LEARNING FROM INTERVIEWERS

Theresa J. DeMaio, U.S. Bureau of the Census

Interviewers are a key and often underrated element in the practice of survey research. They constitute the link between respondents and researchers, and in their direct contact with respondents they can pick up valuable information which may be of interest to questionnaire designers. Although much has been written on the subject of interviewing, the systematic exploration of an interviewer's knowledge has been neglected in the literature.

Two techniques may be employed to elicit information accumulated by interviewers during the routine performance of their job. Data can be obtained from interviewers either through discussions (referred to here as interviewer debriefings) or through written evaluations (referred to here as structured post-interview evaluations). Discussions involve verbal exchanges among a group of interviewers and a discussion leader. Written evaluations, in contrast, entail structured questionnaires asking interviewers about their perceptions of themselves, their respondents, or the questionnaire. These two techniques can also be combined--participants in group discussion sessions may be instructed to fill out questionnaires before or during the session.

In this paper, each of these techniques is discussed and examples of their use are presented.

INTERVIEWER DEBRIEFING

The term "interviewer debriefing" refers to the technique of verbal information exchange between the interviewing staff and the operations staff. Both of these terms are used loosely. The interviewing staff can be comprised of researchers if they happen to be doing the interviewing for an informal test, and the operations staff can encompass either field operations personnel or research personnel as the situation warrants.

Debriefing can be conducted at various points in the life of a survey, from the first stages of developmental testing to the final, large-scale data collection effort. At any or all of these stages, interviewer feedback concerning problems in the structure or wording of a questionnaire can be crucial to improving the survey data. During the developmental stages, results may be useful in revising question wording and response categories, identifying sensitive questions, improving the flow of the questionnaire, and estimating the respondents' ability to answer survey questions. At the end of a survey, suggestions from the interviewers might be used to evaluate the performance of the questionnaire during the survey, to contribute to the analysis of the results, or to recommend future changes in recurring surveys.

The results of the debriefing process are qualitative rather than quantitative in nature. Although it can detect problems in the questionnaire (perhaps isolated among a particular subgroup), the extent of those problems cannot be specified.

While this may be seen as a disadvantage from a statistical point of view, the compensating advantage is that problems that were not anticipated by the survey designers (and thus not included on a form intended for statistical tabulation) may also be uncovered.

Participants

The discussion leader is a critical participant in group debriefing sessions. Several qualities are desirable in a discussion leader, although it may not be possible to combine them all in one person. First, someone involved in the development of the survey will be familiar with issues which were problematic in designing the questionnaire and may note comments that might not seem important from some other perspective. While this has definite advantages, it also has some disadvantages. A discussion leader who has been intimately involved in a survey's development must not be defensive when negative comments are made, as this could discourage interviewers from making constructive contributions. Also, the discussion leader must not lead the interviewers into confirming his/her own preconceived notions about the questionnaire. Second, a discussion leader who is known to the interviewers may promote a more active exchange if this makes the interviewers feel less inhibited in expressing their opinions. This, too, has its drawbacks: field supervisors who are responsible for maintaining standards of productivity and who constantly remind interviewers to read questions exactly as worded may not be the best people, from a research point of view, to lead a discussion of ways in which questions were asked or reasons why interviewers were not able to get accurate responses to survey questions. Third, an experienced debriefing leader should be able to keep the discussion focused on relevant subjects instead of having it drift onto extraneous issues. And fourth, experience and skill are required to obtain participation from timid as well as aggressive interviewers.

In a small-scale field test the questionnaire designers may be familiar to the interviewers, making one of them a logical choice for this assignment. In debriefing interviewers in a national survey, however, it may be difficult to achieve this combination in a discussion leader, since the organization of large-scale data collection requires a separate structure from development of the survey. In this case, a research-oriented staff person as a discussion leader might be preferable to a field supervisor, as long as the researcher possesses leadership qualities which enable him or her to control the group.

When multiple debriefing sessions are held simultaneously, more than one person must obviously be available to serve as a discussion leader. If sufficient numbers of researchers

and/or field staff are not available, another alternative is possible. Representatives of the survey designers or the survey sponsor (i.e., the organization or agency that requests the survey and provides the overall objectives and funding) may be experienced and knowledgeable as discussion leaders. This is particularly important if the discussion leader is not able to approach the discussion from the viewpoint of a person responsible for meeting overall objectives of the survey or as one responsible for the development of the questionnaire. Representatives of the survey designers or sponsors may also attend the sessions as observers of the proceedings, or as participants with a limited role in the discussion. If the observers are identified as sponsors or survey designers, and the importance of their learning about the interviewers' impressions of the questionnaire is explained, their presence should not inhibit the discussion.

The degree to which the survey designers are involved in the debriefing process (as observers or discussion leaders) can determine the extent to which results are incorporated in questionnaire revision or analysis. Too often the investigation of issues relating to the quality of survey responses is conducted separately from the survey itself (perhaps by a different agency or by different personnel within the same agency), and the left and the right hand never get together, so to speak. A close working relationship between all parties involved is suggested for maximum results.

Interviewers comprise the final category of participants in a debriefing session. The number of interviewers may vary according to the type of survey. For an informal test, only five or six interviewers may be involved, while in a formal test or survey, the number might be much larger. Generally speaking, if the number of participants exceeds fifteen, separate groups should be assembled to allow for maximum participation by the interviewers. With smaller groups, more information can be obtained from each interviewer.

Depending upon the geographic area encompassed by the survey, and also depending on constraints of budget and timing, it may be possible to hold debriefing sessions in more than one place. For example, in a national survey or a field test conducted in three areas of the country, two or three debriefing sessions might be arranged in different cities. Increasing the number and location of the sessions has two advantages: 1) it includes reports of experiences with respondents in more than one geographic region, who may have had different reactions to or problems with the questionnaire, and 2) it decreases the possibility that the results (of a single session) may be idiosyncratic due to particular interviewers' skills, supervisors' instructions, or discussion leaders' ability to control the group.

Planning Considerations

In planning debriefing sessions, several elements need to be considered.

1. When to hold sessions

Successful results may be obtained during a field test debriefing either by conducting a single discussion at the end of the test or by conducting them on a continuing basis (e.g., daily). Holding sessions more frequently and implementing changes in the questionnaire throughout the testing period allows a number of versions of a question to be tested, if necessary.

Regardless of the stage of the survey at which the debriefing session is held, it should be conducted very shortly after the end of interviewing. This ensures that the experiences of the interviewers will be fresh on their minds and more accurately reported.

2. How long they should last

The length of a debriefing session depends on the amount of material to be covered. The average session might last two or three hours, but all-day debriefing sessions are not uncommon. Discussions scheduled for longer than a couple of hours should be interrupted by breaks.

3. Outline

To ensure that the discussion covers appropriate, prespecified topics and maintains a proper focus, an outline should be prepared in advance of the debriefing session. The content of the outline can include some topics which are important from the perspective of questionnaire design and some which are not (e.g., discussion of administrative or survey operations procedures). However, only those related to questionnaire design are discussed here.

For our purposes, the content of the topic outline should include those features of the questionnaire about which the designers are most anxious to receive feedback. If different versions of a questionnaire or sections of a questionnaire are being tested, the interviewers' judgment about which version worked best (and their reasons for arriving at that judgment) should be solicited. The extent to which respondents seemed to understand particular words or concepts, had the information or were willing to answer particular questions, viewed particular questions as sensitive, etc., might be included as topics for discussion.

It is generally helpful to provide interviewers with some idea of the topics to be covered during the debriefing. This can be done either by circulating an agenda containing questions for discussion prior to the session, or by handing one out at the beginning of the session. Either method will give the interviewers time to think about the issues and to recall relevant experiences; this promotes more informed discussion during the session itself. It also lets the interviewers know that particular topics will be covered so they will be less likely to interject their views at inappropriate places in the discussion.

Operational Issues

One of the positive features of group debriefing sessions is that the group atmosphere promotes interaction among the interviewers and stimulates them to react to the ideas of others, possibly increasing their own insights and thus the value of the discussion. Initially, interviewers may be hesitant to participate or to express negative comments about a questionnaire that others worked hard to prepare. Or they may not think about problems associated with the questionnaire as much as other types of procedural issues. It is the responsibility of the discussion leader to emphasize the importance of interviewers' input, both positive and negative, and set the tone of the discussion. All parts of the debriefing session will not be equally productive from the questionnaire designer's point of view; some so called "wasted" time should be expected during a session. However, allowing interviewers to vent their frustrations about some topics beyond the questionnaire designer's control will be necessary at some points. Forcing the interviewers to suppress all their complaints may provoke hostility and discourage cooperation with the goals of the session.

Interviewer debriefing sessions are generally tape recorded. Use of a tape recorder may initially inhibit discussion, but participants usually quickly forget that it is there. This practice serves several purposes: 1) it enables a more accurate transcription of discussions that might move too quickly for a note-taker to record; and 2) if the debriefing report is not prepared immediately, it prevents the results from being subject to memory decay.

There is a drawback to this practice, however. The transcription of the debriefing tape is a time-consuming process, often completed after such a long lapse of time that the usefulness of the results in questionnaire revision is diminished. This is particularly true in debriefing interviewers after a field test, when only a short time is allotted to make changes in the questionnaire before the next phase of testing or the final survey.

Even when a tape recorder is used, it is a good idea to have a designated note-taker and to rely on the tape recorder only to review particularly noteworthy parts of the discussion and sections that moved too quickly for accurate note-taking.

During the session, the discussion leader should try to summarize main points at appropriate intervals. This may serve to determine whether the majority opinion, if any, was understood. It may make the note-taker's job easier, as well as the job of anyone who subsequently reviews and summarizes the tapes.

After the completion of all scheduled debriefing sessions, a summary of the main results should be prepared. This should include implications for questionnaire revision if the interviewing is being conducted as part of an informal or formal test, and should be done as quickly as possible. Often, when a questionnaire is revised, the exact changes and the reasons for making those changes are not documented. This has two drawbacks: it prevents others from learning from the experience, and it prevents anyone from knowing whether or not the debriefing results are used.

STRUCTURED POST-INTERVIEW EVALUATION

Structured post-interview evaluations are often referred to as "ratings" and involve administering questionnaires to interviewers after their participation in the entire survey or in a particular stage of the survey has been completed. The attitudes and behavior of an interviewer can influence a respondent's answers. These evaluations contain questions about interviewers' attitudes and perceptions of their respondents, which may provide input concerning potential sources of bias. Do the interviewers feel inhibited in asking for respondents' income? Do they view respondents as cooperative during the interview? Do they think the respondents give accurate and honest answers to the survey questions? How do interviewers feel about the objectives and value of the survey? Factors such as these might influence both the quality of the data provided by the respondents (when they answer the questions) and how often responses to the questions are not obtained.

The ultimate objective of such evaluations is to obtain information about the attitudes and behaviors of the participants in the data collection process that may affect responses to survey questions. In some instances the results of these evaluations can be used to improve a questionnaire draft, in others they can be used to improve a future wave of a survey, and in still others they can be used to give the survey designers or data analysts information about the kinds of errors that may have been introduced during the data collection process. In this last use of post-interview evaluations, the results are more likely to be incorporated as revisions to the procedures for interviewer training or data collection than as revisions to the questionnaire itself.

Participants

The project director for a post-interview structured evaluation program should have enough familiarity with sources of interviewer bias to formulate hypotheses about interviewer effects in the survey (or test) being evaluated, develop a questionnaire that collects data to test those hypotheses, and evaluate the data that are collected. Additional staff may be required to work toward completion of these tasks.

Use of this technique involves a mini-survey of a sort, and requires interviewers to serve as respondents. Most often these "surveys" consist of self-administered questionnaires; if face-to-face or telephone interviews are used instead, additional personnel (i.e., other interviewers or supervisors) are needed to perform the interviewer function.

In practice, post-interview evaluations are generally treated independently of the original field test (or survey) and are often organized and conducted by different groups of people.

This can lead to two problems: 1) lack of coordination between the groups involved in developing evaluation forms and acquiring data for analysis; and 2) lack of incorporation of research results which might improve the survey. These limitations of the method can be minimized by conscious effort and communication between the two groups.

The involvement of all the interviewers who take part in the survey (or test) is generally requested. Because the number of interviewers involved is relatively small to begin with, and because the responses of all types of interviewers are important to the results, it is imperative that interviewers take the evaluation seriously and that all interviewers participate.

Planning Considerations

The content of the evaluation questionnaire depends on the researcher's hypotheses about sources of bias. Several kinds of perceptions can be solicited from interviewers: questions can be asked about the interviewers themselves, about the survey instrument, or about their respondents. When interviewers are questioned about their respondents, a decision about the unit of analysis is required. Interviewers can be asked to complete a separate evaluation for each interview in their assignments, or they can be instructed to make a judgment about their respondents as a whole. Using the first approach, there will be as many evaluations as there are respondents; the second method can be disaggregated only to the interviewer level.

The first method is more cumbersome in planning and execution, but its results are more precise. Using the second method, an interviewer might be influenced in making his or her aggregate ratings by situations that were particularly memorable (as either good or bad experiences) but not typical of the overall assignment. Also, different interviewers have different abilities to generalize, so their estimates of "some," "most," etc., of their respondents may not be comparable.

So far, the description of procedures for obtaining structured evaluations has centered on their use as a separate exercise. Two variations of the technique deserve brief mention. The first is using structured evaluations by interviewers in conjunction with similar evaluations by respondents. The purpose of this operation is to compare the perceptions of the two groups, and promote additional confidence in findings when they both agree. The second variation involves using these evaluations in conjunction with interviewer debriefing. During the debriefing session, interviewers can be instructed to complete a questionnaire containing specific questions (perhaps the same questions that are discussed in more detail during the session itself). In this way, responses to every question can be obtained for every interviewer, which may not be the case in the less structured debriefing session. Another advantage of this technique is that quantitative results are obtained, which can be tabulated to provide a more specific idea of the extent to which specific problems or behaviors are occurring.

Operational Issues

Although the data for post-interview evaluation can be collected either by means of self-administered questionnaires or personal interviews (face-to-face or telephone), it is usually done with self-administered questionnaires. This is less expensive than other methods and more practical, particularly when the evaluation design calls for interviewers to rate each respondent separately.

The evaluation data are obtained during the data collection phase of the test or survey being evaluated. If ratings of each respondent are required, an evaluation form should be completed at the end of each interview, before the interviewer approaches another respondent. If generalized respondent ratings are required, interviewers should complete a single evaluation form at the end of their interviewing assignments.

In most uses of post-interview evaluations, the collection of the evaluation data is part of a larger scheme. The next step is to link the data obtained from the interviewers with information collected in the survey or test itself. To determine whether the interviewers' perceptions had any effect on survey responses, some measure of the quality of those responses is necessary. The importance of this technique in questionnaire design is to learn whether some aspect of the questionnaire, which can be changed, affects interviewers' attitudes.

Two types of response quality indicators are available. One is, obviously, the data collected in the survey itself. The particular data items used to measure response quality can vary according to the hypotheses of the researchers. In general, investigators view interviewer ratings in relation to an indicator of data quality such as item nonresponse or level of reporting. Item nonresponse affects data quality because it affects the amount of imputation or the number of cases that can be used for a particular analysis. It also has the advantage of being easy to measure. Other indicators such as level of reporting require making an assumption about the relationship between that indicator and response quality--for example, the more doctors visits or incidents of illness are reported, the better the data are assumed to be. This may be a reasonable assumption, but it is an assumption nonetheless. Better evidence of data quality (i.e., whether or not the questions were answered truthfully) may be very difficult to obtain. That would involve obtaining independent corroboration of respondents' answers, either through record checks or evidence from another reliable source. This is not always possible, and even when it is possible, it may be quite expensive.

The second type of response quality measure is not directly related to data collected in the survey. Instead, an assumption is made that items contained in the evaluations are indicators of the quality of the data collected

in the survey itself. For example, in collecting data for a consumer expenditure survey, the interviewer's evaluation of a respondent's ability to provide information about expenditures is assumed to reflect how accurately the expenditures were reported. Then, the items included in the evaluation questionnaire are used as the dependent variables in the analysis. Care should be taken in this type of analysis to assure that the assumptions are reasonable ones.

When all is said and done, sometimes the results of this type of research are difficult to apply directly to the operation of a survey. For instance, even if research documents that interviewers with certain types of attitudes have lower response rates or item response rates, how to alter those attitudes may not be obvious. Creative solutions to the problems uncovered by creative research are also a necessary part of the process.

EXAMPLES

Example 1: Interviewer Debriefing on the Consumer Expenditure Survey

The Consumer Expenditure Survey was first conducted by the Bureau of the Census for the Bureau of Labor Statistics, to collect data used in construction of the Cost of Living Index. This survey also provided experience that was used in designing a recurring Consumer Expenditure Survey implemented in 1979. Interviewing for the first survey was done in 1972 and 1973, using a long and extremely detailed questionnaire requesting information about types and amounts of expenditures in all categories of household expenses (e.g., mortgage payments and ownership costs, medical and health expenditures, house furnishings and related household items). The survey was structured to include five personal visit interviews at each sampled household. Data for some types of expenditures were collected in each quarterly interview; other information was collected only in one or two quarters. At the end of each interview, the interviewer told the respondent what types of expenditures would be included in the next interview, and a card or pamphlet was left with the respondent so (s)he could keep track of these expenses.

This example is included here because it illustrates interview evaluations together with interviewer debriefings.

After the first year of interviewing for the survey ended, three debriefing sessions were arranged in various sections of the country (fuller description of this research is contained in Rothwell, 1974). Twenty-one interviewers, most of whom had worked in all five interviewing periods, were assembled and their permission to have the meetings tape recorded was obtained. Discussions were led by members of one of the Bureau's research divisions, who were specialists in the area of questionnaire design. Two staff members conducted each debriefing session; one led the the discussion and the other served in the capacity of assistant.

The introduction given by the discussion leader indicated that the focus of the session was the questionnaire itself rather than pay, working conditions, or supervisory matters. After the introduction was given, the conversation proceeded according to an outline which included discussions of how people recall their purchases, types of questions that annoyed or bothered people, types of questions that respondents had trouble answering, what kinds of probes worked best in eliciting the information, and how many people kept budgets or otherwise kept track of how they spent their money.

During the session interviewers also completed a post-interview structured evaluation form requesting information about the section and item numbers which caused difficulty for respondents.

The discussions lasted approximately four hours. Later, the tape recordings were summarized independently by two researchers, and differences of interpretation were reconciled. After agreement about the content was reached, the summaries were coded for the types of problems identified by the interviewers.

The discussion outline and the evaluation form were designed to provide complementary information. Analysis of the summaries as well as the forms filled by interviewers were included in the final report.

The discussion uncovered problems that interviewers perceived as affecting data quality, which were unrelated to difficulty with any particular question. The most important problem of this type was a double bind perceived by interviewers: they were instructed to ask questions exactly as worded and also to have low item nonresponse rates. These were sometimes incompatible goals.

The discussion pointed out general areas of difficulty for respondents (e.g., the respondents had trouble understanding the vocabulary in the section on home ownership costs, the concept of "consumer unit" in others). The written comments provided specific item numbers that illustrated the problems.

Written comments also provided an estimate of the number of interviewers who reported problems with particular questions.

Substantive contributions of the interviewers, relevant to questionnaire design, fell into the the following general categories:

1. Question wording: for example, interviewers suggested replacing "vehicle registration tags" with "license plates"; it was also suggested that phrases be added to some questions to provide examples and clarify the intent of the question--"did you pay any refundable deposits for this unit, such as a security deposit?"

2. Question sequencing: for example, interviewers suggested combining questions on the same topic that were asked in different interviews--in later interviews some respondents looked up their records and felt trapped or embarrassed when they discovered they had inadvertently answered a question incorrectly in a previous interview.

3. Reference periods: for example, for certain types of items, interviewers felt that the reference period was too long; in other cases, the shift in reference periods was confusing to respondents.

4. Format and physical features of the questionnaire: for example, the cumbersome questionnaire contained many very large (11" x 16 1/2") pages, attached with wire spiral loops across the top. Suggestions were made to increase the size of the loops to facilitate turning the pages, and to print all the pages in the same direction to make the administration of the interview more convenient.

Example 2: Structured Post-Interview Evaluation on the Telephone Health Interview Survey

A research effort was undertaken by the National Center for Health Statistics (NCHS) in 1978 to investigate the feasibility of conducting federal health surveys using telephone rather than face-to-face interviews. This research (Bercini & Massey, 1979) was conducted in conjunction with a cigarette smoking supplement to the Telephone Health Interview Survey, and the indicators of data quality used were overall nonresponse rates and item nonresponse rates for the question requesting names of household members. It is presented here because it illustrates: 1) the use of post-interview evaluations with an experimental design and 2) evaluations obtained about the interviewers themselves rather than about their respondents.

One difference between telephone and face-to-face interviews is that it is relatively easy for telephone respondents to discontinue the interview (i.e., hang up the phone) at any point, whereas once a face-to-face interviewer gets access to the house, it is less likely that the interview will be terminated. The household roster (i.e., the section of the interview in which the household composition, names and demographic information about household members is obtained) is particularly subject to respondents' ending the conversation because of its sensitive nature and its seeming lack of relevance to the stated purpose of the interview.

This section of the questionnaire, therefore, was a suitable subject for investigation concerning ways to reduce nonresponse. Accordingly, an experiment was designed to see: 1) whether obtaining the household roster at the end, rather than at the beginning, of the interview would affect response rates; and 2) whether obtaining the household roster without asking for the names of the household members, rather than including them, would affect response rates.

A 2 X 2 factorial design was employed and four versions of the questionnaire were developed. The four versions were randomly distributed to interviewers, who conducted interviews using more than one version. An alternative approach, randomly assigning interviewers to questionnaire versions, was not feasible without disrupting the continuing survey, although it would have had the advantage of controlling for the effects of interviewers' preferences for one version over another.

After completion of the interviewing, interviewers' evaluations were obtained. They were asked to rate the experimental questionnaires in order of preference and ease of administration. Self-ratings were also obtained of how reluctant they were to ask for names of household members and how persistent they were in obtaining names from hesitant respondents.

The data for analysis included evaluations from nineteen interviewers and the outcomes of attempted interviews with initial respondents at 2,565 eligible households. Three different types of overall response rates and an item nonresponse rate for names of household members were calculated from the survey data.

Results showed that placing the household roster at the end of the interview rather than at the beginning significantly improved overall response rate. They also showed that interviewers preferred the questionnaire versions for which the highest response rates were obtained, and that interviewers' reluctance to ask for names was associated with lower item response rates for the name question.

The results of this research might be used to improve the design of telephone interview questionnaires. The implications of the interviewer evaluation findings, in particular, however, are applicable to interviewer training and selection.

CONCLUSION

Little information is available concerning either the procedures for using these two techniques for learning from interviewers or examples of their use. Possibly as a consequence of this, they are used relatively little in the conduct of survey research. It is hoped that this paper will make questionnaire designers more aware of these constructive tools, and that their use in questionnaire design will be increased.

REFERENCES

Bercini, Deborah and James Massey. "Obtaining the Household Roster in a Telephone Survey: The Impact of Names and Placement on Response Rates," American Statistical Association, Proceedings of the Social Statistics Section, 1979, pp. 136-140.

Rothwell, Naomi D. "Results of Debriefing Interviewers on the Consumer Expenditure Survey," U.S. Bureau of the Census, (Unpublished memorandum, February 11, 1974).