Journal of Official Statistics
Vol. 3, No. 2, 1987, pp. 191–207

Evaluating Survey Questions

Wendy Sykes and Jean Morton-Williams[1]

Abstract: This paper discusses response effects in surveys arising from features of the questionnaire such as the formulation of questions, and their layout.

We describe two techniques, interaction analysis and follow-up interviews, for identifying potential sources of error in questionnaires. Their applications to two attitude surveys are illustrated. We discuss the validity of the techniques and some ways of adapting them to the requirements and resources available to different surveys.

Key words: Survey questions; question development/pretesting; interaction analysis; follow-up interviews.

1. Introduction

One of the most important but least well understood sources of error in surveys is the measurement instrument itself, the questionnaire. This problem has long been recognized. In 1951 Stanley Payne wrote, in *The Art of Asking Questions,*

> "at the present stage of development of the survey method, improvements in question wording ... can contribute far more to accuracy than further improvements in sampling methods can. I don't mean that the sampling experts should stop seeking further improvements, trying to knock a few tenths of a percent off the statistical error. But while they are labouring with tenths of a percent, the rest of us are letting tens of percents slip through our fingers." (Payne (1951, p. 5))

Reiterations of this statement continue to appear throughout the survey literature. A recent example is the admission that, "Despite the importance of individual questions to the whole survey process, there is not much scientific knowledge about questions." (Turner and Martin (1985).)

The widespread recognition of the problem has resulted in the deluge of experimental research on the wording and construction of individual or sets of survey questions.[2] These endeavours have played an important role in the identification of sources of error and in the assessment of their magnitude. Researchers appear confident that the results will ultimately augment traditional textbook instructions on questionnaire design, with their "primary emphasis on common sense and practical experience" (Turner and Martin (1985)), and

[1] SCPR Survey Methods Centre, 35 Northampton Square, London EC1V OAX.

[2] For a review of some of the literature see Kalton and Schuman (1982).

provide a more comprehensive and systematic set of guidelines for questionnaire design.[3]

However, few would maintain that this "best of outcomes" is likely to obviate the need for judgement in writing questions. For this reason, the construction of a questionnaire that causes the fewest problems is likely to remain a multi-stage process. The accumulated wisdom about the causes of bias in questions or the response variability associated with the questions coupled with the researcher's experience may make the construction of a good first draft less haphazard. Nevertheless the testing and refining of questions will remain indispensible in all but exceptional instances.

It is puzzling that pretesting is often handled so casually, given its importance. As Cannell et al. (1985) have pointed out, in many organizations the procedure rests entirely on interviewers' reports on whether or not a question "worked." These reports are usually based on a few interviews carried out by each interviewer and often the interviewers disagree on both the specific questions that caused problems and the nature of the difficulty. "Thus, the critical task of creating a scientific measuring instrument is left to the subjective evaluation of the researcher with little objective information from the pretest experience" (Cannell et al. (1985, p. 1)).

This paper considers methods for improving the identification of potential sources of error in questionnaires. Particularly, we will pay attention to two techniques, interaction coding and follow-up interviews.

2. A Model of the Interview

The familiar concept of survey research interview as a "micro-social system" (Bradburn (1983)) provides a useful framework for considering not only the main problems with survey questions, but also the approaches that can be taken to identify these problems.

The main elements of the system are the interviewer's role, the respondent's role and the task that links them. The investigator determines the characteristics of the task which are then reflected in the questions that are asked, their form, wording, layout, and order. It is the respondent's role to provide full and accurate answers to those questions and the interviewer's to administer the questionnaire in accordance with instructions received through training.

Ideal performance of the respondent role depends on how adequately the respondent processes information. Cannell et al. (1981) suggest a model of the response process in five steps:

- Comprehension of the question;
- Cognitive processing to arrive at an answer, including
  i. assessments and decisions concerning the information needed for an accurate answer,
  ii. retrieval of cognitions (attitudes, beliefs, experiences, facts),
  iii. organization of the retrieved cognition and formulation of the response on this basis;
- Evaluation of the response in terms of its accuracy;
- Evaluation of the response in terms of other goals (e.g., self-image, desire to please the interviewer);
- Giving the response judged as accurate and based on adequate processing.

Survey questions should facilitate this process, setting reasonable cognitive tasks and maximizing respondent motivation. The interviewer's role learned through training or

[3] For example, Schuman and Presser (1981), whose work involved the development of a typology of question forms that would classify all attitude items in terms of a relatively small number of important formal variations. Their aim was to conduct experiments on a sample of questions selected from within each type such that "findings based on experiments with these samples should have some generality" (Schuman and Presser (1981, p. 7)).

explained on the questionnaire is to read questions as written and in the prescribed order, and to deal with responses that are not relevant to the question or are not in the required form.

The content and wording of survey questions should respect the limits of the interviewer's training and avoid generating situations that are difficult for the interviewer to deal with. Where heavy reliance is placed on the interviewer's judgement, response effects due to differences in interviewer behaviour are potentially large (Collins (1980)).

The range of problems that survey questions might present is wide. There are problems with vocabulary level, difficulty or clarity of the concepts employed, the questioning technique used (e.g., open or closed, question batteries etc.), the layout of the question, and other familiar issues that may all affect the comprehension and administration of a question. Similarly, questions may be too difficult given a respondent's recall, ability to organize information (e.g., choosing one from a large number of response categories), or knowledge of obscure ideas or facts. Finally, question design can influence respondents' intrinsic motivation to answer fully and accurately. For example, the wording of questions may affect the respondent's perceptions of how his/her answer will be judged or evaluated and, consequently, may influence how he/she responds. Pretests should evaluate what the task demands of both the interviewer and the respondent. "Can the information be reported? What level of effort is required to retrieve and organize the requested information? Can respondents readily organize their feelings or experiences into a form specified by the researcher? Can respondents summarize complex feelings or perceptions into one of the answer categories provided?" (Cannell et al. (1985).)

How can we discover whether a survey question, or set of questions, is functioning as intended? One approach is to focus on the *responses* that are elicited, evaluating them in terms of one or a number of validity or reliability criteria. For example, there may be information external to the survey which provides the opportunity for checking the respondents' answers. Alternatively, expectations based on past evidence, theories about cultural norms or theories about relationships *between* variables can be used to assess responses. Estimates of interviewer variability and repeated measures (as in a panel survey) can also be used to check the reliability of a question.

Practical and theoretical obstacles militate against using responses as the sole indicator of whether or not survey questions are problematic. External validation of a response is rarely possible and is suitable only for factual questions. When responses fail to match expectations based on past experiences or on theories about relationships between variables, there is always the possibility that it is the expectations which are erroneous. Reliability indicators tell us only whether a measure is repeatable and nothing about what is being measured. Furthermore, the main concern in pretesting is to identify the *nature* of the problems contained in questions, and the methods mentioned have only limited capacity to assist in this.

Instead of focussing on the survey's final responses one can look at the actual *process* of asking and answering the questions. Two approaches based, respectively, on observation and on accounts, offer access to this process. Observation techniques, which do not interfere with the interviewing process, rely on interpretations of what is seen and/or heard in the interaction between interviewer and respondent. The observer is called upon to make his or her *own* judgements about the problems that questions present, based upon what interviewers or respondents do or say in asking and answering those questions. These techniques range from the relatively informal (for instance, the researcher listening to tape-recordings of, or sitting in on, a small number of pilot interviews), to the more structured and systematic. Interaction analysis which is described in detail later, is an example of a more formalized observation technique.

The main limitations of these observation methods are that they depend on outward manifestations of mental processes (which are not always in evidence) and on the matching of behaviours with their underlying cognitive processes. These problems are added to the difficulties of making reliable decisions on the occurrence of a particular behaviour.

One way of detecting problem questions and still avoid the above shortcoming is to look at *accounts* of the interview. The accounts are given by the respondent and/or the interviewer (either during or after the interview), and consist of detailed reports of the experience of asking or answering particular questions or sets of questions. For example, respondents may be asked about their interpretation of particular words or phrases, their thought processes while answering, and their perception of the question's purpose. The information derived from these methods can provide valuable insights into the *reasons why* certain questions are problematic. Interview accounts are also a useful check on the validity of the measurements, assuming that validity "refers to whether measurements produce results consistent with conceptual intent" (Turner and Martin (1985)).

Examples of these methods range from the traditional debriefing of pilot interviewers to the random probe technique described by Schuman (1966), protocol analysis (Loftus (1984)) and the intensive follow-up interview technique, developed by Belson (see Belson (1981)). This paper is concerned with follow-up interview techniques since they provide the most systematic and detailed exploration of question performance.

Clearly each approach to pretesting survey questions (observation, accounts or assessments of the answers) has certain strengths and weaknesses. Ideally there should be a combination of the approaches using methods appropriate to the scale, resources and needs of the survey under consideration. The next section describes, in detail, the techniques of interaction analysis and follow-up interviews. Illustrations of their use are given, focussing in particular on their complementary roles. This is followed by a section assessing the validity of the methods themselves and the paper concludes with some ideas for adapting the methods for surveys with different requirements and resources.

3. Interaction Analysis

Interaction analysis is a method of analyzing behaviour in the survey interview adapted from category analysis introduced by Bales (1950) in the field of social psychology. Its application to survey research was pioneered by Cannell et al. (1973, 1975, 1981) and developed by other researchers (e.g., Morton-Williams (1979), Brenner (1980, 1982), Dijkstra and Van der Zouwen (1982)). Interaction analysis entails the systematic coding of both interviewer and respondent behaviour. It also provides information about the various aspects of the interview process, for example: the extent to which interviewer behaviour conforms to training instructions; the effect of different interviewing styles on respondent behaviour; and the problems that different kinds of questions pose for interviewers and respondents.

It is the last application which is of central interest here. Certain categories and combinations of interviewer and respondent behaviour indicate the difficulties which one or both participants have in carrying out their tasks. The patterns of the occurrence of these behaviours within the interview are then used to help identify individual questions or groups of questions which do not appear to function as intended, and to suggest reasons for their failure.

3.1. Indicators of possible question design problems

The procedure discussed here involves the development or adoption of a code-frame. The code-frame's degree of detail depends on: the aims of the research, the financial and temporal constraints, and the type of behaviour that is typical for a survey interview. The identification and evaluation of relevant interviewer and respondent behaviour is founded upon what might be termed the current paradigm of the ideal survey interview and upon an understanding of the ways in which the question/answer process can break down.

According to the model, the interviewer communicates with the respondent through the medium of the question that is read as worded on the questionnaire. Provided the respondent understands the question and is both able and willing to answer it (the main message of Cannell's model of the response process), a codable, relevant answer is obtained. This answer is then recorded by the interviewer before moving on to the next question. Breakdowns in this ideal process which are attributable to question design can occur at any stage in the asking and answering of items.

Different kinds of respondent behaviour that might indicate the kinds of difficulties described include:

- Requests for clarification about the meaning of a question (e.g., direct questions about the meaning of key words or concepts, questions about the form of response that is wanted, requests for repetition of all or part of the question etc.);
- Responses that are inadequate for coding or recording (e.g., responses which clearly indicate misunderstanding of the question meaning or response task);
- Expressed problems with a question (e.g., comments on the inappropriateness of the response categories offered, comments that the question is difficult, etc.);
- Elaborations, qualifications or explanations of given responses;
- Digression to topics unconnected to the question;
- Refusals to answer.

Interviewer behaviour that might contribute to a breakdown in the interview would include misreading or other maladministration of questions. However, a particularly high incidence of other behaviour that is contrary to training instructions (e.g., interpretations of the question, inappropriate or inadequate probes, etc.) might also point to the need to re-assess the design of a particular question.

The number of interviewer and respondent behaviours associated with the asking and answering of a question may also be a useful indicator of problem items. A question that has three, at most, expected interactions (e.g., question read, response given, response acknowledged by interviewer) but obtains significantly more interactions may need further examination.

4. Follow-up Interviews

The follow-up interview method developed by Belson (1981) is a tool for investigating how respondents answer the questions asked in surveys. Whereas analysis of interviewer and respondent behaviour provides information about the overt working of the questions – whether they are delivered in the intended way and whether they are answered correctly and without difficulty – follow-up interviews provide information about the covert 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. Their main value, however, is in diagnosing the nature of any problem that has been identified through interaction analysis.

The technique works in the following way. Shortly after a survey interview has been conducted with a respondent, a second, intensive, semi-structured interview is carried out by a different interviewer. Respondents are taken carefully through selected questions asked in the first interview and are encouraged to recall how they interpreted and understood the questions, both overall and in terms of particular words and phrases. They may also be asked how they arrived at their answers, how accurately they felt the answer given actually reflected their views and how important they had felt it to be to arrive at an accurate answer. In this way we obtain information on the frames of reference and ranges of interpretation used by respondents.

The following section provides some illustrations of the use of the two techniques, interaction analysis and follow-up interviewing.

5. Illustrations of the Use of Interaction Analysis and Follow-up Interviews

The examples in this section are drawn from two surveys. The first, an experimental survey of Attitudes to Issues of Current Importance, was carried out by SCPR's Survey Methods Centre in 1981. The second, the 1984 British Social Attitudes Survey, was one of the annual series of British Social Attitudes Surveys instigated by SCPR in 1983. The experimental survey consisted of questions on a range of issues topical in 1981: crime, the cost of living, noise, and air pollution and so on. The 1984 British Social Attitudes Survey likewise covered a variety of topics but in more depth and detail. Examples include unemployment and inflation, defence, sexual mores, health, and education.

It should be stressed that the questionnaires used on both surveys had received an unusual degree of attention during the development phases. The experimental survey consisted of questions which not only had been used on previous surveys, but which also had acquired something of the status of "standard" questions in their respective fields. For instance, the first example given in the following section is an item which has been regularly employed in the United Kingdom in studies of people's perceptions of their environment. It was taken from a survey of road traffic and the environment. Many of the other included items had been used repeatedly and apparently without the detection of any major problems.

Similarly, the 1984 British Social Attitudes Survey questionnaire was compiled by a team of researchers with input from academics specializing in the different topics covered by the survey. Again, a number of the items included had been used on other studies. The questionnaire was tested in a pilot survey and amendments to certain questions were made following the pilot interviewer debriefing.

5.1. 1981 Experimental Survey

The survey of Attitudes to Issues of Current Importance took place in Birmingham and Bristol and was conducted among "heads of households" and "housewives" from a random sample of addresses selected from the electoral register. A special feature of this survey was that the questionnaire carried a large number of split-ballot, question wording experiments. These included experiments with open and closed questions, different length rating scales, the use of visual aids (show cards) and so on. In addition, interviewers were allocated to areas in groups of four, each group having interpenetrating assignments, so that interviewer variability could be estimated.

5.1.1. Interaction analysis

For the purposes of interaction analysis, twelve of the interviewers working on the survey were each asked to tape-record eight interviews. Eighty-nine tapes of adequate quality for transcription were obtained. Application of the coding frame involved: coding the interviewers' behaviour in asking each question, the respondents' in answering it, and any subsequent behaviour exhibited by either participant. Seventy-seven questions were coded in this way.

Identifying potentially difficult questions in the field requires that one has criteria to judge the questions by. From the interaction analysis coding frame, behaviours were chosen that were considered indicative of problematic questions.

These behaviours are listed in Table 1. The next step was to decide how frequently each type of behaviour must occur for a question to be judged as causing problems. One approach would have been to set an arbitrary standard of, say, 5 % for all the behaviours. This approach has the merit of being comparable for all behaviours, but takes no account of the frequency of the different behaviours. An alternative (and preferred) approach was to set a level for each behaviour in relation to the mean incidence of that behaviour across all the questions coded either by adding to the mean some constant multiple of it (e.g., mean x 0.5) or by adding some constant number. Although the former – the addition of a

Table 1. Behavioural Indicators of Possible Problem Questions – Survey of Attitudes to Issues of Current Importance

	Mean incidence of behaviour (over 77 questions) (%)	Criterion of question failure (%)	No. of questions failing (out of 77)*
Interviewer behaviour			
(Based on all question reading behaviour)			
Failed to read the question as worded	3	8	12
Respondent behaviour			
a) (Based on total first response to the question asking)			
Requests clarification of question or response task	5	10	9
Answer inadequate for coding (misunderstanding of question or response task)	12	17	19
Interrupted the question reading	4	9	9
b) (Based on all units of respondent behaviour)			
Elaborating answer	5	10	4
Digressing from the response task	6	11	10

On precoded questions, more than two interactions between interviewer and respondent was taken as indicating a possible problem (i.e. more than three items of interviewer behaviour or more than two items of respondent behaviour).

* Sixteen questions failed on two or more criteria, 42 on one or more.

multiple of the mean – would impose standards that were of equal stringency in relation to the mean, the latter method was selected as it allowed a less stringent criterion to be set for the item of behaviour that occurred most frequently (the respondent giving an answer that was inadequate for coding). It can be argued that this is the most clear cut and important indicator of a problem question for respondents. It directly reflects their inability or unwillingness to carry out the response task, uninfluenced by other factors (e.g., personal diffidence) which might affect their asking for clarification. The standard set for each behaviour was the behaviour's mean incidence plus 5 %.

The first column in Table 1 shows the mean incidence of each behaviour for all questions. The second column shows the criteria of question failure, i.e., the frequency of behaviour indicating a problematic question.

Some general discussion of these findings and comparable results from the 1984 Social Attitudes Survey is presented in Section 6. The discussion in this section is confined to a small number of illustrative examples of the identification of problem questions.

5.1.2. Follow-up interviews

Sixteen of the interviewers working on the experimental survey were each asked to arrange for six follow-up interviews. The follow-up interviews were carried out by specially trained interviewers and took place the day after the first interview. These interviewers were equipped with both the questionnaire from the first interview and an outline of the main topics covered in the follow-up interview. The interview started with enquiries about the respondent's reactions to the first interview and proceeded to a more detailed investigation. The respondent was questioned in great detail about his/her understanding of seven questions in the first interview and was questioned in lesser detail about six other questions.

Not all of the interviewers managed to set up all of the required follow-up interviews. Those that were obtained were tape-recorded and there was some further loss due to poor recording quality. In total, 60 interviews were recorded, transcribed and used in the analysis.

5.1.3. Examples of problem questions

One example of a problem question illustrates the use of the techniques in uncovering respondent difficulties in comprehending the *response task.*

The item discussed below is frequently encountered in both social and commercial surveys. This particular question is standardly

How do you feel about the amount of noise, in your area, from cars and lorries or other road traffic?

Show and explain card (at the top of the ladder is someone who feels that the amount of noise from cars and lorries is definitely satisfactory, whereas at the bottom is someone who feels it is definitely unsatisfactory)

		%
Definitely satisfactory	1	23
	2	12
	3	12
Neither satis. nor unsatis.	4	25
	5	11
	6	3
Definitely unsatisfactory	7	13
(Don't know)	8	1
Total		100 (544)

employed in studies designed to evaluate people's perceptions of their environment and had been taken from a survey of road traffic and the environment:

This kind of question is favoured because it enables long scales to be used but with fewer problems of ambiguity or mixed dimensions than are entailed in lengthy verbal rating scales. The respondents are expected to answer with a number selected from the scale.

Many respondents failed to understand the task posed by the question. Ten percent of the first utterances of respondents following the question reading were requests for clarification about the response task or the question's meaning. A further 17 % of first reactions were inadequate for coding, which suggested a misunderstanding of the question or response task. The distribution of answers to the question indicated that this item was not functioning as anticipated. As the distribution above shows, there were marked clustering of responses at each point on the scale where a verbal description was given.

Analysis of the follow-up interviews with respondents revealed that half of those who were re-interviewed said that they had not really understood the purpose of the numbers and had simply used the points on the scale that had words beside them.

The concept of the scale was clearly difficult for some respondents to grasp and no doubt this was communicated to interviewers fairly early in the survey. It is not unreasonable to suppose that, finding the suggested explanation of the scale inadequate, interviewers quickly developed their own individual method of delivering the question. The transcripts of the interviews revealed a number of idiosyncratic explanations of the scale, each particular to a single interviewer. One example was:

> "On this card is a scale between one and seven. If you were living in the country and you weren't bothered by noise you would probably choose number one. But if you were on a busy main road you would probably choose number seven. If I asked you to tell me where you fitted on that scale, where would you place yourself? What number?"

The effect on the data of individual interviewers each adopting their own styles of administration was made clear in the analysis of interviewer variance. This was one of the few questions on the survey which showed statistically significant variation among the interviewers.

A second example of a question with possible weaknesses is one which was used to elicit general opinions about trends in the relative wealth of the countries of the world:

> "Would you say that the gap in wealth between the richer countries and the poorer countries is, on the whole, getting wider, getting narrower, or remaining about the same? Even if you aren't sure I'd like you to tell me what you think."[4]

This question received a high proportion of requests from respondents for clarification of the meaning of the question. Twenty-six percent of respondents' first reactions to this question were a request for clarification and a further 22 percent of first responses were inadequate for coding. Suspicions that many other respondents may not have fully understood the question but made their own interpretation of it without indicating that they had a problem were confirmed by the follow-up interviews.

The following summarizes the problems that the respondents had with the *meaning* of the question.

> The use of the comparative adjectives, "richer" and "poorer" left it to the

[4] The guide used by interviewers in following up on this question is given in Appendix 1 of this paper.

> respondent to decide which countries to put into each category, a task which many of them found difficult; they felt the need to identify at least some as "richer" and "poorer" in order to give them a frame of reference in which to answer the question.
>
> The term "wealth" in this context was also interpreted in different ways as the assets or resources of the country or the standard of living of the people.
>
> Understanding of the ways that the gap in wealth between countries could get wider or narrower included references to the amount of aid to poor countries and whether they used it wisely. Others thought more in terms of economic changes; some countries were getting poorer while others were becoming richer through, for example, the discovery of mineral wealth.

The concepts used in this question were seen as complex, imprecise and confusing. Many respondents clearly were not able to handle them at the level of abstraction perhaps hoped for by the researcher. A large number of those in the follow-up interviews said that they did not feel that they were competent to answer yet very few said "Don't know," partly because this option was not presented to them, and partly because in the final part of the question they were told "Even if you aren't sure I'd like you to tell me what you think." This was correctly interpreted as meaning that they should indicate their general feelings even if they were unsure.

A final example is a question that was asked in two forms on the survey; the first inviting respondents to choose between just two categories and the second offering a third, mid-point category.

> Do you think that giving priority to buses at traffic signals would increase or decrease congestion (or would it make no difference)?

A large percentage of first responses to this question consisted of requests for clarification or repetition of the question. (22 % on the version which did not offer a mid-point and 17 % on the version which did.) As with the previous question almost twice as many respondents gave the mid-point answer when it was offered than when it was not explicitly presented.

We also obtained some insight on this question from the follow-up interviews. We thought that the word "priority' might be unfamiliar to some and would cause problems. This proved not to be the case. However, respondents did have considerable difficulty with the total concept of "giving buses priority at traffic signals." It was not clear to them how this might be done or how it might affect other traffic. The extensive use of the "make no difference" option when it was offered would suggest that it was a way of avoiding grappling with a question that was largely meaningless.

5.2. *1984 British Social Attitudes Survey*

For the 1984 British Social Attitudes Survey, 68 tape-recordings of interviews were obtained and coded but no follow-up interviews were carried out. In addition to the behaviour codes targeted in the previous study, attention was paid to whether or not interviewers engaged in unprescribed behaviour inbetween reading the question and hearing the respondent's reply. We were also interested in "deviant sequences" of behaviour – sequences other than those which might be expected in the straightforward delivery and answering of a question.

Examples of unprescribed behaviour might include unprompted repeats of questions, offers of clarification and so on. It was felt that such behaviour might reflect real problems that interviewers had identified either because previous attempts to administer a question had consistently run into difficulties or because of non-verbal cues – puzzled expressions, hesitations, etc. – received from respondents.

With regard to sequences of behaviour, the very detailed coding frame used on this study allowed for the identification of a number of entire sequences of interviewer and respondent behaviour likely to be representative of the straightforward, unproblematic asking and answering of questions. For example,

Question asked as worded
|
Answer given adequate for coding
|
Response acknowledged by interviewer

Or more elaborately,

Question asked as worded
|
Answer given adequate for coding
|
Interviewer confirms response given
|
Respondent repeats answer
|
Response acknowledged by interviewer

It was then possible to identify questions that had an unusually low percentage of straightforward sequences.

Table 2 on the following page shows the indicators of potential problems which were used on this study. It also gives the average incidence of each indicator on this survey.

A question identified as likely to cause difficulties asked about Britain's relationships with other countries. It formed the second part of an item that began with the following question.

> Do you think Britain should continue to be a member of the EEC – The Common Market – or should it withdraw?

The question of concern here was:

> And do you think Britain should continue to be a member of NATO – the North Atlantic Treaty Organisation – or should it withdraw?

Nine percent of the readings of this question were interrupted. Perhaps because respondents with well formed views recognized the format of the question from the previous item and volunteered their answers before the question had been fully delivered. Other than this, the question reading seemed to present few difficulties.

However, only two thirds (66 %) of first responses were answers adequate for coding. Half of the uncodable answers indicated that the respondent had misunderstood either the question or the response task. The remainder were answers which, although codable in the strict sense that a response category was selected, were accompanied by remarks suggesting they had been randomly given. For example:

> "Yeah. Well I don't know much about that one ... say continue."

One explanation of this might be that the question dealt with issues about which many people have little information and few opinions. This would be consistent with the high percentage of "Don't knows" received from the total sample on the survey – 9 % compared with an average for all questions of 3 %.

The second example drawn from this study was an item the wording of which reflects the great care which had been taken to define the scope of the concepts involved:

> There is a law in Britain against racial discrimination, that is against giving unfair preference to a particular race in housing, jobs and so on. Do you generally support or oppose the idea of a law for this purpose?

Once again interviewers had few problems in delivering the question but 12 % of first

Table 2. Behavioural Indicators of Possible Problem Questions – Survey of British Social Attitudes 1984

	Mean incidence of behaviour (over 151 questions) (%)	Criterion of question failure (%)	No. of questions failing (out of 151)*
Interviewer behaviour			
Failing to read the question as worded (including small changes which do not materially alter the question)	4	9	13
Intervening interactions (as percentage of question readings)	3	8	15
Respondent behaviour			
(All percentages based on total first response for the question asking)			
Requests clarification of question or response task	2	7	6
Requests repeat of the question	2	7	9
Answer inadequate for coding (misunderstanding of question or response task)	4	11	20
Other answer inadequate for coding	2		
Answer suggests response invalid	2	9	10
Can't answer	2		
Interrupted the question reading	1	6	2
"Deviant" sequences	32	37	63

* Fourteen questions failed on two or more criteria, 50 on one or more.

responses were inadequate for coding because of apparent misunderstanding of the question, 12 % were requests for a repeat of the question; 4 % were requests for clarification and 4 % were possibly invalid answers. An above average number of interactions between interviewers and respondents and that well over half of the question-answer sequences associated with the item were not straightforward are further indications of the problems this question presented.

Unlike the previous example, very few respondents seemed to feel unable ultimately to grasp the import of the question and to give a response. However, it is not unreasonable to speculate that a question which poses so many initial comprehension problems will be interpreted in very different ways. Some support for this is provided by the results of a panel study, now in its third year, which is being conducted with part of the sample from the first, 1983, British Social Attitudes Survey (Lievesley and Waterton (1985)). One would expect to find fairly stable views over only three years on the issue of a law against racial discrimination, however, the correlations between responses given by individuals in year one and those given in subsequent years are low, around 0.37. This result is believed to reflect the inadequacy of the question as a measure of support for anti-discrimination laws. Indeed, when asked in 1984 if they felt their views on this issue had changed since 1983, almost all respondents said "no" although substantial numbers gave different answers on the two occasions. The correlation between perceived change and actual change was effectively zero.

6. The Validity and Utility of Interaction Analysis and Follow-up Interviews as Tools for Pretesting

Interaction analysis and follow-up interviews were employed on two surveys to help identify weaknesses in questions which had not been pinpointed during the standard question development and pilot phases of the surveys. To a certain extent the two techniques command a "face validity" which arguably requires no further testing. Questions consistently maladministered by interviewers or eliciting frequent requests for clarification, words repeatedly interpreted in ways not intended by the researcher and so on, all directly suggest the need for some re-design of that item.

Furthermore, when used together the methods frequently provide insight into the *nature* of the problem to be tackled and indicate which changes should be made. For instance, the first example from the 1981 experimental survey described in Section 5.1 revealed plainly that many respondents did not understand what they were required to do. The need for a clearer set of instructions to respondents and/or some alternative scaling method was apparent. In this respect, the value of employing both methods simultaneously was pointed out by the first example in Section 5.2 concerning Britain's membership in NATO. Through empathy and imagination we saw that the problems some respondents had in answering this question were linked to a lack of familiarity with the main issues. Follow-up interviews would have helped to confirm or reject this view.

For the examples discussed above, other evidence was also available to support the indications provided by the two methods. Thus, interviewer variability measures, estimates of reliability from a panel study, the actual distributions of answers obtained (including the proportion of "Don't know") and so on, were all used to give credence to initial judgements based on interaction coding and follow-up interviews alone. However, certain items which the two methods identified as problematic showed no correlations with other indicators. Conversely, certain items which, for example, were found to be associated with high interviewer variability, were not pinpointed by interaction analysis and/or the follow-up interviews. In the latter case the causes of interviewer variability may be undetectable using either interaction analysis or follow-up interviews. For example, the problem may lie with respondents' vulnerability, at certain questions, to interviewer characteristics or other processes which are neither manifested verbally nor experienced consciously. It is not surprising that these techniques are unable to pick up all the potential sources of error and neither should it be regarded as discouraging.

Interaction coding and follow up interviews should not be judged by the fact that other indicators (e.g., interviewer variability) sometimes fail to confirm the problems identified by the methods. Certainly, not all questions which, for example, have a high incidence of requests for clarification or question repeats will result in unreliable answers, but the *class* of such questions is likely to be vulnerable in this respect (e.g., Collins (1980)).

We may even learn something which simply improves the flow of an interview: for example, a battery question on the 1981 experimental survey obtained a very high percentage of interrupted question readings which did not appear to affect the quality of answers as measured by interviewer variability (although this could mean interviewers were getting uniformly "bad" answers). The problem was related to the interviewers' being briefed to read the response options for each of the six items on the battery. Since these were not difficult to remember, respondents began to interrupt after the first few items. Prescribed (and thereby controlled) abbreviations after the first interruptions would have reduced the likelihood of respondents becoming irritated, bored or otherwise disaffected by the repetitious questioning.

To reiterate, interaction analysis and follow-up interviews are particularly valuable if the problem is manifested verbally or is clearly perceived by the respondents. To some extent, techniques such as random probes or protocol analysis can achieve much that is accomplished in follow-up interviews and the information provided by these techniques is more immediate. Both methods are likely to be attractive alternatives where resources are constrained, where interviewers trained in unstructured interviewing are not available, or where large numbers of questions are to be covered. (Follow-up interviews tend to be lengthy and do not allow for a thorough treatment of many items.)

The application of the interaction analysis and follow-up interviews as described in this paper demanded considerable time and money. All tape-recordings were transcribed which is costly, and the coding of transcripts for the interaction analysis was a lengthy process.

In our two studies the input was justified by the wider aims of the research which entailed a thorough exploration, using interaction analysis, of the entire survey interview process. Transcripts were necessary to develop the code-frame and check its reliability. The "marked-up" transcripts enabled us to examine both the identified *boundaries* of behaviour units and the actual *codes* assigned to those units. Both kinds of data were used in comparing the application of the code-frame to the same interviews by different coders.

Finally, as was demonstrated, the transcripts were an easily accessible, valuable source of information. We referred to the transcripts when a problem question had been identified so that the exact nature of the problem could be assessed.

Question evaluation is important both before and after the questionnaire reaches its final form. Once the survey is completed, the results can be used to mediate the survey findings. Arguments for the use of interaction analysis and follow-up interviews for these purposes are self-evident in relation to large-scale, costly surveys including repeat surveys, and those which are crucial to major policy decisions. For example, the techniques described here were used on the 1984 British Election Study and provided valuable information about respondent understanding of key questions and the functioning of important scale items (Heath et al. (1985)). These data were relevant in interpreting the results of the survey as well as in the design of future election studies. The code-frame used on the British Election Study is given in Appendix 2 of this paper.

However, the use of such techniques need not be limited to such "costly" surveys. Simplified methods may be useful even on *ad hoc,* low budget studies. In these instances the methods may be adapted in ways which still lend a degree of rigour and systematization to routine procedures.

For example, tape-recordings of pilot interviews may be quickly coded by either a researcher or the interviewers themselves, using a simplified code frame and working directly from the tape-recordings. Even a few interviews can yield substantial amounts of useful information.

Alternatively, interviewers conducting pilot interviews can be trained to identify and record the occurrence of particular types of problems. For instance, when they have to repeat a question; when they receive requests for clarification; indications of misunderstanding or difficulty; refusals or reluctance to answer, and so on. They can also be asked to write down their general comments on the flow of the questionnaire, respondent interest, etc. Similarly, with telephone interviews conducted from a central location a supervisor "listening-in" to interviews can carry out a simple coding exercise on a number of interviews during just one interviewing shift.

7. References

Bales, R.F. (1950): Interaction Process Analysis. Addison-Wesley, Cambridge, Mass.

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

Bradburn, N.M. (1983): Response Effects. In P.H. Rossi and J.D. Wright (Eds.), The Handbook of Survey Research. Academic Press, New York.

Brenner, M. (1980): Assessing Social Skills in the Survey Interview. In W.T. Singleton, P. Spurgeon and R. Stammers (Eds.), The Analysis of Social Skills. Plenum, New York.

Brenner, M. (1982): Response Effects of 'Role Restricted' Characteristics of the Interviewer. In W. Dijkstra and J. Van der Zouwen (Eds.), Response Behaviour in the Survey Interview. Academic Press, London.

Cannell, C.F. (1973): Some Experiments in Dyadic Interactions and Response Accuracy in Survey Interviews. ISR, University of Michigan.

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975): A Technique for Evaluating Interviewer Performance. ISR, University of Michigan.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981): Research on Interviewing Techniques. ISR, University of Michigan.

Cannell, C.F., Kalton, G., and Fowler, F.W. (1985): Techniques for Diagnosing Cognitive and Affective Problems in Survey Questions (Unpublished).

Collins, M. (1980): Interviewer Variability: A Review of the Problem. SCPR Methodological Working Paper, No. 19.

Dijkstra, N. and Van der Zouwen, J. (eds) (1982): Response Behaviour in the Survey Interview. Academic Press, London.

Heath, A., Jowell, R., and Curtis J. (1985): How Britain Votes. Pergamon, London.

Kalton, G. and Schuman, H. (1982): The Effect of the Question on Survey Responses: A Review. Journal of the Royal Statistical Society, Series A, Vol. 145, No. 1.

Lievesley, D. and Waterton, J. (1985): Measuring Individual Attitude Change. In R. Jowell and S. Witherspoon (eds.), British Social Attitudes: the 1985 Report. Gower, London.

Loftus, E. (1984): Protocol Analysis of Responses to Survey Recall Questions. In T.B. Jabine et al. (Eds.), Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. National Academy Press, Washington D.C.

Morton Williams, J. (1979): The Use of Verbal Interaction Coding for Evaluating a Questionnaire. Quality and Quantity, pp. 59–75.

Payne, S. (1951): The Art of Asking Questions. Princeton University Press, Princeton, N.J.

Schuman, H. (1966): The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review, Vol. 31, No. 2, pp. 218–222.

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

Turner, C.E. and Martin, E. (Eds.) (1985): Surveying Subjective Phenomena. (2 vols.), Russell Sage, New York.

Received June 1986
Revised March 1987

Appendix 1

Questions and Probes Used in Following-up the Question on the Gap in Wealth Between the Richer and Poorer Countries

Would you say that the gap in wealth between the richer countries and the poorer countries is, on the whole, getting wider, getting narrower (or remaining about the same)? Even if you are not sure I'd like you to tell me what you think.

Here is another question that you were asked yesterday ... (READ Q.3a SLOWLY) ... What did that question mean to you *yesterday*?	PROBE FULLY: E.g. What did you think the interviewer was trying to ask? How do you mean ...? ALWAYS ASK: Was there *anything else* s/he was trying to ask? – Anything more?
What did the word "wealth" mean to you in this context? Which particular countries did you have in mind as .."richer countries"? .."poorer countries"?	PROBE AS NECESSARY. CHECK THAT REFERRING TO YESTERDAY AND PROBE TO ESTABLISH ANY DIFFERENCE OF INTERPRETATION NOW EMERGING.
Do you remember what answer you gave yesterday?	IF DOESN'T REMEMBER OR REMEMBERS WRONGLY, REMIND FROM QUESTIONNAIRE AND PROBE REASONS.
How did you arrive at your answer yesterday – what went on in your mind?	PROBE FULLY ALWAYS ASK: What did the interviewer say to help you arrive at your answer? PROBE ANY CHANGES IN INTERPRETATION.
The interviewer said "Even if you are not quite sure I'd like you to tell me what you think". How did this affect you in arriving at your answer?	PROBE AS NECESSARY Why do you think that phrase was put in? What was its purpose?
Did you find that you were able to answer the question quickly? – Or did you have to think about it a bit? ...	PROBE TO ESTABLISH WHETHER S/HE – had any *difficulty* in answering. – was concerned to answer accurately.
Did you realise *yesterday* that the interviewer was putting your answer into categories that s/he had printed on the questionnaire?	PROBE: E.g. How did/do you feel about this? – were the categories the right ones? – did you feel you had more to say?

Appendix 2

Example of an Interaction Coding Frame for Evaluating Questions

Codes *Introduction to question (if applicable)*

01 Introduction misread/altered
02 Introduction partially read
03 Introduction not read

First asking of question

11 Question misread/altered on first presentation
12 Question partially asked/all precodes not presented
13 Question not asked (in error)
14 Question asked in error
15 Question asked in error, then retracted

First response

21 Asked for repeat of question/silence requiring intervention
22 Asked for clarification/definition/ speculated about meaning
23 Not answered adequately for coding/partial answer
24 Answer indicates misunderstanding
25 "Don't know" because answer categories don't fit case
26 Refused to answer
27 Answer given/prompted by other person present
28 Answer not given strictly in wording of question, but sufficient for coding
29 Respondent initiates some other digression
30 Answered adequately for coding but respondent made an additional comment

Follow-up interviewer

31 Repeated question and precodes
32 Repeated question only
33 Repeated precodes only
34 Correct use of probes/prompts/ stonewalling on queries
35 Amplification given/question repeated in altered form
36 Prompted/incorrect use of probes/ partial use of probes
37 Failed to probe when should have done

Second response

As for first response but prefix 4 etc.

Second follow-up

As for first follow-up but prefix 5 etc.

Coding/recording

91 Wrongly coded
92 Coding omitted
93 Open ended information, not recorded in full
94 Open ended information, not recorded verbatim