Balnaves, Mark, Tom O'Regan, and Ben Goldsmith. "The Panel and the Survey." *Rating the Audience: The Business of Media*. London: Bloomsbury Academic, 2011. 49–73. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781849664622.ch-003>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 11:07 UTC.

Copyright © Mark Balnaves, Tom O'Regan and Ben Goldsmith 2011. You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

3 The Panel and the Survey

The claims that have been made for political opinion polling have not been modest. George Gallup has said that the ‘modern poll can, and to a certain extent does, function as a creative art of government’, and refers to political opinion polls as a ‘sort of American equivalent of a vote of confidence in the government’.

Catherine Marsh (Marsh 1984: 32)

There is no logical difference between the study of voting or of buying. In each of these areas, the final goal is the discovering of regularity in social life. Hans Zeisel (Zeisel 1968: xvii)

National television ratings should not be confused with public opinion polls or election forecasts or surveys of how much of what consumers plan to buy. Ratings are better than such estimates, because they do not have to deal with confusions introduced by phrasing of a question, or difficulties in predicting who will vote, or the chance that economic conditions will change. They seek only to measure only a very simple piece of behaviour at or close to the time that it occurs.

Martin Mayer (Mayer 1966: 23)

In Chapter 2 the authors argued that audience ratings are not only a measure; they are a convention that brings together many parties in its creation and use. The most important aspect of this convention is that audience ratings have a *dual identity*. They are a public interest vehicle showing the public preference for programmes or media channels and a market mechanism for making decisions on the media dollar. Ratings are part of politics because they represent the public and the market at the same time. In this chapter the authors will show how the ratings intellectuals created the surveys and panels that have become part of the audience-ratings convention. We will also show the changing pressures on the panel as demands increase to represent increasingly demographically defined publics. These demands are a part of broader survey and sampling wars. The US Census, for example, under-represents its minorities, and the role of sampling and the survey has become the subject of battles over ways to count them, to the point where the Republicans vetoed the commissioning of a Census Director’s appointment because of his association with previous plans to use statistical sampling to represent minorities in

the Census. Audience ratings panels are based on establishment surveys using ... the Census.

The Ratings Intellectuals

George Gallup and Archibald Crossley gave speeches after the 1936 US presidential election in which they stressed that techniques used in political polling were applicable to market research. They did not say this as a commercial plug, pushing their own commercial interests that they had in developing and applying public polling techniques. Both made the link between public opinion polling and markets because, like other contemporary experts in methodology, they had an intense interest in the use of advanced survey research techniques in enhancing democracy, and markets, for them, were an integral part of democracy. The idea of surveys as genuinely being able to represent the 'public vote' was no accident. The dual identity of audience ratings as public vote and market measure was a part of the ethos of Gallup and Crossley's intellectual contemporaries. At the heart of this dual identity was 'scientific sampling' and methodological and popular understanding of what it was and its limitations. As Crossley wrote:

In the early days of sampling there was, and it may be noted that there still is occasionally, a tendency to think of cases taken 'at random' as being typical or representative of a universe, although the term 'universe' was not used. When the transition was made from 'at random' to true 'randomization', the lily was gilded with the phrase 'scientific sampling'. This gilding, I would say, was done at the time of the introduction of the national polls on political issues in the mid-thirties. Actually, however, reasonably reliable stratified or quota sampling methods were in use in marketing research ten to fifteen years before this. Already at that time the effort was to establish what George Gallup later so aptly called a 'micro-America', by selecting the sampling so it would be representative of all possible breakdowns, as shown by the latest census. To the best of my recollection this use of the word 'scientific' was the first of a long string of competitive catchwords which has so thoroughly characterized claims about sampling and interviewing methods. While the bruited word 'probability' did not achieve this fanfare until the 1948 nadir of the polltakers, the word had been used considerably earlier, and many probability principles had been used for a long period – e.g. rotation and randomization of blocks, road segments, homes, and individuals, and the assignment of specific locations to interviewers. (Crossley 1957: 162)

Crossley's point about 'competitive catchwords' taps directly into concerns from survey research methodologists of the time about how an essentially probabilistic

approach might be used to deliver precision statements. Statistics delivered from samples are inferential and have an element of *error* associated with them. Calling them 'scientific' does not make that error less certain. Using decimal points in polling figures or ratings figures as exact when there is a possibility of 5 per cent sampling error, for example, is ludicrous, but advertisers and marketers do exactly this, regardless of sampling error estimates. This problem exploded in the 1960s with the US Congressional hearings on ratings, which the authors deal with in the next chapter.

Crossley was also highlighting the major transition from 'random' as an attempt to remove bias in earlier studies and 'randomization' as the mathematical attempt at representation, a major advance in polling and ratings and methodology and little understood in the industry (something Bill McNair and George Anderson knew very well in their attempts to sell randomization and the extra costs that it might incur).

Gallup and Crossley are what the authors call 'ratings intellectuals'. They were not just academic in their work, publishing academic papers, but interested in all things ratings or polls – working within the industry itself, developing the very techniques they critiqued and publishing seminal papers along the way. They also participated in the fierce politics involved in the application of the techniques and any public arguments about them. C.E. Hooper co-authored a book outlining the deficiencies of Crossley's telephone recall method but also took the argument to industry forums and the industry press. Bill McNair and George Anderson in Australia, likewise, saw themselves as academically interested in the methodologies and not mere technicians. They also saw themselves as taking on the broader role of informing the industry and the public on how audience ratings worked.

Crossley worked with Gallup on public opinion polling after leaving audience ratings in the 1940s, and had maintained his interest in public opinion polling while working with CAB. Gallup himself, of course, was not only interested in public opinion polls. He launched the Audience Research Institute (ARI) in 1940 in the West Coast headquarters of Young & Rubicam; 'a cozy reciprocity' (Ohmer 2006: 79).

ARI conducted a quarterly study that it called the 'continuing audit of marquee values'. The audit asked the public whether seeing a particular person's name on a theater marquee would in itself induce them to buy a ticket to the film. ARI compared these studies to the ratings that Archibald Crossley had carried out for the Cooperative Analysis of Broadcasting, so these audits represented a direct transference of personality ratings from radio to film. (Ohmer 2006: 79)

It was not Gallup or Crossley, however, who developed the underlying sampling techniques that came to dominate audience survey research, survey methods generally and, especially, longitudinal research. Paul Lazarsfeld studied the effects

of radio in 1937 and discovered that radio listening created no public records, such as circulation records in the case of print. He took the public opinion poll and ratings methodology and by the multivariate analysis of responses developed ways to measure the impact of radio on attitudes. 'This transformation of the opinion poll into multifaceted survey research constitutes one of Lazarsfeld's major accomplishments' (Sills 1987). But Lazarsfeld's work on longitudinal methods itself fed back into audience ratings methodology. The *panel* was a revolution in collecting data about the same people over time.

A major finding of Lazarsfeld's research on radio listening is the tendency of audiences to be self-selected; that is, to tune in to programs that are compatible with their own tastes and attitudes. Accordingly, in order to sort out the causal sequences of such problems as the effect of listening upon attitudes versus the effect of attitudes upon patterns of listening, a method of determining the time order of variables was required. Drawing on his research in Vienna with the Bihlers, in which repeated observations were made of the same children over time, as well as on the earlier research of Stuart A. Rice among Dartmouth College students and Theodore M. Newcomb among Bennington College students, Lazarsfeld developed what he called the panel method, in which a sample of respondents is reinterviewed at periodic intervals. The panel method is a form of longitudinal research; it is essentially a field experiment in which a 'natural' rather than an experimental population is studied. Although Lazarsfeld cannot be said to have invented the panel method, it was his imaginative use of it, and particularly his innovative ways of introducing control groups into the analysis of panel data, that made him its earliest and most effective exponent. (Sills 1987)

Lazarsfeld

Paul F. Lazarsfeld ... virtually created the fields of mathematical sociology, multivariate survey analysis, and the empirical study of both voting behaviour and mass communications.

David L. Sills (Rogers 1994: 244)

Theodor W. Adorno, Hans Zeisel, Herta Herzog, Robert Merton, Bernard Berelson and Leo Lowenthal were all famous colleagues of Lazarsfeld. All conducted research on media and new methods emerged from the research. In the case of Merton, a lifelong friend and associate of Lazarsfeld, his focused interviews and focus groups technique became a standard part of industry data collection as well as of qualitative methods generally. Adorno provided the bridge with European critical approaches

and drew the famous distinction between ‘administrative research’, which served government and industry, and ‘critical research’, which sought to critique the whole capitalist enterprise and the role of ideology. While Adorno saw Lazarsfeld’s work as primarily of the former rather than the latter kind, Adorno and Lazarsfeld shared a common interest in making research methods democratic, transparent and hybrid – making sure that qualitative research informed quantitative research and vice versa. Lazarsfeld, Berelson and Gaudet’s seminal work *People’s Choice* (Lazarsfeld, Berelson and Gaudet 1944) is a classic example of hybridity and the application of the panel method to assist with explanatory research.

The 1940 presidential election in the United States presented to Lazarsfeld an opportunity to undertake a study of voting intentions in new ways. Voting records were until then the only material used for analysing voter behaviour. These records made it possible to study the geographical distribution of voting results as well as individual votes. Public opinion polls went a step further and correlated political opinion to the characteristics of the individual voter. Gallup and Crossley’s opinion polls, however, were normally conducted with different people and *not the same people over time*. What Lazarsfeld decided to do was to trace a person’s voting intentions over the course of a political campaign. To do this he created a ‘panel’ technique using the same set of voters over the period of the political campaign and election.

Lazarsfeld’s broader research aim was straightforward: ‘We are interested here in all those conditions which determine the political behaviour of people. Briefly, our problem is this: to discover how and why people decided to vote as they did. What were the major influences upon them during the campaign in 1940?’ (Lazarsfeld *et al.* 1944: 1). Within this broader aim, Lazarsfeld asked, ‘What is the effect of social status upon vote?’ ‘How are people influenced by the party conventions and the nominations?’ ‘What role does formal propaganda play?’ ‘How about the press and the radio?’ ‘What of the influence of family and friends?’ ‘Where do issues come in, and how?’ ‘Why do some people settle their vote early and some late?’ ‘In short, how do votes develop? Why do people vote as they do? By inference and by direct accounts of the respondents, we shall try to show what influences operated’ (Lazarsfeld *et al.* 1944: 6–7).

Lazarsfeld and his colleagues had a specific *sampling frame* and *variables* in mind for measurement. Erie County was chosen for the study ‘because it was small enough to permit close supervision of the interviewers; because it was relatively free from sectional peculiarities, because it was not dominated by any large urban center, although it did furnish an opportunity to compare rural political opinion with opinion in a small urban center’ (Lazarsfeld *et al.* 1944: 3). Three thousand people were chosen to represent the county and a *systematic* sampling technique

adopted. Trained local interviewers, mainly women, visited every fourth house in the county. Four groups of 600 persons were selected by *stratified sampling* and of these only three were re-interviewed once each. These were control groups to test the effect that repeated interviewing might have on the panel (Lazarsfeld *et al.* 1944: 3). The fourth group, the *panel*, was interviewed once each month from May to November 1940.

Operationalization of 'social and economic status' gives the reader an insight into how the study went about quantification. The interviewers were trained to assess the homes, possessions, appearance and manner of speech of the participants. The interviewers then classified them into a stratum according to a set quota: 'The people with the best homes, furniture, clothes, etc, i.e., the ones with the most money, would be classed as As; and the people at the other extreme would be Ds. In Erie County, the quota was approximated in the following distribution: A, 3 percent; B, 14 percent; C+, 33 percent; C-, 30 percent; and D, 20 percent' (Lazarsfeld *et al.* 1944: 17). This was in an attempt to reflect the broader sampling frame. The study used simple and more complex measurements for analysis. For magazine reading, the respondent was asked about several specific articles appearing in current issues of magazines. The resulting index was a count of the number of articles on political affairs that a person said that they had read. In the October interview each person was asked whether they agreed with eight arguments then current in the political campaign, on a scale of -8 to +8. One of the eight arguments was 'Roosevelt has great personal attractiveness, capacity for hard work, and keen intelligence'.

One of *People's Choice's* main findings was that media appeared to have little influence on changing people's voting intentions. Media had a reinforcement not a transformative effect. 'The first thing to say is that some people *were* converted by campaign propaganda but that they were *few* indeed' (Lazarsfeld *et al.* 1944: 94). Interpersonal influence emerged as a key factor. 'Whenever respondents were asked to report on their recent exposures to campaign communications of all kinds, political discussions were mentioned more frequently than exposure to radio or print' (Lazarsfeld *et al.* 1944: 150). The voters most likely to change their voting intentions in the political campaign 'read and listened least' (Lazarsfeld *et al.* 1944: 95). The study also found that opinion leaders had a dramatic influence on people around them. Print and radio did not change behaviour because 'the people who did most of the reading and listening not only read and heard most of their own partisan propaganda but were also most resistant to conversion because of their strong dispositions' (Lazarsfeld *et al.* 1944: 95).

Lazarsfeld and his colleagues' study has all the elements used in modern broadcast ratings, including not least the panel. However, before proceeding to

discuss issues in sampling in more detail, it is worthwhile revisiting how quantitative methods and probability sampling work. The authors italicized some of the concepts that are basic to quantitative methods. Operationalization is an overarching concept that describes quantitative, variable-analytic, method. Wanting to operationalize 'gender' as a nominal variable measured by the categories 'male' and 'female', for example, is unlikely to encounter significant resistance. But not all constructs or phenomena are this easy to measure or indeed to reach an agreement on. Social scientists use language to describe and to define the phenomena that they measure and this language has to be accepted by others to be valid. The social scientist tries to measure constructs described by language that they argue corresponds to the phenomenon of interest.

The Very Idea of Measurement

In the case of audience ratings *exposure* is the phenomenon of interest. Have audiences been exposed to the radio channel and the radio programme, the television and the television programme? This construct might be described in natural language as: 'The stations and programmes listened to, the times that they were listened to, and the demographics of the audience.' This description might then be broken down into various variables that represent real, observable, measures, such as those outlined in Frank Stanton's 1935 doctoral dissertation:

- 1 When does the listener use his receiver?
 - 2 For how long a period does he use it?
 - 3 To what station or stations does he listen?
 - 4 Who listens (sex, age, economic and educational level)?
 - 5 What does he do while the receiver is in operation?
 - 6 What does he do as a result of the program?
 - 7 What are his program preferences?
- (Stanton 1935; cited in Webster, Phalen and Lichty 2005)

How these questions are operationalized as variables becomes very important. For example, question 2 could be operationalized as a nominal/categorical variable or at a higher level of measurement:

'For how long a period do you use your radio receiver?'

A long period

Not at all

This answer, you might tell the authors, is inappropriate. It gives the analyst too little information. You might propose a set of answers like:

‘For how long a period do you use your radio receiver?’

Monday

Tuesday

Number of Hours

Number of Hours, etc.

The definition of the construct and the level of measurement of the question are analytical choices of the researcher, but they must be useful to the research and have validity. If you have noticed, Stanton’s questions 5–7 go beyond exposure in our definition of the construct. These are the kinds of decisions Archibald Crossley had to face. What definitions and measurements of exposure would ratings intellectuals and the industry accept? Which measures were seen as more objective than others?

Figure 3.1 summarizes the relationship between construct and measure. If the authors were to define exposure as ‘how engaged the listener is with the radio programme’ then a whole new set of variables would be needed. Is ‘engaged’ related to length of time that a person listens to the programme or the amount of attention that they show? Crossley, not surprisingly, tried to steer clear of this question of attention. Exposure was measured by whether the listener said that they were listening and for how long, not ‘how much attention the listener paid to the radio programme’. The problem of engagement versus exposure is a contemporary debate but, as the authors want to show you, it is also an old debate and one faced by the early ratings intellectuals. The comment by Martin Mayer at the beginning of this chapter, that ‘they seek only to measure only a very simple piece of behaviour at or close to the time that it occurs’, is highly misleading as it is not a simple behaviour. For example, Nielsen would contest that the problem with self-report, the listener telling you when they turned on the radio set, is itself not reliable and thus his mechanical measure was more valid than self-report. Even the ‘simple behaviour’ can become contestable.

Audience ratings therefore do not escape the same problems faced by any social phenomenon. The construct itself requires qualitative description, in natural

Figure 3.1 From construct to measure

Source: Balnaves and Caputi 2001

language. That construct has to reflect a phenomenon in the real world that can be observed. The methods and the measures chosen need to reflect the phenomenon accurately. Agreement on the measure is not necessarily easy. A construct like 'child abuse' could have many measures, but each could also be contestable. For instance, 'smacking a child five times a week' might be defined by a researcher as child abuse, but the question might be asked, 'Of what does the smack itself consist? And how can it be observed?' In television viewing, having the station switched on is no guarantee that anyone is in the room. Or they might be in another room listening to the television, partly. Or they might be in the television room, talking to other people and not listening to the programme at all. As we can see, this is the type of trade-off that Crossley had to make in his pioneering efforts to create the measure. What do I count? And how do I count it?

The 'construct' therefore is your idea about the phenomenon that you want to measure. The operational definition is the statement about how you want to measure the construct. 'The construct "delinquency", for example, might be operationally defined by "being arrested more than once prior to the age 18". In a questionnaire you might have the question (your variable), "Have you been arrested more than once prior to the age 18? Yes. No." This is a nominal level question. This is one variable. It is also possible to imagine other definitions and operational definitions of the construct that might be useful' (Balnaves and Caputi 2001).

In general, social science assumes that attributes of a phenomenon are measurable. 'Male' and 'female' are, for example, attributes of 'sex' and as categories can be assigned the numbers 1 and 2. Notice, however, that we can only derive frequencies from this measurement, called a nominal or categorical measure. We cannot derive an average of males and females on this measure. If, however, one was to ask each male and female their ages and they gave them to me as '17', '18' and '36', then one would be able to derive averages from this higher level of measurement. Variables embody both constructs that we want to define and the numbers that we use to represent them.

Variables can be operationalized at various levels of measurement, and traditionally textbooks have listed four levels of quantification or measurement – nominal, ordinal, interval and ratio measurement. *Nominal or categorical level* measurement consists of categories that can be given a name or a number. *Ordinal level* measurement has the same properties as nominal scales with the additional property that the categories can be rank-ordered. In audience research, getting people to rank their preferences of programmes is an ordinal scale. Note, however, that the distance between intervals is not necessarily equal. A person who ranks their programmes 1, 2, 3 may like number 1 significantly more than 2 and 3.

Interval level measurement assumes equal intervals on the scale represent equal amounts of the quantity being measured. If we are measuring household income in dollars, then the difference between annual income of US\$60,000 and US\$65,000 is the same amount as the difference between someone who earns US\$70,000 and US\$75,000. One could ask viewers to rate the content of three television programmes in terms of humour using a 5-point scale where '0' means 'not at all funny' and '5' means 'extremely funny'. If we can demonstrate that the difference (in terms of the amount of humour represented in the variable) in ratings between '0' and '1' is the same as the difference between ratings of '3' and '4', then we are using an interval scale. Clearly, establishing this property in this case may be difficult. It is possible to try to establish what properties the observations should have in order to lead to the interval scale measurement and then to investigate whether these properties are in fact met by the observations. Cliff argues, however, that in the social sciences we can only achieve ordinal level measurement (Cliff 1996).

Ratio level of measurement has all the properties of ordinal and interval measurement with the additional property that equal ratios between numbers on the scale represent equal ratios of the attribute being measured. Height measured in centimetres (cm) is an example of ratio measurement. Someone who is 180 cm tall is twice the height of someone who is 90 cm tall. The levels of measurement represent the mathematical possibilities available for quantitative analysis such as adding, subtracting, multiplying and dividing – when it is decided how one wants to define the phenomenon one wants to study; how to measure one's observations. Definition, as we have seen, precedes measurement.

It takes students of statistics in the social sciences and media a little time to get their minds around the idea that there are in fact different levels of measurement and that choice of measurement must relate directly back to their definitions of the constructs. There may be logical limits on which measures correspond to which defined phenomenon. Modern sampling, no less, creates similar difficulties. Modern sampling requires a *sampling frame* and this frame (population or universe) and the sampling technique(s) to be used are decided upon by the researcher, as you saw in the case of Lazarsfeld in *People's Choice*.

Constructs and variables are defined by the researcher. Populations and sampling frames are also operationally defined by the researcher. A sampling frame must be accessible and quantifiable, and related to the purpose of the research. 'All households in New York' is a definition of a sampling frame, with households as the unit of analysis. But, one would correctly ask, what counts as a 'household'? Is it any dwelling, including the boat in the backyard that one's relative lives in? If it is, then this example needs to be included in the definition of the population,

the sampling frame. If a sample of households was drawn from New York, then 'New York' also needs to be defined. Is New York defined by local government boundaries? Is it defined by census boundaries?

When one has decided on all these definitions then every household that meets one's definition is on the list. That list is called a *sampling frame*. It is called a sampling frame because it constitutes a defined universe or population of elements that is to be studied. Without this kind of defined list modern probability or random, sampling selection procedures cannot be used. 'Probability' is perhaps better than 'random' as a term for modern sampling, as Crossley pointed out earlier. It is called random because it is supposed to be non-biased in the selection of the sample. In the popular mind, however, 'random' means 'arbitrary', which modern random sampling certainly is not. This confusion has affected the media industry perception of sampling because 'random' has never been well understood. McNair and Anderson in Australia had to continuously argue, reargue and explain modern sampling techniques, because industry commentators or ratings users found the idea that a thousand households could represent a million households almost impossible to accept.

You can number each household in New York and put them in a large hat. If you drew out ten households then you would have a *Simple Random Sample* (SRS). You know that there was no bias in your choice and each element, each household, had an equal chance of being chosen. You would be more likely to use a table of random numbers and select 50 households out of 1,000 households. In this case you would number each household from 000 to 999 and then select 50 numbers from the table of random numbers. In Table 3.1 below, a shortened version, you start anywhere in the table of random numbers and then, proceeding up, down or diagonally, select the 50 numbers needed. In this case, the selection starting from 284 downwards would be 284, 361, 779, 176 and so on. Each of these numbers would represent a real household.

By chance you might have selected 50 households that only included people who lived in boats in the backyard in New York. You have not represented *all* households in your sampling frame, only boats. An additional probability technique

Table 3.1 Table of random numbers

2	8	4	9	8	8	9	9	5	5
3	6	1	9	0	3	0	1	1	1
7	7	9	9	7	8	3	3	3	2
1	7	6	9	2	4	1	8	6	7

might be required – stratified random sampling. You can break the sampling frame into different strata and then conduct a simple random sample with each stratum. The strata might be divided, for instance, into single-level households, double-level household, boats and other classifications. There may be very few boats-as-households in the sampling frame, let us say 2 per cent. If we wanted to represent the 2 per cent in our sampling frame, then 2 per cent of your stratification would be boats. This is called a proportionate stratified random sample representing each type of household proportionately. A simple random sample from each separate list of types of household would be required. Sometimes we might want to over-represent a particular stratum, especially if there are very low numbers, and this is called disproportionate stratified random sampling. In audience ratings, which often use census data for creating establishment surveys, marginal groups are often under-represented because they are not in the census. In the United States, Hispanics have traditionally been under-represented in the census and consequently severely under-represented in the audience ratings panels.

Large-scale national studies use different types of units of analysis in their sampling frame, or their list. In cluster sampling, for instance, we select groups or categories. In Figure 3.2, representing New York, we could break the list up into suburbs and randomly sample those suburbs. We could then break up those suburbs into census collectors' districts and then randomly select from those districts. We could then select streets and *systematically* sample those streets, as Lazarsfeld did in his study. A systematic sample is when every n th unit of analysis is selected – every fifth house in the street, or whichever interval the researcher thinks is best. Once the researcher has selected the sample of households, then they might select a demographic quota from those households (people of a certain age or sex and so on). In order to randomly select the participants the researcher could again use a table of random numbers. For example, the interviewer could ask, 'How many people are there in your home aged 15 or older?' If the first participant says 'three people', then according to the table at Figure 3.3, the second oldest person is chosen.

Simple, stratified and multi-stage cluster sampling are forms of probability sampling, although there is sometimes debate about whether systematic sampling is genuinely random. When probability sampling is not possible then alternative sampling strategies can be used. For instance, it would be hard to get a list of 'heroin users who watch television'. There is no Association of Heroin Users Who Watch Television (although such an association would constitute a valid list). Going to a shopping mall and talking to someone believed to be a heroin user who watched television is *not* random sampling but it is a form of non-random, non-probability sampling. A 'mall intercept' like this might be useful for research, but the result could not be generalized to the population from which it is drawn, firstly

Figure 3.2 Multi-stage cluster sampling – following the census tracts

No. of people in household					
	1	2	3	4	5
Person to interview 1		2	2	4	3
		1	1	2	5
			3	1	4
				3	1
					2

Figure 3.3 Random selection from households for interview

because there is no sampling frame and secondly because defining the population is not simple either.

The argument of modern researchers is that good sampling reduces the chance that we have picked the wrong people, reduces error and maximizes external validity. Internal validity deals with the logic and construction of your research questions, variables, and so on. External validity deals with the construction of the sampling technique. Statistics that make judgements about the relationship between samples and their sampling frame are called *inferential statistics*. Modern audience ratings figures are inferential, relying on samples drawn from sampling frames.

As might be guessed, advertisers and others had and still do have some difficulty in swallowing the argument that 1,000 homes could represent over 50 million people, which is the national television ratings panel size that Nielsen ran in the United States in 1966. Because a probability sample is a sample, and not the full sampling frame, there is likely to be an element of error. Understanding this error is important because a rating point is a probabilistic estimate, not a certain figure. If Nielsen interviewed each of the 50 million or so television viewers, then it would have a census, a full population. This, of course, is too costly to do as well as time consuming. There is also an element of *intrusion*. A probability sample or panel could limit the amount of intrusion into households. A panel could collect longitudinal data and minimize privacy intrusion and increase the speed by which ratings could be measured. In 1966 the Committee on Nationwide Television Audience Measurement (CONTAM) that represented industry and government conducted a major study to see whether sampling theory was correct. Could 1,000 households properly represent over 50 million households with minimum sampling error?

Ten programmes were selected from the rating weeks, ranging from *The Beverly Hillbillies* (a rating of 50.7 in the 56,386 diaries ARB received in the March 1963 ratings period), *Dr Kildare* (a rating of 37.0) through to *Eyewitness* (7.1) and *Voice of Firestone* (3.0). Eight hundred samples were drawn in computer table of random numbers – 100 samples each of 25, 50, 100, 250, 500, 1000, 1,500 and 2,500 diaries. There were then 8,000 samples, 800 for each of the programmes.

Statistical sampling theory says that about 5,460 of these 8,000 samples should produce results falling within the first standard error. The study produced 5,475 within that range, and the appropriate proportions for the other standard errors. The standard error is a statistical measure of dispersion: under normal circumstances approximately 68 per cent of all cases will fall within one standard error of the mean, about 95 per cent will fall within two standard errors, and 99.7 per cent will fall within three standard errors.

CONTAM had shown that sampling theory does apply to television viewing behaviour for a universe or sampling frame of 56,386 homes and that similar results

would be obtained for a sampling frame of over 50 million homes, although this remained to be demonstrated. The part of sampling theory that is most difficult for people to understand is that the accuracy of results obtained by sampling is independent of the size of the sampling frame or the universe from which the sample is taken. A sample of 1,000 is just as accurate for a population of 50 million, the number of television homes in the United States in 1966, as it would be for 50,000. To demonstrate this, CONTAM again took 8,000 samples, 100 each of 8 sizes for 10 programmes from 4,927 diaries from one section of the United States, shrinking the universe to see if the sampling produced the sample results as those produced for sampling the whole universe.

The 8,000 samples from the large population produced 5,475 results within one standard error of the true result and the sample from the shrunk universe produced 5,528 cases. Bar charts were produced to show the actual results from the samples of 1,000 for each of the 10 programmes studied. On each chart the normal curve of the distribution around the true value was superimposed to show the degree to which the bar chart met the predictions of the theory. Each test showed that there was no unusual factor about television viewing that made sampling less effective than in other domains. A sample of 1,000 homes could represent 50 million homes.

In the United Kingdom in the lead-up to the broadcasting of Independent Television (ITV) for the first time in September 1955, and in the period immediately afterwards, the most important issue for the new public regulatory body, the Independent Television Authority, was one of *exposure*: how many homes were equipped with sets able to receive ITV programmes. Television sets in Britain in the mid-1950s either had to be fitted with a new station selector switch or viewers had to buy new sets, and many homes needed new aerials in order to receive a signal from the ITV station in their region. In the period immediately following the commencement of the ITV service, a controversy blew up over differences in figures for ITV's share of the evening audience in multichannel homes in the London region.

Table 3.2 Panel result differences between providers

	BBC Audience Research	Gallup TV Research	TAM Machine Panels
All Week	44%	49%	59%
Monday to Friday	41%	45%	53%
Saturday and Sunday	52%	57%	73%

Source: TV Research (Gallup Poll) Ltd, 'Enlightenment on the Controversy following Publication of BBC and ITV Research', ITA Archive, Audience Research – Gallup Poll 301/4, 6 December 1955

Research in December 1955 by the BBC Audience Research Unit, under Robert Silvey, was broadly consistent with independent research by the Gallup group in findings about the average number of viewers for each network. Research by Television Audience Measurement Ltd (TAM) indicated higher viewing figures for ITV than those recorded by Gallup and the BBC, particularly on the weekend, when TAM recorded a 20 per cent jump in ITV's share compared with weekdays. Gallup and BBC also recorded a rise in ITV's share at weekends, but much smaller in size. Both the BBC and Gallup used personal interviews and aided recall, almost 5,000 in total per week, to ascertain results. Gallup also used the telephone coincidental method to check audience figures. TAM used a panel of about 100 homes, fitted with machines to record the station to which the television was tuned.

In commentary on the discrepancies between the figures, Henry Durant, the director of the British Institute of Public Opinion (BIPO) which conducted the Gallup polls, noted that unlike the BBC and Gallup, which offered no inducement to their respondents, TAM 'as a minimum' maintained the television sets in panel homes to ensure that they could receive the ITV signal. This was considered potentially likely to produce figures that were unduly favourable to the ITV programmes, or as Durant put it, 'Where a man's treasure is, there is his heart also'. Durant also pointed to the potential for the ideological leanings of the respondents to skew the machine panel figures, noting that, 'It needs precious few of [panel members] with a feeling for "free enterprise rather than the BBC monopoly" to produce a slant in the figures' (Durant 1955).

Despite these criticisms, the ITV companies, advertising agencies and advertisers agreed at an early stage to use one source of statistical information about television, and selected TAM with its metered system, supplemented by diaries, over a rival meter system operated by Nielsen and a variety of survey and interview systems operated by BIPO and other organizations. In January 1957, TAM was awarded the industry contract for an initial period of five years. The contract was extended by the Joint Industry Committee for Television Advertising Research (JICTAR). The issue of discrepancies between BBC and TAM/Audits of Great Britain figures for audience share arose on various occasions until the creation of the Broadcasters' Audience Research Board (BARB), jointly underwritten by the BBC and ITV, in 1981, although as Ehrenberg and Twyman noted in 1967:

Stories of gross disagreement between the BBC's results and Tamratings would appear to have been due largely to the frequently ignored difference in technical definitions (the two systems set out to measure slightly different things and express the results in rather different ways) together with the incidence of sampling errors which in regional comparisons can be particularly large. (Ehrenberg and Twyman 1967: 15)

The audience-ratings convention around sampling and panels is well established, including the formats for reporting audience results. There is basic agreement that probability sampling works. All of the ratings figures can now be produced quickly in standardized formats with a range of permutations possible. You might think that this is the end of the ratings intellectual and that all is well with sampling and surveys in the audience-ratings convention. But this is far from the case. The modern ratings intellectuals, like Gale Metzger and Peter Miller with his colleagues at Northwestern University, while not in the milieu of Lazarsfeld, Zeisel, Crossley, Gallup and others, still have a major role of critique of both the industry and the methodology. There are additional biases that might affect sampling and panels. This includes the problems of multiplication of channels and programmes as well as non-cooperation of participants.

Single Source: ‘The Holy Grail’

Audience ratings panels and survey research generally have always had problems with non-response or non-participation in research, even in the days of Archibald Crossley and Paul Lazarsfeld. However, there has emerged an erosion of the ability to get people to participate in surveys or to give more data about themselves at the very time when there is today an explosion in demand for more measurement. Project Apollo is the paradigm example of an attempt to capture everything people do in one panel. When Nielsen and Arbitron joined forces to set up the experimental Project Apollo in 2005 they expected to capture all of the everyday behaviour of audiences, from reading papers through to using mobile phones and buying food. To their surprise they found that people did not want to participate. The more they were asked to do and to provide, the more they resisted and refused. The current situation involves many separate panels for separate activities, for example, a panel for outdoor advertising, a panel for mobile phones and so on, as we found in Chapter 1. There is also access to ‘buyer graphics’ such as actual purchases of, for instance, books.

In television ratings the longitudinal aspect has been fundamental to the operation of the business of ratings as it has provided the users of ratings with the capacity to map audience flow and to get a sense of what is happening to audiences over time. The process of selection of panels and their retention over time is therefore key to the quality of data that are retrieved. Television ratings panels may obtain an appropriate sampling frame but that is not the end of it. If not all the ‘basics’ – those on the list – do not want to participate, as Nielsen found in the 1990s, then the alternates, those who say yes, may not be representative of the panel originally chosen. In the case of Nielsen, the ‘alternates’ ended up having fewer televisions and were heavier television viewers than the ‘basics’. There end up being key trade-offs if an appropriate statistically representative group cannot be found.

Table 3.3 from Peter Miller's work provides a summary of those trade-offs. The ultimate in survey research is gaining a Data Rich and Case Rich outcome. Data Rich means that the information coming back from the audience is extensive and has qualitative depth. For example, information on Facebook is Data Rich – there is significant information posted by people on to their sites that can be used to understand audiences and what they think or do. However, it is very difficult to establish from Facebook an appropriate sampling frame or probability sample that allows researchers to generalize their findings. At the time of writing, Facebook has 500 million users. These 'users' do not form into well-defined and easily accessible lists. There are, as a result, problems in establishing or verifying identities of users and creating an acceptable list where a random draw or stratification can be made. Facebook, therefore, is Data Rich but Case Poor as an audience and as a source of information. The traditional panel for television ratings, by contrast, is Case Rich but Data Poor.

The Data Poor nature of the television ratings measured by exposure – who watches, where, for how long – was made up by the fact that users of the ratings found them practical and acceptable as a form of currency (a trade off). Miller would call this part of the 'coordination rule' in audience ratings. There was in the past confidence that the television ratings panel indeed reflected the population from which it was chosen. Through peplemeters the television ratings panels have provided information about an audience's *exposure* to television. If a company or a researcher wants more information about whether the audience actually like programmes then customized Data Rich studies are often required. Modern syndicated ratings research therefore has often been complemented by customized studies seeking more detailed information about audiences.

If there is significant non-cooperation in television panels, however, it is only so far that a trade-off can be accepted. Miller conducted a major independent study, the first of its kind on non-response bias, for Nielsen in 2009, and its results give an insight into the role of the modern audience in survey research and the limits the audience itself puts on participation. The study consisted of 2,300 basics,

Table 3.3 Trade-offs in audience survey research

	Case Rich	Case Poor
Data Rich	Single source (the Holy Grail)	Custom
Data Poor	Syndicated	Quick and dirty

Source: Adapted from Miller 2009

with 1,000 responding households (a 95 per cent return rate) and 1,300 refusing households (a 62 per cent return rate). Special in-person follow-ups were made with those households that did not respond. Few studies have followed up with non-responding households on why they have not participated or given data, and this is what makes this particular study important. Table 3.4 shows the difference between responding and refusing households by day-part. The p values are significant for the 8–11 pm day part, with the possibility of differences in the 5–9 am day-part.

Table 3.5 in comparison shows refusing households by channel, with significant differences for Cable News Network (CNN) and Home Box Office (HBO), with MTV (originally Music Television) and Fox approaching significance.

Table 3.4 Day-part comparisons

Day-part	Per cent Response	Per cent Refuse	P value
5–9 am	41.9	45.8	.11
9 am–4 pm	56.1	56.3	.94
4–8 pm	67.4	69.7	.46
8–11 pm	71.2	76.2	.04
11 pm–2 am	39.3	39.7	.89
2–5 am	10.2	12.3	.18
12 am–12 am	91.5	90.1	.41

Source: Miller 2009

Table 3.5 Channel comparisons

Channel	Per cent Response	Per cent Refuse	P value
ABC	70.3	73.7	.19
CBS	68.7	73.2	.12
Fox	68.4	72.7	.11
NBC	71.2	73.3	.38
Univision	7.6	6.7	.52
BET	11.3	12.5	.52
CNN	37.7	45.2	.00
HBO	17.4	27.9	.00
MTV	15.5	19.1	.09

Source: Miller 2009

Table 3.6 Equipment comparisons

Device	Per cent Response	Per cent Refuse	P value
Big Screen	22.6	41.2	.00
Cable	61.9	69.1	.00
Satellite	24.7	28.5	.18
DVR	9.7	20.1	.00
Hi Speed Web	72.9	75.3	.30
Web TV	63.6	69.5	.15
3+ TVs	46.1	60.2	.00

Source: Miller 2009

However, it is the figures in Table 3.6 that start to show why many people do not want to participate in modern audience panels and audience ratings survey panels especially. Nielsen found that technology-rich households did not want to participate in research – the very demographic that modern marketers want to grab. There are various reasons for this, ranging from an unwillingness of households to allow increased intrusion into their lives across a range of technologies and, of course, the simple fact that the technologies themselves need to be intruded upon in order to gain data (for example, additional wiring) and participants do not want their expensive technology tampered with in any way.

In summary, Case Rich studies are those that use statistical sampling methods to derive their samples from sampling frames. Case Poor studies are those that have not deployed statistical sampling methods and therefore their samples cannot be used to represent the populations from which they might have been drawn. Data Rich studies are those that yield detailed data or accounts from participants. In the case of television, for example, an interviewer living with a television household for a month is getting fine-grained qualitative data. Data Poor studies are those that gain minimal information back for their purpose. In the case of television ratings, exposure gives the media industry basic information about who is watching, when and for how long. Those studies that do not gain an appropriate probability, statistical, sample, and do not gain particularly rich and detailed information, are both Data Poor and Case Poor. The modern problem for syndicated research is that participants do not want to participate, especially those who are of most interest to media researchers. If panels are failing to adequately represent the modern audience, then they are Case Poor. The data, as a result, are not generalizable to the defined population.

Gale Metzger also sees the pursuit of single source as impossible for very human reasons that do not only include non-participation – the limits on human patience:

Before the peplemeter, the ‘people’ part of the measurement was based on a diary. Nielsen had a meter-diary combination measurement. The set tuning came from the meter and the persons-per-set from a diary. And you combined those two to obtain a projected audience. The peplemeter brought the people measurement into the meter itself and people had to push buttons to register themselves as viewers. We did a lot of experimentation with passive measurements, voice-based systems or even passive recognition systems. The proliferation of outlets and fragmentation of the TV medium has several measurement implications. It is easier to measure a big thing than a small thing. It is easier to measure a 30 per cent market share than a 3 per cent market share. A relatively small and crude sample can provide a reasonable estimate of a dominant phenomenon. But when you have a single digit share, larger and better samples are required for reasonable precision. Network ratings have gone from the 20s and 30s to twos and threes. Today, there are many stations and outlets that are unmeasurable.

Today, samples of tens of thousands of respondents are needed rather than one thousand – the US national sample size for decades. For our RADAR [Radio’s All Dimension Audience Research] service, we went from a sample of 4,000 in 1969 to – I think we ended up in 2001 – a sample size of 16,000. We progressed in stages from 4,000 to 8,000 and then to 12,000. The price for the service related to sample size. We sold the increases to the industry by demonstrating the loss in precision of audience estimates due to audience fragmentation.

RADAR technique: the RADAR interview involved a one-week period where we would contact a person ideally eight times. Each time a respondent was asked about his/her activity from the time of the last contact up until this contact ...

The easiest way to get out of the interview is to say, ‘I didn’t do anything. I’m done.’ That makes for a very quick interview. If people learn that, the result is under-reporting. The objective is to cultivate complete and accurate reporting. That is a core reason why you don’t want to throw a lot of other peripheral stuff on to a media measurement. By doing so, it creates fatigue factor; you cause people to be discouraged to ever say they do anything. They come to know that if they say ‘yes’, they get more questions. Once they learn that, you’re in trouble as a researcher. So you want to keep the burdens light. By doing so, a better measurement is achieved. If a respondent

is subjected to a two-hour interview and then you come back the next day and want to talk again, a likely response is, 'Are you nuts? I just gave you two hours yesterday.' If you can complete an interview in three or four or five minutes, it's a different matter. You can give them some small incentive and they'll stay with you.

A truism in research is that the more you ask, the less you get on any one subject. By that I mean that a properly defined survey dedicated solely to the measurement of television yields one result. In contrast, if an attempt is made to measure four or five media and product information, a natural result is to find less usage of each component ... In general, more questions per subject will yield more activity; fewer will yield less. That is one element.

The other element is the so-called fatigue factor.

A respondent may begin a survey by wishing to be helpful. As more and more is asked, the mind set changes from 'how can I be helpful' to 'how can I get out of this' ... it makes life easier to say 'no' than 'yes' – in order to get to the end of the process. Multimedia measurements are not as accurate as single subject surveys – assuming all else is equal. When recruiting someone to be part of a survey, you necessarily have to tell something about what is coming. If you tell them you need to move in for a few days to get all your questions asked versus proposing a 15-minute survey, different responses result. So there are many things that mitigate against multi-faceted studies.

We have learned that the so-called single-source data where you try to get it all will not hold up over time. The quality of information is not up to what it should be, or not up to standards. So to this day, notwithstanding the fact that in this country we've had 30, 40 maybe more measured efforts at doing multimedia product media surveys, none have succeeded. One of the more recent ones was Project Apollo by Arbitron – which one of my friends called 'Project Appalling'. (Metzger 2000)

Metzger is highlighting what the authors call underlying survey wars. Everyone in the industry knows that non-participation in surveys has become a major issue and that attempts at single-source data are fraught with methodological difficulties (simply fusing the data from a sample from one panel with a sample in another panel is not going to solve the problem). Creating more measurements of exposure does not create single source data. Simple categories that showed a critical mass of an audience are now competing with thousands of categories attempting to demonstrate a market. This puts the original agreements on exposure as the standard under pressure, as it does the Mint Master, that is, the organization that produces the audience ratings.

For ratings intellectuals like Leo Bogart the single source pursuit and the role of computers have displaced the role of the ratings intellectual; computation has come to replace good judgement.

The universal use of the computer in business prompted a desire to crowd the maximum amount of information onto a single spreadsheet, and thus to squeeze an ever-greater variety of acts about product and media consumption out of a given database. Advertisers have become more and more obsessed with the idea of the 'single source' – the dream of getting all the information on what individuals buy and on all their media experiences from one giant research service, so that cause and effect can be clearly measured. Presumably such a service would end the deplorable present necessity of putting together the data derived from different research organizations, with all the inevitable inconsistencies that result. The underlying premise is that the quality of information from a single source could be equivalent to what can be obtained by specialized studies that focus on just one subject at a time. The premise is incorrect, but good judgement, experience, knowledge, and common sense are commonly overruled by the desire for convenience. (Bogart 2000a: 132)

Summary

Strange are the uses of radio statistics. Last week in *Variety* two double-page advertisements appeared. One tried to prove by Cooperative Analysis of Broadcasting figures that CBS [Columbia Broadcasting System] was the leading network of the land. The other, also using C.A.B. figures, advanced the proposition that NBC's Red Network led all the rest. Not responsible for interpretations of its figures is C.A.B. Having used them against each other, the networks last week united in using C.A.B. statistics against their common foe ASCAP, the cooperative which controls most U.S. music. The networks claimed that according to C.A.B. figures radio listening had not decreased since their contract with ASCAP acrimoniously expired Dec. 31. Promptly they were jolted with the news that in its regular report C.A.B. had checked the popularity of 91 evening programs, discovered that since the New Year 52 were down, 35 up, four unchanged. Anon. (*Time* 1941)

The 1941 *Time* story tells us a lot about how broadcast ratings can be used as a tool to influence the media industry and, indeed, public perceptions about what is happening with networks and programming alike. CAB was the Cooperative

Analysis of Broadcasting, an advertiser-led consortium that produced the first radio ratings. At first glance, the use of CAB ratings looks like manipulation of industries and audiences. In this case, the ratings are used to argue that one network is better or more popular than another or that network popularity has not suffered because of unfortunate contract events in their programming. This is perhaps the most common type of example that people are familiar with in arguments about ratings. But ratings have always been more than that. They have a close resemblance with public polling and voting. This is not surprising. As we have seen the ratings intellectuals who designed the ratings also created the first opinion polls. Audience ratings as a particular form of humanistic social research also adopted the revolutionary panel-based survey research technique using probabilistic representative samples.

The ratings intellectuals have been and remain essential to the audience-ratings convention. They originally had a bridging role, working across private and university sectors. They were creators of methodology as well as users of the methodology, fierce critics and often businesspeople with vested interests in the very markets they studied. Most importantly, however, they anticipated the world of multiplication of publics and a world where large populations needed to express their preferences or voting intentions. Modern discussions on the multiplication of publics, like those happening now on monitory democracy (Keane 2008) and the role of the internet in gift economies and participation, have their precursor in the practical work of the ratings intellectual. Public polling and audience ratings were not created simply to be a capitalist tool of domination. Quite the opposite. The early ratings or survey intellectuals, especially those like Hans Zeisel and Paul Lazarsfeld, who followed a strong socialist tradition, worked purposefully in the media field to develop methodologies that enhanced, from their point of view, democracy. Likewise Gallup and Crossley did not see their activities as purely commercial but intimately tied to developing methodologies to enhance the 'public vote'. Samples and panels were methodologies developed to ensure representation of the public through small samples. This dual identity of the ratings, rarely acknowledged, is undeniable.

The modern era in audience ratings, however, is marked by underlying survey wars. Traditional audience research has relied on statistical sampling that provided Case Rich foundations for syndicated ratings. Exposure has been the key measure, even though it is Data Poor compared with sites like Facebook. However, the traditional panel is under threat because of problems of non-response bias, among other things. This situation is unsustainable because in the long term the media industry will end up with Case Poor and Data Poor delivery of ratings. The alternatives, like drawing information from sources like Facebook, are also not sustainable. These alternatives would most likely be Case Poor and Data Rich. The idea of having non-representative population data for ratings would be an unacceptable trade-off from

the point of view of currency. It is one thing to make the trade-off with exposure as a measure and keeping the study Case Rich and another to have data that does not represent statistically defined populations.

In the meantime, the demand for more measurement is increasing. Collecting data from proprietary set-top boxes, for instance, is growing, but that faces all the difficulties of privacy and intrusion that are emerging, apart from the fact that each proprietary system will have its own measurements. Google has combined with Nielsen to try to capture the tail of small audiences in television and to auction off those audiences. What all these attempts at new measurement are encountering is what the authors have identified in this chapter – the audience. The audience is an integral part of the audience-ratings convention and its assent is essential to the overall working of ratings industry.