DISCUSSION

N. D. Rothwell, Bureau of the Census

I begin with an observation that Mr. Kalton did not stick to the subject of survey questions and he probably was right in not doing so. He extended the scope of his paper to cover survey procedures employed to measure, control or correct response error. Bounding to control for telescoping, diary keeping to aid recall, training interviewers to speak slowly and give feedback considered appropriate, instructing respondents and requesting commitments from them--are all modifications of survey procedures which do not necessarily involve question wording or sequence.

In order to make more of Mr. Kalton's and my remarks appropriate, I'd like to redefine our subject. Survey questions are a subset of respondent tasks and, if all of the tasks are considered, more of our remarks can be seen as pertinent. In addition - sampling, interviewer training, respondent selection, and other aspects of survey research also become pertinent to the discussion when they are viewed as contributing alternative explanations for outcomes otherwise attributed to question wording or sequence.

My second general comment deals with classification of survey questions and the consequences of it. In planning a Census Bureau course about questionnaire design some time ago we wanted to start with a common vocabulary. We listed 11 ways of classifying the question itself; 4 ways of classifying the way respondents perceive the question; 9 ways of classifying the questions in the context of an interview; 3 ways of classifying the environment of the interview; 9 ways of classifying the physical aspects of a questionnaire, and 2 ways of evaluating the question. That first effort to develop a taxonomy illustrated the large number of dimensions which can be used to describe survey questions. The exercise had another value. By indicating the many other features of survey questions, the taxonomy deemphasized the fact-versus-opinion dichotomy and suggested reasons why attributes, commonly ascribed to factual questions, are so often lacking. At the margin, subjectivity or arbitrariness determine not only what is defined as a room - which Mr. Kalton mentioned - but also whether a crime has been committed, who are members of a household, what is a person's ethnicity, whether a house is habitable, and whether a firm is a wholesaler or retailer. Moreover, there is a wide range of factual questions: Attributes, like age or sex; aggregates, like costs or sales; practices, and knowledge (which in our taxonomy was classified as factual!). Some subjects, like intentions, can't be classified easily as factual or opinion but can be validated.

Finally, the other feature Mr. Kalton ascribed to factual questions, the ability to validate by record checking, exists for such a small fraction of them that it can rarely serve as a test to distinguish between factual and opinion items.

Now, if the ability to validate by record checking is rarely possible, even for so-called factual studies, is survey research left only with the second choice Mr. Kalton mentioned? He described that as selecting the question which provides the response most closely resembling an apriori assumption about the general direction of response errors. The answer to that long question is a short no. There are other alternatives. An example of one is described in the Census Bureau Technical Paper 11 by Neter and Waksberg. In their research for the Survey of Residential Alterations and Repairs, external bench marks were not employed nor could they have been, except for a possible small subset of the many kinds of expenditures covered by the survey. Neither were assumptions made that bigger was better or smaller more beautiful. Instead, an optimal data collection system was imagined; one in which memory was taxed least, telescoping was eliminated, conditioning was minimized and ideal respondents were designated for interview. Using rotating panels and alternative compatible procedures, results from such ideal interviews were then compared for identical time periods with results which depended on longer memory, unbounded interviews, later interviews in the cycle, and interviewing any available household respondent. Deviations from the responses considered ideal were measured at the same time that a satisfactory data collection system was determined. For what it is worth, the ideal procedure in that survey produced estimates of expenditures somewhere between the highest and lowest generated by the alternative procedures and no apriori assumptions about the size of estimates would have led to the observed conclusions or recommendations.

You may have noted that only one of the experimental procedures studied in the Neter-Waksberg paper involved question wording changes--1 month, 3 months, 6 months recall. The others were bounding, measurement of conditioning, and respondent selection. So, to repeat, I have been talking about related phenomena, not necessarily survey questions. And, I'd like to continue talking around the subject after a digression to the heart of the matter--some research Mr. Kalton mentioned which the Census Bureau has been doing with self enumerative questionnaires.

In the first results of classroom experiments, a carry over effect was identified. Now, we didn't discover the existence of carry over, position or order effect. Mr. Kalton referred to it and Charles Turner will also talk about it. Survey researchers have long tried to reduce context effects, particularly by putting complex, difficult, or sensitive items at the end of the interview. In discussions about whether an item is sensitive or difficult, whether it matters where it is put, or whether it is included at all, we have heard some researchers suggest that their interviewers have such superior training, skill, and understanding that, for example, they can collect income information at the beginning just as well as at the end of the interview. And they may be right. Even split panel test outcomes may differ according to skill levels of interviewers. A staff of well trained and motivated interviewers

and accurate recorders may obtain information from a question while amateurs or indifferent interviewers fail or misreport, thereby affecting observed differences between that and an alternatively worded question.

Because of confounding between question and interviewer effect, self enumeration provides a better test of the effect of the question or task itself on the response. That leads me to assert that the experiments conducted by the Census Bureau provide the first unequivocal evidence of carry over effects in surveys--although, as Charles Turner correctly warned, there is no assurance that the observed effects would carry over from self enumerative to personal interview surveys.

Here is a description of an experiment: Matched questionnaires differed in whether respondents were requested to report ages of household members only in writing versus in writing and machine readable position marking. The difference in the task is shown on the first page of the handout*. The questionnaires were randomly distributed to three groups which consisted of about 100 night high school students, 100 participants in a job training program half of whom had completed high school, and about 100 members of a U.S. Army Military Government Reserve Unit comprised principally of lawyers. Three hundred and four forms provided equal numbers for each experimental treatment and the second page of the handout shows the results*: Asking respondents to do position marking of age had a depressing effect on response to subsequent questions.

The design for the experiment was completely randomized so that all differences in the items printed after the age question were balanced, no interactions were measured, and the observed differences can be attributed only to the questions about age.

Hypotheses about the reason for this carry over effect were developed in a different setting with different people. Unlike the experimental sessions which started with the briefest possible introductions, observation began with group discussion of how people get mail, impressions of the census envelope, the appearance of its contents, and what participants might do with the form if they were home. Participants were then encouraged to continue talking aloud about the task of filling the form, which each one did as an observer watched and listened. Observation led to the belief that frustration and a sense of failure discouraged some of the less well educated from continuing while some of the better educated were antagonized by what they perceived to be the unnecessary complexity of the form.

Now I'd like to resume discussion of phenomena frequently associated with question wording and suggest some alternative explanations for them, beginning with an explanation for what has been considered social desirability or prestige bias. A few years ago we collected evidence that questionnaires may sometimes be blamed incorrectly for inflated estimates of socially desirable behavior which is really due to nonresponse or, by conjecture, undercoverage. In a small survey conducted in Camden, New Jersey at the time of the city-wide census of April 1977, respondents were asked whether they had mailed back their census questionnaires. Like voting, mail back has been considered a civic duty and some over-reporting was expected. Fifty-four percent of those who answered the question reported that they had mailed the forms back. There was a high proportion of no answers and the survey estimate only approached being significantly higher than the Bureau's official estimate of 50% mail back. The survey which produced the estimate of a 54% mail back of census forms had a response rate of 83%, poor by Bureau standards but certainly not unheard of in survey research. We had the unusual opportunity of being able to learn from the Census forms themselves the behavior and characteristics of survey non respondents. A record check showed that a significantly smaller percentage--only one-third of the non respondents--had mailed back their questionnaires. Thus, the difference between the mail back rate reported by respondents and the official estimate was explained by the lower mail back rate of nonrespondents.

The appearance of prestige or other response bias can also be created by sample design: household samples which omit people occupying nonconventional living quarters, telephone samples which omit households without telephones, and any sample plan which permits substitution, or drops recalcitrant or hard to reach people. Sample universe limitation and nonresponse contribute to an image of survey research respondents as the friendly, public spirited majority; people who respond to requests for cooperation from government, universities or any public interest organization. They may, in fact be more likely to vote, read books, contribute to charity and engage in socially acceptable behaviors than non respondents or other populations. Spilt panel tests might help to distinguish between response and nonresponse bias but observation of them has often proved discouraging. The first respondent who is asked version A doesn't understand it and soon someone is offended by version B so the interviewer solves the familiar double bind (ask questions exactly as written but always get replies!) by using version C for almost everyone. Since interviewers don't have identical experiences and do have differing abilities at question wording, version C has lots of variants.

As I suggested earlier, hope for dependable spilt panel tests lies in studies of self enumeration. We can also hope that computer-assisted telephone interviewing or CATI research will improve questions and control for interviewer effects--though expectations of CATI achievements may be too optimistic, since reports from SRC now suggest that interviewer effects are more subtle and stubborn than anticipated.

Although I just suggested that sample design and nonresponse can create only the appearance of response error, I certainly do not deny the existence of response error. I have, however, become increasingly impressed with two other potential contributions to it. The first is the interview setting which Mr. Kalton mentioned and the second is respondent belief about surveys. The friendly majority who respond to surveys can be divided into people who are intrinsically friendly, people who are merely mannerly, and people persuaded by the interview situation or interviewer to act as if they were. Observation of survey interviewing generally shows a polite exchange. Even when they are not themselves middle class, interviewers are trained in middle

class manners--to be deliberate, nonjudgmental, affable, and to behave as if they were guests in respondents' homes. Moreover, a good interview questionnaire is logical and orderly. Orderly presentation of ideas for consideration tends to dampen emotional affect and create blandness. The one-to-one relationship of a successful, self-assured, and authoritative interviewer and a cooperative respondent keeps the interview on track and the respondent compliant.

What makes me sure that the environment of an interview affects response is the great difference we find between household survey and focus group interview results. In the latter, people are often paid; they assume that the reason they are being paid or brought together is to be critical; they reinforce each others' criticisms if only because they are a group and out-number the discussion leader; they free associate, ignore logical order, and feed each other ideas. There are a number of reasons why the image of the Oakland 1978 Census obtained from a survey the Bureau conducted and from focus group interviewing conducted for the Bureau are unrecognizably different. It is hardly exaggeration to suggest that, if replies to attitudinal questions were related to behavior, almost everyone interviewed in Oakland would have mailed back the form and hardly any of the participants in the focus group interviews would have done so. The actual mail back rate was about 55%. Selectivity of participants, bias of auspice, as well as question wording can, of course, account for some of the wide discrepancy but differences in setting seem paramount to an observer of both.

Respondent beliefs about surveys also deserve attention as a source of response error. Answers to these questions may provide explanation of results sought in exploration of question wording: How prevalent are the ideas that surveys are indistinguishable from inquiries like credit card, job, and housing applications or loyalty investigations which lead to decisions about the individual who is reporting? How common is the perception that the interviewer is a police interrogator or teacher administering a test? Do proposals like those of giving instructions and requesting commitment fit or are they incongruous with preconceptions about the interview? How can a person believe he is anonymous and yet feel his replies are valuable? How can a person object to being depersonalized or "reduced to numbers" at the same time he objects to having his privacy invaded? How are promises of confidentiality understood and what are the bases for any trust in or skepticism about such promises?

Thus far, I have spoken only from my own perspective. I gave Mr. Kalton's paper to Catherine Baca, an anthropologist on our staff, and her ideas for an anthropological study of the interview as a social phenomenon constitute a step beyond the mere statement of a problem. To do her proposal justice would have taken all of my allotted discussion time and I wanted to save at least a minute for the psychologists on the staff who also read Mr. Kalton's paper and believe that an area of their discipline may have been overlooked in efforts to understand, predict, and control question effects. Here I quote Jeffrey Moore who suggests that, in some circumstances, cognitive structures and memory processes may play a larger role than do the social interactive aspects of the interview in producing biased or unreliable responses. The manner in which experiences are perceived, stored in, and retrieved from memory may affect survey responses. There may be differences in the internal complexity of the memory store for different events, in the degree of interrelatedness or overlap with other memory stores, or in the frequency with which events are retrieved. These differences could account for differing susceptibility to question effects.

The Census Bureau's Center for Social Science Research is contributing to a workshop sponsored by the Crime Survey Research Consortium. Headed by Albert Biderman, the Consortium is responsible for the research and development aspects of the National Crime Survey and its workshop "Applying Cognitive Psychology to Recall Problems of the Survey" will bring together academicians and government researchers including psychologists, survey research practitioners and subject matter experts. I hope that, by the next ASA and APA meetings--because the exchange should be mutually stimulating--there will be results worth reporting about the workshop and outgrowths of it.

---

*Space does not permit reproduction of handouts. Interested readers are referred to page 401 of the Journal of Marketing Research August 1979 Volume XVI for the illustration and page 404 for the data.