1

The Invention of Survey Research

Survey research has roots in centuries of census taking, intelligence and psychological testing beginning in the late nineteenth century, research on attitudes from the 1920s, and 'social surveys' of the conditions of the urban poor, pioneered in England by Charles Booth in the 1880s and Joseph Rowntree in the 1890s, and extended to many other countries in the first third of the twentieth century (Bulmer et al., 1991). Modern surveys did not evolve directly from these ancestors, however. Instead, aspiring to represent entire populations, to ask questions on almost any topic, to gather data in a timely manner, and to yield quantitative results, surveys developed out of the public opinion polls initiated in the mid-1930s by American market researchers, notably Archibald Crossley, Elmo Roper and George Gallup (Converse, 1987: 87ff.). Surveys' successful arrival was signalled by their correct prediction of the 1936 US presidential election.[1] In Roper's words, 'advertising men … are to be credited with … the early development of the technique which has been evolved for sampling public opinion' (1940: 325).

Academics were no strangers to the public opinion industry. In 1937, psychology professor Hadley Cantril of Princeton University and Frank Stanton, then research director and later president of Columbia Broadcasting, obtained a Rockefeller Foundation grant to study the psychological and cultural effects of radio. To direct the project they hired sociologist Paul Lazarsfeld, whose first American publication, 'The art of asking why', appeared in the 1935 inaugural issue of the *National Marketing Review*. The first description of longitudinal surveys, by Lazarsfeld and Fiske (1938), is in the second volume of *Public Opinion Quarterly*, established in 1937. During the Second World War, the two US government organizations responsible for surveys were led by Elmo Wilson, an associate of Roper, and academic psychologist Rensis Likert. The monumental surveys of the US Army – over half a million soldiers were surveyed with more than 200

[1] For an interesting argument that modern social science *creates* public opinion, see Osborne and Rose (1999).

questionnaires – were led by Harvard sociologist Samuel Stouffer (compiled as *The American Soldier*[2] by Stouffer et al., 1949). According to Lazarsfeld, the development of survey research

> might be dated from the appearance of 'The American Soldier' after World War II. In this work, a large body of data was made coherent and meaningful by careful statistical analysis. 'Survey analysis' … became the language of empirical social research, possessing its own rules for forming basic concepts and combining them into meaningful propositions. (Cited in Rosenberg, 1968: vii)[3]

The next three sections of this chapter describe the development of survey sampling, questionnaire design and data collection until the early 1950s. This work established the conceptual core of modern survey research, but 60 years later no longer serves as a practical guide.

About Survey Sampling

The fundamental idea of applied survey sampling, which is that a properly selected random sample can accurately represent any population, no matter how large and diverse, dates to the late nineteenth century when Anders Kaier employed his 'representative method' to survey the entire Norwegian population. First, Kaier divided the country in two, separating urban and rural areas. In the urban 'stratum' he selected all of the five largest cities and eight smaller cities, to represent medium and small communities; then in each of the 13 selected cities he divided streets into groups according to size and selected a sample of streets; and finally he selected a fraction of the dwellings on each selected street. Counts from the Norwegian Census were used to calculate the appropriate number of selections in each community. In rural areas, a sample of municipalities was chosen on the basis of their main industry. Because Kaier calculated the probabilities at each stage of sampling to give every dwelling in Norway the same chance of selection, characteristics of the population could be estimated directly from the sample without weights (Bethlehem, 2009: 10ff.; Kuusella 2011: 91ff.). In modern parlance, the sample was 'self-weighting'. In other samples selected by Kaier, parts of the population

[2] For an extraordinary account of how Elmo Roper convinced General Eisenhower to agree to survey soldiers, see his *You and Your Leaders* (1957: 233–234). Roper is cited by Hyman (1991: 69), who also provides a detailed account of the programme of *Army Survey Research*. For an assessment of the role of *The American Soldier* by one of the researchers, see Williams (1989).

[3] C. Wright Mills (1959) challenged the hubris of Lazarsfeld's assertion that survey analysis had become *the* language of empirical social research, labelling it 'abstracted empiricism'. In his American Sociological Association presidential address Herbert Blumer described it as 'the scheme of sociological analysis which seeks to reduce human group life to variables and their relations' (1956: 683).

had different probabilities of selection and weights were used to produce unbiased estimates of population characteristics.[4,5]

Kaier's sample of the Norwegian population is like a modern multi-stage probability sample, except that the municipalities were not selected at random, but rather 'purposively', on the basis of his knowledge of their characteristics. The key idea is that a representative sample of a complex population can be obtained using two or more *stages* of selection – selecting communities, then streets within the communities, then dwellings on the streets – as long as the probability of selection at each stage is known. 'Systematic selection' is still used routinely, in place of strict random sampling and when the sample size is very small (for example, selecting a small number of communities) there is a good argument for selecting a 'purposive' sample based on a deep knowledge of the units, rather than a random sample. Kaier's methods are close to the modern ideas of 'balanced' samples, discussed in Chapter 5.[6]

Skip forward more than 30 years and modern survey sampling begins with the publication of Neyman's 1934 article, 'On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection', which demonstrates conclusively the risk of bias when a sample is *not* selected at random.[7] Also, he demonstrated the value of sample

[4] Kruskal and Mosteller (1980: 174ff.) provide a fine discussion of the concept of representativeness and describe the evolution of Kaier's ideas between 1895 and 1903 and responses at the International Statistical Institute, which was the statistical 'establishment' of the time. To someone schooled only in contemporary sampling statistics, what's remarkable is that the obvious principles of sampling were discovered and not universally welcomed at the time.

[5] Between Kaier and Neyman, the major intermediate figure was Arthur Bowley. In 1906 he argued that the precision of sample estimates could be predicted from Edgeworth's central limit theorem, which showed that sample means were normally distributed, irrespective of the variable's distribution (Converse, 1987: 42). In his 1915 book, Bowley proposed the use of the probable error (rather than the 'standard error' based on the *squared* errors) as a measure of sampling error. Also, he noted that systematic sampling yielded smaller errors than random selection (Kruskal and Mosteller, 1980: 184). Bowley did not, however, conclusively favour random over purposive sampling. Fisher's work was important as well – not because it contributed directly to survey sampling, but because it established the critical importance of randomization. Interestingly, Neyman gives credit only to Bowley, and not Kaier, for the 'representative method'.

[6] The story of survey sampling does not end with the ascendancy of conventional probability samples. Groves and Lyberg (2010) write: 'For years, survey samplers have argued about the merits of deeper stratification permitted in a systematic sample on an ordered list, yielding biased estimates of sampling variance, versus a paired selection design from fixed strata, yielding an unbiased estimator for the sampling variance. Most practical samplers forgo the unbiased estimate of the sampling variance (and error measurement) for the "assumed" lower magnitude of the true sampling variance' (p. 873). In other words, a sample with greater precision is better, even if the error cannot be calculated exactly.

[7] Neyman examined Gini and Galvani's non-random sample of 29 of the 214 geographical units of an Italian national census. Although the 29 units were chosen because they were closest to the overall population average on a number of variables, the resulting sample was shown to be *dis*-similar to the population on other measures. In addition, restricting the selection to units that were close to the average resulted in underestimates of the variation in the population. Neyman's paper is fascinating and not difficult to read.

stratification, whereby the population is divided into two or more sectors or 'strata' and a separate sample is selected in each stratum, not necessarily with the same probability.

Neyman established *the* criterion for the precision of an estimate of a population characteristic: for a random sample, the *confidence interval* is the range of possible values of a population characteristic, with a specified probability. So, the conventional 95 per cent (or any other) confidence interval refers to the range of values, *computed from the sample itself*, that includes the true population value with 95 per cent probability. Important for researchers who need to design a sample before collecting any data, the expected confidence interval of an estimate can be computed from the sample design parameters – the structure and size of the sample – and (except for small samples with non-normal distributions) does not depend on the distribution of the variable of interest.

Neyman also showed how to compute confidence intervals for cluster samples, where the first stage involves the selection of a random sample of *groups*, and then individuals ('elements') are selected within the groups. The most important application of this idea was to 'area probability' samples where geographical areas, such as municipalities, rural districts or city blocks, are selected first, and then a random sample of households is selected in the areas. The huge advantage of area probability samples is that a list of the entire population is *not* required, only a list of all the geographical areas into which the population is found, taken from a census or government records. Neyman's 1937 lectures at the Graduate School of the US Department of Agriculture (published as Neyman, 1938) led to the design of the Sample Survey of Unemployment, the first modern labour force survey, soon renamed the *Current Population Survey*[8] (Frankel and Stock, 1942; Fienberg and Tanur, 1983: 136).

According to Smith, 'The only major features of current survey design that he [Neyman] failed to introduce were multi-stage sampling and variable probability (p.p.s.) sampling, but these followed logically from his work' (1976: 185). The development of area probability samples required two further steps. First, working at the US Bureau of the Census, Hansen and Hurwitz (1943) showed that the most precise estimates of population characteristics were obtained with 'paired selection' – at each stage of a sample where clusters are selected (for example, communities, census tracts or city blocks), *two* clusters should be selected at random, with probability proportional to their size. Second, in the mid-1960s Leslie Kish and his colleagues developed a method called 'balanced repeated

[8] The Sample Survey of Unemployment did not employ a strict probability sample. Urban counties outside the largest cities were stratified into 27 cells, based on the cross-classification of three population size groups, three geographical areas and three economic levels. From each cell, one county was selected, except for the largest cell where two were selected (Frankel and Stock, 1942): 'The selection of counties was at random except that a deliberate effort was made to maximize state coverage' (p. 79). Also, the decision to 'maximize state coverage' by selecting just one county (a cluster) per cell did not allow the computation of errors, because the variation between clusters cannot be estimated.

replication' to estimate the precision of estimates of complex statistics, such as differences between means and regression coefficients, from multi-stage samples (Kish and Frankel, 1970).[9]

The first standard texts on survey sampling appeared in the 1950s (Smith, 1976: 186), but the practical methods for estimating errors in complex samples only came into view in the 1970s (Kish and Frankel, 1974) and they were not incorporated in standard survey analysis software until the mid-2000s.

For studies of political attitudes and market research, the adoption of probability samples was much slower. In 1944, Stock wrote:

> A stratified random sample may be entirely selected in the central office, in which case the interviewer's quota will consist of a specific list of names and addresses; or the stratification alone may be determined by the central office, in which case the interviewer's quota will consist of a set number of interviews with each of the various *types* of people. With this method the individuals representing each type are selected 'at random' by the interviewer. The first method, widely used by government agencies, is more accurate but also more expensive. The second method is relatively inexpensive and accurate enough for most public opinion research. It is used by the vast majority of opinion research agencies today. (p. 142)

The 'various *types* of people' from which interviewers were to select specified numbers of survey respondents 'at random' were identified by their 'colour', age, sex and economic status. Bias could arise from mistakes in classifying people on sight, but also there was 'reluctance of the typical middle-class interviewer to approach people in the lowest economic brackets' (Rugg, 1944a: 149).

Berinsky describes the rationale for quota samples in the public opinion research of the 1930s and 1940s as follows:

> Gallup and Roper did not trust that chance alone would ensure that their sample would accurately represent the sentiment of the nation. Through the selection of particular interviewing locales and the construction of detailed quotas for their employees conducting interviews in those locales, these researchers presumed that they could construct a representative sample. (2006: 502)

While this suggests a distrust of the fundamental principles of probability, the strategy has some merit in light of the cost constraints of the public opinion industry, the small sizes of the sample in each community, and the vagaries of survey fieldwork of the time, particularly the interviewers' difficult-to-control avoidance of poorer dwellings and poorer-looking and less cooperative respondents. Without the guarantee of unbiased estimates that comes with probability samples, to a degree the success of non-probability methods relies on luck, which eventually fails.

[9] For a fascinating interview with Leslie Kish, who fought in the Spanish Civil War before becoming a leading statistician, see Frankel and King (1996).

Pre-election polls in US presidential elections first legitimized and then undermined quota sampling. After successfully predicting the winners of the four elections between 1920 and 1932, the *Literary Digest* magazine's 1936 poll mistakenly projected the election of Landon over Roosevelt, based on the 25 per cent return of more than 10 million 'ballots' sent out to its readers and to names taken from car registrations and telephone books. The failure is attributed to the over-representation of the middle and upper class among the magazine's subscribers, car owners and households with telephones, compounded by similar bias in the response rates of people who did receive a ballot. Also, changes in political support over the course of the campaign may not have been captured, because many ballots were returned early in the campaign (Squire, 1988; Cahalan, 1989), a problem that plagues election polling to this day. Surveys using quota samples by George Gallup's *American Institute of Public Opinion* and by Archibald Crossley correctly predicted Roosevelt's victory (Crossley, 1937), apparently vindicating their sampling method.

Then in 1948 polls by Crossley, Gallup and Roper all incorrectly predicted the victory of Dewey over Truman in the US presidential election. An investigative committee appointed by the Science Research Council and headed by statistician Frederick Mosteller did not fault quota sampling in principle, although it concluded that: 'It is impossible to separate the error introduced by the quotas set from that arising in the process of selection by interviewers' (Committee on Analysis of Pre-Election Polls and Forecasts of the Social Science Research Council, 1948: 608). The fault was seen to lie in the design of the particular samples – setting quotas that did not match the voting population – or the misapplication of quotas by interviewers. Nevertheless, this Report effectively ended the use of quota sampling for academic studies and led to its slow demise in market research.

About Survey Questionnaires

Long before modern surveys, censuses and the social surveys included extensive questions about individual demographic characteristics and the economic condition of households, and early twentieth-century 'intelligence' tests employed questionnaires of a kind. Modern surveys covered a much broader range of topics, beginning with attitude studies by psychologists, market research and election polling, then extending to a wide range of research on personal experience and perceptions of life.

The first book on question design and the culmination of this period was Payne's *The Art of Asking Questions*; his

> 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 … 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. (1951: xi)

Modesty did not leave Payne short on concrete suggestions, including 'a concise checklist of 100 considerations' for question design, detailed consideration of question formats and an annotated checklist of 1000 common words.

It was recognized that questions on subjective topics were more ambiguous and prone to bias. Abstract concepts and greater detail could make a question more difficult to answer, increasing measurement error and non-response, and the use of response categories with vague boundaries (such as 'agree' versus 'strongly agree') was unavoidable. Also, the validity of answers to subjective questions could not be established by comparison to records or other concrete measures (Cantril and Fried, 1944: 23; Connelly, 1945).

Cantril and Fried's list of the pitfalls of question design is perfectly contemporary. Questions could be 'too vague to permit precise answers', 'obscure in meaning', 'getting at some stereotype or overtone implicit in the questions rather than at the meanings intended', or 'misunderstood because they involve technical or unfamiliar words'. The alternative answers might be too numerous, too long or not exhaustive, or a question might be 'concerned with only a portion of the population and therefore meaningless to many people' (1944: 3).

Payne's book ends with the advice that 'Controlled experiment is the surest way of making progress in our understanding of question wording' (1951: 237). This involved printing two versions of a questionnaire, with some questions worded differently in each version, called a 'split ballot'. Roper included experiments in his surveys from the mid-1930s and Cantril followed shortly. One of the first systematic treatments of question wording, by Rugg and Cantril (1942), is based largely on experiments, although they caution that 'there is seldom any way of determining which presentation is the more valid … evaluations of the relative merits of different presentations of an issue must rest on *a priori* considerations.'

From their experiments, Rugg and Cantril concluded that:

- '[O]n issues where people were uncertain, it was possible to produce sizable effects by biasing [the formulation of a question], but where opinion was well crystallized, biasing statements had relatively little effect' (1942: 491). Intentionally biased questions could therefore be used to measure the stability of public opinion, on the basis of comparisons to the responses to neutrally worded questions.
- The number and wording of the responses to a question affect the distribution of responses. Respondents tend to choose only from the responses offered to them explicitly, even if this resulted in bias: 'Where a genuine intermediate step exists … distortion inevitably results when answers are forced into a dichotomy' (1942: 479).
- The tone of a question could affect the answers. For example, Rugg (1941) found that 62 per cent of Americans would 'not allow' speeches against democracy', but only 46 per cent would 'forbid' them. For other topics, however, a comparison of the same two words could result in a smaller effect of wording, or no difference at all.

For more complex topics, there was extensive debate over the use of 'open' and 'closed' questions. 'Closed' questions had respondents choose among fixed answers, while ‘open’ questions did not offer any answers and the verbal response was recorded verbatim. An example might be a question about the most important problems facing the country. The issue was whether the greater cost of open questions, in terms of time required to answer, the need for better trained interviewers, and the need to classify the responses after the interview resulted in better answers.

This became a dispute between the two groups conducting surveys in the wartime US government. Naturally, the division led by Elmo Wilson, whose background was in commercial political polling, favoured closed questions, but it is harder to understand why psychologist Rensis Likert strongly supported open questions (Converse, 1984; Hyman, 1991). Indeed, based on his earlier research on scales, Likert’s name is given to questions asking for respondents’ opinions on statements, using a scale from ‘strongly agree’ to ‘strongly disagree’. For reasons of cost and timeliness, the commercial polling firms and their associated academics, including Lazarsfeld and Cantril, firmly sided with closed questions.[10]

This carried over into support of open questions at the University of Michigan’s Institute for Social Research (ISR), which was Likert’s post-war academic destination, and for closed questions at the University of Chicago’s National Opinion Research Center (NORC), which was tied to the political polling firms. Eventually, the conflict was resolved decisively in favour of closed questions, although Converse observed that: ‘The open/closed debate was shaped in good part by institutional needs and capacities, and by ideologies *about* research, remaining largely untouched *by* research’ (1984: 279).

Cantril and Fried’s view, from 1944, is close to the current consensus:

> The major advantage of the open-ended or free-answer question is obviously its ability to record opinion which is catalogued to the minimum degree by the investigator. When issue has become fairly clear-cut, however, or where common sense and experience have shown that meaningful alternatives can be posed, there is little advantage to an open-ended question from the point of view of its faithfulness in reporting opinion. There is even, on the contrary, a considerable disadvantage in the open-ended question from the point of view of reporting precise trends, keeping costs down, and avoiding bias in the coding of answers for statistical treatment. (p. 10)

Perhaps this makes a virtue of necessity, because the development of national public opinion polls in the 1930s was predicated on the ability to sell timely reports

[10] Lazarsfeld (1944) attempted to mediate this dispute by proposing a division of labour for major survey projects. First, an open survey would be used to develop closed questions – so that the answers offered in the closed questions included the full range of responses. Those questions would then form the basis of a larger survey using closed questions. Third, a more qualitative survey using open questions would be used to confirm the interpretation of results from the larger survey. As a general practice, this lengthy and complex strategy was simply impractical.

of the findings to major newspapers and magazines, which was much easier using closed questions.

Researchers of the time understood the potentially multi-dimensional character of attitudes and they distinguished between the answers to individual survey items and more fundamental traits underlying them. In a survey conducted in the USA in 1941 for example, Harding (1944b) used about 30 questions on 16 separate topics to measure civilian morale, and he employed factor analysis (an arduous manual calculation in the time before computers) to identify three underlying dimensions of morale.

About Data Collection

The method of data collection of the new survey research was face-to-face interviewing, and the main concern was the effect of interviewers on survey response. By the mid-1940s there were studies of:

- non-response bias due to respondent refusals – although it was not perceived as a substantial threat, at this time when response rates approached 90 per cent (Harding, 1944a);
- the effect of interviewer training on the quality of survey response – more training did not seem to have much effect on the quality of the data, perhaps because most interviewers were well educated (Rugg, 1944b);
- the effect on survey response of the presence of an interviewer – it was found that differences could arise between questions answered on a confidential paper 'ballot' and the answers given to interviewers directly (Turnbull, 1944);
- whether the interviewer's own opinions led to bias in responses – evidence of bias was found by comparing respondents' and interviewers' answers to the same questions (Cantril, 1944: 107ff.); and
- the impact of the interviewer's social class and race – for lower income respondents, Katz (1942) found that working-class interviewers found higher levels of support for labour, while middle-class interviewers found more conservative views; also the combination of interviewer's and respondent's race affected survey responses – to questions about the living conditions of African-Americans, black and white respondents expressed different views to black and white interviewers, with a much larger racial gap in a survey conducted in Memphis than in New York (Hyman, 1991: 39).

Conducting surveys required organizations able to conduct face-to-face interviews on a national scale. Although censuses were just huge face-to-face surveys, their enormity, infrequency and restricted content were the opposite of the agility needed to conduct timely surveys on a variety of topics. Government agencies did develop the capacity to conduct surveys, initially labour force surveys to measure unemployment, but market research and advertising agencies were the first to develop modern survey infrastructure and their methodology was ripe for franchising. George Gallup, who founded the American Institute for Public Opinion in 1935, established the British Institute of Public Opinion in 1937 and French, Australian and Canadian affiliates in 1937, 1938 and 1942, respectively.

The two leading American academic survey centres also date from this period. The National Opinion Research Center was established at the University of Denver in 1941 by a close associate of Gallup, before moving to the University of Chicago in 1947, and Rensis Likert left the US government to establish the Institute for Social Research at the University of Michigan in 1946 (Converse, 1987: 305). By the early 1950s, there were dozens of Gallup affiliates. Within each country, at a time when communication was largely by mail, the economies of scale favoured the emergence of large oligopolistic survey organizations.

Conclusion

Modern survey research emerged between 1935 and 1940, with sample designs capable of representing almost any population, questionnaires covering a wide range of objective and subjective topics, and the development of procedures and establishment of organizations for large-scale face-to-face surveys. Each element was necessary, but it was the combination of sampling, questionnaire design and data collection that constituted the invention of survey research.