Questionnaire Design: From Art into Science

Paper Delivered at the Fifth International Conference on Social Science Methodology

Cologne, Germany

October 3, 2000

Norman M. Bradburn

National Science Foundation

It is a great honor to have been asked to speak about developments in questionnaire methodology, especially in a series that has been billed informally as talks by the "founding fathers." It is a particular pleasure for me personally to be giving this talk here in Cologne because much the genesis of my work comes from the year I spent here in Cologne as a Von Humboldt fellow at the Zentral Archive, just a few blocks from here.

Since I have been asked to reflect on the developments in the field during my active professional life, please forgive me if I am overly autobiographical. But I think something general can be learned about scientific progress by reflecting self-critically about the evolution of my thinking about survey methodology. Recognizing that generalizing from a sample of one is very dangerous, I shall nonetheless try to draw from general lessons from my own intellectual journey.

When people ask me to introduce myself and tell them a bit about what I do, I often say: "I study error." This usually provokes a laugh or at least some feeling that I must an eccentric because everyone knows that scientists search for the truth. Perhaps because error is everywhere with us and one never wants for data, I have managed to make a good career out of trying to understand errors. I was trained as a psychologist rather then a statistician which may have something to do with the fact that I have always been more interested in measurement error than in sampling error, more interested in bias and than in variance, more interested in validity than in reliability. For the past thirty years I have been studying response errors in surveys. Much of the work has been done in collaboration with the late Seymour Sudman. Since our contributions to survey methodology have been so intertwined, this occasion is as much as a memorial to him as is it an honor to me.

Since the beginning of scientific surveys, researchers have been sensitive to the fact that changes in the wording of questions could produce dramatic changes in the distribution of responses. Already in 1941 Rugg (1941) had shown that changing the wording of a question from "Do you think the United States should allow public speeches against democracy?" to "Do you think the United States should forbid speeches against democracy?" changed the proportion of respondents favoring free speech from 21% in favor to 39% in favor. Analogous findings were published from time to time and Payne's classic book, The Art of Asking Questions was filled with cautionary tales and good advice about pitfalls in question wording. But there was no question in practitioner's minds that questionnaire design was still an art and not a science.

I came to survey research from a background in experimental psychology. In designing psychological laboratory experiments one is constantly aware of the possible biasing effects of the order in which stimuli are presented. Experimentalists are careful to counter-balance the order of presentation of stimuli so that any order effects are spread equally over the conditions and thus are controlled for. When I first began to design survey questionnaires, I brought this training to my work and began to introduce experimental procedures for changing the order of questions.

When I went to write up my first methodological article two things surprised me. First, when I reviewed the literature on order effects in surveys, I found the literature very sparse. Second, perhaps related to the first was that the standard journals were not interested in the topic. Thus my first methodological article on questionnaire order effects was publish in the Journal of Marketing Research, which was just getting started and looking for articles. Fortunately for my academic career, my appointment at the time was in the Univ. of Chicago Business School, so a publication in a marketing journal was viewed positively. If I had been in the sociology or psychology departments, I might not have been so lucky.

At Chicago I was also working at the National Opinion Research Center (NORC). NORC had a long tradition of methodological research on interviewer effects, but had done relatively little on questionnaire design. When I became director of NORC in 1968, I decided that we should turn our attention to studying total error, including measurement error. At that time, the received wisdom was that the most important source of measurement error was the interviewer. Seymour Sudman, who at that time was NORC's chief statistician, and I began a systematic review of the literature on measurement error in surveys. The 1970-71 academic year I spent here in Cologne working on that review, which was eventually published under the title Response Effects in Surveys: A Review and Synthesis. (1974). During that year I visited a number of survey organizations in Europe and collected a number of published and unpublished methodological studies that we incorporated into our review.

As we were developing a conceptual framework to do the review, I was particularly impressed with the work of the Prof. Elisabeth Noelle-Neumann at the Institut fuer Demoskopie/Allensbach. She called my attention to the 1934 work of Bingham and Moore (1934) who conceptualized the survey interview as a "conversation with a purpose." Building on this idea, Sudman and I proposed a simple model for understanding the sources of response effects. The model started with the view of the survey interview as a special kind of conversation between individuals who speak the same language. It has a definite structure and particular rules that differentiate it from other conversational forms such as a casual conversation between friends or an instrumental conversation between colleagues.

We viewed the survey interview as a micro-social system consisting of three roles held together by a common task reflected in the questionnaire. One person (the interviewer) has the role of asking questions, a second person (the respondent) has the role of answering the questions and a third person (the researcher) defines the task. The researcher has a particular purpose in mind, that is, to obtain information to answer some general practical or scientific questions. The interview is in some ways like all conversations, in some ways like many conversations, and in some ways like no other conversation.

By conceptualizing the survey interview as a social system with several roles united by a common task, we were able to identify three primary sources of response effects---those stemming from the interviewer, those coming from the respondent, and those coming from the task itself, particularly the questionnaire and the context within which it is perceived. The meta-analysis we did in the review suggested that the largest effects were associated with the task itself, that is the question asking and answering process.

In the 1970's Sudman and I undertook a series of studies to look at wording effects on the accuracy of reporting various behaviors that could be validated with external data. More or less at the same time Howard Schuman and his collaborators at the University of Michigan (Schuman and Presser, 1981) were engaged in a series of split-ballot experiments on attitude questions. All of this work built on the earlier and continuing work at the Allensbach Institute in Germany (XXX) and that of William Belson in England (1981, 1986). Together these lines of work indicated that the most important part of the process was the wording of the question. The wording of the question is what the respondent must process in order to understand what information is being sought, or for that matter, that it is information that is being sought and not, say, an expression of affect or some other type of communication, as might be the used in another conversations.

This point is so fundamental that its importance is often overlooked. To take a vivid but trivial example, the questions: "How old are you?" and "In what year were you born?" have only one word in common but are easily understood as calling for the same information. Estimates of the age distribution of the population based on responses to the two questions from a sample of the population will differ, but not by very much. On the other hand, if the questions were: "When were you born?" and Where were you born?" the responses would be so different that the answers could not be compared, even though the questions share all the same words except one. The point is that the question's meaning is what is important, not how many words are in common. As Bateson (1984) pointed out: "A survey trades in meanings, and meanings are embodied in language. A survey consists of a transfer of meanings between the participants through the medium of language."

Interviews entail much more than just questions, of course, including a lot of nonsemantic information. For example, there is the social context in which the interview takes place. It may take place in the home of a respondent or in the workplace of the respondent who is an employee of an organization about which the questions are asked, or in a researcher's laboratory with respondent as a "subject" in an experiment that is part of a study whose purpose is perceived to contribute to scientific knowledge or to learning the craft of research.

In addition characteristics of the interviewer and the respondent, such as race or sex, may influence the willingness to express attitudes or affect relations in ways that are independent of the interview and reflect larger social beliefs. The mode of administration of the questions, such as in person, by telephone or self-administered or using various forms of computer assistance, may also have important consequences for the responses. These sources of response effects, however, have proven to be of smaller magnitude than those effects coming from question wording and context.

While the work that Sudman and I did in the response effects book had some effect on the type of methodological work that was done in the 70's, progress toward making questionnaire construction more of a science began in earnest a little over 20 years ago with a seminar held in 1978 by the British SSRC and the Royal Statistical Society on problems in the collection and interpretation of recall data in social surveys (Moss and Goldstein, (eds.). The Recall Method in Social Surveys. London: NFER Publishing co., Ltd., 1979). Two important events occurred in the United States in 1980. The first was a workshop convened by the Bureau of Social Science Research in connection with its work in the redesign of the National Crime Victimization Survey sponsored by the National Institute of Justice and conducted by the Census Bureau. This workshop brought together cognitive scientists, survey statisticians and methodologists to discuss the contributions cognitive scientists could make to understanding response errors in reports of behavior.

The second was the establishment of a panel on the measurement of subjective phenomena by the Committee on National Statistics at the National Academy of Sciences (ref). This panel produced two large volumes that stimulated a considerable amount of research on response errors involved in the measurement of subjective phenomena which complements the work that had been stimulated by the earlier seminars on measuring behavior or more "objective" phenomena, that is, phenomena that have external validity criteria. One of the recommendations of the panel in its final report was that there should be more interdisciplinary research involving cognitive and social scientists working with survey methodologists on problems of measurement of subjective phenomena.

I was a participant in the BSSR workshop, and I found the discussion with cognitive psychologists very stimulating. For a number of years before the workshop I had been teaching introductory psychology and had become acquainted with recent work in cognitive psychology. As a result of teaching that course I was "primed", as cognitive psychologists might say, to be receptive to the ideas presented by the cognitive scientists. The major stimulus, however, came in 1983 when the Committee on National Statistics with funding from the National Science Foundation, organized a 6-day seminar on Cognitive Aspects of Survey Methodology. (Jabine et al., 1984) The conference was extraordinarily fruitful and has led to a whole new field of research in survey methodology both as applied to objective and subjective phenomena.

We were not alone in seeing that a partnership between cognitive scientists and survey methodologists would lead to new knowledge. Norbert Schwarz and his colleagues at Heidelberg and Mannheim were also working along these lines. A chance encounter when I giving a lecture in Mannheim led to further workshops and a visit by Sudman to Mannheim. When Schwarz moved to the University of Michigan, collaboration became closer with a yearly workshop devoted to building bridges between various subdisciplines of psychology, linguistics and survey methodology as well as a considerable program of original research by scientists who participated in these activities and their students. The culmination of the work that Sudman, Schwarz and I did was published in our book Thinking About Answers (1997).

It might be useful to reflect for a moment on the conditions that gave rise to this burst of creativity. While it is impossible to designate any one person as the father of all of this work, it is clear that an important person in making it happen was Murray Aborn, the head of the Methods, Measurement and Data Resources program at the US National Science Foundation. Aborn was the intellectual entrepreneur who fostered the idea of bringing cognitive psychology together with survey methodology and, through selective funding of developmental work and workshops, he was able to get the right scientists together to make it happen. Scientists working at several important research centers such as NORC at the University of Chicago, the Survey Research Laboratory at the University of Illinois, ZUMA in Mannheim, and the Survey Research Center at the University of Michigan took up the cause. These scientists were mostly trained as psychologists and statisticians but working in centers that specialized in problems of survey methodology. They were actively involved in designing or conducting surveys and survey questionnaires. They were also interdisciplinary centers so that other ideas from other disciplines could and did influence their work. And, of course, there had to be sufficient funding to support the work. This funding was provided most notably by the U.S. National Science Foundation, the

Alexander Von Humboldt Stiftung and the Deutsche Forshungs Gemeinschaft. The ideas, the people and the tools all came together at a time when sufficient funding was available to make possible an exciting period of creativity that I had the privilege to be a part of.

There has been so much work done in this area that it would be impossible for me to summarize it here. What I would like to do in my remaining time is to take one area of concern to give you an idea of how our thinking about these problems evolved. This area concerns context effects in attitudinal surveys

Order Effects

Let me start with the old problem of order effects. There have been many examples of order effects in the literature, almost from the beginning of surveys. (Am. Marketing Assoc., 1937; Cantril, 1944) The results from research on order effects, however, had been so confusing and contradictory that it was tempting to conclude that the positive studies demonstrating order effects were really artifacts of some sort.

As I mentioned, from my training in experimental psychology I knew that the order in which stimuli were presented in experiments could influence the responses. A well established principle in psychology was that in the visual mode, there was a primacy effect, that is stimuli coming at the beginning of a series had a perceptual and cognitive advantage and, in the auditory mode, a recency effect, that is the last stimuli in a series had an advantage. Also, learning and memory experiments dating back to Ebbinghaus had shown that words at the beginning and end of lists were remembered better than words coming in the middle. Primacy and recency effects are pervasive phenomena in human perception and cognition. It therefore seemed highly likely that they would be operable in survey questionnaires which frequently present respondents with lists of items about which they have to make some sort of judgments such as agree/disagree or like/dislike.

In interviews that are conducted in person, questions have to be asked in some order, although they need not be asked in the same order for everyone; but some order, usually predetermined, nonetheless. Since interviews are conducted in different modes, simple serial order effects will differ by mode of administration. In telephone interviews, recency effects will predominate because the interview in entirely oral. In personal interviews printed cards with response categories are often used or now computer-assisted personal interviewing frequently allows or requires the respondent to view the lists on the screen. In such cases primacy effects should predominate. Self-administered paper-and-pencil questionnaires in which respondents can see the entire set of questions are one time appear to be less susceptible to simple serial order effects because the questions and response categories are there all at once for respondents to refer to back and forth.

Simple serial order effects for lists and response categories, however, are not all there is to order effects. One of the factors that inhibited our progress in understanding order effects was our lack of a theoretical structure within which to investigate the mechanisms by which they might occur. In the 80's a more general framework built on information processing was developed that had an important role in moving forward our understanding of order effects, or more generally as it became clear, of context effects. There were several models developed which built on some earlier work by Cannell at Michigan, one by Tourangeau and Rasinski at NORC and the other by Strack and Martin at ZUMA. These models conceptualize the question/answer process as one of stages, which require different types of cognitive processing and thus might be susceptible to different types of context effects. For convenience, I shall organize my discussion around the Tourangeau and Rasinski model of four stages, interpretation, retrieval, judgment and response.

Interpretation: Let me start with interpretation. One theme in studying order effects has been the way in which preceding questions affect the interpretation respondents give to questions. A good example of how preceding questions can radically influence the interpretation of what appears on the surface to be a fairly unambiguous question has been given by Jaak Billiet, a Belgian sociologist. Consider the question: "How many children do you have?" When this question was asked of teachers preceded by questions about their own families, it was interpreted as referring to their own biological children. When it was asked following questions about their classroom, it was interpreted as referring to the number of children in their class. In the first case, one got replies on the order of magnitude of 0 to 4; in the other case, one got answers on the order of magnitude or 25-40.

One of the more well established findings of research on order effects is that when a general question and a more specific related question are asked together, the general question may be affected by its position, while the more specific question is not. In the absence of any theoretical framework we had been unable to do more than document the occurrence of the phenomenon. If we view the survey interview as a conversation, however, we can ask whether there are general rules about conversations that might help us understand some of these effects. Schwarz has used the conversational rules developed by Grice to understand this type of order effect. One of the rules of conversation is that questions should not be redundant, that is the questionnaire does not ask the same question twice in a row. Thus when a general question is asked first, as for example: “Taking all thing together, how happy are you?” it is presumed to include all aspects of life such as work, marriage, social life, etc. If it were followed by series of specific questions about different aspects of life, such as marriage or work happiness, these questions would not be redundant. If the specific questions are asked first, however, the general question is interpreted as being restricted to those areas of life that have not already been asked about so as not to be redundant.

The effect of the order on the responses will depend on the relation of the two questions for the particular respondent, which is why the effects are often so elusive. For example, consider the pair of question about general happiness and marriage happiness. If the marriage happiness question is asked first and the general question second, the general question will be interpreted as asking for a judgement about happiness in all aspects of life except marriage. If the respondent’s marriage is very happy, then the respondent may report less happiness in general than if the general question came first and the respondent had included marriage happiness in the response to the general question. But if the respondent’s marriage is not so happy, then the respondent may report more happiness in response to the general question than if the general question had come first. In the population as whole, most people have happy marriages. Therefore, when used in a population survey, the overall effect is to have lower reported general happiness when the general question comes after the marriage happiness question than when it precedes it. But for some individuals, the effect goes in the opposite direction

Retrieval: I now turn to retrieval. Priming is one mechanism studied by cognitive psychologists. The effect of priming is seen in the faster retrieval of the primed information from memory. Under standard survey conditions in which respondents are asked to answer questions on a number of topics under fairly rapid time pressure, particularly if it is a telephone interview, we might expect that questions (or introductions to groups of questions) will have a priming effect on memories related to the topic under discussion. Thus questions later in a section of related questions might be influenced by those earlier in the questionnaire through their priming effect on related information. Such effects may be particularly important for questions about behavior that require retrieval of behavioral episodes from memory. Indeed, Sudman,Bradburn and colleagues (1979) and Cannell, Marquis, and Laurent (1977) have shown that longer introductions to behavioral questions result in fuller reporting of behavior, as effect that we interpret as due to priming of related memories.

The priming of memories through activation of semantic networks is thought of as an automatic process. Individuals, however, will have somewhat different networks or hierarchies, so the effect of activation will not be simple. If we think of attitudes as organized semantic structures in memory, priming may cause some respondents to retrieve new information or it may just make some information more available than other information so that it comes to the fore more readily than the non-primed information.

Tourangeau, Rasinski and Bradburn reported studies that suggest preceding context questions can prime schemata that lead to different responses. Of particular interest is the finding that some attitudes, e.g. about welfare, consist of schemata that are organized along a pro/con dimension of support, while other attitudes, e.g. about abortion, consist of schemata that are not so organized. Knowing which schema is activated, however, does not give us any clear picture of what the resulting judgment will be. It depends on teh values associated with the schemata.

Judgment: After questions are interpreted and the relevant information is retrieved from memory, respondents must still make a judgment about how the retrieved information can be used to answer the question, that is they must still map their answers onto the response categories offered by the interviewer. Hippler and Schwarz suggested a way in which the length and form of response alternatives can affect responses. While some observed effects of position on long lists is due to primacy and recency effects which may work at the retrieval stage, some of the effects of response alternatives appear to come about by the implicit information in the response alternatives that allows respondents to fit their own attitudes or behavior onto a comparative scale and make a judgment about where they belong. Questions often explicitly, but sometimes implicitly, ask respondents to rate themselves in relation to an average value or along a dimension that has implicit distributional properties, e.g. the response categories require that respondents rate themselves as high or low or very high or very low, etc. on some dimension such as agreement, satisfaction or support.

Evoking norms explicitly or implicitly may affect responses at the judgement stage. Question order may evoke norms implicitly. For example, Sheatsley found that Americans were more willing to let communist journalists work in the United States when the question was preceded by a question about Russians letting American journalists work in Russia, than when the question about Russians allowing American journalists to work in Russia followed the question about communist journalists working in the U.S.   The question about Russians letting American journalists work in Russia primed a norm of reciprocal rights that carried over into following question. Numerous examples of this type of effect have been found.

The general point here is that both preceding questions and response categories may evoke some sort of norms or standards of reference that respondents then use to match the ideas and thoughts retrieved with the answer alternatives being offered by the interviewer. We expect that the effects are somewhat more pronounced in closed-ended questions where the response categories are fixed and respondents are forced to pick among a limited array of possibilities than in open-ended questions where respondents can answer in their own words. Even here, however, respondents will use contextual information about norms and standards of references.

Responses: After respondents make their judgements about the answers that fit the question, they must actually produce the answers so that the interviewers can record them. It is at this stage that issues such as self-presentation and consistency may come into play, which might alter answers that are reported. I noted earlier that that the interview situation is a social system, and, as such, is subject to the general norms of social behavior. Even though efforts are made to minimize the cues for norms of ordinary social discourse, and respondents are asked to respond truthfully, there may be a conflict between the role of a good respondent and the respondent’s own norms of self-presentation. For sensitive questions, such as those about drug use, sexual behavior, or illegal behavior, this conflict may be quite acute and lead to respondents distorting their answers in a socially approved direction. Question context or order may make things better or worse. Relatively little research as been done recently on this stage of the question/answer process.

Conclusion

What progress have we made? I believe that we have made some progress over the years since scientific surveys have been on the scene. I would distinguish three stages in our understanding of context effects. The first stage, which began almost with the birth of survey research in the 1930's, consisted of research that demonstrated that question order effects did in fact exist and could alter results. I would call this a demonstration period. Sometime in the 1950's concern began to move toward classifying the types of effects that might result from question order. This was the classification stage. In my first paper devoted to the problem of order effects (Bradburn and Mason, 1964) I distinguished four sources of order effects: redundancy, salience, consistency and fatigue. It now looks to me as if the first three might map fairly easily onto the effects at the stages just discussed. Redundancy effects look like what we see at the interpretive stage when the conversational rules are operative so that respondents interpret questions as not including any information already given. Salience is similar to priming whereby certain cognitions are activated so that they are more easily accessible to respondents in answering subsequent questions. Consistency look as if it might be either at the judgment or response stages depending on whether it comes from priming norms used to form judgments or is seen as part of a self-definition that pushes respondents to appear to the interviewer as consistent individuals. Fatigue, which refers to the placement of questions in the total interview, is only a factor in very long interviews. Respondents may be less motivated to fully answer questions that come at the end of the interview because they are just plain tired.

Finally, the development of an information processing approach to studying response errors in surveys has allowed us to move from classification to explanation and to see that order effects are just a special case of context effects. We now have a theoretical framework within which we can investigate order and other context effects and expect that our knowledge will be cumulative. We have begun to understand the mechanisms that produce different types of response effects. We have much more to learn, but when I look back over the 40 years I have been active in this field, I am pleased at our progress and proud of the part I have played in it.