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. Honville 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).

© 1982 Royal Statistical Society

0035–9238/82/145042S2.00
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 concerns 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, recalling 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 marshal 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, $10000, $15000, $20000 and $25000; form 1 started with $5000 and took increasing values of X until a "no" answer was given, while form 4 started with $25000 and took decreasing values until a "yes" answer was given. With form 1 37.5 per cent of respondents reported family incomes of $15000 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 tying 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


F. 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...
Discussion of the Paper by Dr Kalton and Dr Schuman

[Part 1,

will need to stand the test of time and to explain the non-existence of effects in other situations. The topic of form resistant correlations is one in which no clear pattern has yet emerged and no theoretical justification exists. To illustrate my point, several times this evening we have heard about specific methods that "this often occurs but not always".

If this topic is a science rather than an art form, then we clearly need a much stronger framework in which to measure the effects which we observe and there are one or two places where the possibility of such a framework looks like a distant promise. In particular, the work of Cannell and others on the effects of question length and the completeness of responses is interesting. The authors give a number of suggestions as to why questions which have been padded out with redundancies may encourage a more complete response and it is not difficult for amateur social psychologists to leap in with both feet and suggest others. If a framework is to be developed for such topics, then we need a way of coding questions and, indeed, whole sections of questionnaires, probably in psychological terms, to yield a measure for the question. For example, in Cannell's work a measure of redundancy or fluency or time lag between the introduction of a concept and the need of a response, or whatever the relevant characteristics of the question design are thought to be. If such a measure is shown to be related to the completeness of the responses over a variety of questionnaires in form and content, then we would have the beginnings of a framework against which future questionnaire designs could be measured. It is this development to get beyond ex-post explanations of what has happened, to predict what effects will exist in future questionnaires, and to provide guidance to survey methodologists which is the most urgent need.

Finally, may I take issue with the authors on an assertion which they make in several places throughout the paper and particularly Section 5. Here I feel that there is tacit support for the view that, since marginal results are difficult to measure, the concentration on comparisons is acceptable. Of course comparisons are valid in their own right but I have recently returned from New Zealand, where press coverage has been dominated by the recent Springbok tour. A great deal of emphasis was placed on the survey finding that 54 per cent of the population was against the tour taking place. It seems to me that the desire to derive some measure of the attitude of the population—to a Springbok tour in this case—is perfectly valid and can influence the political and social life of a country. If the current techniques for measuring this are inadequate, then there is a need to derive valid methods to measure the proportion supporting a particular issue. The alternative is to demonstrate that a simple proportion is an inadequate measure of opinion in such a complex issue and to provide an alternative entirely new set of measuring instruments; otherwise survey methodologists will be cast into the role of saying that they cannot measure what is needed.

May I close by complimenting the authors on a well-written and a well-presented review. It gives me great pleasure to propose a vote of thanks on behalf of the Society.

Mr B. Hedges (Social and Community Planning Research): I should begin, perhaps, by declaring a special interest, no doubt shared by many of the audience, in the subject of this paper. I spend my time primarily in conducting sample surveys or on matters connected with them. I therefore welcome this paper as an excellent review of what is known about the effect of question wording on responses, and it is one that I shall find very useful in my work.

Kalten and Schuman remind us that a great deal of work has been done on this topic, and that we have learnt a lot from it. It is not uncommon to meet people who approach the task of designing their first questionnaire with perfect confidence in their ability to perform what after all appears to be an ordinary, everyday business—the asking of questions. Everyone does that all the time; it is simple enough, then, to devise a string of questions that will produce good data ready for sophisticated analysis. However, this paper should make such people think twice. Designing questions is not easy, and is full of pitfalls.

Although a lot is known as a result of the work that has been done, what is not known is far greater. Anyone who tries to design a questionnaire, and in that process refers to the literature for guidance, is likely to be left with a large quantity of unsolved problems. I think that this will continue to be so for a long time. It is true that the rate of experimentation on, and interest in, question experiments is speeding up. In spite of that, and even if the integrated research the authors advocate is adopted, rather than the piecemeal research we have had in the past, I still think that systematic experiment will solve only a small proportion of the questionnaire designer's problems.

Reference to what is known from well-planned experiments is clearly the best basis for question design, but what about all the cases in which such knowledge is not available? My belief is that much could be learnt from much more open discussion and criticism of questions, even in the absence of specific experiments. There tends to be an implicit assumption in the literature that in the absence of hard
evidence there is little point in speculating or reasoning about question effects because any opinion is likely to be as good as any other. That is, if you have knowledge, you have knowledge; if you do not, you have only opinion, and opinion is not a particularly valuable commodity.

It seems to me that large numbers of bad questions are asked in surveys. I have been asking bad questions for 26 years, and I know that I am not the only one to do so. How are we to improve our questionnaire design standards? Certainly, as far as we can, by further systematic experiment. But also, I think, by bringing questions under much more critical discussion.

This is a very sensitive issue. It is unsatisfactory to invent questions to discuss, but it is also unsatisfactory to discuss questions that have actually been used, because such discussion can appear to be a personal criticism of the researcher concerned. I know only too well how unfair such criticism can be, because the researcher whose name appears on the report in which a certain question appears is not necessarily the person who invented it. For example, it is often necessary to repeat a question which has been used in the past simply for the sake of comparability.

It is, of course, very easy to be critical after the event—for example, when a question has produced a clearly unsatisfactory result. That does not mean, though, that all criticism is illegitimate. Many questionnaire errors are in all probability unavoidable, and would escape detection by the scrutiny even of a panel of researchers, however eminent. But many could be avoided. I have myself often polished a draft questionnaire to the stage at which I think it is entirely satisfactory for its purpose. I have then submitted it to colleagues for discussion, only to find that they can quickly spot many flaws, because my own perception of what I was asking was incomplete. In some experiments with which I myself have been concerned the problem has been the lack of clarity with which the concept that the question is intended to embody has been formulated. It is all too easy to finish up with a form of words intended to embody an idea which superficially seems precise, but in fact is not. The person devising the question can very easily overlook possible misinterpretations. If it is submitted to other people's criticism, these may well be discovered.

Discussions of questions is undoubtedly a useful way of weeding out those open to misinterpretation. When I read the question literature, even the part of it that contains experiments on questions and question effects, I often feel that that kind of discussion seems not to have taken place.

My plea, then is not only for more experiments but for much more open discussion of questions, and much more willingness to debate their structure and to expose possible pitfalls.

It gives me great pleasure to second the vote of thanks for an excellent review paper.

The vote of thanks was passed by acclamation.

Dr John Bynner (Open University): The plea at the end of the paper for a proper theory of and systematic research into question effects is one to endorse. Until we have such a theory of question effects and its accompanying structural model then survey practitioners must by necessity go on with the procedures of piloting and pre-testing in the context of a particular survey rather than look to any general guidelines to help them. This raises the question of whether we model bias in the mean brought about by inaccurate measurement and/or bias in measures of relationships brought about by correlated errors, which may be a more appropriate distinction than the standard one proposed in the paper between "factual" and "opinion" questions.

This raises the question: When is a bias not a bias? As Kalton and Schuman point out, many apparently factual questions can serve either the purpose of elucidating facts or elucidating perceptions of a personal condition. In many cases such questions can serve both purposes in which case the bias in relation to the first purpose is of substantive interest in its own right in relation to the second. If we ask a child whether he smokes or not and he denies it, or claims to have given up, when from the nicotine stains on his fingers we can see this to be patently not true, then this points to an ambivalence in relation to factual status and consequently possibly a propensity to change behaviour (of central; interest in, say, a health education survey). It seems to me important to decide, in line with Cronbach's generalizability theory, whether our focus of interest does lie in such individuals before we dismiss their responses as biased. For in such cases it is precisely the questions that will bias the mean that may well do most to help us illuminate ambivalence, i.e. through the investigation of their relationship with other questions.

This leads to my second question: Who is producing the Bias and Why? with one exception in the printed paper, the National Crime Survey (Cowan et al., 1978), there is no mention of sub-group and individual differences in response to different question forms; nor is there any mention of attempts to elucidate the reasons for them within the context of the experiment itself. If we take the typical experiment
Discussion of the Paper by Dr Kalton and Dr Schuman

Comparing two forms of the same question, we find that some people conform with the null hypothesis and some people do not, but very rarely do we see any further detailed investigation of these differences either at the group level or at the individual level to see what else defines them. This is of course asking for more interviewing, this time about the interviewing itself, but without such an approach to the problem of bias, it is difficult to see where any satisfactory theory to account for its origins is going to come from.

Mr F. E. Whitehead (Social Survey Division, Office of Population Censuses and Surveys): I would like to add my congratulations to the authors on a useful and comprehensive review which will be profitably read by all those who, like those in my organization, are concerned with the conduct of interview surveys. Questionnaire design and the wording of questions are certainly key skills of the survey researcher but, as people have already commented this evening, they remain much more of an art than a science. In designing even apparently straightforward factual questions the questionnaire designer has to steer between many uncharted hazards. Short and simple questions are often conceptually vague and hence unreliable, and as soon as we start building in definitions and qualifications the whole mouthful becomes incomprehensible and we look for ways of breaking it down.

We have simple concepts—at least, we think that they are simple. Income is a concept thought to be simple, but in one of our big surveys it takes 20 or 30 questions actually to disentangle it. A lot of survey research is concerned with breaking down concepts into their component parts and asking questions on each one of them.

I want to give an example from the early days of the General Household Survey of question wording effects which seems to me quite striking. It concerns a question asked in 1971 to establish the prevalence of chronic sickness and long-term incapacity. This was an important topic at the time, just after we had published the results of a very detailed survey into the handicapped and impaired; and we were looking for ways of monitoring chronic sickness over time. The question that evolved in 1971 was: "Do you suffer from any long-standing illness, disability or infirmity which limits your activities compared with most people of your own age?"

In 1972 the question was changed, and it was broken up into two separate questions, the first part of which was: "Do you suffer from any long-standing illness, disability or infirmity?" If the answer to that question was "yes", we then asked "Does it limit your activities compared with most people of your own age?"

We obtained quite different prevalence rates from these two questions for far fewer people answered "yes" to both parts of the question in 1972 than answered "yes" to the single question in 1971, as the following table illustrates.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Limiting, longstanding disability rates per 1000 population</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–4</td>
<td>31 (1971) 17 (1972)</td>
</tr>
<tr>
<td>5–14</td>
<td>55 (1971) 36 (1972)</td>
</tr>
<tr>
<td>15–44</td>
<td>89 (1971) 62 (1972)</td>
</tr>
<tr>
<td>45–54</td>
<td>250 (1971) 190 (1972)</td>
</tr>
<tr>
<td>65–74</td>
<td>412 (1971) 329 (1972)</td>
</tr>
<tr>
<td>75+</td>
<td>484 (1971) 390 (1972)</td>
</tr>
</tbody>
</table>

Later we changed the question further because we suspected that the phrase "most people of your own age" was having an effect on the answers. We now ask the question by using the phrase "in any way", instead. You will note that the drop in prevalence between 1971 and 1972 occurred over all age groups, but happened to be relatively greater for children than for adults. This is a rather detailed example, and has been published in the 1972 report on the General Household Survey. I think that this is quite a good example of what Professor Kalton described as a "judgement" question: is the existence of long-term limiting capacity a fact or an opinion?

I want to finish by saying what I think survey organizations can do—and, in fact, do—to take account of some of the effects that have been described. As an organization we try to detect and eliminate as many faults as we can in questionnaire design and question wording at the piloting stage. Many surveys, government ones included, are not adequately piloted because of insufficient allowance of time and money. Of course, many question design effects cannot be properly answered without controlled experimentation.
of the kinds described, but in my view it is often possible to weed out questions which are patently not working properly through consultation with other researchers and feedback from interviewers at the pilot stage.

I feel that many problems in interpreting replies to attitude questions arise because they have been included without appropriate prior questions being asked about informant’s knowledge of the subject under investigation. Many years ago when I was at the Department of Health and Social Security we commissioned a survey of public attitudes to Social Security arrangements and to certain aspects of the Health Service. The level of knowledge of the services and the way they were financed was appallingly low, but the public willingly gave opinions on these matters and, in fact, knew a lot of things about the Health Service and Social Security which just were not so. It is important to distinguish between the informed and uniformed opinion.

Professor Martin Collins (SCPR Survey Methods Centre and The City University): In seeking solutions to the various questioning problems raised here, it may help to distinguish between four categories:

1. **Imprecision.** Imprecision in framing a question will lead individual respondents, or sometimes interviewers to apply their own interpretations. We need to develop questions that minimize this variability, acknowledging that they *may* not be the questions that come closest to the truth. To quote from Deming (1944): “Many biased methods . . . show smaller variability than so-called unbiased methods. If the relationship between biased and unbiased results were known . . . the biased techniques would sometimes be preferable.” We need, then, a reliable question with a known relationship to the truth—an objective that at least narrows the field.

2. **Method effects.** The problem with effects such as yea-saying is one of bias among all or part of a sample and of patterning that can easily be mistaken for a substantive finding. The solution here is to document and classify such effects, as in the current paper, so that they become better-known and more predictable (given that they may be unavoidable).

3. **Respondent task problems.** These problems include those of recall or frankness. We know a good deal about such problems because the necessary documentation and classification has been performed. Hence, we have developed ways of avoiding—or at least coping with—problems that are predictable.

4. **Asking different questions.** The fact that we obtain different responses according to question form is not always a problem. It can mean that we have asked two different questions, the answers to both of which are relevant to the research objectives. This applies to many of the examples cited, including perhaps the question-order effect for two questions concerning attitudes towards abortion.

These are effects that we do not necessarily want to remove. Indeed, we should more often criticize the “opinion-poll” approach to a problem that does not deliberately introduce such variation. The comparison of answers to two different questions may well be more informative than either set of answers in isolation. It would be wrong in our methodological work to over-state the importance of any one question—or even any one survey. Any given response distribution arises in the immediate context of a full survey and in the broader context of background knowledge, some but not all of it formalized or data-based.

A prerequisite of a solution to each of these four problem types is the need to know of their existence, in general and in the particular case. This points to some desirable developments in survey practice:

- more and better pre-testing of questions, with formal analysis;
- survey design and analysis that exposes failures;
- deliberate redundancy in a questionnaire in readiness for such failures;
- more use of good standardized questions.

Professor Louis Moss (Birkbeck College, London University): This paper contributes to a very necessary discussion but does not offer any very direct guidance about design. This may be because, as the authors note, there is not yet any theoretical framework to which their results can be related but I wonder whether enough attention has been given to what we do know already and can use in designing studies. It is not always clear what is fact or opinion or judgement. We know, for example, that after an election some people change their supposed vote (how they say they voted) from the losing to the winning party. And it has been shown that appreciable proportions who say they voted have not in fact done so. But participation in elections up until recent years in this country has been socially desirable, and so is being
on the winning side. Similarly when people offer opinion and judgements when invited to do so on issues with which they are unfamiliar maybe they are only doing their best to cooperate in response to the urgent prompting of interviewers to do so. As the authors note however many, when given the opportunity to opt out by an offer of “no opinion” or “don’t know” status will in fact opt out. If it seems important to give all sections of the sample equal opportunities of responding out of their own experience or knowledge the wording of questions would have to be simplified or generalized accordingly. People can only respond out of their own experience of the subject matter under review or their perceptions and beliefs about it. The interviewer is acting on behalf of researchers whose interests and patterns of thinking may be in many ways very different. The results discussed suggest to me that in designing questions we have to be more willing to recognize the nature of this social event—the contact of respondent and interviewer. The relationship achieved in the interview will affect the response. It therefore seems important to identify in some way the climate of the interview and perhaps to incorporate it as an explicit variable in future experimental work. This would be a logical conclusion to draw also from the experimental work of Charles Cannell. As I understand it he showed that the relationship between interviewer and respondent was one of the most important variables influencing the outcome of interviews.

Clearly respondents do sometimes express apparently contradictory attitudes e.g. agree with both sides of an issue. But on most matters only small minorities may be expected to voice completely consistent opinions. This means that in any chosen dimension the response to one question alone may mislead. It is therefore safer to construct scales based on a cluster of questions which between them invite responses to different expressions of the same dimension. This makes the “validity” of response to any one question a less urgent matter. Each question used to help construct a scale may also be much simpler if, by itself, it does not have to represent the whole of a dimension. The simpler the individual questions the more likely it is that respondents will interpret them in the way the investigator desires and this will make a further contribution to validity.

So in some ways we can already build on what we know. But this does not in any way reduce the need for the kind of critical review the authors have given us.

Dr W. A. Belson (Survey Research Centre, London): The paper by Kalton and Schuman is, I believe, a very useful contribution to our developing technology of question design for survey work. It is against the background of that general statement that I am putting forward what I consider are some of its weaknesses.

In the first place, the printed paper appears to accept at face value the adequacy of the many findings reviewed. I find it very different in this respect from that devastating appraisal of postal survey methodology research made to this Society by Christopher Scott some years ago. Scott dealt with defects like the use of volunteer subjects, the use of students as a basis for generalizing to the total population, the use of very small samples, the inadequacy of research design, the absence of statistical evaluations, and so on. I want to be assured that the findings utilized in this review have passed the sharply critical appraisal of its authors. And I need to be warned about the weaknesses of the findings they do present.

I am less than happy about the way in which opinion-type questions appear to be graded as non-factual questions. Through opinion questions we are seeking facts about a person’s opinions. If a woman rates some public service as fairly satisfactory, that is meant to present a fact as far as her feelings are concerned. Moreover, from a validation point of view, we can go a long way, using intensive techniques, towards finding out if she has correctly stated this fact about her position. I do not think it helps us in testing the validity of opinion type questions, to call them non-factual.

On the other hand, there is a distinction in the paper that I think the authors might have made more of than they did. It is the distinction between open-ended and closed questions. Indeed it seems at one point as if they are equating the functions of the two. That of course would be entirely incorrect. The principal function of the genuine open-ended question is exploratory in character. It tells us the different kinds of ideas that people have on some issue—their different kinds of behaviour—their different kinds of beliefs. To get them listed fairly fully, we go on adding more survey respondents till the number and variety of those different items ceases to grow. We are then in a position to build those items into a head counting system—a closed questioning system—that has been validated. It is a mistake to equate the functions of those two methods—or even to risk equating them. The distinction is fundamental.

I think it is unfortunate that the authors referred to a question that repeated some of its instructions and asked for careful attention as being a “long” question. That is not what we normally mean by “long”. A long question is more often one that goes on adding more and more information
which the respondent is required to absorb in order to know what it is that the questioner wants to know about. A long question may also be fairly complex in its structure. Of the finding itself, there is nothing surprising really. Anyone who is in the fortunate position of being able to repeat the question and to somehow inject interest into the respondent, is likely to get more considered replies.

But what we must keep in mind when we send interviewers out to deliver these repetitive questions, is that many of them are likely to rip the question back to its essentials. This sort of interviewer deviation from instructions is a major consideration in realistic question design, and I am rather surprised that the authors don't refer to it in presenting an otherwise interesting result.

There are many other issues that could and should be raised in relation to this paper, but I must leave these to others. Having said that, let me conclude by adding that I have particularly appreciated the opportunity to read this review, for it opens up again the very important but much neglected field of question design.

Mr N. Webb (Gallup): Certainly, I add my own welcome to the research presented. If we look through the literature of the Royal Statistical Society on surveys, we find all kinds of material about multistage samples, concomitant variables, disproportionate stratification, and other mathematical points—then we find that it is all to do with the study of the rice crop in a province of India, which does not help us practitioners at all. To a large extent, the literature ignores the blunt end (I am at Gallup)—that is, the interface between the interviewer and the respondent. Quite incidentally, a friend of mine suggested to me that one very good way of piloting questions was to mumble the questions to a slightly deaf respondent. This would actually reproduce what was happening out there in practice rather more than we might have thought! I commend that as an idea.

Validation occurs, shall we say, fortuitously, quite often. For instance, our material is validated by the actual vote in elections, whatever people may see in the cheaper newspapers. Again, we have found that the degree of interest in a forthcoming general election was a better indicator of turnout than a question about whether they would turn out or not. That is odd, but we have validated that.

We had a complicated study on skin conditions in which it was possible to validate the frequency that people claimed to have gone to the doctor about their skin condition by means of doctors' records, which gave us a great deal of faith in other data such as their grandmother's advice—which was what we were really trying to find out about—when they did not go to their doctor.

I would like to try to relate a near horror story. We were conducting an experiment on television research, which would probably cost about £1 million nowadays—this was nearly 20 years ago. Although I was in co-charge of the research, I piloted some of it. We did our pilot work in Cambridge because I wanted to visit my old College. The first question after ringing at the door was to ask whether they were watching television when they heard the doorbell ring. That seemed to be all right; a perfectly straightforward, simple, factual, easily identifiable situation. A lady said that she was, and we went on with the questionnaire. After we had finished I asked her another question—I told her that she had a picture window through which I could the kitchen, and as I rang the bell I saw that she was in the kitchen washing up, yet she said that she was watching television when the doorbell rang. She replied that she was watching television when the doorbell rang, but she had just slipped out to wash up a few cups. We suddenly realized that what we thought was a precise question was not a precise question. Since this whole study was carefully timed with a view to measuring the audience to commercial breaks, among other things, the whole project would have been destroyed had we not had the lucky chance in our piloting to visit a home built in such a way that we could observe what was going on.

I would like to mention two points in this area. First, we do a lot of international research: that is, research is done simultaneously with the same questions, say, in France, Germany, Italy and Great Britain. There are problems with translation, correspondence in meaning and so on. I know that there has been some research done on whether verbal scales have the same meanings in different languages, but this is a problem which we find very arduous to solve. The more we do this—we are doing these surveys almost every week—we find that it is not the problem between German and English (that is no problem because we know that we are dealing with a problem), but the problem is between Venezuelan Spanish and Chilean Spanish, Belgian French and French French, or American English and English English, where we do not necessarily recognize the problem. It occurs to me that maybe within countries as various as Britain, with Scotland, Wales, the North, the South-West and so on, we may think that there is a culture that is sufficiently uniform to have a national questionnaire, but that there are problems of cultural variation which mean that this questionnaire need not be valid throughout a nation. The same
may be true in America, with the West Coast, the East Coast, the South and Mid-West. Nobody seems to have addressed himself to that question. It may be a very real one.

The title of the paper is “The Effect of the Question on Survey Responses ...”, so there is another point I want to raise. We have a responsibility to develop a questionnaire on a general issue to cover that issue fully, or adequately. This cannot always be done. If you happen to be an anti-abortionist pressure group, you will commission us to do a survey on the problems of doctors and nurses involved in abortion in terms of their professional career and professional conscience. However, if you are pro-abortionist, you will ask us to do a survey of only fertile women—and nobody else—to see what they think. Ethically, we are doing the right thing in both cases, but we know that to some extent, by the simple choice of the topic area, we are being manipulated or utilized in some way which worries me from time to time.

We feel that there is also a question of money involved—perhaps the Right and the rich can finance opinion polls, or social surveys, whereas the Left and the poor and the weirdos cannot. An important ethical problem is raised here, and it is something to which we should in fact address ourselves in a much wider context.

Professor T. M. F. Smith (University of Southampton): It gives me particular pleasure to see Graham Kalton at the Society again and to add my congratulations to him and Howard Schuman on their paper. Graham and I have shared many common features in our careers not the least of which is our research interest in the theory and practice of sample surveys.

Whenever I lecture on sample surveys I always start by listing the possible sources of error. Having gone through a massive list of non-sampling errors I arrive finally at statistical sampling errors. I then have to admit that despite the fact that the majority of errors are non-sampling errors the majority of my lectures will be devoted to sampling errors which relatively are unimportant. I then have to justify this on the grounds that it is only for sampling errors that we have a satisfactory statistical theory; or should I say theories.

This paper is welcome because it reminds us about the importance of question wording and concept formulation, in other words, of measurement, in the practice of survey sampling. It is quite clear that these errors could easily swamp sampling errors. The disappointing aspect of the paper is that at the end we are still lacking an agreed theoretical basis for determining the wording of questions. We have to rely on that uncommon quality called common sense, and it is hard to teach this.

In this paper the authors draw attention to Gray and Gee’s excellent evaluation of the errors in the 1966 census. Am I alone in not being satisfied with this year’s census form? I found some questions hard to interpret and I must admit to making an error in my responses to the occupation questions. On further enquiry I was told that a very high proportion of forms were incorrectly completed first time and had to be corrected subsequently by the enumerator. This is surely not satisfactory by any standards. Is it worth asking such error prone questions?

Another example of interest to me is aircraft noise surveys. A major problem is the measurement of the respondent’s annoyance to aircraft noise. Aubrey McKennell has devised several scales of measurement which include a simple self-rating scale and some more complex Guttman scales based on activities disturbed. Aubrey and I used the latter scales in our analyses but I still feel that the simple self-rating scale based on the question “how annoyed or disturbed are you by aircraft noise?” may have been just as appropriate as a measure of annoyance although it lacks the interpretation of the activities disturbed scale.

My questions to the authors are the following:

(i) What criteria should we have adopted for choosing between the alternative scales for measuring annoyance?

(ii) Would Graham be prepared to give Mr Heseltine and Members of Parliament a lecture on question wording bias before we have compulsory referenda on local authority spending imposed on us.

May I express my thanks to both authors for their excellent review paper.

Dr F. H. Hansford-Miller (Inner London Education Authority): I would like to have seen rather more discussion on the order of questions. There is a short reference to it in Section 3(i)—“Order of alternatives”—whilst in Section 2 the authors refer to employment interviews. One would expect that the candidates in these would be lined up in alphabetical order—that is my experience. What effect does that alphabetical order have on the result of the interviews?
Table 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Expected</td>
<td>( \chi^2 = \frac{(O-E)^2}{E} )</td>
</tr>
<tr>
<td>A</td>
<td>363</td>
<td>3.57</td>
<td>17 24</td>
</tr>
<tr>
<td>B</td>
<td>964</td>
<td>9.47</td>
<td>83 64</td>
</tr>
<tr>
<td>C</td>
<td>833</td>
<td>8.19</td>
<td>50 56</td>
</tr>
<tr>
<td>D</td>
<td>476</td>
<td>4.68</td>
<td>26 32</td>
</tr>
<tr>
<td>E</td>
<td>233</td>
<td>0.29</td>
<td>19 16</td>
</tr>
<tr>
<td>F</td>
<td>342</td>
<td>3.36</td>
<td>22 23</td>
</tr>
<tr>
<td>G</td>
<td>497</td>
<td>4.88</td>
<td>35 33</td>
</tr>
<tr>
<td>H</td>
<td>804</td>
<td>7.90</td>
<td>54 54</td>
</tr>
<tr>
<td>I</td>
<td>73</td>
<td>0.72</td>
<td>8 5</td>
</tr>
<tr>
<td>J</td>
<td>290</td>
<td>2.85</td>
<td>14 19</td>
</tr>
<tr>
<td>K</td>
<td>268</td>
<td>2.63</td>
<td>27 18</td>
</tr>
<tr>
<td>L</td>
<td>433</td>
<td>4.26</td>
<td>24 29</td>
</tr>
<tr>
<td>M</td>
<td>1041</td>
<td>10.23</td>
<td>54 69</td>
</tr>
<tr>
<td>N</td>
<td>196</td>
<td>1.93</td>
<td>11 13</td>
</tr>
<tr>
<td>O</td>
<td>143</td>
<td>1.41</td>
<td>6 10</td>
</tr>
<tr>
<td>P</td>
<td>531</td>
<td>5.22</td>
<td>43 35</td>
</tr>
<tr>
<td>Q</td>
<td>22</td>
<td>0.22</td>
<td>1 1</td>
</tr>
<tr>
<td>R</td>
<td>483</td>
<td>4.75</td>
<td>41 32</td>
</tr>
<tr>
<td>S</td>
<td>900</td>
<td>8.85</td>
<td>57 60</td>
</tr>
<tr>
<td>T</td>
<td>413</td>
<td>4.06</td>
<td>18 28</td>
</tr>
<tr>
<td>U</td>
<td>31</td>
<td>0.30</td>
<td>2 2</td>
</tr>
<tr>
<td>V</td>
<td>70</td>
<td>0.69</td>
<td>4 5</td>
</tr>
<tr>
<td>W</td>
<td>709</td>
<td>6.97</td>
<td>56 47</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>0.01</td>
<td>0 0</td>
</tr>
<tr>
<td>Y</td>
<td>47</td>
<td>0.46</td>
<td>3 3</td>
</tr>
<tr>
<td>Z</td>
<td>12</td>
<td>0.12</td>
<td>4 1</td>
</tr>
<tr>
<td>Total</td>
<td>10,175</td>
<td>100</td>
<td>679</td>
</tr>
</tbody>
</table>

Letters with \( E < 10 \)

<table>
<thead>
<tr>
<th>( \chi^2 )</th>
<th>( n = 20 )</th>
<th>33.28</th>
<th>( n = 19 )</th>
<th>16.41</th>
</tr>
</thead>
<tbody>
<tr>
<td>( = 2.5% )</td>
<td>( P = 5% )</td>
<td>32.85</td>
<td>( P = 5% )</td>
<td>28.87</td>
</tr>
</tbody>
</table>
This question of order is something that I feel needs to be brought in. Of course, there is the work of Brook and Upton, relating to various aspects of bias in elections due to the candidate's position on the ballot paper (Brook and Upton (1974) Upton and Brook (1974) and Upton and Brook (1975)). They show in their paper “The Determination of the Optimum Position on a Ballot Paper” that first position is almost uniformly less favoured than those positions which follow it, and suggest that good strategy for a candidate named Smith would be to change his name to B. Smith.

It was in 1977 that I was struck by the fact that of the 40 Fellows elected in that year to the Royal Society there were three surnames beginning with the letter A, 8 with B and 5 with C. I felt intuitively that there seemed to be a rather greater emphasis on the higher reaches of the alphabet than on those lower down, although that year there was in fact a Z. I did nothing about that until this year, 1981, when I found that once again the Fellows elected to the Royal Society seemed in their surname distribution to show similar evidence of bias, with no A's, 7 B's and 4 C's, again in a total of 40 names.

I have now checked on FRS elections for the twenty years 1962–81, a Total of 679, and these are shown in Table 1, column 4, distributed according to the initial letters of their surnames. How do we find the expected frequencies with which to compare them? My method was to take sample Telephone Directories from various geographical locations in the United Kingdom (British Telecom Telephone Directory, 1980/81), in particular Birmingham, Canterbury, Exeter, Leeds, London and Manchester from England, Glasgow from Scotland, Cardiff from Wales and N. Ireland, and from this total of 10 175 pages of subscriber's names to count the number of pages devoted to each alphabetical letter. These are shown in Column 2, with the corresponding percentages in Column 3. The Expected number of FRS per surname category can then be easily calculated and these are shown in Column 5 of the Table.

I then carried out a chi-squared test on the data, although some might consider other tests more appropriate, and to do so had to group together the small frequency letters I, Q, O, U, V, X, Y, Z. Chi-square amounts to 33.28, which is greater than the P = 5 per cent value of 30.14 and also greater than the P = 2.5 per cent value of 32.85. This suggests that FRS elections are indeed subject to initial surname letter bias and in this connection it is interesting to note that the letter B, as in Brook and Upton's research, is again the most over-represented category, followed by K, R and W. The least popular letters, with under-representation, are T, M and A. Although it is perhaps reassuring that the augst election processes of the Royal Society appear to be subject to the same influences to bias as the more humble citizens voting in an election, I do suggest that if nominations for Fellowship are processed in alphabetical order I have certainly made a case for some more random process to operate in the future.

Before leaving this matter I felt that I would like some confirmation, or otherwise, of my Telephone Directory method of allocating surname letters. I therefore did a similar test on the surnames of Members of Parliament sitting in the United Kingdom Parliament in 1981 (Whitaker, 1981). The figures for this test are shown in Table 1, columns 8–11. The result is some confirmation of the validity of my method in that, with chi-square for M.P.'s names of 16.41 and P = 5 per cent of 28.87, there is no suggestion of bias in M.P.'s selection by names, unlike that with the Fellows of the Royal Society.

Finally, as we seem to be approaching a time when we shall having voting in elections by proportional representation in this country with a single transferable vote, it is worth recalling the work of M. J. Mackerras in Australia where they already have this system. He has shown that the "donkey vote", that is people voting 1, 2, 3 4 down the list without any knowledge of the candidates, accounts for about 4 per cent of the voting. With voting compulsory this could be in part a protest vote but even so to supporters of P.R. like myself it is rather chastening that in a closely fought election this "donkey vote" could be enough to decide the issue.

Mr Brian Allt (Head of Market Research, Mirror Group Newspapers): As a guest, I would like to thank the Society and the speakers for a very interesting paper.

We have heard reference to the need for research into all the various factors that may affect response to questions. In the meantime, though, we have to go on designing questions—those of us who work in the field. A framework which I have found helpful to the way in which one thinks about the problem of question design involves an initially small change in the title of the paper. What I shall say is, I believe, operative for relatively straightforward behavioural areas where people have perfect recall of the occasion. Whether or not it is true in that extreme case, I think that it is certainly applicable in all the more complicated things, such as behaviour a long time ago, attitudes, political opinions, etc.

It seems to me that traditionally we have tried to impose the structure of the physical sciences on what we do in survey design as if there is a clear-cut observable entity there, and that if we could only break open the rock, we would find this fossilized fact, behaviour or attitude. It just does not correspond with
any of my experience. We make the mistake of talking about questions and responses, whereas we should talk about stimuli and responses. All our questions, even on the most simple things, are stimuli. What we get back are responses. The only objectification of what we think that we are measuring lies in the wording of the question—an imposed structuring.

Mr Hedges has already referred to the gap between the concept that the researcher thinks he is asking about and how he actually writes the question. We then have the gap between that and the way the interviewer administers it, followed by another gap between what the interviewer administers and what the informant understands. If we go from the researcher to the informant there is an enormous distance, and the sample's answer to any question, I would hypothesize, must represent the net result of a range of interpretations of the stimulus. If we get differences when we change a question or ask the same question in a different way, probably what that represents is a difference in the distribution of interpretations, and we are looking at the net result of that.

This may sound fairly pessimistic, but it suggests what one or two other speakers have referred to, that people should be asked much more about what they understand by the questions—what they think we want to know. Certain methods of piloting never really tackle that; they tot up the answers to one wording or another wording, but never get at why results differ. Piloting can be useful or useless, depending on how it is done. In a method I use, which I call structured depth interviewing, where I want to quantify, I ask virtually the same question several times for about 15 or 20 minutes, very slightly differently worded. I have found that different people start to "give", start to "flow", at different points in this sequence. That ties in with the point about longer questions and what Dr Belson said—give people a chance to think about the topic and you get different replies. If we flash a light in somebody's eye, we do not expect that everyone will blink at exactly the same number of seconds after that flash—there is a distribution of responses to the flash, and likewise of interpretation of questions.

For really important issues I think we may have to face up to something which is ostensibly very non-scientific, perhaps asking people how they would like to talk to us about the subject in which we are interested. Standard questions are not necessarily the best way to obtain equally accurate information from all people. Is it more important to have identical stimuli which satisfy some shackle-like goal of "scientific" rectitude or identically-based answers? In that sense, I think that perhaps we should sometimes reappraise our targets when surveys are being designed—what it is that we are trying to do and for what purposes, i.e. "just ask questions" or guide problem-solving.

The following contributions were received in writing, after the meeting.

Dr Aubrey McKennell (University of Southampton): The authors are to be congratulated in providing us with an account of the perplexing variety of question wording effects which is at once wide-ranging, succinct and insightful. Many people are going to want to refer to this paper for that reason alone. On its innovative side, two themes in the paper stand out for me. One is the emphasis, where non-factual or at least opinion questions are concerned, on multiple indicators; the other is the avowed aim to build and test theoretical structures. These two aspects are not only related but, in my view, need to be developed together. I am inclined to doubt whether satisfactory theories about the operation of single opinion items can be established otherwise.

It is now approaching forty years since the psychometrician Quinn McNemar (1946) published his classical review of experiments on the wording of opinion questions—an abundant literature even then—and came out firmly with the recommendation that single-item opinion gauging be discarded in favour of a multi-item scaling approach. Broadly speaking, that advice has been ignored by the fraternity of professional large-scale survey researchers, but has been followed and been the basis of many developments by academic researchers, notably psychologists. Attitude scaling methods have become increasingly sophisticated since McNemar's day. Techniques such as the semantic differential and smallest space analysis, for example, allow subtle variations in the connotations of verbal items to be precisely mapped. On the face of it, these and other developments in psychometrics would seem relevant to the study of issues in the wording of survey questions but it is surprisingly hard to trace any influence in the way such issues have customarily been investigated.

It could be that the traditional split ballot approach with its focus on one question variant at a time has obscured the possibility of contributions from the discipline of psychometrics. The psychometric focus is on series of items and on the item intercorrelations rather than their marginals, it is true, but the basic notions of true and error scores would seem transferrable. It ought to be possible in principle to compare question wording variants for reliability, in the technical psychometric sense, or even for
validity, in terms of their relative loadings on an underlying factor. Identification of the best question form by this means would achieve something which, as the authors note, split ballot experiments usually fail to do.

The application of psychometric principles in new areas is beginning to look more feasible with recent developments in structural equation modelling (cf. Alwin, 1979). It may be possible to construct models in which the overall bias in marginals is reflected in method effect parameters (that is in terms of correlated error), and to do this even where the assumption of “form resistant correlation” holds. The type of multi-trait multi-method experimental design required has already been applied with some success to the analysis of subjective social indicators. It proved possible to order the variants in rating scale format commonly used in surveys of subjective well-being in terms of their validity, random error and correlated error (method effects) components ( McKennell 1980). The findings help to counter at least some of the difficulties, mentioned by Kalton and Schuman, of comparing social indicators across a series of surveys. Work I have in hand with the SCPR Methods Centre is aimed at seeing if the new methodology can be combined with the split ballot technique and generalized to question wording effects. There are special difficulties which would take too long to outline here. Format effects are not quite the same as wording effects, and it remains to be seen if the latter can be investigated with a similar kind of experimental design.

There is going to be no royal road to the study of question wording effects. I suspect that progress will depend on breaking out of the mould into which investigations have been set by the first researchers in the field. The initial pioneering wave of intensive investigation petered out more than a generation ago as it became evident that understanding was not being furthered by yet more split ballot trials of single items. The research tradition established was remarkable for being so atheoretical. It is not just that past experiments have been isolated form each other; they have also, as a whole, been peculiarly insulated from developments in allied fields. Our present authors show sensitivity to such developments. They refer to recent work on cognitive theory and the slightly older psychological literature on acquiescent response sets. Besides these, and the developments I have myself cited, there must be many others which a more active scanning within such disciplines as psychology, sociology and linguistics would show to have a bearing. Older developments ignored by the pioneering generation of survey generation researchers might be looked at afresh. (One thinks for example of the knowledge accumulated by the early psychophysicists about judgements in an interval of uncertainty—in their experiments on differential thresholds, and on the way internalized frames of reference operate to determine a distribution of responses over verbal categories.) The authors deliberately choose not to extend their research efforts in such directions. Rather they start with a specific question wording effect and follow a piecemeal building-block approach toward its explanation in broader terms. It remains to be seen how viable this “frustratingly slow” approach will prove over the long term. The alternative approach would be to invest effort initially in the search for relevant explanatory frameworks established elsewhere, and then attempt to model examples of question wording effects in the light of these. It is perhaps too much to expect existing theories to be directly testable immediately, and I am not sure that the work of deriving testable models by this route is any quicker or less arduous than the approach adopted by the authors. But the dearth of good theories is a principal reason why research in this field languished in the past, and a more determined search for explanatory principles may be a precondition of introducing a fresh, sustainable momentum. In any case the two approaches are not inconsistent. The choice between them is perhaps best represented as one of balance or relative emphasis. I would hope that authors in their search for cause and testable hypotheses relating to the particular wording effect they have singled out for attention will study neighbouring fields more intensively than have previous researchers, and be prepared to consider experimental designs that go beyond the split ballot approach.

Dr Chris Scott (World Fertility Survey): Without criticizing the authors for what is a very useful review, I would like to raise a basic issue. It appears to me that in this, as in most, discussions of response error a key element has been ignored, namely that of the question objectives.

A central assumption in drafting questionnaires is that the question means the same to respondents as it means to its author. When we learn that the number of respondents agreeing with A may be substantially larger than the number disagreeing with not-A, this assumption is thrown into doubt, for in the language of the author of the question these two must be the same. This is not a matter of response variability or instability, although it is revealed by such instability. The problem is whether the questioner has succeeded in asking the question he wanted to ask. If he has not, then even if the response variance is zero and the relationships are perfectly form-resistant he is in trouble (or ought to be).
There is a tendency to overlook this issue in a systematic way. For a minor example in this paper, see the discussion on offering a middle alternative. The issue is presented as if we had two alternative ways of presenting the same question, with different results depending on the way selected, so that the matter is seen as one of question design. But if question authors were clear headed enough they would not regard the two questions as being the same and the issue would be one of their intention: do they want the expression of opinion with the middle view excluded, or not?

In much the same way, the attempt to overcome question instabilities by the introduction of multi-question attitude scales ignores the question of content: even if it works it yields an improved measuring instrument only at the expense of greater vagueness about what is being measured. That these issues are not generally perceived as crucial is perhaps a reflection of the fact that question authors usually don’t much mind exactly what they ask. This leaves the methodologist free to treat the issue as one of question design. But if the question author really doesn’t mind which of two versions he asks then his needs are so imprecise that he is unlikely to mind much about the variation between their responses discovered by the methodologist. Perhaps this explains the generally low level of interest in, and support for, methodological research from the survey users.

There are, of course, some users—election pollsters and household budget economists, for example—who do know almost exactly what they want. For them, research on question design is highly relevant.

Professor Seymour Sudman (University of Illinois): This is an excellent and thorough review of a very important topic. Response effects are typically the largest source of survey errors, and it is encouraging to note that the problem has obtained increased attention in recent years. I have no disagreements with what is said. My comments are intended to supplement some of the authors’ discussion.

Kalton and Schuman describe the use of filter questions to screen out respondents without opinions. Another alternative is to precede the attitude question with one or more knowledge questions on the same topic. This procedure is not new. It was developed and has been used extensively by George Gallup.

The use of properly constructed closed questions to obtain measures of attitudes is described as more valid than the use of open questions. It is more important, however, that this need not be the case for factual questions, especially if the topic is sensitive. Bradburn, Sudman and Associates (1979) found substantially higher levels of reported use of beer, wine and liquor and of sexual activity when open questions were used, rather than closed questions.

The authors note that, as expected, the use of a middle alternative clearly increases the proportion selecting a neutral view when three point scales are used. If one wishes to allow for a middle alternative, but to avoid bunching in the center, one solution is simply to increase the number of points on the scale to 5, 7 or 9.

One topic that is not covered since it is slightly tangential, but is of frequent concern to users of survey data, is the effect of method of administration. The generally encouraging findings are that responses do not differ significantly by method of administration for either factual or non-factual questions if the topic is not sensitive. For questions dealing with socially desirable behavior, there is some indication that less personal methods, mail and telephone, produce more valid information than do face-to-face interviews. For socially undesirable behavior, however, the threat of the question generally swamps any differences in method of administration. Even the use of sealed ballots and the other techniques mentioned by Barton (1958) do not produce satisfactory answers. Two methods that seem worthy of additional experimentation are the use of group administration of anonymous forms and the use of respondents to report about the behavior of close friends who remain anonymous.

The authors replied later, in writing, as follows.

We thank all the discussants for their valuable contributions. We are in broad agreement with many of their remarks, and will therefore confine our response to some brief comments on the issues raised.

Several discussants (Professor Holt, Mr Hedges, Dr Belson, Professor Smith and Professor Sudman) comment on the long way we have to go in understanding questioning effects, and we welcome their support for our view that the subject is a difficult but important one. We are pleased that Professor Holt, Dr Bynner and Dr McKennell emphasized the need to develop more general theoretical frameworks to explain the effects, which we consider to be the essential direction for future research. We are convinced that systematic research programmes are required for this challenging task.

The task ahead, however, should not obscure the fact that we have learned a good deal from past research, as Mr Hedges and Professor Moss note, even if it is mainly warning of major pitfalls. Yet, despite the guidance that past research provides, obviously poor questions are still prevalent on survey
questionnaires. These questions ought to be eliminated at the design stage but, as Mr Whitehead points out, insufficient time and money often lead to inadequate piloting. We feel that often too little attention is given to questionnaire design, and that many surveys would be much improved if more care were taken in this crucial stage of the survey process. We recognize Mr Hedges's point that the designer of a question often has a blinkered interpretation of it, and see value in his suggestion that a questionnaire should be reviewed by a panel of researchers. We also believe, like Professor Collins, that more thorough piloting is often needed, perhaps along the lines of the structured depth interviewing described by Mr Ailt, where that is practicable. A telling point in Mr Webb's account of his discovery of the problem with the television viewing question is that he carried out some of the piloting; a good case can be made that researchers should routinely take part in piloting, thus obtaining firsthand experience of how their questionnaire works in practice rather than relying on (or even neglecting) the reports of interviewers. All this developmental work must take into account that the questionnaire will be used by interviewers in uncontrolled and sometimes difficult settings; Mr Webb's suggestion of piloting questions by mumbling them to a slightly deaf respondent has much to commend it! As Dr Belson observes, realistic questionnaire design must also recognize that interviewers often deviate from instructions. Although more careful questionnaire design may help to reduce considerably the number of clearly bad questions, it will not address the kinds of effects discussed in our paper. Understanding of these effects can, we believe, come only from experimental research.

We are in full agreement with Dr McKennell that those attempting to develop theories for questioning effects should search the work of related disciplines like psychology and socio-linguistics to see what they have to offer. However, in our work we have not yet uncovered any ready-made theories that were directly transferable, and for this reason we do not anticipate speedy progress from such cross-fertilization. Perhaps a more fruitful line is to interest researchers in these disciplines in long-term collaborative research on questioning effects, to see if they can profitably apply their general theoretical structures and research methodologies to our problems. Our approach to theory development for questioning effects starts from known effects and attempts to identify theories to explain them; when a possible theoretical explanation arises, the next step is to test it by seeing whether its predictions hold, sometimes with other questions. As Dr McKennell notes, one could start instead with a theory from a related discipline, and construct questions to test its applicability in the questioning context. We prefer our approach partly because of the lack of any clearly relevant theory in other disciplines and partly because we want to stay close to the survey researcher's problems; however, it may be possible to make headway with the other approach. Another point Dr McKennell makes is that research on questioning effects may benefit from the application of psychometric methods in conjunction with structural equation modelling. There are some attractive possibilities here, and we look forward to seeing the results of this research he is conducting in collaboration with the Survey Methods Centre at Social and Community Planning Research.

Professor Collins, Professor Moss and Dr McKennell comment on the need to use a number of questions on an issue to minimize the effects of response artifacts. We fully concur that this should be done for the key issues investigated; however, there remains the problem of cumulative effects from items having the same form and also the problem of incidental issues on which the researcher can include only one or two questions. On the latter point, Professor Holt gives the specific example concerning the Springbok tour of New Zealand. It would be valuable to be able to estimate the proportion of the population supporting the tour, but care is needed in assessing the answers to a single question since both substantive and nonsubstantive changes in wording (or even placement) can shift univariate results markedly. Multiple items reduce the risk of misinterpretation and also handle the oversimplification of complex issues to which Professor Holt refers at an earlier point in his discussion. Given that several items are included in the questionnaire, there remains a question of how they are used in the analysis. Professor Moss and Dr McKennell argue in favour of scale construction, but Dr. Scott is negative about scaling on the grounds that it creates greater vagueness about what is being measured. We acknowledge that scale scores are more abstract than the answers to individual questions, but we believe that, with careful interpretation, the dimensions charted by scales can be fairly precisely identified, and therefore side with Professor Moss and Dr McKennell; however, we feel that due attention should also be given to the responses to the individual items, for these provide a down-to-earth feel for the data. There is no need to rely exclusively on either scales or separate items.

Professor Smith asks how one should choose between scales. The considerations we would suggest are reliability and construct validity, the latter being that on theoretical grounds the scale scores should relate to other variables in clearly defined ways. The chosen scale is then a reliable one that conforms best
to the theoretical expectations. However, in practice, useful tests of construct validity are extremely difficult to devise.

It is tempting to view Professor Holt's Springbok tour example as a sample referendum, but caution is needed here. As Mr. Whitehead says, and some results in the paper illustrate, many people are willing to give opinions on issues about which they know little or nothing. It is unclear whether these people would vote in a referendum and, if they did so, whether they would first acquaint themselves with the issue. One can, of course, filter out many of those who have no opinions by the procedures suggested in the paper or by knowledge questions, as Professor Sudman points out, but this is a matter of degree since different types of filters remove varying proportions of respondents. In addition, there may be a sizeable gap between an off-the-cuff survey response and a referendum vote. For a similar reason, care also needs to be taken in attempting to transfer Dr. Hansford-Miller's example concerning the possibility of an alphabetical bias in the election of Fellows to the Royal Society to the survey context; the election of several Fellows also differs from the usual survey tasks of selecting a single category or choosing an unrestricted number of all the categories that apply. (In passing, it may be noted that alternative explanations to the one offered by Dr. Hansford-Miller may be hypothesized to account for the association he reports.)

Mr. Whitehead provides a good example of how changes in a question can affect responses. The change from 1971 to 1972 of splitting the question into two parts appears to be a question form variation, but the later wording change from "limiting activities compared with most people of your own age" to "limiting activities in any way" is clearly a substantive change with a different question being asked. We need to distinguish between these two situations, as Professor Collins observes.

Dr. Scott argues that all question form effects in fact arise from substantive changes, since at least some respondents must be interpreting the questions differently. We agree that this is often the case, but certainly not always: effects due to the order of responses (such as Payne, 1951, gives) can scarcely be avoided by clarifying the investigator's goals. More generally, the distinction between question form effects and substantive effects is useful, even though blurred in places. The researcher can—or believes he can—identify the reason for a substantive effect, and hence he can choose the question to meet his objective, but this does not hold for a question form effect. An example where the distinction becomes blurred is the issue of offering or not offering a middle alternative, to which Dr. Scott refers: since the apparent question form change of including the middle alternative effectively modifies the substance of the question, the change may reasonably be viewed as a substantive one. As Payne (1951) advised long ago, it is probably better to offer the middle alternative if the objective is to discover definite convictions, but not to offer it if the objective is to find out which ways respondents are leaning.

As noted in the paper, the offer of the middle alternative can lead to a sizeable proportion of respondents opting for the neutral view. Professor Sudman points out that this bunching at the middle of a three-point scale can be reduced by increasing the number of scale points to five, seven, or nine. There still remains, however, the question of whether to use an even or an odd number of scale points (see Kalton et al., 1981, on the comparison of four and five point scales).

Dr. Bynner observes that biases in responses to factual questions may be of substantive research interest in their own right. It should be noted, however, that the biases themselves have to be measured—that is both the respondents' reports and the true values have to be obtained—if they are to feature in the survey analysis. Dr. Bynner's second point concerns the value of identifying subgroup differences in questioning effects as a way to obtain a greater understanding of the effects. This is an important point, and in our work we routinely look for differences by basic socio-demographic variables and also often by variables included in the experiment specifically for subgroup analysis—for instance, a measure of the salience of the issue. Disappointingly, however, subgroup analyses rarely yield positive findings, and it is for this reason that the paper reports so few examples of differential subgroup effects. But this absence also lends some support to the form-resistant correlation hypothesis that many survey investigators implicitly accept.

Dr. Belson argues that response to open questions should be used only to find out the range of answers that people give, with the alternatives identified then being built into closed questions for headcounting. In practice, however, responses to open questions are often classified and counted, and for this reason it seems useful to find out whether the distributions and classifications arising from the closed and open question forms coincide. Indeed, the absence of such comparisons has long made the debate over open vs. closed questions a matter of personal preference rather than scientific judgement. Professor Sudman draws attention to an interesting example where open questions obtained greater reporting than closed questions on the frequencies of certain activities which might be expected to be underreported, possibly
because the frequency classification offered with the closed form suggested the norms for the activities. It would be useful to conduct additional experiments to further explore this point.

Dr. Belson raises two issues of terminology. In the first place he is unhappy about our classification of opinion questions as non-factual questions. Since responses to opinion questions cannot be validated against an external source, it seems natural to classify them as non-factual questions according to our usage of the term. In this connection we might take the opportunity to emphasize that we used the factual/nonfactual distinction as a simple division of questions for purposes of organizing our material; we were not attempting to devise a comprehensive question classification which would be a substantial task in itself (see, for instance, the discussion of Rothwell to Kalton and Schuman, 1980). Dr. Belson also considers it unfortunate that we use the term ‘long’ to describe the questions lengthened by redundancies, as employed in the experiments by Cannell and his co-workers. It seems to us, however, that ‘long’ is the appropriate adjective to describe such questions, whereas the kinds of question Dr. Belson would describe as “long” should more accurately be described otherwise, perhaps as “complex” or “long and complex”.

Mr. Webb draws attention to the possibility of regional variation in the understanding of survey questions. Like him, we are unaware of research on this topic, and we agree that it warrants investigation. Some of the interest in “Black English” in the United States suggests similar problems along ethnic or social lines.

In order to bring out the fact that survey responses are affected by a range of factors in addition to the question asked, Mr. Allt recommends that we talk about stimuli and responses rather than questions and responses. We acknowledge the importance of interviewer and other effects on responses, but we still consider it useful to study the question separately, as one component in the data collection process. Mr. Allt goes on to pose questions of whether we should have identical stimuli or identically-based answers. While we can see that a case may be made for some limited variation in questions to obtain more valid factual responses from different classes of respondent (a procedure made more feasible by computer-assisted telephone interviewing), we consider that the procedure would be used extremely cautiously and only after preparatory research has documented the comparability of responses. In statistical surveys, and particularly with opinions, it seems to us essential to insist on a high degree of standardization of stimuli so that response can be aggregated in a meaningful way for quantitative analysis. We fail to see how identically-based answers could be obtained by survey interviews using different stimuli, and how to assess the comparability of responses obtained this way.

Professor Sudman usefully draws attention to one consoling and somewhat surprising finding in this difficult area; namely that for nonsensitive issues responses do not seem to differ significantly by mode of administration. This is an important point in view of the current interest in mixed modes of data collection.

In conclusion, we would like to provide Dr. Belson with the assurance he requested that we have critically appraised the material that we presented in this paper. It was not possible to supply detailed accounts of all the studies cited, and for these the reader must turn to the original publications. We can, however, say that we relied mainly on large, well-conducted, studies, and the findings have often been reinforced by one or more replications.

REFERENCES IN THE DISCUSSION


BRITISH TELECOM TELEPHONE DIRECTORY (1980/1981). Birmingham Area, Section 211; Canterbury Area, Section 282 (Alpha); Exeter and North Devon, Section 293 (Alpha); Leeds Area, Section 282 (Alpha); Exeter and North Birmingham Area, Section 211; Canterbury Area, Section 282 (Alpha); Exeter and North Devon, Section 293 (Alpha); Leeds Area, Section 232 (Alpha); London Postal Area, Sections 101–104; Manchester/NE and City Centre, Section 264; Glasgow Area, Section 275, Cardiff and SE Wales, Section 301 (Alpha); and N. Ireland, Section 241 (Alpha).


Discussion of the Paper by Dr Kalton and Dr Schuman


As a result of the ballot held during the meeting, the following were elected Fellows of the Society.

<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAVIES, Peter M.</td>
<td>INGRAM, David M.</td>
<td>TURNER, Keith</td>
</tr>
<tr>
<td>DHARMALINGHAM, Thiru M.</td>
<td>RUSKIN, Heather J.</td>
<td>WALLEY, Peter</td>
</tr>
<tr>
<td>FREEMAN, David R.</td>
<td>THOMAS, Roger K.</td>
<td>WHITAKER, John J. M.</td>
</tr>
</tbody>
</table>

As a result of the ballot held during the meeting of October 14th, 1981, the following were elected Fellows of the Society.

<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALI, Asghar</td>
<td>ELLERAY, Elaine A.</td>
<td>NESS, Mitchell R.</td>
</tr>
<tr>
<td>ALLOTEY, Charles A.</td>
<td>FOCHTMANN, John A.</td>
<td>NICHOLAS, Timothy R. M.</td>
</tr>
<tr>
<td>ANDERSON, Bernard G.</td>
<td>FOTOPoulos, Stergios</td>
<td>PEWSEY, Arthur R.</td>
</tr>
<tr>
<td>ANGHILERI, Roberto A.</td>
<td>FRANE, James W.</td>
<td>POON, Fun C.</td>
</tr>
<tr>
<td>BIRKHEAD, Brian G.</td>
<td>GRAY, Christopher T.</td>
<td>RAFFE, Najib M.</td>
</tr>
<tr>
<td>BROWN, Peter J. B.</td>
<td>GRIFFITHS, Caroline L.</td>
<td>REGAL, Ronald R.</td>
</tr>
<tr>
<td>BUTLER, Ronald W.</td>
<td>HILL, Barry</td>
<td>SANDBACH, Jonathan</td>
</tr>
<tr>
<td>CHAPPELL, Roma L.</td>
<td>HILL, Peter R.</td>
<td>SMEDLEY, Peter J.</td>
</tr>
<tr>
<td>DARZENTAS, John</td>
<td>KIDD, Eileen P.</td>
<td>STEPHENS, Helen J.</td>
</tr>
<tr>
<td>DAVIES, Martin V.</td>
<td>KIRBY, Simon P. J.</td>
<td>WATSON, Gordon J.</td>
</tr>
<tr>
<td>DAVIES, Peter T.</td>
<td>KROLL, Mary E.</td>
<td>WHARTON, Ann</td>
</tr>
<tr>
<td>DAVISON, Anthony C.</td>
<td>McGIVERN, Kevin</td>
<td>WILSON, Ian S.</td>
</tr>
<tr>
<td>EASTBROOK, Gillian A.</td>
<td>MORRIS, Alfred C.</td>
<td>WORRALL, Leslie</td>
</tr>
<tr>
<td>EL-HELBawy, Abdalla T.</td>
<td>MYLVAGANAM, Arunthathi</td>
<td></td>
</tr>
</tbody>
</table>

This content downloaded from 140.180.251.64 on Wed, 8 Oct 2014 14:13:27 PM
All use subject to JSTOR Terms and Conditions