrimental to the quality of the information collected and this has to be pointed out (it could be an unsatisfactory ambition level of the testing, poorly constructed questions that may not be changed because they are part of a time series and so on).
5. The laboratory plans to explore methods for using a panel of respondents connected to MED via terminals and printers. Optimally, at least fifty subjects should be included in the panel and used for questionnaire pretesting, instructions and questionnaires to be given by telecommunication means. However, so far the preliminary results from using such methods at Statistics Sweden for the collection of survey information from the general population have not been very encouraging. A major problem has been a very high nonresponse rate, especially for older people, with a possible strong over-representation of computer-oriented persons among the respondents (Lindeberg 1993). Hence, people with difficulties in understanding questions might be underrepresented among the respondents, an extremely serious problem in pretesting.
6. It is an important task for MED to monitor problems and explore facilities arising from the switch from paper and pencil techniques to computer-assisted data collection. CATI, CAPI and electronic questionnaires for self-administration are increasingly used at Statistics Sweden. Surveys are designed using the new data collection methods, and recurrent paper and pencil surveys are often adjusted to computer-based data collection methods. This conversion is not free of problems. For instance, apparently innocuous changes in layout can sometimes affect the results (Bergman, Kristiansson, Olofsson, and Säfström 1994). The computer-based data collection techniques put new demands on questionnaire testing, but also provide new opportunities for experimenting with questions since the administration of differently formulated questions becomes much simpler. We want to take advantage of new technologies and their spin-off benefits.

It is hoped that with the use of modern networks we will be able to share experiences, problems and to some extent experiments with other agencies. However, language is an obvious barrier, since questions often are difficult to translate accurately. Nevertheless, this is an interesting area for exploration. The process of translating questions and of discussing problems caused by language may also teach

us how respondents experience questions. At least this was our experience in a joint questionnaire development project undertaken by Swedish and Canadian researchers, both groups using the resources of their country's labour force survey.

7. References

Bercini, D. (1989). Pretesting Questionnaires in the Laboratory: An Alternative Approach. Invited paper presented at the 1989 EPA/A&WMA Symposium on Total Exposure Assessment Methodology, Las Vegas, Nevada, November, 1989.

Belson, W. (1981). The Design and Understanding of Survey Questions. London: Gower.

Bergman, L.R., Kristiansson, K.E., Olofsson, A., and Säfström, M. (1994). Decentralised CATI Versus Paper and Pencil Interviewing: Effects on the Results in the Swedish Labour Force Surveys. Journal of Official Statistics, 10, 181–195.

Bergman, L.R. and Wärneryd, B. (1982). Om Datainsamling i Surveyundersökningar. [Data Collection in Survey Research] Stockholm: SCB, Liber. (In Swedish.)

Björnsson, C.H. (1968). Läsbarhet. [Readability] Stockholm: Liber. (In Swedish.)

DeMaio, T.J. (1983). Approaches to Developing Questionnaires. Statistical Policy Working Paper 10. Washington, D.C.: U.S. Government Printing Office.

Ericsson, K.A. and Simon, H.A. (1984). Protocol Analysis: Verbal Reports as Data. Cambridge, Mass.: MIT Press.

Esposito, J.L., Campanelli, P.C., Rothgeb, J., and Polivka, A.E. (1991). Determining Which Questions Are Best: Methodologies for Evaluating Survey Questions. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Featherston, F. and Moy, L. (1989). EXAMPLES. Designing Questionnaire Items. Successes and Failures. Unpublished manuscript, Washington, DC.: GAO.

Fienberg, S.E. and Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. Science, 243, 1017–1022.

Forsyth, B.H. and Lessler, J.T. (1991). Cognitive Laboratory Methods: A Taxonomy. In P.P. Biemer, R.L. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman, (eds.), Measurement Errors in Surveys. New York: John Wiley.

General Accounting Office (GAO) (1986). Developing and Using Questionnaires. Washington, DC.: United States General Accounting Office. Program Evaluation and Methodology Division.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley.

Hunt, S. D., Sparkman, R.D., and Wilcox, J. B. (1982). The Pretest in Survey Research: Issues and Preliminary Findings. Journal of Marketing Research, 19, 269–273.

Jobe, J.B. and Mingay, D.J. (1989). Cognitive Research Improves Questionnaires. American Journal of Public Health, 79, 1053–1055.

Kasprzyk, D. (1988). The Survey of Income and Program Participation: An Overview and Discussion of Research Issues. R & D Report 1988: 14, Statistics Sweden.

Lessler, J.T., Tourangeau, R., and Salter, W. (1989). Questionnaire Design in the

Cognitive Research Laboratory. National Center for Health Statistics, Vital and Health Statistics, Series 6, No. 1.

Lindeberg, A. (1993). TERES-projektet, TEleguide i RESvaneundersökningen. [Project TERES: Teleguide in a Survey of Traveling Habits] Statistics Sweden. [In Swedish.]

Lyberg, L. and Dean, P. (1989). The Design of Pilot Studies: A Short Review. R & D Report, 1989:22, Statistics Sweden.

Nathan, G. and Sirken, M.G. (1986). Response Error Effects of Survey Questionnaire Design. Paper presented at the Joint Statistical Meetings, American Statistical Association, Section on Survey Methods Research.

Nelson, D.D. (1985). Informal Testing as a Means of Questionnaire Development. Journal of Official Statistics, 1, 179–188.

Oksenberg, L., Cannell, C., and Kalton G. (1991). New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349–365.

Platek, R. (1985). Some Important Issues in Questionnaire Development. Journal of Official Statisitics, 1, 119–136.

Presser, S. (1989). Pretesting: A Neglected Aspect of Survey Research. Proceedings of the Fifth Conference on Health Survey Research Methods , 35–37.

Presser, S. and Blair, J. (1993). Survey Pretesting: Do Different Methods Produce Different Results? Unpublished manuscript, Survey Research Center, University of Maryland.

Royston, P., Bercini, D., Sirken, M., and Mingay, D. (1986). Questionnaire Design Research Laboratory. Proceedings of the Section on Survey Research Methods, American Statistical Association, 703–707.

Schuman, H. and Presser, S. (1981). Questions and Answers in Attitude Surveys. New York: Academic Press.

Sirken, M.G. (1986). Error Effects of Survey Questionnaires on the Public's Assessments of Health Risks. American Journal of Public Health, 76, 367–368.

Tanur, J.M. and Fienberg, S.E. (1992). Cognitive Aspects of Surveys: Yesterday, Today, and Tomorrow. Journal of Official Statistics, 8, 5–17.

Thorslund, M. and Wärneryd, B. (1985). Testing/Assessing Question Quality - Some Swedish Experiences. Journal of Official Statistics, 1, 159–178.

Wärneryd, B. (1986). Att fråga [To Ask] Stockholm: SCB, Liber. [In Swedish.]

Received November 1993
Revised August 1995