7

Question Design and Evaluation Issues in Perspective

The quality of data from a survey depends on the size and representativeness of the sample from which data are collected, the techniques used for collecting data, the quality of interviewing, if interviewers are used, and the extent to which the questions are good measures. Methodologists have a concept that they call total survey design (Groves, 1989; Horvitz & Lessler, 1978). By that, they refer to the perspective of looking at all sources of error, not just a single source, when making survey design decisions. The quality of data from a survey is no better than the worst aspect of the methodology.

When Sudman and Bradburn (1974) looked at sources of error in surveys, they concluded that perhaps the major source of error in survey estimates was the design of survey questions. When Fowler and Mangione (1990) looked at strategies for reducing interviewer effects on data, they, too, concluded that question design was one of the most important roads to minimizing interviewer effects on data. Moreover, although the design of surveys often involves important trade-offs, improving the design and evaluation of survey questions is one of the least expensive components of the survey process. Compared with significantly increasing the size of a sample, or even the efforts required to improve response rates significantly, improving questions is very cost effective. Thus, from the perspective of total survey design, investing in the design and evaluation of questions is a best buy, one of the endeavors that is most likely to yield results in the form of better, more error-free data.

The book has covered many issues, some big, some small, that affect the quality of questions as measures. In this final chapter, we attempt to summarize the main points to provide some perspective on the most important issues to which to attend.

Factual Questions

Almost certainly, the biggest problem with questions designed to measure facts and objective events is the failure to make the step from

the question objective to a set of questions that people can answer. Too often, questions are simply a repetition of the question objectives.

The key principles are straightforward:

1. Ask people questions they can answer.
2. Make sure that all the key terms and concepts are clearly defined, so people know what question they are answering and they are all answering the same question.
3. Provide a context in which people will see answering questions accurately to be the best way to serve their own interests.

One further point should be made about interviewer-administered surveys. Attention must be paid to the fact that the survey instrument is also a protocol for an interaction. Attending to the sequence of questions and the way that answers to prior questions will affect the subsequent question-and-answer process can be a key part of improving the standardization of data collection and making the interview a positive data collection experience.

Measuring Subjective States

The primary problem for designers of measures of subjective states, like those of objective phenomena, is defining the objectives. A clear statement of what is to be measured is one key to the solution of many question design problems. Most often, the specification of measurement of objectives will take the form of wanting to place the respondent on a continuum or place the respondent's perceptions of something else on a continuum.

Once the objectives are specified in a clear way, the three key standards for subjective questions are that:

1. the terms of a question be clear, so everyone is answering the same question;
2. the response task is appropriate to the question and is relatively easy for most people to perform; and
3. the response alternatives are designed so that respondents who differ in fact in their answers will be distributed across the response alternatives.

In addition to these basic principles, it also is valuable to maximize the extent to which answers to questions provide measures for all respondents, not just a subset. Careful examination and pretesting of

questions, to identify those that have hidden contingencies in order for them to be meaningful questions, can greatly improve the quality and efficiency of survey measurement. In the tradition of personality testing, when testers could include extraordinarily long inventories of questions, it may have been valuable to include items that provided useful information about small segments of respondents. However, respondent burden is a major concern in general-purpose surveys. Although multi-item measures can greatly improve the measurement process, particularly for subjective phenomena, investigators also have a responsibility to minimize respondent burden and to place people on continua as efficiently as possible. The information contained in the answers to 20- or 30-item scales can virtually always be reproduced with a small subset of those items, if they are carefully chosen. In this context, choosing items that provide the most information about each respondent is the efficient, and indeed ethical, way to proceed for measures of this sort.

Finally, having respondents place rated items, themselves or others, on scales, rather than using an agree-disagree format, will almost always provide better measurement both from the point of view of the simplicity of the task and the amount of information derived from each question.

Testing Questions

Focus groups, group discussions, cognitive interviews, and field pretests that include coding interviewer and respondent behavior should be a standard part of the development of any survey instrument.

The most important three premises for the evaluation of survey questions are:

1. Questions need to be consistently understood.
2. Questions need to pose tasks that people can perform.
3. Questions need to constitute an adequate protocol for a standardized interview, when interviewers are involved.

These goals seem so self-evidently valuable it is hard to believe that all survey questions do not meet these standards. However, they do not. In one study of 60 questions drawn from government and academic survey instruments, a clear majority were identified as failing to meet one or more of these basic criteria (Oksenberg, Cannell, & Kalton, 1991). On average, over a third of all survey questions are subject to

significant interviewer effects on results (Groves, 1989), and there is clear evidence that questions that require probing and clarification by interviewers are most likely to be affected by interviewers (Mangione, Fowler, & Louis, 1992).

Cognitive interviews and behavior coding field pretests provide reliable, replicable information about question problems. The problems identified can be corrected, and the results are better data (e.g., Fowler, 1992; Oksenberg et al., 1991; Royston, 1989).

There is still work to be done to refine these procedures, to develop better and clearer standards for question problems, and to improve the generalizations about how to solve the problems that are identified with these processes, yet one of the important realities for students and researchers to grasp is that many of the worst question problems can be identified with simple, informal testing. Try questions on friends, parents, or children. Have them answer the test question, then describe in narrative form how they understood the question and how they arrived at the answer. Although rigorous, routine testing is necessary to advance survey science, better questions and better measurements result whenever researchers take steps to critically evaluate how consistently people can understand and answer their questions.

Evaluating the Validity of Questions

Around 1970, Robinson and associates published a critical evaluation of common survey measures of social psychological states and political attitudes (Robinson, Rusk, & Head, 1968; Robinson & Shaver, 1973). Those books were embarrassing testimony to how little attention was given to the assessment of how well commonly used questions performed as measures.

Twenty years later, progress has been made. A recent book by Robinson, Shaver, and Wrightsman (1991), which covers ground similar to the earlier volumes, finds many more standard measures that have been carefully evaluated. McDowell and Newell (1987) review common measures of health status and quality of life, again finding some encouraging trends with respect to the studies that have been done, particularly of more recently developed measures. A recent book by Stewart and Ware (1992) provides a kind of prototype for systematically developing measures of important health concepts.

Increasingly the word is out that particularly when scales and indices are used, validation studies are necessary. On occasions, measures are referred to as if being "validated" was some absolute state, such as

beatification. Validity is the degree of correspondence between a measure and what is measured. Measures that can serve some purpose well are not necessarily good for other purposes. For example, some measurements that work well for group averages and to assess group effects are quite inadequate at an individual level (Ware, 1987). Validation studies for one population may not generalize to others. Kulka et al. (1989) report on a set of items to measure mental distress that differentiated extremely well between mental patients as a group and the general population. However, when those same items were used in a general population sample, they correlated very poorly at the individual level with independent clinical assessments of psychological problems.

The challenges at this point are of two sorts. First, we need to continue to encourage researchers routinely to evaluate the validity of their measurement procedures from a variety of perspectives. Second, we particularly need to develop clear standards for what validation means for particular analytic purposes.

Conclusion

To return to the topic of total survey design, no matter how big and representative the sample, no matter how much money is spent on data collection and what the response rate is, the quality of the resulting data from a survey will be no better than the questions that are asked. In 1951, Stanley Payne titled his landmark book, *The Art of Asking Questions*. We now know we can do better than that. Although we can certainly hope that the number and specificity of principles for good question design will grow with time, the principles outlined in this book constitute a good, systematic core of guidelines for writing good questions. In addition, although the development of evaluative procedures will also evolve with time, cognitive testing, good field pretests, and appropriate validating analyses provide scientific, replicable, and quantified standards by which the success of question design efforts can be measured. In short, at this point, there is no excuse for question design to be treated as an artistic endeavor. Rather, it should be treated as a science.

Unfortunately, there is a long history of researchers designing questions, in a haphazard way, that do not meet adequate standards. Moreover, we have a large body of social and medical science, collected over the last 50 years, that includes some very bad questions. The case for holding on to those questions that have been used in the past, in order

to track change or to compare new results with those from old studies, is not without merit. However, a scientific enterprise is probably ill served by repeatedly using poor measures, no matter how rich their tradition. In the long run, science will be best served by using survey questions that have been carefully and systematically evaluated and that do meet the standards enunciated here. There is work to be done so that researchers routinely build in the kind of pretest and question evaluation procedures necessary to ensure that their questions are good. Such processes are increasingly being used, and it is to be hoped that this book will make a contribution to the further development and improvement of the question design and evaluation process.