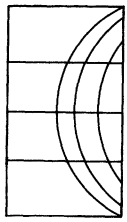
Questionnaire Design

Current Sociology 1998 46: 7
DOI: 10.1177/0011392198046004003



Questionnaire Design

More than three decades after its publication in 1951, Stanley Payne's *The Art of Asking Questions* was described as 'the best book on question wording that there is' (Marsh, 1982: 168). This 'classical' tradition of questionnaire design, discussed in the first section of this chapter, is more than a historical curiosity for it remains the foundation of current survey research. Not until the 1980s was there a second generation of survey methods texts, notably Seymour Sudman and Norman M. Bradburn's (1982) *Asking Questions* and Jean Converse and Stanley Presser's (1987) *Survey Questions: Handcrafting the Standardized Questionnaire*, which incorporated methodological research initiated in the mid-1970s. The second section of this chapter describes this effort to establish a stronger empirical basis for questionnaire design. Beginning in the mid-1980s, a third generation of survey methodologists began to apply cognitive psychology to survey design. The publication of Sudman, Bradburn and Schwarz's (1995) *Thinking About Answers* signals the maturation of this newest approach and it is the subject of the last section of this chapter. These titles mark the changing conception of questionnaire design, from a skilled art, through an earnest, research-based craft to a more complex, but less definite, contemporary approach. Like any intellectual periodization, this demarcation sets sharp lines on the fuzzy boundaries of slowly changing practice.

Craft-Based Questionnaire Design

Survey Items and Respondents

For the first generation of survey researchers, *individual* questions measured real mental states, and answering questions involved the *retrieval* of pre-existing, stable information from respondents' minds. Their expertise lay in identifying and formulating questions about the shared mental dimensions of the survey population and offering answers corresponding to the range of

Current Sociology, October 1998, Vol. 46(4): 7–47 SAGE Publications (London, Thousand Oaks, CA and New Delhi)
[0011–3921(199810)46:4;7–47;005650]

positions on this dimension. Of course, not everyone had an opinion on every subject, and the amount of non-response was routinely reported. Payne offers a touching and modest self-assessment of his trial-and-error-based craft:

> . . . this little book was not written by an expert in semantics, not even by a specialist in question wording. The author is just a general practitioner in research. Having made more than my share of mistakes in phrasing issues for public consumption and feeling the need for a book on the subject, I found that it was necessary to write it myself . . . the reader will be disappointed if he expects to find here a set of definite rules or explicit directions. The art of asking questions is not likely ever to be reduced to easy formulas. As it stands, this book consists of some observations of human behavior, a few principles of wording, many exceptions to these principles, several unexplained oddities, and numerous unsolved dilemmas. It is undoubtedly richer on the how-not-to side than on the how-to side. (Payne, 1951: xi)

Skill and experience had their limits, however, and Payne advised: 'One last recommendation . . . Controlled experiment is the surest way of making progress in our understanding of question wording. Never overlook an opportunity to employ the *split-ballot* technique' (Payne, 1951: 237, 73–4). Split-ballot experiments in which (usually just two) different versions of a question were asked of different parts of the sample provided an empirical means for solving problems in questionnaire design. Experimentation was not confined to academic research but was used more routinely, for example in Elmo Roper's polls for the American business magazine *Fortune* in the late 1930s.[1] Small-scale, qualitative methodological inquiries were also common, particularly using follow-up questions to probe how questions were interpreted. Cantril's (1944: 4) discussion of questionnaire design is based on '40 careful and intensive interviews made by three especially skilled interviewers'.

The 50th anniversaries of the *American Association of Public Opinion Research* and the *International Journal of Opinion and Attitude Research* were in 1995 and 1997, respectively. And, about 50 years ago, a combination of craft experience, experiments and small-scale studies had produced the commonsense guidelines that remain the core of good questionnaire design.

Commonsense Guidelines

Careful Wording Perhaps the early survey researchers' most important realization was that many common words were easily misunderstood or ambiguous. Even if interviews had conversational style, survey questions required much greater precision than everyday conversation. Seemingly inconsequential changes in a question, it was found, could affect the distribution of answers; and there was often remarkable variation in the answers to differently formulated questions about the same issue (Blankenship, 1940; Cantril, 1944; Payne, 1951: 7). For example, 'adding a law to the [American]

Constitution' was much less threatening, and public approval was 10 percent higher, than 'changing the Constitution' to achieve exactly the same result (Cantril, 1944: 44). Different versions of a question about allowing defence production workers to strike over working conditions produced between 17 and 45 percent positive responses (Rugg and Cantril, 1944: 46).

Even innocuous efforts to clarify a question could change the distribution of responses. Attaching a political leader's name to a government policy, using stereotyped phrases ('going to war' rather than 'using the army'), and putting questions in terms of newspaper cliches all affected responses. While such changes typically resulted in only small shifts in the response distributions, perhaps by 5 or 10 percent, the 'tone' of a question could have a much larger impact. A well-known example is Rugg's (1941) finding that 21 percent of Americans would '*not allow* public speeches against democracy', although 39 percent believed that such speeches should be '*forbidden*'. Rugg and Cantril (1944) report numerous other examples.

Payne's book is largely an exposition of how to craft questions to minimize ambiguity. An entire chapter is devoted to the meaning and potential misunderstandings of individual words, another provides a list of 1000 common words indicating which were most troublesome, and a third chapter is a detailed discussion of the formulation of a single question. The elementary technical requirements of question design were also well understood at the time. Researchers were advised against 'double-barrelled' questions (which dealt with two different issues), that responses should cover the entire range of the possible answers, and so on.

Question Form Survey researchers developed a variety of ways to ask questions: direct questions with yes/no or agree/disagree answers; 'sentence completion' questions where the respondent would choose among two or more responses (would you say that. . ., or that. . .); rating scales such as 'Likert' items (with answers strongly agree, agree,. . .); evaluative scales (excellent, good,. . .); and numerical scales. The different formats were used according to circumstance and there was no consensus on their relative merits.

The kind, number and range of response categories affected the distribution of answers. For example, respondents were much more likely to choose a 'middle' position – supporting the status quo or neither agreeing nor disagreeing – when this option was mentioned explicitly, rather than being recorded only if brought up by the respondent. A greater number of response categories produced more reliable measurement, though the choice of response categories and the order in which they were presented could also affect the responses (Cantril, 1944: 30ff.).

Whether a question was put abstractly or 'personalized', to sound like the respondent's own opinion, also affected responses. For example, Rugg and Cantril (1944: 47) found that 30 percent of Americans thought that

people 'should object' to reporting their income in the census and another 6 percent had no opinion, but only 22 percent objected to reporting their own income and just 1 percent had no opinion.

The magnitude of the differences between alternative formulations of a question was known to be a function of the topic and the characteristics of the respondent. Offering a 'status quo' answer in addition to polar responses, for example, *generally* encouraged respondents to temporize, but to a much greater extent if there was a smooth distribution of opinion than if opinion was sharply divided. Also, the form of the question was thought to have less impact on respondents with 'well crystallized' opinions (Rugg and Cantril, 1944: 33).

Intensity There was *some* recognition that respondents' principal evaluations of a question – whether they agreed or disagreed, approved or disapproved, and so on – might not fully capture their opinions. Katz (1944) referred to the 'intensity' of opinion, by which he meant how firmly respondents held to their views. Intensity was thought to be particularly important in political polls, where the momentary allegiance of voters might be less important than the proportions of durable supporters of each party's position. Respondents with stronger opinions were found to give more consistent answers to logically related questions.

To measure intensity Katz used both survey questions and interviewers' judgements of the respondents. One such measure was just the standard choice among five responses – strongly approve, mildly approve, no opinion, mildly disapprove and strongly disapprove – implying that attitudes are unidimensional, and measuring their 'intensity' requires only a graduated alternative to yes/no answers. The other measures, however, differentiated a respondent's evaluation from the intensity of her or his opinion. One question asked respondents how strongly they felt about a question, using a 100-point 'opinion thermometer', calibrated from 0 for 'don't care' to 100 for 'very strongly', with the mid-point, 50, labelled 'fairly strongly'.

Open and Closed Questions Later disagreements over the relative advantages of open and closed questions are a weak echo of an earlier, heated debate over what were essentially different *kinds* of research. Converse (1984c) gives the following description of the wartime surveys directed by Rensis Likert,[2] whose 'fixed question/free answer' technique:

> . . . was designed to be as close as possible to a natural conversation, featuring most questions in open format that (1) *suggested no alternatives*, answers to which were recorded as (2) *verbatim transcription*, or as close to that as interviewers could get. The (3) *probing for detail and clarification* was the most artful feature of the technique, requiring detailed knowledge of questions, which in turn made (4) *interviewer training* especially important. Training was adapted

> from Rogerian nondirective clinical psychology. . . . Interviewers charged with the open style were generally (5) *well educated*, with college or professional degrees. All the abundant narrative material they collected required (6) *coding* of a more elaborate kind than closed questions required. (Converse, 1984c: 271; emphases in original)

In a noted 1944 article, entitled 'The Controversy over Detailed Interviews – An Offer for Negotiation', Lazarsfeld was sympathetic to open questions. But, in the long run the use of open questions declined, largely due to their cost (on average, three closed questions can be asked in the time required for each open one), the greater demands on interviewers and the slowness and cost of processing and classifying the responses. Converse (1987: 367) notes that, in the USA, the more academic Institute for Social Research (ISR) at the University of Michigan was more favourable to open questions than the more commercial National Opinion Research Center (NORC).

For the most part, the relative advantages of open and closed questions were considered in a more limited context. Open questions were mainly used when there were too many answers to read out, where the range of answers was not known beforehand, and in pretests in order to determine which response options should be used in closed questions for a following, larger survey. It was not routine to 'probe' beyond an initial answer and the responses were usually classified into categories, rather than being used to gain insights unavailable from closed questions. Quotations from open questions were sometimes used to illustrate quantitative survey results. While two-thirds of the questions used in the landmark ISR study 'Americans View Their Mental Health' were open, Gurin et al.'s (1960) 400-page report does not include a single quotation, only the results of classifying the responses into categories.[3]

There was also concern with the *quality* of the responses to open questions. Payne (1951: 51ff.) cites interviewers' difficulty in uniformly recording the answers, the greater number of ambiguous responses, the tendency of some respondents to seize on peripheral issues, variation in the complexity of answers and the number of points made in a response, and the difficulty of classifying the answers. Little was made of Cantril's (1944: 38) interesting observation that, 'when opinion tends to be vague and unstructured on a difficult question' the *distributions* of answers to open and closed questions could be quite different.

Question Order Since the questions in a survey must be placed in some order, discovering that the order of questions affected responses seriously threatens the enterprise. Researchers were aware of the potential for 'order effects' and tried to discover when they were likely to appear. Cantril (1944: 28–9) described and provided illustrations of three such circumstances: consecutive questions that invoked a norm of reciprocity; situations where earlier

questions provided information relevant to later ones; and questions with overlapping content. Whether the order of questions had any effect, however, depended on the circumstance, 'It should not be concluded, of course, that the placement of a question *invariably* affects the response. Many negative instances could be cited, even of questions closely related in subject matter' (Cantril, 1944: 29). There was, however, no systematic investigation of this problem.

Summary By the mid-1940s survey researchers had developed commonsense approaches to the central problems of questionnaire design. Good design required, first, careful attention to the wording of questions, especially avoiding ambiguity. A stock of experience within the profession, much of it uncodified, provided examples of the kinds of questions that 'worked' and, perhaps more importantly, cautionary tales about questions and approaches that did not. There was an understanding of the need for a series of questions, not just one, to measure opinion on a topic adequately. The main risk, researchers believed, was not that a single, poorly formatted question would produce misleading results or that the format of the question would affect the response in an unforeseen way, but that just one question could not address the variety of ways of thinking about almost any topic.

There were quasi-rules of question design, of the kind now found in elementary methods texts. These had mainly to do with avoiding mistakes and did not add up to serious advice about exactly how to write good survey items: questions should not be too long and should deal with only one topic, alternative answers should be exhaustive and mutually exclusive, and so on. Among the methodological issues addressed in later research, only non-response was not seriously considered. It was not assumed that a respondent could answer every question, nor were questions with low non-response necessarily superior (Payne, 1951: 23–4).

Where generic advice and past experience were not sufficient, split-ballot experiments provided the means to resolve design problems. Which issues in questionnaire design were regarded as problematic can be seen in the following recommendation by Rugg and Cantril:

> The split-ballot technique should be used wherever possible in order to test the stability and consistency of opinion by noting the effect of (1) variation between free and prescribed responses, (2) variation of alternative answers presented, (3) variation of contingency surrounding the issue, (4) bias by an introductory statement or deliberately forced wording, (5) prestige introduced by a word, name or phrase, (6) explicit rather than implicit presentation of alternatives, (7) stated deviation from currently accepted practice [of question design], (8) variation of the content provided by other questions on the ballot, or (9) personalization of the issue. (Cantril, 1944: Ch. II, 50)

Experimentation is not only prescribed as a means to design better questions.

Rugg and Cantril also thought of differences in the answers to alternative formulations of a question as additional information about the survey respondents.

Survey Items, Scales and Disciplinary Perspectives

In a lengthy and influential critique of survey research, the prominent American psychologist Quinn McNemar made this distinction: 'The typical attitude study involves a scale or battery of questions for ascertaining attitudes whereas the typical opinion, particularly public opinion, study leans heavily on a single question for a given issue' (McNemar, 1946: 290). Because the reliability of individual questions could not be measured and the risk of bias prevented a literal interpretation of the proportion of respondents giving each answer, McNemar called on researchers to abandon 'opinion' research based on single items in favour of multi-item 'attitude' scales (McNemar, 1946: 316). Even with scales made up of several questions, he believed that response errors limited measurement to rank ordering individuals, rather than locating them absolutely. McNemar was also very critical of the disregard for the external validation of scales (McNemar, 1946: 310) and criticized the excesses of the 'pollsters' of his time, Roper and Gallup in particular.

Many survey researchers of the time did not see the incompatibility between interpreting individual questions as literal expressions of opinion and compensating for idiosyncratic elements of individual questions by forming scales from a group of questions. Taking on McNemar's argument, Lazarsfeld (1954: 349) differentiated the 'procedure of itemized tests' from the detailed analysis of individual survey questions. As conceptualized by psychometricians, individual items were imperfect indicators of underlying *continuous* traits revealed by analysis of the correlations between items. The variance in each survey item could be divided into *common* ('trait') and *unique* variance. In Lazarsfeld's view, the unique variance, which is the information that an item does *not* share with other items, is just as important as the common variance, taken by psychologists to measure a trait. Though perhaps not explicitly, in emphasizing the analysis of the answers to individual questions sociologists dissented from the psychometric view that less stable and thoughtful 'opinions' on particular topics could be understood in terms of a small number of more fundamental traits.

In considering procedures for combining survey questions to obtain a more reliable and valid composite measure, a distinction can be drawn between Thurstone (1928) and Guttman (1944) scaling methods and psychometric methods such as Likert (1932) scaling and factor analysis. Thurstone's and Guttman's procedures assigned respondents to discrete, substantively meaningful categories. Thurstone scaling involved the formulation of a finely graduated series of positions on an issue (often 20 or more), which were placed in order on the basis of experts' ratings. A respondent's

position on the scale scores was simply the rank order of the statement closest to her or his position (the median of the ranks, if the respondent agreed with two or more statements).

In Guttman scaling, the ordered categories of a single dimension of opinion are derived from the pattern of responses to a series of items of varying 'difficulty'. Respondents were assigned a scale score, which was actually an ordinal rank, on the basis of the most difficult question that they answered correctly, or strongest statement of opinion that they endorsed. The survey data, rather than external raters, were used to design and evaluate the scale. In order to decide whether a set of items actually represented an underlying *single* dimension, it was necessary to determine whether respondents giving a positive response to one item also responded positively to 'easier' items. Invariably, some respondents were 'non-scale types', who 'incorrectly' answered at least one question that was 'less difficult' than the most difficult question they answered correctly. If the scale properly represented an underlying dimension of ability or opinion, the proportion of such respondents was small. The more difficult problem, that was beyond the statistics of the time, was how to specify and test a 'latent trait' model that could separate random errors in the answers to questions (resulting from question ambiguity, carelessness and so on) from errors due to the inclusion of inappropriate items in the scale. Thus respondent error limited the number of items that could be used in a Guttman scale, and so the precision of a scale (since the number of separate categories in the scale was just the number of items plus one).

Likert's method of scaling also began with the correlations between the survey items (to compute Pearson correlations ordered verbal responses were simply assigned scores of 1, 2, 3, etc.). After removing any unsuitable items, identified by their low correlations with the other items, each individual's scale was computed by adding the scores.[4] One desirable feature of Likert scales was the tendency of their distributions to become normal, as the number of items in the scale increased (thanks to the central limit theorem). Researchers set about designing Likert scales for many, many 'traits' and large catalogues of measures for particular areas, such as work experience, were published. This work was uneven: scales were seldom validated with independent evidence; often items were selected and scale reliabilities measured for small and atypical samples; and little attention was paid to the possibility that the correlations between items, and hence the measured reliability of a scale, might be inflated by similarities in the format of items, respondents' desire to answer consistently, or carelessness. The sometimes inappropriate emphasis on the relative positions of respondents, rather than their actual experience or expressed opinions, gave us a generation of articles about which kinds of respondents were low, medium and high on various 'dimensions'.

For survey researchers using individual questions to locate respondents

in discrete states, a simple and appropriate means of data analysis is cross-tabulation. Rosenberg's (1968) influential *The Logic of Survey Analysis* is concerned entirely with cross-tabular models.[5] Percentage differences between groups were used to measure the effects of variables and χ^2 significance tests were used to determine whether observed effects could have resulted from random error. Without a statistical model of contingency tables, provided by log linear analysis many years later, it was not possible to test whether the relationship between two variables (say education and income) was significantly different for two groups (say women and men). Also, a more appropriate strategy for scales computed from a series of survey items was analysis of variance and regression.

By the mid-1950s interest in the general principles of question design had largely disappeared. Hyman et al.'s (1954) study of interviewing could serve as the bookend to this period. This is testimony to the achievements of the very small number of researchers who, by the end of the 1940s, had created *the* method of survey research, and provided a practical basis for its institutionalization.

Systematic Experimentation and Survey Error

In the early 1970s Howard Schuman and Stanley Presser initiated a series of experiments aimed at providing a systematic empirical basis for question design.[6] The idea was to discover regularities in the effects of different elements of the format of questions. Using over 200 separate 'split-ballot' comparisons between alternative questions spread over about 30 surveys, they replicated some of Cantril and Rugg's experiments from the 1940s and added many new examples comparing open and closed questions, questions with and without explicit 'no opinion' and middle responses, and questions with a different 'tone', and so on. Schuman and Presser were not content to find unusual examples demonstrating the effects they predicted. For each characteristic of a survey question investigated, they employed a variety of questions on different topics, in order to determine whether their results were contingent on the question topic and to discover any variation in the magnitude of the methodological effects. Other work in this genre includes Bradburn and Sudman (1979) and Kalton et al. (1978).

Formal Models of Survey Response

Before describing this experimental work, an account of the new, *statistical* perspective on survey data is necessary; it is also relevant to discussion of data collection in the next chapter. One begins with the distinction between sampling error, which is what classical statistics is about, and *non-sampling error*,

which arises out of the process of survey measurement. Non-sampling error is additional to the sampling error that arises from random and systematic differences between the sample and population. Building on more limited models, such as Fellegi's (1964) conceptualization of interviewer effects, different sources of non-sampling error can be treated as *components of the variance* of each survey item. Each respondent is characterized by a 'true' value of the characteristic that the question is designed to measure. Over the population, the true values have a 'true variance', sometimes called the 'validity'. Each respondent's answer to the question (the *observed* value or 'observation') is made up of a combination of her or his true value of the characteristic and *non-sampling* error.

Groves (1989) broadens this model to the conception of 'total survey error', in which the non-sampling error in the responses to each survey question is conceived as the sum of the effects of the format of a question, the mode of data collection, the interaction between interviewers and respondents, respondent error and processing of the responses. Non-sampling error related to the survey question, in turn, arises from the tone of the question, the range of answers, the order in which the questions are presented and so on.

Heise and Bohrnstedt (1970) divide the non-sampling error in each survey question into three components: bias, random error and *in*validity. Questions affected by *invalidity* partly or even entirely measure a different trait than was intended by the designer. Invalidity is unaffected by the sample size and may give rise to seriously distorted findings. For example, if the respondents reporting a high level of participation in an election campaign are a mixture of actual participants, for whom the variance is valid, and people who feel they ought to have participated but did not, for whom the variance is invalid, then the resulting measure of 'participation' includes two different traits. Of course, this statistical framework involves a restricted concept of validity, since it is based only on finding unpredicted factors in a preselected group of variables, with no external criterion or even comparison to other variables in a survey.

Estimating the effect of each component of the survey error requires survey data for which the component varies. For example, in order to measure the non-sampling error that arises from interaction between interviewers and respondents one must compare the results of interviews conducted by at least two interviewers; and to estimate the errors arising from different formulations of a question one must compare groups of respondents who answered the different versions. Since it is generally not practical to estimate all the components of the variance at one time, the effects of individual sources of non-sampling error are usually estimated on the assumption they are independent. From a wide review of the literature, Groves (1989) reports on the magnitudes of non-sampling error arising from different sources.

The development of latent variable models by Jöreskog and Sörbom (discussed in more detail in the chapter 'Quantitative Analysis of Survey Data') was a powerful stimulus to this statistical approach to non-sampling error. An elegant application of these 'LISREL' models (for *li*near *s*tructural *rel*ations) is the analysis of 2115 questions from six different surveys by Frank Andrews (1984). Following the logic of Campbell and Fiske's (1959) 'multi-trait multi-method' technique, the surveys included several measures of several different traits (to 'identify' the models, three or more different measures of three or more traits are required) and the analysis involved comparing the correlations between different types of questions *on the same topic* with the correlations between questions about different issues *in the same format*. To the extent that the 'same-topic' correlations are greater than the 'same-format' correlations, one assumes that the answers to a question reflect the topic and not how the question is asked. Andrews then examined the relationship between the validity, reliability and error in the individual survey items, estimated from the LISREL models, and their format, as measured by the number of response categories, whether a 'don't know' response was offered explicitly, the length of the question and its position in the survey, the topic of the question, the method of data collection, and so on.

Andrews found that an average of 66 percent of the variance in the responses was valid variance, which measured the desired trait, and 28 percent was residual error – random in the context of the model, but related to imprecision in language, respondents' varying interpretations of the boundaries between adjacent response categories, and so on. Only about 3 percent of the variance was related to what the researcher actually controls, which is the format of the survey question. This suggests that improvements in the format and style of survey questions have relatively little impact on the quality of survey data. Of course, there were statistically significant, mostly common-sense, effects of question format. For example, on average, moderately long questions had greater validity than shorter or longer questions. Probably, the shorter questions provide respondents with too little information to produce a consistent understanding of the issue, while longer questions tend to be difficult and confusing. These results make for sensible, general guidelines, but these effects are not large enough to override substantive considerations, in this case researchers' judgements about how much needs to be said to ask a meaningful question. There is also a way in which Andrews' findings cannot capture the importance of good question design. The questions that he analysed were of many different types, but all had passed through the rigorous design process and pretesting of a premiere survey research organization. So his figures for the level of error must come close to the best results that can be achieved consistently. Less skilled and careful design would produce worse results. Also Andrews' figures for the validity of questions are likely overestimates, because they do not involve external comparisons – only a

determination of the amount of response variation when the same question is formulated in different ways.

Molenaar (1991) employs a more limited meta-analysis strategy in which he characterizes a large pool of items from Dutch surveys in terms of their response distributions, including the amount of non-response and the mean and standard deviation of the responses. These outcomes are then regressed on a large number (34) of the characteristics of the items, including their substantive content, the complexity and other aspects of their wording, and their position in the survey. His results, such as the finding that more positive reference to one side of a question affects respondents, are reasonable, but quite obvious.

Opinions and 'Non-Opinions'

Because of the impact of random guesses on survey results and also because having an opinion may be as significant as the polarization on different sides of an issue, determining which respondents have any opinion on an issue is important. Based on analysis of a US election survey, Philip Converse (1964) argued that only about half the population could rely on even the most general ideological framework in answering policy-related questions; and that other half answered in an idiosyncratic, if not completely random, manner. The ensuing debate centred on the interpretation of inconsistencies in the response to substantively related questions, either for two related questions from the same survey or from longitudinal surveys in which the same question is asked on two or more occasions. Ambiguities in the wording of questions, his opponents argued, could account for enough measurement error to result in the low correlations between questions, which Converse had interpreted as the effect of respondents answering inconsistently.[7]

In a number of experiments, Schuman and Presser (1981: Ch. 4) found that many more respondents will say they have no opinion on a subject than will give a 'no opinion' answer if this alternative is not mentioned explicitly. It was not unusual for 20–30 percent of respondents to say they had no opinion, even though only 5–10 percent did not answer if asked the question directly. This is consistent with the idea, though not proof, that respondents avoid 'no opinion' answers because they do not wish to appear ignorant or uncooperative.

Following up on Gill's often cited finding that 70 percent of respondents voiced an opinion on the fictitious 'Metallic Minerals Act', Schuman and Presser (1981: Ch. 5) asked Americans about the real, but obscure 'Agricultural Trade Act' and 'Monetary Control Act'. About one-third of the population were prepared to offer an opinion on these Acts, with less educated respondents *more* likely to answer. Modifying the question to conclude with the phrase 'or do you not have an opinion on that issue', however, decreased the number of respondents with an opinion to 10 percent, and

stating an opinion was then unrelated to education. When asking questions that respondents are not likely to be able to answer, one should ask whether they have opinions. But it is not obvious what this implies for more typical questions, which are included in a survey because most respondents *are* expected to have answers.

In extensive overviews of a large body of research, Smith (1984b) and Krosnick (1991) find consistent evidence that non-respondents tend to be less knowledgeable about the topics of questions, less certain about their knowledge, and lower in formal education and other measures of cognitive skill. Faulkenberry and Mason's (1978) regression analysis of the number of 'no opinion' responses in three surveys showed that non-response is inversely related to education and interest in the issues in the survey. This also suggests that efforts to 'filter' out respondents with no opinions is a good practice that will produce better data.

There is a fly in this ointment, however. The assumption is that respondents who can be dissuaded from answering a question by a polite query about whether they have an opinion do not actually have opinions and would otherwise have answered randomly. Therefore, one would expect that the correlations between questions dealing with related topics would increase if the random responses were eliminated. But Schuman and Presser (1981: 128ff.) find that the *correlations* between survey questions *do not necessarily increase* when 'filter' questions were used to remove respondents who think they have no opinion. McClendon and Alwin (1993) computed two different versions of three multi-item scales, only one of which included the additional phrase 'or don't you have an opinion about that', which increased non-response from about 5 percent to 20 percent. Again discouraging responses from what are assumed to be uninformed respondents with the 'filtered' questions did not result in more reliable scales.[8] So, respondents who say they have no opinion apparently answer questions as consistently as those who say they do!

Sniderman et al. (1991) have an explanation of how respondents with little or no information about a topic can answer questions consistently:

> . . . Citizens frequently can compensate for their limited information about politics by taking advantage of judgmental heuristics. Heuristics are judgmental shortcuts, efficient ways to organize and simplify political choices, efficient in the double sense of requiring relatively little information to execute, yet yielding dependable answers even to complex problems of choice. . . . Insofar as they can be brought into play, people can be knowledgeable in their reasoning about political choices without necessarily possessing a large body of knowledge about politics. (Sniderman et al., 1991: 19)

Explicit in Converse's work and implicit in much of the research on non-response is the conception of attitudes as things that respondents do or do not have. What Sniderman et al. suggest is that respondents are also

differentiated in terms of *how* they answer survey questions, and that different strategies are employed according to how much the respondent knows about an issue. At a minimum this suggests the importance of measuring respondents' knowledge about a survey as well as her or his opinions, but it also shows the need for a model of how respondents answer questions.

The Tone of Questions

Repeating Rugg's (1941) experiment in 1974 and 1976, Schuman and Presser (1981: 281, 283) found that Americans were still more likely to 'not allow' than to 'forbid' speeches in favour of communism; in 1976 the difference was 25 percent. Questions about 'speeches against democracy', 'showing X-rated movies' and 'cigarette advertisements on television', however, produced quite different results. While about 15 percent more Americans were willing to 'not allow' than to forbid speeches against democracy, for X-rated movies and cigarette advertisements the difference was only about 5 percent. Of course, this pattern can be interpreted in terms of the particular issues: forbidding and not allowing cigarette advertising amount to the same bureaucratic imposition of government regulations on television stations; but forbidding speeches hints at police action, arrests and violence, while 'not allowing' speeches implies polite and successful deterrence.

Schuman and Presser reasoned that variation in the tone of questions would have less impact on more educated respondents, who would be better able to separate minor variations in terminology from conceptual differences. And, for 'speeches against democracy' there is indeed a steep decline in the difference between forbidding and not allowing such speeches, from 30 percent for respondents with 12 years of education or less, to 15 percent for 13–15 years of education, and down to 8 percent for 16 years of education or more. For speeches in favour of communism and X-rated movies, however, the difference in the response to the two forms is unrelated to education, and for cigarette advertisements the results are ambiguous.

There is also evidence that bias does not invariably result from using 'loaded' terms in survey questions. Schuman and Presser (1981: 186) report that Americans' opinions about abortion were unaffected by the inclusion of a reference to 'end[ing] the life of her unborn child' and opinions about gun control were unaffected by whether it was said to 'interfere too much with the right of citizens to own guns'. Having made up their minds on these controversial issues, respondents become insensitive to the exact formulation of the question.

Certain terms and phrases have the capacity to induce bias, but whether they do depends on the topic and context; and sometimes the impact of particular question formulations is related to respondents' level of education, though why is not clear. If broad generalizations about the effect of the tone of questions cannot be made, how to ask more neutral questions is more

properly an aspect of research in the different fields of social research than a general problem.

Open and Closed Questions

By the 1970s, the debate about open and closed questions had been long settled in favour of the latter (see Converse, 1984b). Still, no one had thought to compare the answers to open and closed versions of the same question until Schuman and Presser (1981: 79ff.) examined three questions dealing with the 'most important problem facing this country at present', what the respondents would 'most prefer in a job' and the 'most important thing for children to learn to prepare them for life'. These questions admit a very wide range of potential responses and they are somewhat ambiguous (what is the 'present'? how old is a 'child'?) and difficult (is there any, single most important lesson for life?) – exactly the circumstances in which open questions are used. There were remarkable inconsistencies in the answers to the corresponding open and closed questions, even accounting for the tendency for open questions to result in more missing data. For the question about jobs, furthermore, the difference remained after a careful effort to align the responses offered in the closed question with the most frequent answers to the open question. The inconsistencies in the response distributions of the open and closed questions suggest that respondents go about answering them in different ways, though larger differences between open and closed questions are much more likely to appear when the range of answers is wide than when alternatives are few and self-evident.

In the face of the greater efficiency and data quality (especially measured by the amount of non-response) of closed questions, these interesting findings had little impact on the bias against open questions. Open questions can only become more appealing when it becomes possible to analyse the text of answers in more detail, rather than merely classifying the answers into a small number of categories.

For 'retrospective' questions about previous events, there is strong evidence of the superiority of closed questions. For example, Sudman (1980) found that respondents who were given a checklist of alternative answers said they had read more newspapers and magazines in the previous day than respondents who were merely asked how many newspapers and magazines they had read. Similarly, researchers studying participation, in sports or community activities for example, invariably find that asking respondents to indicate the activities in which they had participated, read from a checklist, uncovers some activities that respondents forget when they are asked a single, general question. Not only is unassisted recall much more difficult for respondents than recognizing items from a list, but the list also describes which specific items qualify for inclusion, rather than leaving the respondents

the sometimes error-prone task of interpreting what may be a very general question.[9]

Closed Questions: Format and Responses

Schuman and Presser examined the effect of even-handedness on responses to questions. Was there any difference, for example, between asking whether respondents agreed with a statement – yes or no – and asking whether they agreed *or disagreed*? None, they found. Response alternatives that incorporated reasons for each answer, however, did affect the distribution of responses. For example, support for a law requiring people to turn down the heat in their homes at night in order to conserve energy was much lower when the negative response included a phrase saying that such a law would be difficult to enforce (Schuman and Presser, 1981: 187). What constitutes a substantive argument, however, can be subtle. Payne (1951: 7–8) found that 63 percent of respondents agreed that 'most companies that lay off workers during slack periods could arrange things to avoid layoffs and give steady work right through the year', but only 35 percent agreed when offered the alternative, 'or do you think that layoffs are unavoidable'. The innocuous, defeatist cliché shifts the question from whether companies are *capable* of avoiding layoffs to whether they are *likely* to do so. The questions about abortion and gun control, noted earlier, show that the impact of substantive response alternatives depends on the question. Differences in question formulation have more impact for issues that are not salient or on which public opinion is unformed.

A common situation in which the form of a question might affect survey responses involves which side of an issue is put forward for evaluation, the concern being that respondents are biased in the direction of agreement, referred to as 'acquiescence' or 'agree response bias'. A notorious example of this form of bias is the 'F-scale' measure of authoritarianism (see the account in Brown, 1965: 477ff.), in which respondents are presented with a long series of statements, many in emotional terms and involving broad generalities, to be evaluated using the categories of a 'Likert' scale ranging from 'strongly disagree' to 'strongly agree'. As the title of Schuman and Presser's (1981) chapter on 'The Acquiescence Quagmire' suggests, the empirical evidence on this point is quite mixed. Acquiescence appears only in some circumstances, particularly for difficult questions,[10] and *may* be related to respondent characteristics, particularly education and interest in the topic.

The inability to predict when acquiescence will affect survey questions has prompted survey researchers in many areas to switch from agree/disagree questions to the 'sentence completion' style, where respondents choose among alternative phrases for completing a sentence. It is mainly psychologists, and survey researchers in substantive areas where psychologists are prominent, who continue to employ scales based on agree/disagree ratings of

long lists of statements. Often this is because there is little concern with the responses to individual questions, so long as the items can be added to give scale values indicating the relative positions of respondents. This strategy does not, however, guard against the possibility that respondents are *differentially* subject to acquiescence. Respondents who find the questionnaire more difficult, who are more intimidated by the task (with weaker reading skills or less education?) or who are less interested may be more prone to this form of bias.

More Qualitative Response Alternatives Even when the alternative responses reflect a clear underlying dimension, the *range* of answers provided to respondents can change their meaning. Schuman and Duncan (1973–4) describe two versions of a question about Americans 'making changes in the way our country is run'. The first version gives four alternatives, asking if the respondent would: (1) *rarely*, if ever, make changes; (2) be *very cautious* of making changes; (3) *feel free* to make changes; or (4) *constantly* make changes; while the second version offers only the middle two alternatives. Logically, respondents who would 'rarely make' changes should prefer being 'very cautious' over 'feeling free' to make changes; and respondents who would make changes 'constantly' should prefer 'feeling free' to make changes over being 'very cautious'. When all four alternatives are presented, 32 percent of the respondents wanted to feel free to make changes and another 24 percent wanted to make changes constantly; but with just two alternatives, only 37 percent of the population preferred to 'feel free' about changes over being 'very cautious'. So, even for clearly ordered categories, respondents are affected by the number and range of the alternatives. Thus, rather than formulating a response to a question and then locating the answer that is closest, the meaning of the question is inferred from the combination of the question *and the alternative answers*.

With qualitative responses, especially where the alternative answers are complex, a concern is whether the order of the answers affects the response. The terms 'primacy' and 'recency' refer, respectively, to the tendency of respondents to choose the first and last answers in a series. Again, the results of experimentation were quite mixed. Statistically significant, though usually not large effects of response order were found in four out of 12 experiments (three recency effects and one primacy effect), but Schuman and Presser (1981: 56ff.) were unable to identify a consistent principle to identify the circumstance or even the direction of the effect.

Finally, there is evidence that the order of questions can affect the responses to complex and long questions, such one about parental values described by Krosnick and Alwin (1987). As survey questions become more abstract and difficult, there is a greater danger that respondents will make unforeseen connections between different items and that their answers will

be affected by the question format. The pressure on respondents to answer quickly exacerbates these effects as well as contributing to random error arising from misunderstanding and haste.

Quantitative Response Alternatives A question about television viewing is the numerical equivalent to the last example about making changes in the country. Schwarz et al. (1985) found that presenting respondents with alternative answers covering shorter periods of time (under 30 minutes, 30–59 minutes, etc.) resulted in lower estimates of the amount of time respondents watched television than if the lowest category was very wide (under two hours per day). What are intended to be 'neutral' response categories are again used as additional information by respondents, especially if the question is difficult or ambiguous (how much attention must a respondent pay to the television to be 'watching'?) or if the respondent is not very interested. Low response options can be taken to imply that the events in question are infrequent and that respondents should use stricter criteria in deciding what qualifies. Even the numbers attached to responses for data processing purposes and ranges from rating scales (negative numbers are more *negative)* can affect responses. The experiments with different response categories are clever, but the implication is that questions about durations, numbers of events and so on are best asked directly, without response categorized responses. Failing that, one ought to choose more, detailed response categories over fewer, broader ones.

A similar problem arises from the use of verbal responses to describe quantitative categories, terms such as 'often', 'sometimes' and 'many'. Bradburn and Sudman (1979: 157–9) first asked respondents how often they had a variety of different experiences in the previous month 'not too often, pretty often or very often'; and then asked just how many times the experience had occurred. They found that quite variable meanings were given to the categories, since respondents with the same number of events in mind often described themselves as being in different verbal categories.[11]

Unlike concrete questions, for example about the number of times an event has occurred, there is no standard, natural scale (or 'metric') that survey respondents can use to describe their attitudes. Instead, each question must also present a series of response categories from which a answer is chosen. Having too few categories will lump together respondents whose opinions are actually different; but having more categories than respondents can differentiate may result in more response errors and non-response. There are a few comparisons of the impact of the number of response categories for individual items, but only Andrews (1984) and Alwin (1992) conduct analysis combining items from a number of surveys (see also Ramsay [1973], and Cox [1980] and Molenaar [1982] for reviews). The consistent finding is that the reliability of survey questions increases as the number of response categories

increases, but at a diminishing rate. Above ten categories, there is little gain in reliability. The exception is that two-category questions are more reliable than three-category questions.[12] For non-numerical categories, the mode of data collection limits the number of response categories. Particularly for telephone interviews, respondents' ability to remember the alternatives sharply limits the number of possible responses (also, in face-to-face interviews 'show cards' can be used to present the responses).

Most of the work on questionnaire design is not done by survey methodologists, but by researchers developing solutions to their own problems. For example, many 'quality of life' researchers favour numerical scales with seven points or more, for responses to measure satisfaction with 'life in general', jobs, marriage and so forth (see Andrews and Withey [1976] for an exhaustive discussion). And political scientists studying political attitudes, efficacy and trust in government favour four-point scales (disagree strongly, disagree somewhat, agree somewhat, agree strongly) perhaps because they believe that a middle category might provide too much of a temptation to avoid thinking about the questions.[13] Such local research cultures often provide practical answers to specific problems, but may isolate researchers from relevant methodological findings.

Question Order

Schuman and Presser (1981) distinguished two different ways in which the order of questions can affect responses. 'Part-part' effects involve two or more questions at the same level of generality that invoke a norm of reciprocity or consistency. Replicating a 1948 experiment by Hyman and Sheatsley, they found that 55 percent of Americans agreed that 'the United States should let Communist newspaper reporters come in here and send back to their papers the news as they see it'; but 75 percent agreed if they were first asked the equivalent question about 'American newspaper reporters . . . [in a] . . . Communist country, like Russia'. There is nothing mysterious about this: having claimed a right for American reporters, respondents were less likely to deny it to Communist reporters. Consistency is in the mind of the survey respondent, however. In a small-scale survey, I found that respondents were *more* likely to say that trade unions were too powerful if they had previously been asked a question about the power of large corporations, and vice versa. Thinking that both unions and corporations are too powerful implies a consistent opposition to 'big' institutions, but it is not more logical than believing that corporations are too powerful and unions too weak, or vice versa.

'Part-whole *contrast* effects' on the answers to consecutive questions involve interpretations of survey questions that were not intended by the researcher. Schuman and Presser (1981: 36ff.) give the example of two questions dealing with abortion for women who are 'married and [do] not want

any more children' and for women who have learned that 'there is a strong chance of a serious defect in the baby'. Respondents were less likely to support legal access to abortion in the first case if the second, more specific question was asked first. Respondents tend to *exclude* the specific circumstance mentioned in the earlier question, so that the second question is interpreted as a situation in which the married mother of what is known to be a healthy foetus does not want any more children. Not only are effects of this kind extremely rare (Kalton et al. [1978] report another), but the opposite pattern may appear. Schuman and Presser (1981: 42) also describe part-whole *consistency* effects, in which asking a more specific question first tends to result in respondents answering a more general question in a similar manner. Unlike the part-part effects, which can appear regardless of the order of questions, part-whole contrast effects can be eliminated by placing more general questions before specific ones.

Alongside these interactions between specific, related questions, the general order of questions can affect how respondents answer questions. One such effect involves capturing respondents' attention. McFarland (1981) observed that survey respondents said they were more interested in politics after being asked a series of questions about their political views. Andrews (1984) showed that response errors were greater for the first one or two items in a group of questions in the same format, presumably as respondents try to understand the questions, and after about the fifth item in the series, as carelessness sets in. Using a variety of question formats reduces response error, but very frequent changes can produce confusion and lower reliability.

Finally, Smith (1991) has reassuring news for researchers who might be concerned that order effects were so pervasive that inadvertent juxtapositions of items would often introduce bias. He looked for order effects in about 500 different items on a variety of topics whose positions in the US General Social Survey were changed as part of a general procedure to rotate items. Only 12 statistically significant effects were found, and the average effect involved a 7.5 percent shift in responses (it was possible to detect effects that changed the response distribution of about 3 percent). This suggests that the main risk of order effects involves questions whose content is directly related.

Respondent Characteristics and Response Effects

Ideally, survey responses depend on the content of a question, but not on exactly how it is formulated. It is one thing to find, say, that education is related to having a particular opinion, but another to find that the effect of education depends on how the question is formulated. The most wide-ranging examination of the impact of the format of questions on different kinds of respondents is a by-product of Andrews' analysis of question formats. He found that the impact of the question format was unrelated to respondents' education or race, but increased sharply with age (Andrews,

1984: 434). Random error, which involves respondents answering more erratically but not favouring one or the other side of a question, was greater for older, less educated and black respondents (though most of the effect of race disappears when education is held constant).

Another possibility is that respondents' interest in a topic affects the impact of question formats on how they answer. Reviewing the evidence on this issue, Krosnick and Abelson (1992) find mostly negative results, consistent with the findings, cited earlier, that removing respondents with weak opinions by emphasizing a 'no opinion' option does not increase reliability. It is possible that less interested respondents are more likely to misunderstand questions and answer carelessly, while very interested respondents pay too much attention, discerning unintended nuances in questions.

Researchers have devoted considerable attention to the idea that interviewees might give answers designed to cast themselves in a good light. An independent measure of this attribute, known as 'social desirability', could be used to control for its effect. Since Edwards' (1957) first effort to provide such a measure, however, there is no consensus about how to measure social desirability (usually the Crowne–Marlowe scale [1964] is used) or even whether it exists, as DeMaio's (1984) extensive summary indicates. Andrews' (1984: 434) highly qualified assessment is that the Crowne–Marlowe scale 'produced results in the expected direction: respondents who scored relatively high on this concern had a modest tendency to give data that were below average in validity and above average in residual error'.

Validity Issues

A serious weakness of the research on questionnaire design is the lack of attention to the validity of survey questions. This is more understandable for attitude questions, where it can be difficult to identify and measure an 'external' criterion to estimate the validity. There is also an argument that attitudes – both at the theoretical level and in respondents' minds – are related to, but not the same as, intentions to act. Even for 'objective' questions where there is external evidence, Presser (1984) remarks how little effort is devoted to validation. In his analysis of a 1949 survey of Denver, Colorado he found that the validity of survey questions was affected by the topic and the characteristics of survey respondents. Searching for comparable analysis turned up only two studies, both with highly atypical samples.

Wentland and Smith (1993) have made an extensive search of American validation studies and conducted a meta-analysis of the relationship between the characteristics of questions and the validity of responses. They examined overreporting of socially desirable events, such as voting, and underreporting of undesirable events, such as criminal activities and declaring bankruptcy. Their findings present some interesting similarities to the internal analysis reported above. For example, questions involving just two response

categories provide more accurate data, except for difficult questions when a large number of categories appears to stimulate more accurate recall. More sensitive questions, about illegal activities for example, resulted in greater *in*accuracy, though multi-category responses produced less misreporting. Also, there was more of a tendency for respondents to overreport desirable activities than to underreport undesirable activities.

Some Conclusions

Schuman and Presser's experiments achieved a great deal, but not their intended goal of creating a systematic basis for question design.[14] Aided by new telephone survey technology, their revival of split-ballot experiments was the basis for a broad exploration of question design. The effects they found were highly variable and depended on the topic of a question. Even when there were *consistent* effects, their *magnitude* was highly variable.[15] More optimistically, Andrews' findings suggest that the format of questions has relatively little impact on responses, relative to the 'valid' variation reflecting true differences among respondents. In other words, the quality of survey questions depends more on content than style. Lest this seem to make the craft of questionnaire design disappear, remember that the methodological research described herein involves comparisons between questions that are carefully designed by expert researchers. Especially in a long questionnaire and for less experienced survey researchers, errors and misjudgements are much more a worry than designing optimal questions.

The complicated and sometimes contradictory findings of Schuman and Presser's experiments must reflect the complexity of the mental processes required to answer surveys and the variability in how questions about different topics are answered. But their behaviourist research strategy, in which the format of questions is manipulated without a theory of how respondents answer questions, prevents them from arriving at convincing generalizations. They conclude:

> What is needed most is theoretically directed research, but exactly what this means is not so clear. It is sometimes suggested that research on survey questions should draw its theory from cognitive or social psychology, from linguistics or psycholinguistics, or from one of the other basic social sciences. We are sceptical of this recommendation, having tried it ourselves. . . . Theorizing in survey research will have to begin by formulating problems that arise more directly from its own data, methods and ideas. (Schuman and Presser, 1981: 313–14)

Further gains in understanding question design required both a theoretical conceptualization of survey response and better means to examine how respondents answer questions and to observe the interaction between respondents and interviewers. To the extent these could draw on existing

social science, the answers lay in the disciplines of psychology and social psychology.

These findings and the research strategy of split-ballot experiments can be seen in a quite different way. If the impact of different formulations of a question depends on what the question is about, then experiments with systematically varied questions might provide a substantive research tool. The routine use of experimentation would involve a shift from the *optimal design* of survey questions to experimentation where *intended* variation in survey questions becomes a research method. Echoing Payne, some years after the experiments described above, Schuman and Bobo (1988) argue that:

> Every survey question or set of questions must be regarded as a single treatment in an incomplete experiment. There are always other substantively important ways to ask and order the questions. . . . The most effective way to deal with such possibilities is to carry out . . . experiments in which question form, wording and context are varied. Although there are always many such variations possible, in most practical cases only a few are serious candidates for experiment.

Interestingly, they make no mention of Peter H. Rossi's (1979; see Rossi and Anderson, 1982) development of what he called 'factorial surveys'.

Setting out to understand judgements about the social standing of individuals and families described in terms of a variety of different attributes, Rossi developed a broader and richer experimental strategy for surveys. First, he recognized that two or more aspects of a single question could be varied simultaneously. In a study of perceptions of fair pay, for example, respondents might be asked to evaluate a series of 'vignettes' describing women and men in different occupations. The impact of the incumbent's sex on the ratings could then be separated from the impact of occupations.

Social judgements reflect complex contexts. Judgements of the seriousness of a crime, for example, might involve the characteristics of the offender, the crime and the victim and so the number of unique combinations can be in the thousands – more than any one survey respondent could consider. The second, critical step in Rossi's development of factorial surveys was his recognition that it was not necessary for each respondent to evaluate all the possible vignettes. Instead, each respondent could rate a sample of them. To determine the impact on the respondent's judgement of the different, randomly varied factors, it was only necessary to regress the ratings on characteristics of the vignette. By including the characteristics of respondents in the model, it is also possible to measure individual differences in the basis of social judgements (see, for example, the essays in Rossi and Anderson). Data of this type are an obvious candidate for multi-level models, described in the chapter 'Quantitative Analysis of Survey Data'. Sniderman and Grob (1996) provide a very fine review of and argument for experimentation in attitudes surveys.

Applying Cognitive Psychology to Questionnaire Design

Organized efforts to reshape social science are rare, and the developments in survey research described above are no exception to our incrementalist tradition. This is why the Advanced Research Seminar on Cognitive Aspects of Survey Methodology, the CASM project for short, is so striking. Aborn (1991: 173) comments, 'Few interdisciplinary innovations in social science have been launched with as clear an issuance of purpose, delineation of objectives, and strategy for implementation.' The 'project' was actually two seminars organized by the US Committee on National Statistics in 1983 and 1984, 'whose goal was to foster a dialogue between cognitive scientists and survey researchers and to develop ideas and plans for collaborative research' (Jabine et al., 1984: 1). Between 1986 and 1990 the newly established Social Science Research Council Committee on Cognition and Survey Research held a series of workshops on topics such as 'the semantics of interview questions', the 'effects of theory-based schemas on retrospective data' and 'the structure of the survey interview'.[16]

The genesis of the initiative dates to the mid-1970s, when government researchers in some countries felt increasing demands to conduct research in areas, such as health, crime and the quality of life, that had traditionally been left to academic researchers. There was also growing concern that the answers to the 'objective' questions about employment and income, the staple fare of official surveys, were subject to significant response error. The direct impetus for CASM goes back at least to a 1978 British conference on retrospective data in surveys (Moss and Goldstein, 1979), and more explicitly to the 1980 Panel on Survey Measurement of Subjective Phenomena, convened by the US Committee on National Statistics and motivated by broad concerns about the quality and influence of survey research among the US government agencies and in the American scientific community.[17] In nearly 500 pages, the 1982 report of this panel (published as Turner and Martin, 1984: Vol. 1) provides a wide-ranging and thoughtful overview of survey research, with a distinct stress on precision in measurement. For example, considerable attention is paid to the generally small and hitherto almost completely neglected differences in the results obtained by different research organizations (called 'house effects').

The proposal for the CASM project came from statisticians Stephen Fienberg and Myron Straf (Jabine et al., 1984: 149) and its clients were US government agencies. The survey chosen to provide a substantive focus for the seminars was a large-scale health study, which employed face-to-face interviews. According to Jobe and Mingay (1991: 176): 'Government agencies in three industrialized countries (Britain, Germany and the US) took a lead role in promoting these scholarly conferences and helped to support much of the subsequent scientific research.' The priorities of policy-makers

are also evident in the choice of research topics. Most prominently, because many government surveys involve such 'retrospective' reports, policy-makers were concerned about the ability of survey respondents to report reliably the frequency and characteristics of events such as obtaining medical care, getting jobs and being the victim of a crime. The CASM strategy was also set in the organizational and intellectual context of large government surveys, characterized by explicit planning over long time frames, relatively narrow and well-defined goals, a high degree of division of labour in the survey organization, and a focus on point estimates and minimizing bias.

The CASM project aimed to bring the research traditions and tools of cognitive psychology, specifically the combination of simple, fairly mechanical theorizing of mental processes with small-scale experimentation, to bear on questionnaire design. Instead of experiments with ordinary questions in surveys of representative samples of hundreds of respondents, cognitive psychologists were accustomed to presenting esoteric tasks to much smaller, usually unrepresentative samples in laboratory settings where the interviews could be carefully monitored and recorded. The disinterest in sampling reflected the assumption that the psychological processes under study did not vary systematically across the population, or at least that any such variation was secondary. While its proponents described the potential for a mutually enriching relationship between the survey research and cognitive psychology, the dominant thrust of the CASM movement was to make survey research into a problem in cognitive psychology, rather than to use surveys to conduct cognitive research outside laboratories.

Even if the institutional factors were critical, they accelerated a transformation for which the intellectual ground had been laid. By 1980 split-ballot experimentation was at a dead end. The strategy was expensive, because surveys were required for the experiments and the absence of a theory of survey response left researchers to post hoc explanations of their results. Jobe and Mingay date the paradigm shift to at least the early 1980s when 'some survey researchers . . . recognized that cognitive psychology might offer an array of theories and techniques which could be used to improve the reliability of the information obtained through the survey method' (Jobe and Mingay, 1991: 176)

The main academic disciplines contributing to cognitive science are psychology, artificial intelligence, linguistics, philosophy, neuroscience and sometimes anthropology; but not sociology (Gardner, 1985: 6). Although the cognitive scientists attracted to surveys were largely psychologists (for computer modelling, though, see Graesser et al., 1996), survey researchers, rather than the major figures in cognitive psychology, put it forward as the new methodology for surveys. The emphasis on theorizing the survey response process, of course, involves exactly what distinguished cognitive science from the earlier behaviourist paradigm, which was the commitment to building

models of mental processes, even if the processes could not be observed directly, but only inferred from behaviour. The emphasis on rational thinking and the computer-like interpretation of memory and mental processes sets the cognitive approach apart from other anti-behaviouralist traditions in psychology: 'Though mainstream cognitive scientists do not necessarily bear any animus against the affective realm, against the context that surrounds an action or thought, or against historical or cultural analyses, in practice they attempt to factor out these elements to the maximum extent possible' (Gardner, 1985: 41). More optimistically, Hastie (1987: 65) comments that 'One of the most important consequences of efforts to extend the [information processing approach from cognitive psychology] to survey research will be to force the theory to develop explicit principles to characterize the relationship between motivation, affect and cognition.'

The Model of Survey Response

Without direct access to mental processes, the application of cognitive psychology to survey response has proceeded by breaking the respondent's task, in answering a question, into simple steps. The standard cognitive model, described by Sudman et al. (1995: 57), has five sequential components: 'interpreting the question, retrieving information, generating an opinion or a representation of the relevant behaviour, formatting a response, and editing it are the main psychological components of a process that starts with respondents' exposure to a survey question and ends with their overt report'.[18] Of course, this sequence and the mechanism animating it are understood in computer-like terms, with a central processor looking to a differentially accessible store of interconnected ideas.

The cognitive model has been successful in introducing order into question design and interpreting what previously seemed inconsistent findings. The effect of differences in question wording, for example, may be seen in terms of stimulating access to particular elements of memory (see Dovidio and Fazio, 1992). Another useful idea is that respondents choose between two strategies at the second, 'retrieval' stage of answering questions: if the question is recognized as pointing to a pre-existing element of memory, the answer is 'retrieved'; but if no match is found, the respondent must 'compute' a response, by bringing together ideas stimulated by the question. Of course, the steps in the response process approximate a process that may not be so orderly in practice. For example, there may be loops in the response process, as tentative answers are checked back against the initial perception of a question.

A nice example of the use of cognitive ideas is Wilson et al.'s (1996) assessment of questions that ask respondents *why* they have an opinion. Such questions have low validity, they argue, because the reasons for attitudes often cannot be retrieved from memory in the course of an interview, and

maybe not at all; and low reliability, because of the difficulty of answering questions and their high 'reactivity', which refers to the tendency for answers to be affected by earlier, related questions. More generally, as respondents answer a sequence of questions, previously retrieved elements of memory become more accessible and tend to be reused, especially by respondents who are less informed and less sure of their opinions, and so have more difficulty finding appropriate new elements of memory.

To some extent, cognitive researchers returned to the problematic of the first generation of survey researchers and away from the concerns with response errors and 'method effects' of their immediate predecessors. There was much more of a concern to understand the impact of the format of questions in the context of the question topics and less of an interest in – and perhaps belief in the possibility of – developing general rules of questionnaire design.

Experiencing Surveys

Cognitive researchers have used the two different ideas to conceptualize the process of completing a survey from the respondents' perspective. The first is to think of completing a survey as a single, complex task in which respondents accumulate information and skills as they proceed. In attempting to master the situation, however, respondents may incorporate irrelevant information and make unpredicted connections between questions. This is how the response categories of a question, merely designed to make the question easier for respondents to answer, and even numbers used for computer coding, can be taken as information about what a question means and/or what constitute norms for the population.

Complementing the instrumental interpretation of the interview task is the idea that survey interviews have conversational aspects.[19] Clark and Schober (1992) provide a lovely introduction to this idea. Cognitive researchers have especially been drawn to the work of H. P. Grice (1975), who sets out five principles – of speaker's meaning, utterance design, accumulation, cooperation and grounding – that govern conversation. The principle of speaker's meaning, for example, specifies that participants in conversations *assume* that they are being understood. The point is not that respondents may misunderstand questions, but that respondents *and* interviewers may not be aware of this. Grice's scheme is further elaborated,

> This principle [of cooperativeness] can be expressed in the form of four maxims. A *maxim of quality* enjoins speakers not to say anything they believe to be false or lack adequate evidence for. A *maxim of relation* enjoins speakers to make a contribution relevant to the aims of the ongoing conversation. A *maxim of quantity* requires speakers to make their contribution as informative as required but not more informative than required, and a *maxim of manner* holds that the contribution be clear rather than obscure, ambiguous, or wordy. (Sudman, Bradburn and Schwarz, 1995: 62–3; emphases in original)

In these terms, the order in which questions are asked can affect their answers because respondents attribute a 'conversational' coherence to the survey, sometimes connecting the answers to unrelated questions. And respondents may avoid the most obvious answer to a question because they think that a less common answer will be more interesting and so the *quality* of their answer will be higher. These social aspects of the survey interview operate alongside cognitive mechanisms, so that order effects can also be seen to arise from earlier questions having increased the accessibility of certain elements from memory, either consciously or unconsciously (on the latter, see Banaji et al., 1996).

Probably, it makes more sense to think of survey interviews as conversation*al* than, literally, as conversations. Standardized survey interviews are too stylized, inegalitarian and lacking in spontaneity to fit the definition of ordinary conversation, not to mention that in a sense the 'conversation' is between *absent* researchers and the respondent, with the interviewer a highly constrained intermediary.

Cognitive Survey Design

Cognitive researchers were no less concerned with the details of questionnaire design than their predecessors. Indeed their best work has a meticulous quality motivated by an acute awareness of the impact of minor variations on survey response. This is complemented by an awareness of the larger task-related and conversational context of the survey interviews and of the importance of understanding what respondents think they are doing when they answer surveys. In this sense the cognitive approach represented an important advance over the many discrete concerns about question design that had previously been investigated. To those practical questions, about whether to include a middle response, to ask respondents if they had an opinion and so on, no better, general answers were found. To understand the cognitive approach more clearly, it is necessary to look at the particular problems that were addressed.

Memory The study of past experience, termed 'autobiographical memory' to distinguish it from memory in general, is critical in many research contexts. Surveys on employment, education, expenditures, the use of government and medical services, leisure and many other topics must rely on retrospective reports, often covering long periods of time. Setting out the purpose of the CASM seminars in 1984, Jabine et al. write 'how to improve recall was perhaps the central question of the seminar' (Jabine et al., 1984: 14). Two of the three substantive chapters in the special issue on surveys of the journal *Applied Cognitive Psychology* deal with memory, as do five of the 12 substantive chapters in Tanur's recent volume; a series of papers from a conference devoted to memory appear as a volume edited by Schwarz and Sudman (1994; also see the review in Sudman, Bradburn and Schwarz, 1995: 197–226).

Ideally, a cognitive model of memory would combine a description of the 'contents' of autobiographical memories – both in terms of what is stored in respondents' minds and what it is about events that is remembered – with an explanation of the mechanisms used to access memories in the context of responding to a survey. There is certainly no good, unified theory of the contents of memory, but there are some helpful ideas. For example, Brewer (1994) distinguishes the 'personal memory' of an experience, which involves the ability to reimagine a specific event with fair accuracy but often not to locate it accurately in time, from 'autobiographical facts' which involve the exact recollection of particular events. Barsalou (1988) argues that particular memories are embedded in sequences of events and are accessed by first locating the larger sequence. There is also a critical debate over the extent to which memories are copies, as opposed to reconstructions, of experience – since reconstructions are likely to be much more malleable and can even be altered by the effort of remembering them (see Neisser and Winograd, 1988). These ideas provide clues, but certainly not a formula, for designing survey questions.

In extracting information from the complex store of memories, survey respondents are seen to use two different processes, depending on the frequency of the events and whether the memory has been stored reasonably exactly. Memories of more exact, discrete and infrequent events are recalled individually, a process that Tulving (1972) refers to as 'episodic recall'. When this process cannot be used, because memories are not very exact or exact recall is too laborious, respondents resort to what Blair and Burton (1987) call 'direct' estimation, that is they make approximations, for example by considering a typical day or week, implicitly averaging and throwing out unusual events. Asked how often an event had occurred in a given period of time, Burton and Blair (1991) found that: respondents almost always used episodic recall for events that occurred three times or less; they estimated the number for events that occurred ten or more times; and about half used each strategy for events that occurred about five times. Direct estimation results in increased response error, because respondents discount unusual events, but in doing so they may actually produce more reliable aggregate data.

The formulation of survey questions and the style of data gathering (not only the kind of data collection, but, for example, whether interviewers hurry respondents) can encourage a particular style of response. For example, Loftus et al. (1992) showed that respondents had a great deal of difficulty remembering individual medical visits. For a one-year period, the estimate of the number of visits was only 39 percent of the number recorded in the respondents' medical records. Furthermore, efforts to assist respondents, by suggesting that they work forward in time, or backward in time, or just recall events freely in no particular order, produced no improvement. There was a striking improvement, however, to 87 percent of recorded visits, when

respondents were asked to *estimate* how many visits they had made. In another study, Skowronski et al. (1994) show that respondents are better able to fix the dates of events than to say how long ago they occurred.

Another common problem with estimates of the frequency of events is not *under-* but *over*counting. The accepted explanation is that such bias is due to 'forward telescoping' in which distant events appear more recent. Bradburn et al. (1994) account for telescoping with a simple and elegant model based on the idea that respondents make larger *errors* in estimating the time of more distant events, though *on average* they are no more likely to think that events occurred before rather than after they actually occurred. This explains why a strategy invented by Neter and Waksberg (1964), called 'bounded recall', produces better estimates. The idea is to ask respondents how often events of a particular type occurred for two different overlapping intervals, with *the longer one* first. For example one might first ask how often respondents attended sports events in the last year and *then* ask about the last six months. Responses to the first (one year) estimate suffer from telescoping, but telescoping is reduced for the second (six months) estimate.

Survey researchers are also interested in measuring respondents' states at particular times in the past, such as their income at some time, characteristics of their parents and families when they were growing up, whether they voted in elections and their attitudes at an earlier time. Strube's (1987) model suggests that the answers to such questions are affected by the way that past events were originally 'encoded' in memory *and* by 'restructuring of the representation' over time. In answer to survey questions, memories of this type tend to be 'constructed', rather than retrieved. Reviewing a large body of research, Pearson et al. (1992) characterize the process in which respondents combine an estimate of their *current* state with an 'implicit theory' of how they have changed over time, as follows:

> . . . responses to retrospective questions are a function of an interplay between the past, the present, and implicit theories of change, stability, and relationships among attributes. If respondents assume that their state has not changed in a significant way, they can construct their past from their current state (and will be inclined, unless otherwise motivated, to do so). If respondents assume that they have changed, their construction of their earlier state is likely to be guided by implicit theories about changes . . . the research findings indicate the existence of two forms of systematic bias. . . . In some studies people exaggerate their consistency over time. . . . In other studies people overestimate the extent to which their present state differs from an earlier state. (Pearson et al., 1992: 83)

In the numerous examples cited by Pearson et al., the more common form of bias involves overestimating the similarity of the past and present, under the influence of an 'implicit theory' of stability.

For some topics, these biases appear to be compounded by 'social desirability', that is respondents' conscious or unconscious efforts to cast

themselves in a good light. US studies using record checks reveal that about one-quarter of survey respondents who claim to have voted have not done so. Abelson et al. (1992) report on what turn out to be unsuccessful efforts to decrease the tendency to overreport having voted in past elections. Neither the technical 'fix' of bounded recall techniques nor formulations of questions designed to remove the stigma of not voting substantially diminish over-reporting. These findings are consistent with the conventional wisdom that retrospective questions about attitudes and perceptions are not a credible means for knowing the past, which is why longitudinal studies have largely replaced retrospective enquiry as a means of studying change.

While this cognitively based theory and empirical research on memory are also not easily reduced to rules of question design, a great deal is known about retrospective questions and their limitations – this section has described only a fraction of the material. In gathering retrospective information, the main problem for survey researchers is not how to ask the best questions, but the fundamental limitations of memory.

Attitudes Attitudes have long been a central interest of survey researchers (Converse, 1984a). In the classical view, attitudes themselves were not problematic, but *measuring* them was. Researchers had only to identify mental objects of interest – 'traits' or 'factors' – and provide respondents with a means of evaluating them. More abstract mental objects (authoritarianism, support for women's rights, etc.) might not be accessible to direct questions, but could be measured by combining the answers to a series of more specific questions, based on analysis of their correlations. If the correlations between questions designed to measure the same trait were often quite low, this was because of the difficulty of formulating exact responses, response errors and perhaps the inclusion of respondents with no attitudes. The validity of traits was commonly assessed by reading the items ('face' validity), rather than with external measures.

The CASM movement brought two broad bodies of research to bear on these issues. The first deals with the dimensionality of individual attitudes. The title of Krosnick and Abelson's (1992) article, 'The Case for Measuring Attitude Strength in Surveys', speaks for itself: they argue that,

> Although it is very common to see attitudes measured in surveys, it is rare for a survey to measure the strength of attitudes. And yet it seems patently obvious that not all attitudes are alike. Some are strong, in the sense that they have profound effects on individuals' cognition and behaviour, and resist even the strongest pressures towards change. And other attitudes are weak, vulnerable to situational pressures, and with little if any impact on an individual's thinking or action. (Krosnick and Abelson, 1992: 177)

And this is what Katz said in 1944:

> . . . if we are able to measure the intensity of opinion as well as its direction, we

> shall know a great deal more about the individuals whose opinions we are studying. We shall know, for one thing, something of the permanence of opinion, its crystallization, the extent to which it is structured, and the degree to which an individual may be expected to be suggestible. (Katz, 1944: 51)

But there is still no consensus about what the 'strength' of an attitude is. Krosnick and Petty (1995: 5) describe more than *15* different aspects of the strength of attitudes, including their magnitude, ambivalence, salience, embeddedness, accessibility and importance. Nor is there consensus that these dimensions are reducible to simpler underlying constructs. Krosnick and Abelson (1992: 179ff.) advise that five dimensions of attitudes are most central: the extremity of opinion, which they define as the 'degree of favourableness or unfavourableness of an individual's evaluation'; its intensity, defined as the strength of *feelings*; the extent to which an individual is certain of her or his opinion; the importance of the attitude to the individual; and the amount of knowledge about the attitude object.

This work suggests a reorientation of attitude research towards groups of questions about different aspects of a smaller number of topics and away from the strategy of asking questions about many different topics, which might later be factor analysed to identify the underlying dimensions. Narrowing the range of questions is especially unattractive to policy researchers who often want to test a wide range of alternatives and for large-scale opinion surveys whose content is shared among researchers.

A more explicit application of cognitive models to attitude research, raised at the 1983 CASM seminar by Tourangeau (in Jabine et al.,1984; also see Strack and Martin, 1987), focuses on the distinction between 'retrieval', and the much more variable processes used to 'compute' opinions about less salient topics. In the extreme, topics may be so unfamiliar that this 'computation' becomes nearly random. As questions become more abstract, respondents switch from retrieving to computing responses. Also, respondents differ in the size and composition of their stock of retrievable answers and the context affects whether they choose to retrieve or compute an answer. Time pressures encourage respondents to retrieve responses, also increasing the probability that they will retrieve an attitude that is not a good match to the question asked. Computing responses is a more complex process, which is subject to greater random error and to bias arising from the method of measurement. One strategy that has been used to investigate questionnaire responses is to measure how long respondents take to answer each question, which is known as the 'response latency'. Retrieval is generally assumed to be a faster process than computation (see Bassili, 1996).

This cognitive research on attitudes remains largely within the conventional paradigm: attitudes are understood to be things (in cognitive terms, the things are elements of memory) and good survey questions direct respondents to find the right answer in a very large store of possible answers. For

questions that are unusual, complex or abstract, and for respondents with a smaller stock of memories or less skill in retrieving the relevant ones, finding the most appropriate elements of memory is more difficult and prone to error. There is, however, a radically different model of attitudes. From the 'constructivist' perspective, attitudes are *created* in answer to questions. Rather than retrieving attitudes from a pre-existing store, respondents choose among complexes of linked ideas (Tesser, 1978; Wilson and Hodges, 1992). 'Cognitive psychologists have begun to recognize the importance of relatively high-level conceptual structure, referred to variously as scripts, ... schemata ... and frames.... These structures are organized packages of beliefs, feelings and knowledge about classes of situations or things' (Tourangeau, 1987: 154). He goes on,

> ... many attitude issues are represented cognitively as scripts. For one person, an issue like Welfare may activate a whole set of interrelated images, stories and feelings involving Welfare Queens, fraudulence, and a sense of injustice. For another person, the same issue may activate a completely different script involving people down on their luck and feelings of obligation. For some issues, multiple scripts about an issue are available within a society, with some people subscribing to one script and others to different scripts. ... For the script theorist, the key step in the comprehension of an attitude question is the activation and application of the appropriate script.

The constructivist perspective suggests a dramatic reconceptualization of attitude research towards identifying and describing scripts and away from the idea that attitudes constitute a fixed mental stock. This process of metaphorical reasoning contrasts sharply with the rationality and logic of the standard cognitive model. It also suggests a more qualitative orientation to attitude research, focusing on how respondents understand and then respond to questions and, implicitly, less on the large-scale comparisons between demographic and socioeconomic groups. A more agnostic approach might be to ask how questions about different topics, formulated in different ways, are answered by different kinds of respondents who choose whether to 'retrieve' attitudes, compute attitudes from relevant ideas, or invoke scripts.

Certainly, this work raises serious questions about what exactly is being measured in the everyday studies of attitudes – on education, technology and so many other topics – that are a mainstay of public opinion research. At the very least, too little attention is paid to what respondents know about issues, to asking sufficient numbers of questions to understand opinions and their strength, and to assessing the consistency of opinion by comparing the answers to differently formulated questions.

Question and Response Order If there was any area in need of the theoretical clarity of cognitive approaches, the impact of question and response order

qualifies. Numerous experiments had resulted in a commonsense classification, but not a more general theoretical understanding of response effects. Certain configurations of questions seemed liable to affect responses, but whether any effect was observed depended on the topic. Furthermore, there are many examples of statistically significant effects that could be not replicated, suggesting that effects varied across survey populations or could be affected by the context of previous questions. These problems have received a great deal of attention. A 1992 volume edited by Schwarz and Sudman includes ten contributions dealing with question order effects. Sudman, Bradburn and Schwarz (1995) produce a convincing synthesis based on the cognitive model of response.

In cognitive terms, order effects can be seen to arise from the individual steps in question response (for typologies see Strack and Martin, 1987; Tourangeau and Rasinski, 1988; Smith, 1991). In the first stage of response, efforts to understand and resolve ambiguities in the meaning of questions may be affected by inferences from previous questions. A frequently cited example involves consecutive questions about respondents' satisfaction with marriage and with their 'whole life'. When the general question is asked second, respondents infer that it concerns aspects of their life *other than their marriage*, that is they tend to contrast their marriages to other aspects of their lives. On the other hand, asking *a series of questions* about various aspects of their lives will result in respondents' general evaluation taking account of these aspects, but not others.[20]

An element of memory that has been retrieved in order to answer a question becomes more accessible and is likely to be reused in answering later questions, a process known as a 'priming'. The priming process is deactivated, however, when the next question is seen to require different information. The direction of order effects on survey responses thus depends on the respondent's understanding of the relationship between two questions. The perception that the 'targets' of subsequent questions are similar increases the agreement between respondents' answers – researchers refer to 'consistency', 'assimilation' or 'carryover' effects; while the perception that they are dissimilar increases disagreement – termed 'inconsistency', 'contrast' or 'backfire' effects. Interestingly, the effect of previous questions can be suppressed by drawing respondents' attention to the relation between adjacent questions.[21] Finally, respondents may make conscious efforts to answer questions consistently or, alternatively, to appear more even-handed! Tourangeau (1992) describes research demonstrating *all* of the effects just described. A fundamental difficulty is that there is an after-the-fact cognitive explanation for almost any observed pattern arising from the juxtaposition of questions. Furthermore, the relationship between survey questions is not simply established by researchers who decide on the order of questions, but also reflects the associations made by respondents.

Cognitive researchers have also considered how the order of response options affects answers, especially when the response alternatives are long or complicated. Depending on the circumstances, 'primacy' or 'recency' effects may appear. More complex questions, such as questions requiring preference rankings, are also more prone to methodological artefacts. These effects are complex and not well understood. Sudman et al. conclude that 'the emergence and direction of response order effects seems to depend on a complex interaction of serial position; presentation mode; item plausibility, complexity and extremity; and respondent ability and motivation' (Sudman, Bradburn and Schwarz, 1995: 160).

Cognitive Explanations of Respondent 'Error' Without direct measures or a conceptualization of how respondents answered questionnaires, respondent error has largely been understood as a statistical category, of variance associated with respondents' inability to understand and answer questions properly. Maybe because it was easily measured, researchers concentrated on the idea that educational attainment was a good measure of the skills required by respondents, and this has some empirical support. Missing from this approach is any *active* role for the respondent, a problem that Krosnick (1991) has addressed by applying Simon's (1957) economic concept of 'satisficing' to survey respondents. Satisficing is an explanation of why economic actors often settle for good, rather than optimal outcomes in finding jobs, making purchases and trading. This is because the information and time required to improve outcomes are costly and the decision-making itself requires effort. Therefore, after a relatively short search, the expected gains of additional efforts diminish drastically.

According to Krosnick, survey respondents begin with the desire to provide high quality responses, but quickly find it is difficult to give quick and thoughtful answers to a long series of complex questions. He argues that respondents then adopt a satisficing strategy, settling for good answers, but not trying their best. What Krosnick terms 'weak' satisficing involves respondents taking shortcuts by searching memory less carefully and answering more rapidly, with the result that they may access irrelevant information and misinterpret questions. 'Strong' satisficing carries this further, to the point where respondents say they have no opinion or choose answers randomly.

Another approach is to think in terms of respondents' psychological needs: Petty and Jarvis (1996: 232) define the 'need for cognition' in terms of whether individuals 'tend to engage in and enjoy effortful cognitive activity' and the 'need to evaluate' in terms of whether they 'naturally tend to engage in evaluation (that is, whether they think in highly valenced terms about the people, objects and issues around them)'. They cite evidence, though almost entirely from small-scale psychological studies of students, of wide variation in individuals' cognitive styles. This leads to the interesting suggestion that

respondents can try too hard! Respondents with a low need for cognition make the kind of mistakes associated with less education; but a high need for cognition results in respondents using information too well. Such respondents are more likely to be influenced by ideas stimulated by previous questions and to find unintended subtleties in questions. Similarly, respondents with a low 'need for evaluation' are more prone to non-response, but those with a high need are likely to make evaluations even when they have no information.

These ideas are interesting, not least because they get away from the demographic classification of better and worse respondents and deal with what respondents are actually doing when they answer surveys. The paucity of work on the motivation and response styles of respondents and how it might affect their answers to questionnaires reflects the broader orientation of cognitive psychology. A necessary addition to cognitive models is more explicit inclusion of respondents' motivation, which is likely affected by the content and context of the survey (has the interviewer phoned without warning; is she or he impatient; what is the response to queries about the questions?), as well as respondents' previous experience with and knowledge and opinions of surveys.

Success of CASM?

The cognitive approach showed a way out of an impasse in understanding survey questions and response, but it has not led to a new theory of question design. About many everyday design questions – whether to offer a 'no opinion' response or a 'middle' alternative, how many and what kind of response alternatives, and so on – earlier, largely experimental work showing the consequences of different alternatives still stands. The cognitive approach is also compatible with, and provides explanations for, experimental results demonstrating great variability in the effect of particular question formats and the order of questions that is a function of the topic of the question.

As well as many interesting examples to act as models, cognitive researchers have provided concepts for thinking about questionnaire design. The most important ideas include: the central role of differentially accessible elements of memory; the distinction between retrieval and computation of answers; the idea that entire questions, including any responses offered, are tasks; the idea that answering a questionnaire, not just individual questions can be thought of as a task; and the conversational elements of survey interviews.

Sudman, Bradburn and Schwarz (1995: 266) are sensitive to the criticism that much effort has produced little in the way of concrete design guidelines. What they describe as the first guide to questionnaire design based on cognitive principles ends on a somewhat passive aggressive note:

> We suppose that many readers would like to see more specific recommendations for questionnaire design. But such recommendations would be unlikely

> to capture the complexity of the [survey] processes we examined in this book. Unwelcome a task as it may be, questionnaire design problems require analysis of the design issue at hand in light of . . . theoretical principles.

Their point is that there *can* only be limited general principles of questionnaire design. The corollary is that there is no substitute for intensive effort to improve each questionnaire. This fits well with the needs of the government researchers and surrounding academics whose concerns prompted the cognitive movement. Cognitive methods do less to address the dilemmas of producing the good questionnaires with limited resources.

Perhaps because it constitutes a kind of claim staking and grist for more ingenious experiments, cognitive researchers emphasize the impact of methodological effects on surveys. The result is a rather postmodern tinge to work in a speciality that takes pride in bringing a systematic and scientific approach to the earlier muddle of disparate findings:

> . . . context effects are ubiquitous and cause complex interaction effects in attitude measurement, depending on the order of the questions and response categories, the mode of administration, and the information that respondents retrieve from memory . . . any questionnaire reduces the myriad possible contexts in which people may think about an issue. Hence, the best we can do is avoid asking questions in a context that is likely to deviate strongly from the probable context in which an issue will be considered. (Sudman, Bradburn and Schwarz, 1995: 263)

Though he diplomatically avoids questioning the general style of cognitive research, Groves (1996: 400) raises serious concerns about the reproducibility and generality of the results obtained in many small-scale experiments conducted in laboratory or classroom environments with atypical (often student) samples.

Cognitive researchers have made significant contributions to the development of many specific areas of questionnaire design. For example, retrospective questions about individual experiences, day-to-day habits and less common but regular activities are relevant to many survey topics; quality of life researchers can benefit from studies of the impact of question order on evaluations; and political scientists would be foolish not to worry about the impact of the order of asking questions about issues, political parties and individual politicians. In raising questions about the fundamental nature of attitudes, cognitive researchers have done a great service. They have drawn the attention of more sociologically oriented survey researchers to work by psychologists on the linked complexes of ideas and images referred to as scripts and schemata, and they have refocused attention on the multidimensionality of the attitudes. This work suggests a shift towards measuring fewer things more precisely.

The cognitive findings and ideas can also be cast in a more radical *qualitative* light that suggests a shift away from the traditional problematic of lining

up respondents on attitudinal dimensions. Answering questions in a survey can be seen in terms of the *construction* of the meaning and of the complex way that respondents make choices among the rich variety of potentially relevant mental elements. The question is whether this orientation is compatible with the fundamental framework – sampling, standardized surveys mainly with closed questions and statistical analysis – of survey research.

Conclusion

For an antidote to excessively theoretical thinking about question design, it is only necessary to read Tom Smith's (1995) entertaining account of the controversy following the publication of the Roper Organization poll with the question 'Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?'. More than one-fifth of Americans, 22.1 percent exactly, thought it was 'possible' that the Holocaust had not occurred, and a further 12.4 percent had no opinion. This was widely interpreted as a sign of growing anti-Semitism, though some commentators immediately expressed concern that the double-negative in the question introduced serious response bias. Only at considerable financial and emotional expense was the finding discredited by strong evidence that the question was flawed. The need for commonsense in questionnaire design, it is clear, has not changed.

Over 50 years, questionnaire design has matured into a durable practice mostly guided by informal theorizing and in response to concrete research questions. The CASM movement brought order to what were disconnected findings and has led to the development of a more coherent methodological research agenda. The better theorized understanding of survey responses and the much more systematic body of empirical findings appear to have had only an incremental impact on everyday survey practice. The best contemporary survey research is done by people who know a lot more about survey design than their predecessors, but this knowledge must still be combined with the craft skills to create survey questions that people understand and can use to describe themselves. In how many other fields would a lovely, but undeniably dated, 1951 text still be regarded as a definitive guide?

With carefully formulated questions, over half the variance (Andrews estimated two-thirds of the variance in his study) is valid; after that the sources of error are highly variable, including the interpretation of terms in the question, the respondent's mood and events in the last while, her or his perception of the interview, and the impact of previous questions. Perhaps, like statisticians, survey designers are rescued by the central limit theorem. In many circumstances, the sources of error appear to cancel out, contributing random error to the response, *but not bias*.

Notes

1 Cantril (1944: 23) credits the American Institute of Public Opinion for inventing the technique.

2 In England, a somewhat similar, though much less systematic effort was mounted by anthropologists, known as 'mass observation' (see Madge and Harrisson, 1938). See the sympathetic comment in Lazarsfeld (1939) and Marsh's (1982: 32–3) pithy assessment.

3 Converse (1987: 288–92) provides an interesting discussion of the blend of and changing balance between 'unstandardized' open questions and standardized questions in research at Columbia's Bureau of Applied Social Research.

4 Likert advocated this simpler scoring technique when empirical studies showed that it gave nearly identical results to the more complex procedure of assigning scores to response categories on the assumption that the answers represented categories of an underlying but unobservable normal distribution of opinions on each question (Converse, 1984a: 21; McNemar, 1946: 306).

5 From Rosenberg's foreword, 'this book is fundamentally an exposition, extension, and exemplification of Lazarsfeld's approach to survey data analysis. In this area, as in so many other areas of social science methodology, Lazarsfeld's work has been fundamental; the sum of his contributions is monumental' (Rosenberg, 1968: xv).

6 Much of their work appeared earlier in articles in *Public Opinion Quarterly* and elsewhere, but for convenience the citations are to their collection and integration of these pieces in *Questions and Answers in Attitude Surveys* (Schuman and Presser, 1981).

7 For responses to Converse, see particularly Achen (1975), Erikson (1979), Judd and Milburn (1980) and see Converse (1980) for a rejoinder. Summaries of the extensive, important debate include Kinder and Sears (1985) and Smith (1984b); and also Zaller (1992: 31ff.).

8 This exercise pays alarmingly little attention to validity issues. Two of the items that supposedly measure a general trait ask whether 'Lawyers are less honest and ethical than most other professionals,' and whether 'Lawyers charge too much for their services.' McClendon and Alwin take agreement with both statements as a sign of consistency, but there is no logical connection between the statements. There is no inconsistency in thinking that lawyers are no less honest than other professionals, but charge too much.

9 Alternatively, David Northrup suggests that respondents who are read a long list of activities or experiences of some kind may feel under pressure to respond positively at least once.

10 Schuman and Presser's ingenious experiments on acquiescence, which involved experiments, panel data, the addition of open questions asking about why respondents gave their answers, and questions about salience, employed two very difficult questions. One involved permutations of the question 'Which in your opinion is more to blame for crime and lawlessness in this country – individuals or social conditions?' (Schuman and Presser, 1981: 207) and the other question asked 'Would you say that most men are better suited emotionally for politics than

are most women, that men and women are equally suited, or that women are better suited than men in this area?'. Such problematic questions may produce methodological artefacts that would not normally arise.

11 Not to detract from the general point that more precise categories produce more reliable answers, the *topic* of the question is also relevant. It makes sense to ask exactly how often people read newspapers, but Bradburn and Sudman's (1979) questions about how often respondents were excited and bored in the previous month might be better answered with vague quantifiers that do mix perception of experience with the objective character of events. Who can accurately remember how many times she or he was bored in an entire month, especially when asked in a long questionnaire and given a few seconds to reply? Saying that one is 'often' bored, though, is perfectly meaningful.

12 Schuman and Presser (1981: 161ff.) devote an entire chapter to this issue. Much of their efforts show the obvious, that respondents are more likely to choose a middle alternative when it is presented as a legitimate response than if they are required to raise it spontaneously after being asked if they are for or against a proposition. Beyond this point, however, their analysis becomes very confused. They would like to be able to show that respondents choosing a middle alternative are more likely to have less education or that they are less informed, but their data will not cooperate.

13 Andrews (1984: 430) finds that whether a response scale has an explicit mid-point has no impact on the quality of the resulting data, except for the greater reliability of two- than three-point scales, already noted.

14 The 'cookbook' based on their research is Converse and Presser's (1986) short guide to questionnaire design. Its recommendations are generally sensible, but sometimes overstep the conclusions of the methodological research. For example, they recommend against offering respondents a middle position on the grounds that three-category questions tend to be less reliable than two-category questions, without regard to the content of the question.

15 There is certainly no agreement across the social sciences of what constitutes adequacy of explanation. In research traditions dependent on small numbers of 'cases' and/or when each observation is expensive, data are often said to be consistent with theory when the observed effect is in the predicted direction and statistically significant. In traditions where large numbers of observations are the norm, so that statistically significant effects may be of trivial magnitude, there is more attention to the relative magnitudes of the variance 'explained' by the variable of interest, explained by other measured variables and unexplained.

16 There was a more explicit intervention by the Social Science Research Council following the erroneous prediction that Landon, rather than Truman, would win the 1948 US presidential election (see Mosteller et al., 1949).

17 It goes without saying that this national community is dominated by the physical and biological sciences, and by social scientists – mainly in psychology, economics and statistics – with an affinity to them. More detailed accounts of these developments can be found in Jobe and Mingay (1991: 176–8), Tanur (1992b: ix–xii) and Sudman et al. (1995: 11–14). The Committee on National Statistics is part of the National Research Council, which was established by the US National Academy of Sciences.

18 The earliest reference to a similar model appears to be Cannell et al. (1981).

19 The idea of the interview as a 'conversation with a purpose' is credited to Bingham and Moore (1924). Some researchers think of structured interviews as a kind of conversation, literally: 'The survey interview, however, is best considered as an ongoing conversation in which respondents conduct their own share of thinking and question answering in a specific social and conversational context' (Sudman, Bradburn and Schwarz, 1995: 55; see also p. 245). For a fine discussion of the point see Schaeffer (1991).

20 The order of questions also affects their interpretation. Quality of life researchers tend to distinguish affective from cognitive measures of well-being. Questions about happiness and the seven-category 'delighted–terrible' scale (which offers a series of adjectives of this kind) are affective, while questions asking about 'satisfaction' are cognitive, in the sense of encouraging a balanced, more cerebral and longer term evaluation. Even if the question about well-being focuses on this cognitive satisfaction measure, asking it first will encourage an affective response, while asking it after questions about a person's personal relationships, job, income and so forth will encourage a cognitive response.

21 This also provides an instance in which the results of psychological research are directly applicable to surveys: Sudman, Bradburn and Schwarz (1995: 87ff.) summarize interesting evidence that the mood of the respondent and even the weather can affect survey responses. They report that respondents view their lives more positively on nicer days, but only if the survey does *not* mention it. Asking a question about the weather in the respondents' locale at the very beginning of the interview removes the effect.