J. R. Statist. Soc. A (1982),
145, *Part* 1, *pp.* 42–73

The Effect of the Question on Survey Responses: A Review†

By GRAHAM KALTON and HOWARD SCHUMAN,

Survey Research Center, University of Michigan, USA

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the SOCIAL STATISTICS SECTION on Wednesday, September 30th, 1981, Professor G. HOINVILLE in the Chair]

SUMMARY

The paper reviews the effects of the precise wording, form and placement of questions in a survey questionnaire on the responses obtained. Topics discussed include: randomized response; respondent instructions; feedback and commitment; question length; the treatment of "don't knows"; open and closed questions; the use of balanced questions; acquiescence; the offer of a middle alternative; the order of alternatives; and question order and context effects.

Keywords: FACTUAL QUESTIONS; OPINION QUESTIONS; MEMORY ERRORS; SOCIAL DESIRABILITY BIAS; QUESTION WORDING; QUESTION FORM; QUESTION CONTEXT EFFECT

1. INTRODUCTION

THE survey literature abounds with examples demonstrating that survey responses may be sensitive to the precise wording, format and placement of the questions asked. A useful start to examining these effects is to classify questions according to the type of information sought.

A widely-used distinction is that between factual and opinion questions. Questions like "What was your regular hourly rate of pay on this job as of September 30?" clearly fall in the former category, while questions like "As you know, many older people share a home with their grown children. Do you think this is generally a good idea or a bad idea?" clearly fall in the latter. However, not all survey questions can be classified as either factual or opinion ones: other types of question include questions testing respondents' knowledge, questions asking for reasons, hypothetical questions and preference questions.

One further type of question, widely used in survey practice, deserves special comment. These questions, which have a factual component overlaid with an evaluation, may be termed judgement or perceptual questions. Examples are: "Do you have street (highway) noise in this neighbourhood?" and "Would you say your health now is excellent, good, fair or poor?" In many cases the intent of such questions is to obtain factual information, but the approach adopted seeks respondents' evaluations of the facts rather than their measurement according to objective criteria. The use of perceptual questions for this purpose probably results from the questionnaire designer's decision that he could not ask sufficient questions or take the measurements necessary to determine the information objectively; hence he has respondents make the assessments for him. The often low levels of correlation found between perceptions and facts make this use of perceptual questions, although widespread, a dubious one. A different use of perceptual questions is indeed to obtain respondents' perceptions of their situations; in this case the questions are similar to opinion questions.

For present purposes, it will be sufficient to divide questions into factual and non-factual ones (including as factual questions those perceptual questions seeking to ascertain factual information). An important difference between these two types of question is that with factual questions there are individual true values for the information sought which can—at least in

† This paper is a slightly revised version of a paper presented at the American Statistical Association meetings, Houston, August 1980 (Kalton and Schuman, 1980, with discussion by Rothwell and Turner).

 0035–9238/82/145042$2.00

theory—be determined from some source other than respondents' reports, whereas with other questions this does not apply. While it is true that the responses to some factual questions cannot be validated against external sources—for instance, reports of past unrecorded private behaviour—the difference holds in general. As a consequence, validity studies are often conducted to examine how successful factual questions are in obtaining respondents' individual true values, whereas with non-factual questions such studies are not possible.

Although numerous validity studies of responses to factual questions have been carried out in many subject areas, the majority of them have examined only the level of accuracy achieved by a given questioning procedure; they have not compared alternative procedures, as required for making an assessment of how aspects of a question may affect the accuracy of the responses obtained. Many of the comparative studies that have been conducted have avoided the need for data from an external validating source by making an assumption about the general direction of the response errors to be encountered, the assumption adopted being based on evidence from other validity studies. Thus, for instance, it is often assumed from past evidence that certain events such as purchases made or illnesses experienced in a given period will be underreported. Given this assumption, the best question form is then taken to be the one that produces the highest frequencies for the events. On the other hand, a socially desirable activity may be assumed to be generally overstated, in which case the best question form is the one that gives the lowest reported frequency for it. While the difficulties of obtaining validity data make this approach attractive, it does depend critically on the validity of the assumption about the direction of response errors.

With non-factual questions, validation is even more difficult and less certain. The accuracy of responses can often be examined only by means of construct validity, that is by determining whether the relationships of the responses with other variables conform to those predicted by theory. At the current stage of theory development in the social sciences, a failure of data to fit a theory is usually as likely to cast doubt on the theory as on the measuring instruments. Then, even if the observed relationships coincide with their theoretical predictions, this agreement is not a clear confirmation that the responses are valid; it may, for instance, instead be an artifact of the set of measuring instruments employed—a "methods effect".

In view of the difficulties of validating responses to non-factual questions, research on questioning effects with such questions has relied mainly on split-ballot experiments, in which alternative forms of a question are administered to comparable samples of respondents, with the responses to the different question forms being compared for consistency. This concern with consistency rather than validity means that the research usually fails to identify which is the best question form. It serves only to warn the researcher of the sensitivity of responses to question form if the responses differ markedly, or to increase his feelings of security in the results if they do not differ.

A second difference between factual and non-factual questions involves the features studied as possible influences on the responses obtained. Although many of the features potentially apply with both types of question, researchers have been more concerned about some of them with factual questions and others with non-factual ones. Thus research on factual questions has focused on problems of definition, comprehension, memory and social desirability response bias, while that on non-factual questions has concentrated on various question form effects, such as issues of balance, the offer of middle alternatives, and the order of presentation of alternatives. The features primarily studied in relation to factual questions are reviewed in the next section, and those studied in relation to non-factual questions are taken up in the following one. The effects of question order and context, which have received attention in relation to both types of question, are discussed in Section 4.

Before embarking on the discussion of question effects, we should note that we are primarily concerned with surveys in which interviewers ask the questions and record the answers. Responses to self-completion questionnaires may be affected by a number of other features, such as the physical location of a question on the questionnaire, the placement of

instructions, the general layout, and the colours of print for questions and instructions. Reports of experiments on the effects of some of these features are given by Rothwell and Rustemeyer (1979) for the US Census of Population and Housing, and by Forsythe and Wilhite (1972) for the US Census of Agriculture.

2. Question Effects with Factual Questions

The starting point in constructing a factual question is a precise definition of the fact to be collected. It has been shown on many occasions that apparently marginal changes in definition can have profound effects on survey results. Definitions of unemployment and labour force raise a host of issues (e.g. Bancroft and Welch, 1946; Jaffe and Stewart, 1955), but even ostensibly simple facts like the number of rooms occupied by a household pose a range of definitional problems (for instance: Is a kitchen to be included if only used for cooking? Are bathrooms, toilets, closets, landings, halls to be included? Is a room partitioned by curtains or portable screens one or two rooms?).

Once the fact has been defined, the request for it has to be communicated to the respondent. A number of difficulties can arise in this process. In the first place, the need for a precise definition can lead to an unwieldy question which the respondent cannot—or will not make the effort to—absorb. In the quality check on the 1966 Sample Census of England and Wales, Gray and Gee (1966) found that 1 in 6 householders reported an inaccurate number of rooms in the household, which they ascribe mainly to the fact that householders know how many rooms they have according to their own definitions, and they therefore ignored the detailed census definition. To avoid this problem, some looseness is often accepted in survey questions (especially in perceptual questions), but this may well lead to inconsistent interpretations between respondents.

Another aspect of the communication process is to ensure that the respondent fully understands what he is being asked and what is an appropriate answer. At one level he needs to understand the concepts and frames of reference implied by the question (Cannell and Kahn, 1968). At a more basic level he needs to comprehend the question itself. Methodological research by Belson and Speak found that even some simple questions on television viewing were often not perceived as intended by a sizeable proportion of respondents. For instance, the questions "What proportion of your evening viewing time do you spend watching news programmes?" and "When the advertisements come on between two programmes on a weekday evening, do you usually watch them?" were misinterpreted by almost everybody who answered them. With the first question, very few respondents knew what "proportion" meant, and only 1 of the 246 respondents knew how to work it out. With the second, "weekday" was often misinterpreted as either "any day of the week" or "any day except Sunday" (Speak, 1967; Belson, 1968).

To give a correct answer to a factual question, a respondent needs to have the necessary information accessible. Accessibility first requires that he has had the information at some time and has understood it. Then, if the question asks about the past, he needs to be able to retrieve it from his memory. Ease of recall depends mainly on the length of the recall period and the salience to the respondent of the information being recalled (see, for example, Cannell and Kahn, 1968). His success in recalling the information depends on the ease of recall and the effort he is persuaded to make. Many survey questions ask about events occurring in a specified reference period (e.g. seeing a doctor in the last year), in which case the respondent also has to be able to place the events in time. A well-known placement distortion is the telescoping error of remembering an event as having occurred more recently than in fact is the case (see, for example, Sudman and Bradburn, 1973, 1974).

The effects of recall loss and telescoping work in opposite directions, recall loss causing underreporting and telescoping causing overreporting. The extent of these two sources of error depends on the length of the reference period: the longer the period, the greater is the recall loss, but the smaller is the telescoping effect. Thus, for short reference periods, the telescoping

effect may outweigh the recall loss, while for long periods the reverse will apply; in between there will be a length of reference period at which the two effects counterbalance each other (Sudman and Bradburn, 1973). The meaning of "short" and "long" reference periods varies with the event under investigation, depending on the event's salience. The choice of an appropriate reference period needs to take into account the telescoping and recall loss effects, as well as the fact that longer periods provide estimates with smaller sampling errors. This choice has been examined in a number of different subject areas (see, for example, National Center for Health Statistics, 1972; Sudman, 1980).

A technique which aims at eliminating telescoping errors by repeated interviewing is known as bounded recall (Neter and Waksberg, 1965). Respondents are interviewed at the beginning and end of the reference period. The first interview serves to identify events which occurred prior to the start of the period so that they can be discounted if they are then reported again at the second interview.

Three procedures are widely used in survey practice to attempt to minimize or avoid memory errors—the use of records, aided recall techniques and diaries—and each procedure has its own sizeable literature. Where records are available, say from bills or cheque book records, their use can reduce both recall loss and telescoping effects, as well as provide accurate details of the events. Aided recall techniques aim to reduce recall loss by providing the respondent with memory cues; these techniques are widely used in media research, where the respondent would be provided with, say, a list of newspapers or yesterday's television programmes from which he chooses the ones he looked at. In their summary of the effects of aided recall techniques, Sudman and Bradburn (1974) conclude that they do increase reported activity, but point out that this may at least in part represent an increase in telescoping errors. Where the events to be reported are numerous and relatively insignificant, there may be no way to help respondents remember them with sufficient accuracy. In such cases, as with household expenditures, food consumption and trips outside the house, memory problems may be avoided by having respondents complete diaries of the events as they take place. Diaries, however, have their disadvantages: they are expensive, it is harder to gain respondents' cooperation, the diary keeping may affect behaviour, it may be incomplete, and its quality usually deteriorates over time.

Another well-documented source of invalidity in responses to factual (and other) questions is a social desirability bias: respondents distort their answers towards ones they consider more favourable to them. Thus, for instance, it has been well established that a higher proportion of survey respondents report that they voted in an election than the voting returns indicate (for instance, Parry and Crossley, 1950; Traugott and Katosh, 1979). If an event is seen as embarrassing, sensitive or threatening, the respondent may repress its report, or he may distort his answer to one he considers more socially acceptable. There are a number of well-known techniques for eliciting sensitive information, including making responses more private by using a numbered card (often used for income) or a sealed ballot, and attempting to desensitize a particular response by making it appear to be a common or acceptable one. Barton (1958) has provided an amusing summary of these techniques.

A more recent development for asking sensitive questions is the randomized response technique, in which the respondent chooses which of two (or more) questions he answers by a random device; he answers the chosen question without the interviewer being aware which question is being answered. In this way the respondent's privacy is protected, and in consequence it is hoped that he gives a more truthful response. Since Warner (1965) introduced the technique, many articles have appeared developing it, extending its potential range of application, and examining its statistical properties. The main focus of this work has been, however, on theoretical issues, and comparatively little attention has been given to its practical utility. One common, but not obvious, finding from studies in which it has been applied is that it has generally been well received by respondents. In a small-scale experimental study by Locander *et al.* (1976), for instance, only 1 in 20 respondents said it was confusing, silly or

unnecessary; the interviewers thought that about 7 out of 8 understood the use of the random response box, and that a similar proportion accepted the explanation of the box and believed that their answers really were private.

Several experimental studies have obtained higher rates of reports of sensitive information from randomized response techniques than from traditional questioning—for instance, Abernathy *et al.* (1970) and Shimizu and Bonham (1978) on abortion rates, Goodstadt and Gruson (1975) on school students' drug use, and Madigan *et al.* (1976) on death reports in a province in the Philippines. In their validity study comparing the accuracy of reporting of personal interviews, telephone interviews, self-administered questionnaires and a randomized response technique for five issues of varying degrees of threat, Locander *et al.* (1976) found that the randomized response technique was most effective in reducing underreporting of the socially undesirable acts, being declared bankrupt and being charged with drunken driving. However, the use of the technique still led to a substantial amount of underreporting of drunken driving (35 per cent for the randomized response technique compared with an average of 48 per cent for the other techniques).

Any gain in bias reduction with the randomized response technique has to be set against a sizeable increase in sampling error. The use of the technique also hampers analyses of the relationships between the responses to the threatening question and other variables. For these reasons the technique seems useful only for special, very sensitive, issues for which overall estimates are required. It does not appear to provide a widely applicable approach for dealing with sensitive survey questions.

In a programme of research extending over the last two decades, Cannell and his colleagues at the Survey Research Center have developed a variety of new approaches to deal with both problems of memory errors and problems of sensitive questions. Their research has been directed mainly at improving the quality of reporting of health events, but it has wide potential application. In their early work they identified the need to have respondents understand adequately the task expected of them and have them make the necessary effort to retrieve and organize the information into a suitable reporting form. They then developed techniques aimed at meeting these objectives.

One technique stemmed from research on speech behaviour in employment interviews, where an increase in an interviewer's speech duration has been found to result in an increase in the respondent's speech duration. This finding raised the possibility that, counter to accepted survey dogma, longer questions may in some circumstances yield longer, and hence more valid, answers. To test this hypothesis, experiments were conducted to compare responses to a short question with those to a longer question formed by adding redundancies which did not affect the content. One such experiment compared responses to the following two questions:
Short question: "What health problems have you had in the past year?"
Long question: "The next question asks about health problems during the last year. This is something we ask everyone in the survey. What health problems have you had in the past year?"

The experiments did not find that the longer questions produced significantly longer responses, but they did yield a greater number of relevant health events being reported. Since a questionnaire made up of only long questions would be cumbersome, experiments involving a mixture of long and short questions were also carried out: this mixture was found to yield more reporting of health events to both the long and short questions (Cannell *et al.*, 1977; Cannell *et al.*, 1981).

The researchers postulate three reasons for this effect: that by essentially stating the question twice the respondent's understanding is increased; that the time lag between the first statement of the question at the start and the need to answer it at the end allows the respondent the opportunity to martial his thoughts; and that the respondent interprets the length of the question as an indication of its importance, thus encouraging him to give it greater attention.

It should be observed that the longer questions in these experiments were in no way more

complex than the short ones. The usual advice "keep questions short" is probably an inaccurate way of saying "keep questions simple"; in practice the difficulties from long questions probably derive from their complexity rather than their length *per se*.

Other techniques developed by Cannell and his colleagues to improve survey reporting include the use of respondent instructions, the use of feedback and the securing of respondent commitment.

The purpose of including respondent instructions in the questionnaire is to advise the respondent on how he should perform his task. Cannell *et al.* (1981) have experimented with providing general instructions at the start of the interview to ask the respondent to think carefully, search his memory, take his time and check records, and to tell him that accurate and complete answers are wanted. In addition, respondents can be given specific instructions on how to answer individual questions; these specific instructions have the added benefit of lengthening the questions, thus securing the advantages associated with longer questions.

The purpose of feedback is to inform the respondent on how well he is performing. The interviewers are provided with a selection of feed-back statements from which to choose, their choice being governed by the respondent's performance. Examples of positive and negative feed-back statements are "Thanks, we appreciate your frankness" and "Uh-huh. We are interested in details like these" on the one hand and "You answered that quickly" and "Sometimes it's easy to forget all the things you felt or noticed here. Could you think about it again?" on the other.

The theory behind the commitment technique is that if a respondent can be persuaded to enter into an agreement to respond conscientiously he will feel bound by the terms of the agreement. The technique can be applied with personal interviewing by asking respondents to sign an agreement promising to do their best to give accurate and complete answers. In practice Cannell and his colleagues have found that only about 5 per cent of respondents refuse to co-operate. With telephone interviewing, respondents may be asked to make a verbal commitment to respond accurately and completely: a study applying this procedure encountered no problems in securing respondents' co-operation.

The evidence from the various experiments conducted to examine the utility of these techniques suggests that each of them leads to an improvement in reporting, with a combination of all three giving the best results. A concern that high-education respondents might react negatively did not materialize. In a health study, the use of the three techniques together increased the average number of items supplied in answers to open questions by about one-fifth; substantially improved the precision of dates reported for doctor visits, medical events and activity curtailment; increased by about three-fold the checking of data from outside sources; and secured almost a third more reports of symptoms and conditions for the pelvic region (considered to be potentially embarrassing personal information). In a small-scale study of media use, comparing an experimental group interviewed with a combination of all three techniques with a control group interviewed with none of them, the experimental group reported a greater amount for activities likely to be underreported and a lesser amount for those likely to be overreported. Thus 86 per cent of the experimental group reported watching TV on the previous day compared with 66 per cent of the control group; the experimental group listened to the radio yesterday for $2\frac{1}{2}$ hours on average, compared with an average of $1\frac{1}{2}$ hours for the control group; 38 per cent of the experimental group reported reading the editorial page on the previous day compared with 55 per cent of the control group; and the experimental group reported an average of 2·9 books read in the last 3 months compared with 5·3 for the control group.

These experimental results suggest that the techniques hold considerable promise for improving survey reporting. However, at this stage of their development, it seems premature to advocate their general use in routine surveys. They involve significant alterations to questionnaires, interviewers need to be trained in their use, and interviews take longer to complete. Before they are widely adopted, further research is called for, to attempt to replicate

the findings across a variety of survey types in different survey environments, to identify restrictions on their range of application, and to seek improvements in them. An important component of this research should be experiments which incorporate an external source of data against which the survey reports can be validated. The availability of the validity data not only avoids the need for assumptions about the directions of reporting errors, but also means that the extent of any improvements can be assessed against the amount of reporting error still remaining.

3. QUESTION EFFECTS WITH NON-FACTUAL QUESTIONS

As with a factual question, the initial stage in the formation of a non-factual question is the conceptualization of the construct to be measured. By its nature a non-factual construct is usually more abstract than a factual one, and hence more difficult to define precisely in theoretical terms; it is also more difficult to operationalize. Since often no single item can represent the essence of a construct, multiple indicators are needed; each indicator overlaps the construct, with the set of indicators being chosen so that the overlaps between them tap the construct. In attitude measurement, the conceptualization and operationalization of an attitude dimension are often closely interwoven: the initial conceptualization determines the choice of items used to operationalize the dimension, but then the items themselves serve to refine the dimension's definition. With the infinity of attitude dimensions that could be measured merging imperceptibly from one to another, the precise definition of the one being measured must depend ultimately on the set of items used to operationalize it.

We have noted that changes made in factual questions to accommodate marginal changes in definition can frequently have substantial effects on the responses obtained. Survey analysts deal with this instability by carefully specifying exactly what has been measured. In general they are not too disturbed by a variation in the results obtained from different questions because, as a result of the relatively precise definitions, they can usually account for the variation as the net effect of the various definitional changes involved.

With non-factual questions, similar effects occur with marginal question changes: two apparently closely similar questions may yield markedly discrepant results. Sometimes a detailed examination of the questions can identify subtle content differences which provide a convincing explanation for the discrepancies. There are, however, also some cases where no obvious content change can be detected, and yet the results still differ substantially. In particular, certain variations in question form have often been shown to have sizeable effects on results.

In view of this sensitivity of responses of non-factual questions to minor changes and the somewhat imprecise definitions of the concepts under study, experienced survey analysts are wary of taking the marginal distributions of answers to such questions too seriously. Instead they concentrate their attention on some form of correlational analysis, for instance contrasting the response distributions of different subclasses of the sample. This form of analysis is justified by the assumption that, even though different question forms may yield markedly different distributions, the question form effect cancels out in the contrast. Schuman and Presser (1977) have termed this assumption that of "form resistant correlation". Evidence is given below to show that, while this assumption is often reasonable, it does not always hold.

In constructing a non-factual question, the questionnaire designer has to make a number of decisions on the form of the question to be asked. We will briefly review a selection of these decisions, mainly with respect to opinion questions, to see how they might influence the results obtained.

(a) *Treatment of "Don't knows"*

With a factual question a response of "Don't know" (DK) represents a failure to obtain a required item of information; there is an answer to the question, but the respondent cannot

provide it. With opinion questions, however, DK has a different interpretation, for respondents may truly have no opinion on the issue under study.

The standard way of allowing for DK's with opinion questions is the same as that used with factual questions; the option to answer DK is not explicitly included in the question, and interviewers are instructed to use the DK response category only when the respondent offers it spontaneously. The danger with this procedure is that some respondents may feel pressured to give a specific answer even though DK is their proper response. This danger exists for both factual and opinion questions, but it is probably greater for the latter.

Two examples given by Schuman and Presser (1980) illustrate that many respondents will indeed choose one of the alternatives offered for an opinion question even though they do not know about the issue involved. In both examples, respondents were asked for their views about proposed legislation which few, if any, would be aware of, and yet 30 per cent expressed opinions. Bishop *et al.* (1980b) report similar findings about a wholly fictional issue.

As a way of screening out respondents without opinions, some type of filtering may be used. One possibility is to include an explicit "no opinion" option or filter in the response categories offered to respondents—a "quasi-filter"; in the Schuman and Presser experiment, this offer reduced the proportion of respondents expressing opinions on the two laws to 10 per cent or less. A more forceful possibility is a preliminary filter question "Do you have an opinion on . . .?"—a "full filter".

Schuman and Presser (1978) carried out several experiments to examine the effects of filtering. They found that the use of the full filter typically increased the percentage of DK's over those obtained from the standard form by around 20–25 per cent. Bishop *et al.* (1980a) also found in their experiments that the increases were generally in the range 20–25 per cent, but they report a much smaller increase for a very familiar topic and a much larger one for an unfamiliar topic.

In Schuman and Presser's experiments the effect of the variation in question form on substantive results was somewhat unexpected. In the first place, once the DK's had been eliminated (as would usually be done in analysis), the marginal distributions of responses turned out in most cases not to be significantly affected by question form; also the relations between the opinion responses and standard background variables were little affected. However, the associations between the opinion responses themselves did differ significantly in certain cases between question forms: in one case the association was stronger with the filtered form, in another it was weaker.

(b) *Open or closed questions*

When asked a survey question respondents may either be supplied with a list of alternative responses from which to choose or they may be left to make up their own responses. The major advantages of the former type of question—termed variously a closed, fixed-choice or precoded question—are standardization of response and economy of processing. Its major disadvantages, and hence arguments in favour of open questions, are that the alternatives imposed by the closed form may not be appropriate for these respondents, and that the alternatives offered may influence the responses selected.

The main context in which open questions are used extensively is when the potential responses are both nominal in nature and sizeable in number. These conditions occur often with motivation questions, asking for the principal or all reasons for an occurrence, and with questions asking for the choice of the most, or several most, important factors involved in an issue. In such cases the questionnaire designer faces a real choice between open and closed questions.

As part of their research on question form effects, Schuman and Presser (1979) carried out several experiments on open and closed questions, using items chosen for their utility in one form or the other in a major past survey. In all the experiments important differences occurred

between the response distributions to the open and closed forms. The two versions of a question on work values in the first experiment were:

Open: "The next question is on the subject of work. People look for different things in a job. What would you *most* prefer in a job?"

Closed: "The next question is on the subject of work. Would you please look at this card and tell me which thing on the list you *most* prefer in a job? 1. High income (12·4 per cent); 2. No danger of being fired (7·2 per cent); 3. Working hours are short, lots of free time (3·0 per cent); 4. Chances for advancement (17·2 per cent); 5. The work is important and gives a feeling of accomplishment (59·1 per cent)." (Figures in brackets are percentages choosing the alternatives.)

While all but 1 per cent of responses to the closed question fell into one of the five precoded categories, nearly 60 per cent of those to the open question fell outside these categories (important additional codes developed were: pleasant or enjoyable work, 15·4 per cent; work conditions, 14·9 per cent; and satisfaction/liking the job, 17·0 per cent). The open question responses gave rise to five coding categories comparable to those listed above for the closed question form. For the first two, the proportions of respondents choosing the code were almost identical for the two question forms, for the third the proportion was somewhat lower with the open form, while for the last two it was much lower with the open form. The equivalent code for the "Chance for advancement" code was "Opportunity for promotion" with the open form: this code was used for only 1·8 per cent of responses as compared with the 17·2 per cent use of the "Advancement" code. The code corresponding to "Accomplishment" was called "Stimulating work"; it was used for only 21·3 per cent of responses as compared with 59·1 per cent for "Accomplishment".

A possible explanation for these substantial differences is that they were not caused by the change in question form *per se*, but were rather the result of the use of unsuitable response categories. Schuman and Presser therefore conducted two more experiments using a revised set of response categories constructed from the codes developed in the first experiment for the open question. This revised set aimed to represent more adequately the work values that respondents offered spontaneously and also to retain the theoretical goals behind the question. Even with these revised codes, however, the response categories of the closed form covered only 58 per cent of all open responses, and there remained differences between the proportions choosing the five common categories on the two forms of the question.

Besides the differences in marginal distributions, Schuman and Presser also found that the question form sometimes affected relationships of the responses to background variables. In the first experiment, the results from the closed form indicated that men were more likely than women to value "Pay" and "Advancement", but in the open form no such associations appeared. In the second two experiments, there was a substantial downward trend in the proportion choosing the "Security" category with increasing education for the closed question form, but there was no clear relationship between this category and education for the open form. The authors present indirect evidence suggesting that when open and properly constructed closed forms of questions yield different responses, the responses to the closed questions are sometimes more valid in their classification of respondents and in describing relationships of the responses with other variables.

(c) *The use of balance*

In asking for respondents' opinions on an issue, the questionnaire designer often has a choice of the extent to which he presents the alternative contrary opinion. At one extreme, the question may be expressed in an unbalanced form simply as "Do you favour *X*?", with the contrary opinion left entirely unmentioned, while at the other extreme a substantive alternative may be explicitly stated, as in the question "Do you favour *X* or *Y*?". An intermediate position is to use a token alternative, to draw attention to the existence of the

alternative opinion without specifying exactly what it is; questions like "Do you favour or oppose *X*?" are of this type.

A number of split-ballot experiments have been conducted to compare the results obtained using the unbalanced form of the question and those using the form with the token alternative. Perhaps not surprisingly these experiments have generally found only small differences.

On the other hand, large differences have often—but not always—been found between the responses given to questions asked with and without a substantive alternative. In a number of experiments it can be argued that the inclusion of the alternative has introduced new issues, effectively modifying the choice the respondent is being asked to make (Hedges, 1979). Even so, the survey analyst needs to be aware of this effect, since it means that two questions apparently tapping closely comparable issues can yield very divergent results.

(d) *Acquiescence*

A widely used method of attitude measurement is to present respondents with a set of opinion statements with which they are asked to agree or disagree. An issue that arises with this procedure is that respondents might tend, regardless of content, to give "agree" rather than "disagree" responses. This tendency, which has received a good deal of attention in the psychological literature, is often known as acquiescence or agreeing response set (bias). The dominant view now appears to be that the effect is of little importance in psychological testing, but Campbell *et al.* (1960) long ago provided evidence to suggest that this conclusion may not hold for the social survey situation.

In one of their experiments on this issue Schuman and Presser (1981) compared the responses to two statements with which respondents were asked either to agree or to disagree (and also a forced choice version of the question). The two statements were constructed to be exact opposites of each other. The detailed analysis to investigate the presence of an agreeing response bias cannot be adequately summarized here, but we will report one simple result to indicate the magnitude of the effect found.

The two agree/disagree statements were:

A. "*Individuals* are more to blame than *social conditions* for crime and lawlessness in this country."

B. "*Social conditions* are more to blame than *individuals* for crime and lawlessness in this country."

Without a question form effect, the proportion of respondents answering "agree" to *A* should be the same as the proportion answering "disagree" to *B*, and vice versa. In the event, however, 59·6 per cent agreed with *A* and only 43·2 per cent disagreed with *B*, a highly significant difference. Schuman and Presser also found that the variation affected associations between the responses and education and other important variables.

(e) *Middle alternatives*

When respondents are asked their views on an issue, often some may want to choose a middle or neutral response. The problem facing the questionnaire designer is how to allow for this response. Should a middle alternative be explicitly offered? Should a neutral response be accepted only if offered spontaneously? Or should it be actively discouraged?

As might be expected, the explicit offer of a middle alternative often substantially increases the proportion of respondents stating a neutral view. In a series of experiments conducted by Kalton *et al.* (1980), the increases were between 15 and 49 per cent; in a series reported by Presser and Schuman (1980) the increases were between 10 and 20 per cent.

Presser and Schuman observe that in their studies and earlier ones involving three point scales (pro, neutral and anti) the increase in support for the neutral view with the offered question form came proportionately from the polar positions, so that the balance between pro's and anti's was not affected by the variation in alternatives offered. This comforting

finding failed to hold, however, in two of the three experiments with three-point scales reported by Kalton *et al.*

There is little evidence that this question form variation affects associations between opinion responses and other variables. In view of the substantial impact of the question form variation on marginal distributions, however, it seems dangerous to place uncritical reliance on the "form resistant correlation" assumption.

(f) *Order of alternatives*

The responses to closed questions may be affected by the order in which the alternatives are presented. In discussing this order effect, two modes of presentation may need to be distinguished: the alternatives can be presented in written form, as with self-completion questionnaires or when flashcards are used; or they can be presented orally, with the interviewer reading them to respondents, sometimes as a running prompt. When they are presented in written form, there appears to be a slight tendency for the first alternative to be favoured (e.g. Belson, 1966; Quinn and Belson, 1969). When they are presented orally, Rugg and Cantril (1944) provide examples where the last-mentioned alternative is favoured, but Payne also gives several examples where the order effect is negligible. Kalton *et al.* (1978) report the results of experiments on varying the order of orally presented alternatives with four simple questions. In all cases, the evidence suggested that, if anything, the first-mentioned alternative was favoured; the effects were, however, very small (around a 2 per cent increase), and only on the border of statistical significance.

4. General Question Effects

The preceding discussion has been divided into two parts, questioning issues relating to factual questions and those relating to non-factual (opinion) questions. This arbitrary division was made for convenience of exposition to reflect the differences in emphasis of question wording and format research between the two types of question. However, it should not be taken to imply that the effects noted for one type of question do not apply to the other. Thus, for instance, issues of sensitivity can clearly arise with opinion statements, as also would issues of memory if the survey was concerned about changes of opinion (an extremely difficult matter on which to collect accurate information by retrospective questioning). Equally, while many of the question form variations discussed above for non-factual questions are not applicable for factual questions, the latter may also be affected by variation in question form. Locander and Burton (1976), for instance, show how four versions of a question asking for family income, all designed for use with telephone interviewing, yielded markedly different income distributions. All the questions used an unfolding technique for presenting the set of response categories as a sequence of binary choices, but they employed different forms of the technique. Forms 1 and 4, for example, both asked whether the family income was "more that X" for X = \$5000, \$7500, \$10 000, \$15 000, \$20 000 and \$25 000; form 1 started with \$5000 and took increasing values of X until a "no" answer was given, while form 4 started with \$25 000 and took decreasing values until a "yes" answer was given. With form 1 37·5 per cent of respondents reported family incomes of \$15 000 or more; with form 4 the corresponding percentage was 63·7 per cent.

A final, important, questioning effect to be discussed concerns the presence of other questions in the questionnaire, and the position of those questions in relation to the question under study. Question order and context effects may occur with both factual and non-factual questions, but they appear to operate in different ways.

A sizeable number of studies have been carried out to examine the effect of question order on responses to opinion questions. On many occasions no order effect has been discovered, even for questions closely related in subject matter. However, one type of question order effect has been found in two cases and seems worth further exploration. This effect occurs when one of the questions is a general one on an issue and the other is more specific on the same issue. Schuman *et al.* (1981) with two opinion questions on abortion, and Kalton *et al.* (1978) with

two questions on driving standards, both found that the distributions of answers to the more specific questions were the same whether the specific question was asked before or after the general question, but that the distributions of answers to the general questions differed according to the questions' position. (However, Kalton *et al.* also report another such experiment with a contrary finding.) In the Kalton *et al.* experiment, respondents were asked about driving standards generally and about driving standards among younger drivers. When the general question was asked first, 34 per cent of respondents said that general driving standards were lower than they used to be; when that question followed the more specific question about younger drivers, the corresponding percentage fell by 7 per cent to 27 per cent. Further analysis showed that the question order affected only respondents aged 45 or older, where the difference in the percentages was 12 per cent. No definitive reason for this effect has been established, but it may possibly be explained as a subtraction effect: after answering the specific question, some respondents assume that the general question excludes the specific part (e.g. in the driving example, they assume that the general question excludes consideration of the driving standards of younger drivers).

With factual questions, one situation where other questions on a questionnaire may influence the answers to a particular question arises when respondents are asked to respond to a long list of similar items, as for instance in readership surveys where they are taken through a list of newspapers and periodicals to find out which ones they have looked at. Here levels of reporting sometimes tend to be lower when items are placed later in the list. For instance, in studying readership reporting in the UK National Readership Surveys, Belson (1962) conducted an experiment in which he varied the relative position of the different types of periodicals between different parts of the sample. The weekly publications were most affected by the presentation order: when they appeared last their reported level of readership was only three-quarters of what it was when they appeared first.

Another source of evidence on the disturbing influence of other questions comes from an examination by Gibson *et al.* (1978) of the effects of the inclusion of supplements on the results for core items in the National Crime Survey (NCS), Current Population Survey and Health Interview Survey.

In the NCS Cities Sample a lengthy series of attitude questions about topics such as neighbourhood safety, opinions regarding local police, crime trends and news coverage of crime was asked of a random half of the sample of adults in addition to the core NCS questions on crime victimization. Since it was thought that the responses to the attitude questions might be affected by the victimization questions if they were asked after the core items, the attitude questions were asked first. The effect of the prior inclusion of the attitude questions was, however, to substantially and significantly increase the reported victimization rates: on average the rate for personal crimes was around 20 per cent greater and that for property crimes was around 13 per cent greater for the half sample that answered the attitude supplement than for the half sample that did not. Possible explanations for this effect are that the attitude questions served to stimulate respondents' awareness or memory regarding victimization experiences, that they increased respondents' desire to produce what they perceived to be the desired answers—victimization experiences—or that a combination of both these causes operated.

From a further analysis of the NCS Cities Sample, Cowan *et al.* (1978) deduce that the effect of administering the attitude supplement was to increase reporting of the less serious victimizations (such as simple assault, those not reported to the police and those involving a loss of under $50) and to increase reporting among population subgroups experiencing high victimization rates (younger persons, males). They also found that the higher rates were spread throughout the 12-month reference period with no discernible pattern, a factor which argues against an increased telescoping effect stimulated by the attitude supplement. They conclude that the effect of the supplement is to produce better reporting in the reference period, but they suggest that it may be an oversimplification to attribute this effect to memory stimulation.

The findings of the substantial effects that the inclusion of supplements can have on the responses to core items raise a major concern for the comparability of results across surveys. Survey analysts are properly cautious in their interpretations of differences in results between surveys when there are even slight changes in the questions being compared. These findings suggest that they also need to be concerned about differences in the rest of the questionnaires. This conclusion has serious consequences for the replication of survey results because, while it is often fairly easy to replicate individual questions or small sets of them for purposes of comparability, it is extremely difficult to replicate an entire questionnaire. In view of the importance of replication and measures of change in the analysis of survey data, further research in this area is surely called for.

5. CONCLUDING REMARKS

The general conclusion from this review must be that survey questioning is not a precision tool. The survey literature contains ample evidence to indicate that serious response errors can, and do, occur with factual questions, and many experiments have shown that the responses to opinion questions can sometimes be substantially affected by apparently insignificant variations in the question asked. This conclusion on the limitations of current survey questioning procedures may be unexceptionable to survey methodologists, but it is not sufficiently recognized by the wide variety of people who carry out surveys and use survey data for a range of different purposes.

In view of this state of affairs, experienced survey practitioners often treat marginal results on absolute levels with considerable caution, concentrating their attention more on comparisons, either between different subgroups of the sample or between two or more surveys. In interpreting these contrasts as estimates of true differences, they are assuming that there is a constant bias across subgroups or surveys, a bias which thus cancels out in the contrast: this is essentially the form resistant correlation assumption discussed earlier. While this assumption is often a reasonable approximation, it should nevertheless not be relied on uncritically: we have noted several reported examples of its failure with opinion questions, and differential biases can also be expected to occur with factual questions.

A special problem in contrasting the results of different surveys is the question order and context effect discussed in the previous section. The demonstrated sensitivity of responses to both factual and opinion questions to the presence of other questions on the questionnaire must be a major concern to survey researchers, who frequently make great use of comparisons between surveys in their analyses. A good illustration of this problem is provided by Turner and Krauss (1978) who examine the difficulties of comparing subjective social indicators across a series of surveys.

The evidence from methodological research points to considerable room for improvement in the questioning phase of the survey operation. Although a substantial amount of research has been conducted in this area, we remain largely ignorant of the nature of question wording and form effects, of when such effects will occur and when they will not, and of how they operate. In the past, most of the research on factual questions has simply assessed the extent of response errors, and most of that on opinion questions has just examined discrepancies between two or more variants of a question. Present research is beginning to study question effects in a systematic way, with an attempt first to codify the types of effects, and then to understand the psychological processes underlying them. This stage of work is still in its infancy, the task is a daunting one, and progress is likely to be slow with little prospect for major breakthroughs. Rather than the isolated experiments of the past, what is now needed is a series of developmental programs aiming to build and test theoretical structures for questioning effects. Until we have a much clearer understanding of the factors involved in the questioning process and their interrelationships, we will lack the basis for constructing good questions.

One possibility is that an existing psychological theory can explain a broad range of effects of the type we have noted and perhaps also lead the way to prevention. For example, Bishop *et al.* (1981) suggest that recent cognitive theories can provide such a framework and they illustrate this by interpreting a single context effect they obtained by accident. Until the theory can be used in a predictive way, however, or at least applied to a variety of existing examples, it is difficult to know whether much more than a onetime after-the-fact fit has been achieved. A rather different and admittedly slower and more arduous approach can be illustrated by our own recent work, which starts from the context effect with two abortion items mentioned above. The initial effect, again first discovered by accident and then replicated experimentally, involved two adjacent items in NORC's General Social Survey. Our first step was to substitute other questions on abortion that we thought to be conceptually similar to the original items in order to determine whether the effect is limited to the exact wording of the original NORC questions. This was done one item at a time, and we have now found that the context effect does generalize beyond both the original items to other parallel questions dealing with abortion.

A further step was then taken to determine whether the effect requires that the items be contiguous, as in all the experiments thus far, or whether it extends to situations where the two are separated by a number of unrelated items. Initial results support the former requirement, for the effect seems to disappear once contiguity is eliminated, although this is a case where replication is essential and is being pursued. Finally, having obtained a good fix on the degree of generality of the abortion context effect over variants in wording and position, we are attempting to formulate and test hypotheses as to the underlying source of the effect. Thus far an attempt to use open-ended follow-up questions to respondents has not been successful, since the answers did not differ from one context to the other. We are still considering ways of testing rigorously various hypotheses, and as yet are still far from having reached a satisfactory conclusion as to the cause of the original effect.

This whole process of generalization and search for cause has required a series of experiments, each of which attempts to widen and deepen our understanding of the initial accidental finding. If closure is achieved in this one case it will then be necessary to determine whether other cases can be assimilated to it; or if not, to develop separate sets of findings and explanations.

This approach is obviously time-consuming and frustratingly slow. But we suspect that a series of such systematic investigations will be necessary in order to understand the nature of, first, context effects, and then response effects more generally. In other areas of science progress ordinarily involves repeated small steps, and there is no particular reason to believe that our understanding of response effects in surveys can avoid similar efforts. If we are to go beyond merely producing ad hoc instances of effects and ex post facto explanations, this kind of medium level detective work may be essential.

In the meantime, the survey practitioner has one strong defence against the kinds of artifacts that we have described in this paper: use of multiple questions, contexts, and modes of research. Damaging response effects are not rare, but they are also not so pervasive as to occur in the same way with every survey item. By tieing an important concept to at least a few items that differ among themselves in form, wording, and context, the investigator is unlikely to be trapped into mistaking a response artifact for a substantive finding.

REFERENCES

ABERNATHY, J. R., GREENBERG, B. G. and HORVITZ, D. G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, **7**, 19–29.

BANCROFT, G. and WELCH, E. H. (1946). Recent experience with problems of labor force measurement. *J. Amer. Statist. Ass.*, **41**, 303–312.

BARTON, A. H. (1958). Asking the embarrassing question. *Public Opinion Quart.*, **22**, 67–68.

BELSON, W. A. (1962). *Studies in Readership: A Report of an Enquiry*. London: Business Publications.

—— (1966). The effects of reversing the presentation order of verbal rating scales. *J. Advert. Res.*, **6**(4), 30–37.

BELSON, W. A. (1968). Respondent understanding of survey questions. (*International Review on Public Opinion*), **3**(4), 1–13.

BISHOP, G. F., OLDENDICK, R. W. and TUCHFARBER, A. J. (1980a). Experiments in filtering political opinions. *Political Behavior*, **2**, 339–369.

BISHOP, G. F., OLDENDICK, R. W. and TUCHFARBER, A. J. (1981). Question order and context effects in measuring political interest. Paper presented at the 36th Annual Conference of the American Association for Public Opinion Research, May 1981.

BISHOP, G. F., OLDENDICK, R. W., TUCHFARBER, A. J. and BENNETT, S. E. (1980b). Pseudo-opinions on public affairs. *Public Opinion Quart.*, **44**, 198–209.

BRADBURN, N. M., SUDMAN, S. and ASSOCIATES (1979). *Improving Interview Method and Questionnaire Design.* San Francisco: Jossey-Bass.

CAMPBELL, A., CONVERSE, P. E., MILLER, W. E. and STOKES, D. E. (1960). *The American Voter.* New York: Wiley.

CANNELL, C. F. (1977). *A Summary of Studies of Interviewing Methodology.* Vital and Health Statistics, Series 2, No. 69. Washington, DC: US Government Printing Office.

CANNELL, C. F. and KAHN, R. L. (1968). Interviewing. In *The Handbook of Social Psychology. Volume Two: Research Methods* (G. Lindzey and E. Aronson, eds), 2nd ed., Chapter 15. Reading, Mass.: Addison-Wesley.

CANNELL, C. F., MILLER, P. V. and OKSENBERG, L. (1981). Research on interviewing techniques. In *Sociological Methodology, 1981* (S. Leinhardt, ed.), pp. 389–437. San Francisco: Jossey-Bass.

COWAN, C. D., MURPHY, L. R. and WIENER, J. (1978). Effects of supplemental questions on victimization estimates from the National Crime Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1978, pp. 277–282.

FORSYTHE, J. B. and WILHITE, O. (1972). Testing alternative versions of Agriculture Census questionnaires. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 1972, pp. 206–215.

GIBSON, C. O., SHAPIRO, G. M., MURPHY, L. R. and STANKO, G. J. (1978). Interaction of survey questions as it relates to interviewer–respondent bias. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1978, pp. 251–256.

GOODSTADT, M. S. and GRUSON, V. (1975). The randomized response technique: a test on drug use. *J. Amer. Statist. Ass.*, **70**, 814–818.

GRAY, P. and GEE, F. A. (1972). *A Quality Check on the 1966 Ten Per Cent Sample Census of England and Wales.* London: HMSO.

HEDGES, B. M. (1979). Question wording effects: presenting one or both sides of a case. *Statistician*, **28**, 83–99.

JAFFE, J. A. and STEWART, C. D. (1955). The rationale of the current labor force measurement. In *The Language of Social Research* (P. F. Lazarsfeld and M. Rosenberg, eds), pp. 28–34. New York: The Free Press.

KALTON, G., COLLINS, M. and BROOK, L. (1978). Experiments in wording opinion questions. *Appl. Statist.*, **27**, 149–161.

KALTON, G., ROBERTS, J. and HOLT, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician*, **29**, 65–78.

KALTON, G. and SCHUMAN, H. (1980). The effect of the question on survey responses: a review. (With discussion by N. D. Rothwell and C. F. Turner). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1980, pp. 30–45.

LOCANDER, W. B. and BURTON, J. P. (1976). The effect of question form on gathering income data by telephone. *J. Marketing Res.*, **13**, 189–192.

LOCANDER, W., SUDMAN, S. and BRADBURN, N. M. (1976). An investigation of interview method, threat and response distortion. *J. Amer. Statist. Ass.*, **71**, 269–275.

MADIGAN, F. C., ABERNATHY, J. R., HERRIN, A. N. and TAN, C. (1976). Purposive concealment of death in household surveys in Misamis Oriental Province. *Population Studies*, **30**, 295–303.

NATIONAL CENTER FOR HEALTH STATISTICS (1972). *Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents.* Vital and Health Statistics, Series 2, No. 50. Washington, DC: US Government Printing Office.

NETER, J. and WAKSBERG, J. (1965). *Response Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study.* Bureau of the Census Technical Paper No. 11. Washington, DC: US Government Printing Office.

PARRY, H. J. and CROSSLEY, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quart.*, **14**, 61–80.

PAYNE, S. L. B. (1951). *The Art of Asking Questions.* Princeton: Princeton University Press.

PRESSER, S. and SCHUMAN, H. (1980). The measurement of a middle position in attitude surveys. *Public Opinion Quart.*, **44**, 70–85.

QUINN, S. B. and BELSON, W. A. (1969). *The Effects of Reversing the Order of Presentation of Verbal Rating Scales in Survey Interviews.* Survey Research Centre, London School of Economics.

ROTHWELL, N. D. and RUSTEMEYER, A. M. (1979). Studies of Census mail questionnaires. *J. Marketing Res.*, **16**, 401–409.

RUGG, D. and CANTRIL, H. (1944). The wording of questions. In *Gauging Public Opinion* (H. Cantril, ed.), Chapter 2. Princeton: Princeton University Press.

SCHUMAN, H. and PRESSER, S. (1977). Question wording as an independent variable in survey analysis. *Sociol. Methods and Res.*, **6**, 151–170.

SCHUMAN, H. and PRESSER, S. (1978). The assessment of "no opinion" in attitude surveys. In *Sociological Methodology, 1979* (K. Schuessler, ed.). Chapter 10. San Francisco: Jossey-Bass.
—— (1979). The open and closed question. *Amer. Sociol. Rev.*, **44**, 692–712.
—— (1980). Public opinion and public ignorance: the fine line between attitudes and nonattitudes. *Amer. J. Sociol.*, **85**, 1214–1225.
— (1981). *Questions and Answers on Attitude Surveys. Experiments in Question Form, Wording, and Context.* New York: Academic Press.
SCHUMAN, H., PRESSER, S. and LUDWIG, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quart.*, **45**, 216–223.
SHIMIZU, I. M. and BONHAM, G. S. (1978). Randomized response technique in a national survey. *J. Amer. Statist. Ass.*, **73**, 35–39.
SPEAK, M. (1967). Communication failure in questioning: errors, misinterpretations and personal frames of reference. *Occupational Psychology*, **41**, 169–181.
SUDMAN, S. (1980). Reducing response errors in surveys. *Statistician*, **29**, 237–273.
SUDMAN, S. and BRADBURN, N. M. (1973). Effects of time and memory factors on response in surveys. *J. Amer. Statist. Ass.*, **68**, 805–815.
—— (1974). *Response Effects in Surveys.* Chicago: Aldine.
TRAUGOTT, M. W. and KATOSH, J. P. (1979). Response validity in surveys of voting behavior. *Public Opinion Quart.*, **43**, 359–377.
TURNER, C. F. and KRAUSS, E. (1978). Fallible indicators of the subjective state of the nation. *American Psychologist*, **33**, 456–470.
WARNER, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Ass.*, **60**, 63–69.

DISCUSSION OF THE PAPER BY DR KALTON AND DR SCHUMAN

Professor D. HOLT (University of Southampton): It gives me great pleasure to propose the vote of thanks this evening and also to welcome back Graham Kalton, who served the Society so well on Section committees and Council before moving to Ann Arbor. May I say, too, how appropriate it is to have, as authors on this important topic, two members of the Survey Research Center, Michigan, where so much has been done to advance our understanding of the effects of question wording and presentation on survey responses.

The paper itself collects together a selection of results and conclusions on various aspects of the question wording problem in surveys. I use the term "question wording" as a catch-all phrase, although the paper we have heard this evening goes beyond the mere wording of individual questions. It is clear that, in the last twenty years, a great deal of work has been done in this area, and yet at the end of the paper one is left, and I think that the authors share this sentiment, with some feeling of disappointment at the size of the task which remains. This is perhaps not surprising since the issues which arise are extremely complex. Nevertheless, the magnitude of the effects which can arise are not trivial as we have seen tonight. Yet in some situations we seem to be little nearer to the development of an accepted methodology, except in the narrow case of saying "If you want to ask this particular question then do not do this or that." Add to this the conclusion later in the paper that, not only are there question wording effects but there are also contextual effects, and the practising survey researcher may be forgiven for thinking that we are sinking further into a morass of complexity. I offer no criticism, of course, to this evening's authors, since they are right to draw our attention, once again, to these problems but what is clearly needed, as the authors indicate in their final section, is a planned campaign to reach a position where at least some of these effects are so well understood that standard methods will be adopted by all responsible practitioners. If, as the authors indicate, the road is going to be long and painful, I wonder whether it is because we are sometimes treading the wrong path. It seems to me that there is a tendency to trivialize complex issues to produce simple measures such as the proportion "in favour" or "against" a particular issue. It might be that the gross oversimplification of issues is the source of some of the large effects observed.

If questionnaire design and question wording is an art form, then what we need are clearly enunciated principles, which the survey methodologist can apply in designing a questionnaire for a particular purpose and in analysing the results. Some crude principles are already available, such as the avoidance of obviously biased practices, but crude principles are clearly not enough, since significant differences still arise between alternatives which one would have thought were comparable. One hopes for a time when explanations of particular effects are given in terms which go beyond the specific context of the question and I congratulate the authors on offering these in various places. To be convincing, such explanations