Brian Allt
Mirror Group Newspapers
London, UK

3.1 INTERVIEWING AND READERSHIP - SOME THINKING BEHIND THE EXTENDED MEDIA LIST DESIGN

SOME GENERAL OBSERVATIONS ON PROBLEMS OF QUESTIONING

In Britain each year we send out interviewers to have about 28,000 conversations with strangers. The topic is reading. Do we sufficiently face up to the basic problems inherent in trying to base measurements of behaviour upon conversation, and only semi-interactive conversation at that?

In the last two Symposia much valuable work has been reported on the effects of different questioning techniques upon readership estimates. There are also many studies of other types of survey questions and how easy it is to get different answers to what is conceptually the same question.

But in this literature, and particularly in readership studies, there seems to be relatively little discussion of why the differences occur. Of course, when people decide to test a new technique they must have some beliefs about *why* it would be better, but often these are not made explicit. Indeed, it is sometimes hard for the outsider to see why a given change was expected to be advantageous, or why it was tested.

In survey research we are not normally in a position to verify our observations. Without verification, all we can say about an answer is that it is a form of behaviour which is a response to a question. Even if behavioural corroboration is theoretically possible, the only true *objectification* of the behaviour a question measures lies in the wording of the question (and the interview situation).

We can state a simplified antinomy:

The answer a person gives to a question is true.
The answer a person gives to a question is false.

When we look at the responses to differently structured but conceptually identical questions we find that the pattern of answers frequently varies, eg according to the wording of the question (Kalton & Schuman, 1982), the number of answer categories (Tennstädt, 1983), and so on.

This means that the answers arising from different wordings or formats of the question contain different balances of true and false answers.

We attempt research that will reduce or remove the false element. Do we do this within the right frame of reference?

We tend to work within a purposive and directed frame of reference of which the key point is the behaviour to be measured.

I would suggest we might proceed better if our key point of reference was as much or more the questions, rather than the behaviour.

What do we do when we ask a question? We produce a stimulus and insist on a response. In the majority of survey questions put, we offer directly to the informant (or indirectly, to the interviewer) a range of possible responses.

If we asked a sample of informants to select one of four abstract symbols every time we pressed a buzzer or flashed a light, we would expect a random distribution of answers, providing each symbol was similarly devoid of meaning and their positions were completely rotated. On the other hand, if the offered responses were 'yes' or 'no', or the numbers one to six, we would not expect so random a distribution of answers.

I am suggesting that when we consider questions and the response categories offered we should start by thinking of

the act of responding as fundamentally random. The degree of randomness will be lessened according to the strength of any or all of the following factors:

(a) Understanding of the question and task
(b) Interest in answering questions
(c) Interest in the topic itself
(d) Relationship to the topic area
(e) Behaviour in the topic area
(f) Cumulative clarification stemming from previous questions.

In relation to any particular question, or a survey as a whole, our sample will evince a distribution of tendencies to randomness. (A special sub-category of randomness is that kind of uncertainty which can be described as the informant's 'indifference range', eg the difficulty in deciding whether one reads seven or eight issues out of 12).

The basic tendency to randomness will be further varied according to the way in which any particular concept is structured into a question. Thus I believe that an unmeasurable part of the differences in proportions qualifying as readers according to the proportion of answer categories which qualify, as demonstrated by Tennstädt, is due to the influence of the question structure upon the basic tendency to randomness - if two out of six categories 'qualify' more people will select them than will select when only one out of four qualifies. I am not suggesting that all, or the majority of responses are random, but that some probably are.

It is difficult to discuss these problems without almost *'moral'* criteria entering the debate. Some people will argue that giving people a lot of categories to select from should increase their change of responding precisely - others will argue that it simply makes the task more confusing and that the ideal question is dichotomous.

I realise that all I have done is to assert a hypothesis about question and answer problems, and also that it is difficult to conceive experimental tests of the tendency to randomness which could be sufficiently close to a normal interview situation to justify extrapolation. Nevertheless, insofar as our attempts to improve questions depend upon our basic beliefs and thought process models I believe that the hypothesis advanced represents a more positive and realistic basis than the approach which simply says 'we have an unsatisfactory question, let us try another version and then think about why the answers are different'.

The latter approach leads us to talk about variation in results as due to external effects such as desire to finish the interview, to avoid further questioning, to establish higher status, and such things as fatigue. Instead we should apply economy of hypothesis more and first consider all such effects as being partly or totally induced by the questions and interviews we administer to the public.

It should not be assumed from the foregoing that this randomness is purely of a 'do not care' type. Depending upon the nature of the investigation, and the informant's involvement with it (as listed earlier) the informant may well attempt to reduce that randomness *of which he is aware*. He may convince himself that the 'last occasion' was over a week ago rather than in the last week when he is not really sure. He will settle for eight out of 12 when the range in his mind is perhaps 'more than half but less than all'. Such tendencies reduce the conflict inherent in the puzzle-solving task and the actual interpersonal situation.

Some of these attempts by the informant will be beneficial in terms of accuracy, others the opposite.

QUESTIONING IN READERSHIP SURVEYS

A further problem with much experimental research on questioning is

that the topic of investigation appears largely to be seen as individual *questions*, often repeated ones, instead of the interview as a whole.

Let us consider the interview situation. A stranger comes to your door. With a very brief explanatory introduction she starts thrusting prompt-aids under your nose and asking questions of an extremely tenuous and ill-defined nature. Many of the publications you will have never heard of. She, on the other hand, is obviously familiar with all of them. She rattles off such concepts as reading or looking at, reading frequency, and thus imposes a frame of reference within which you are supposed to locate yourself while still trying to understand what it is all about. She is operationally in a position of authority through her knowledge. She presents you with a puzzle and asks you to play a card game with which you are unfamiliar. You do not get a score to tell you whether you are playing well, or even praise for trying, so you are not helped to do 'better'. The actual act of reading papers and magazines is pleasurable and interesting. How much interest is there in answering questions about what you read? You are not being asked how you *feel* about the various titles or why you do not read the ones you do not read.

So we have a situation where the informant is at a social disadvantage while trying to solve a puzzle of low interest. In psychological terms it can be a situation of threat.

Why do people complete the interview? We do not pay them (yet!). We depend for interview completion upon the extent to which the interviewer can make it a pleasant social contact, upon people's innate friendly responses to strangers, often their loneliness, their curiosity, boredom, their feeling that it might lead to something advantageous and, in some cases to a feeling that they have to co-operate with a perceived authority.

When we analyse our data, however, we tend to treat them as if they were a set of uniformly conducted height and weight measurements.

What they are, are the responses to variously understood stimuli where the ostensibly unchanging stimulus (the question can progressively change as the interview proceeds and your understanding of what it means changes.

When you are asked about the first title or set of titles it is in the context of the question. Your response will be the resultant of your relationship with that title and your interpretation of the question. If the relationship is weak the interpretation of the question will be likely to be different from that if the relationship is strong. The fact that you are being even *asked* about titles which mean little to you will structure your view of the level of relationship which is of interest to the interviewer. Some informants will resolve the tension puzzle by saying No! No! No! Others will stretch the range of meaning implied by the question to include marginal or ancient exposure and resolve the tension by saying 'Yes' to some titles. My view is that the latter is the more likely to occur and that that is a major cause of order effects in the old British NRS. There one of the most pronounced results was higher level of claiming for monthly magazines (which are largely minority publications) when 50 or so such titles appeared at the beginning of the interview, as compared with appearing at the end - a ratio of 118 to 83 (100 as overall average) in Ted Whitley's paper, at New Orleans (1981).

Putting it another way, when we started the interview with a lot of titles which most people did not read we were saying in effect 'we think you have probably had *some* contact with these publications even though you are not familiar with most of them'. By the same token, in interviewing you at all we are suggesting that you do read

something. When you are asked your views in an opinion poll it immediately suggests that you *have* a view or damn well ought to!

At the other extreme, most people have contacts with daily newspapers and a fairly close and regular contact. Thus when similar questions are posed about newspapers the balance of the relationship component in the situation to the *question component* is very different. The actual behavioural part of the relationship component will also be more recent. The relationship component is much more clearly structured and unambiguous and hence the question is more easily answered. The pressure to classify marginal relationships into significant ones will thus be lower because the uncertainty is lower and more significant relationships already exist.

In considering the balance between the question element and the behaviour/ relationship element in an interview we may consider them as weights on a physics forces-board.

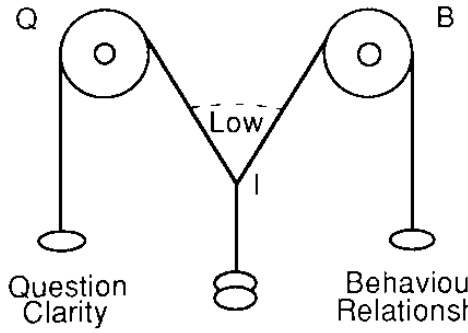
The greater the clarity of the question and/or the greater the clarity of the behaviour, relationship, or opinion in the informant's mind, the greater the accuracy of understanding and consequent response in the interview situation as represented by the angle QÎB. (See chart)

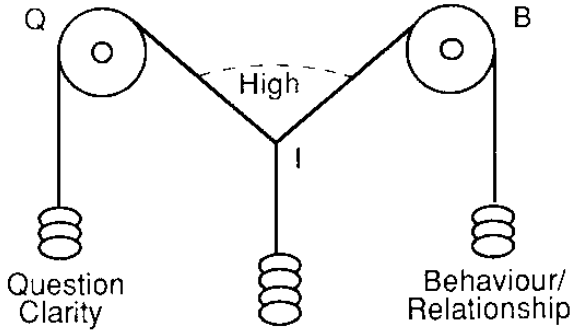
Returning to order effects, what seems likely is that a *number* of position reactions operate. Firstly, the publication group(s) asked about first or early will structure the way the question is interpreted for that and subsequent groups. Secondly, publication groups read by large numbers may have more influence on the interpretation of the question for groups which appear later than will publication groups read by smaller numbers of people.

Thus when magazines were asked about after newspapers an operationally tighter definition of readership may be in informants' minds than when magazines were asked about first.

When we show a title on a specialised topic or a title which is similar to other titles we are saying 'here is one of the so-and-so family of titles, you probably read this one don't you (ie rather than the other(s)?' We tend to call the problem 'title confusion' but the results are probably due only partly to actual *confusion* and partly due to the title which comes earlier being thus given temporal prominence. The element of extra claims due to *genuine* confusion probably cause further confounding as it is likely that people claim the other title(s) as well when it eventually appears.

Readership questions asked in the context of omnibus surveys or marketing surveys usually produce lower levels of claim than those produced by the National Readership Survey. While the interviewers may not be so thoroughly briefed and perhaps more concerned with the rest of the interview I would hypothesise that because the bulk of omnibus questions are about fairly clear-cut behaviour and relationships





the frame of reference for the readership questions is more likely to rule out more marginal contacts with publications - hence lower levels.

Another example of progressive modification of the interview frame of reference emerges from analysis of the pre-1984 frequency questions on the NRS. In 1975 I hypothesised that some informants might carry over the filter time-period from one publication group to the next (Allt, 1975). For example, in answering how many copies of a quarterly were read in the last year, they might answer in terms of number in one month if weeklies had immediately preceded quarterlies. Extensive unpublished analysis of NRS rotations by Tom Corlett produced substantial evidence of what he called the Allt effect. The table below gives an example for daily newspapers.

Overall both the later appearances result in gross levels lower than when dailies are first. This is evidence either of what I have called previous structuring, or fatigue effect, or both. But when we compare dailies following monthlies, where the reference frame was six months, we get a substantially *higher* level than for dailies following weeklies where the preceding frame is four weeks. This difference cannot be due to fatigue or boredom because the average number of titles asked about before was virtually the same.

The effect is even more marked when we look at the first daily title to be asked about - the *Daily Express* in the forward order and *The Sun* in the reverse order. Here the carry-over effect from the previous group is so marked when monthlies precede dailies that the average of the two papers is the *same* as when they open the total interview.

If we accept the possibility of progressive modification of the understanding of the interview as it proceeds then it follows that rotation of questions can by no means be expected to equitably balance out what we normally refer to as order effects - the cumulative effect of A-K upon L and M will be different from the cumulative effect of Z-N upon M and L.

In the old NRS design the sequence was a frequency-filter question asked for all 120 titles, only then followed by the 'when last read' question for all titles which passed the filter. Thus there were two sequences, the same in any one interview, ie 1-120 for all titles in frequency terms, and 1-120 for a dozen or so titles as to when last read. We thus had a potential

Carry-over effect (January-December 1975 JICNARS NRS)

National dailies - Gross levels of claiming any readership in average week and average first dailies

	Gross dailies %	Average first dailies %	Average number of titles preceding	Preceding time period
Dailies first	192.5	40.2	0	-
Dailies after Sundays/Weeklies	176.4	37.1	72	Average four weeks
Dailies after monthlies	185.3	39.7	68	Last six months

(Eight group rotations were in use in 1975)

progressive modification firstly in relation to understanding of the frequency question and once again in relation to the 'when last read' question. While the frequency question was structured by the answer categories being on the prompt card, the when last read question was not prompted. However, where the informant was unclear there must have been a number of occasions where the interviewer would have to ask 'was that within the last month?' or 'within the last week?' thus potentially structuring the frame of reference for subsequent titles.

Asking people one filter question about a lot of things to decide which merit further questioning is common in survey research and rests on an assumption that if people realise early in the interview that a positive response means more work then they will resist by saying 'no' to subsequent filter questions. If this effect really exists I would first question how much it may be an informant effect and how much an interview or interviewer effect. The hypothesis that lower levels for titles near the end of the interview are due to a 'conditioning to negative' cannot be proved nor disproved ultimately. In fact with the old design the *informant* did not know that a positive response would lead to another question.

My belief is that we were progressively re-structuring the interview in such a way as to confuse the informant and make him or her change their frame of reference in an inappropriate way as the interview proceeded.

Also by presenting titles one by one we were depriving the informant of the relevant reference frame for solving the puzzle of which titles had been read. Thus the survey suffered from two confusion components - genuine confusion as to the names of titles read, and a survey-induced confusion resulting from separate sequential presentation of the titles.

Summing up, one has to acknowledge that an informant can never fully know what an interview is about until it is finished. The rate at which people learn what it is about will vary from person to person. What we should aim to do is reduce the person-to-person variance in learning rates and also the discrepancy between early and late states of understanding.

These differences result in a variety of interpretations of any particular question. *Any* survey question is subject to a variety of interpretations. Our tabulated data represent the results of a distribution of interpretations. A classically ambiguous question represents a bimodal distribution of interpretations. We also thus have to try to reduce the variance in the distribution of interpretations.

The more that we can achieve interview learning early in the interview, reduce the discrepancy between early and late learning, and reduce the between-informant variance in interpretation, the more our data will be internally stable.

THE EXTENDED MEDIA LIST DESIGN

All the foregoing therefore, were among the thoughts which lay behind my suggestion in 1981 that we should adopt a grouped titles approach, and in the minds of those others concerned with its eventual development and experimental testing in 1983[3]. It has been operational since January 1984.

There were, of course, other significant considerations. We were under a great deal of pressure to include many more titles in the survey. We believed that it was undesirable to go one at a time through title after title in which most informants had little interest - skin magazines, motoring magazines, teenage magazines, because it was likely to produce boredom and irritation and lessen involvement. This applied to the old list of 120 titles, but would apply even more had

we attempted 200 or so because most of the additional titles would have been specialised and only appeal to small minorities.

We felt that one of the apparent anomalies of the traditional design, namely higher than theoretical probability levels for the lower frequency categories, was due to new readers answering reasonably correctly and some lapsed readers not responding positively at all because they simply forgot about distant reading. Accordingly, we opted for a frequency scale structured in current terms rather than historic terms. We also wished for direct claims as to reading over long periods - up to a year - to set alongside the coverage estimates based on the frequency questions.

We felt that all interviews should begin with groups of titles which most people read, therefore we start half the interviews with daily newspapers and half with Sunday newspapers.

We wished to have a measure of reading on the average day for all titles, and the arguments of Pym Cornish and Michael Brown (1980) convinced us that the best way to achieve this was by direct questioning about yesterday rather than 'when last read'.

EML IN PRACTICE

Stage 1

The interview starts by putting all our cards on the table in a period of two or three minutes. That is to say the informant is handed a pack of 42 cards, each bearing several titles, and asked to sort the pack. Three piles are asked for - those where *any* title on the card was read or looked at in the last year; those where *no* titles were read in the last year; and a 'Not sure' pile. The informant is then asked to explain the problem of the 'Not sures' and these are sorted into the 'Yes' or 'No' piles. Finally, at this first stage the informant is asked to double-check all the 'No' cards again, possibly adding to the 'Yes' pile.

Thus in a very short space of time the informant is made aware of the total universe of publications in which we are interested.

Stage 2

Then the informant is told that we wish to know 'how often you read the papers and magazines you remember seeing in the past year and when you last read or looked at them'.

The first 'Yes' card is handed to the informant but showing the reverse side. This shows the titles again, the frequency scale and the frequency of publication - daily, weekly, etc. (some cards have mixed publishing frequencies. The frequency question is the same for all titles and contains only three positions - 'Almost always' (at least three issues out of four); 'Quite often' (at least one issue out of four); and 'Only occasionally' (less than one issue out of four).

Thus, at last in common with many other countries, *we have the same frequency question for all publication intervals, reducing one element of learning confusion.*

The informant is asked to say from the card *firstly* all titles read 'almost always', *secondly* all titles read 'quite often', and *thirdly* all titles read 'only occasionally'. Titles not thus far mentioned are probed individually and classified on the scale or as 'not read in last year'.

Then, still with the same card the informant is asked to say which, if any, of the titles was read or looked at yesterday and then which in the last week. This is followed by a question asking, for each title on the card not thus far mentioned, 'When did you last read or look at a copy of ...?' and the answers are coded as 'past four weeks', 'past three months', or 'longer ago'.

Thus all the crucial questions are asked one card at a time and we do not have to go over the same titles again at a later stage. The card or 'family' is finished with and the next card taken. The frequency and recency questions are thus not asked about at different stages of the interview and confusion is minimised because as far as possible similarly named and similar topic titles are simultaneously visible on the same card.

Informants can see the pile of 'Yes' cards and thus can quickly grasp the scale of the task and realise how far they are through it, and I believe this to be valuable as a form of feedback and indication that *the task length* is related to their own personal wide or narrow range of reading interests.

Other papers at this Symposium go into the findings in more detail. Broadly speaking, compared with the previous NRS method we have greatly reduced order effects and title confusion. The levels of readership overall for titles covered by both methods are broadly similar so we have not produced lower levels of claiming via the new design even though a lot more titles are being covered. In my opinion this casts grave doubt on some of the traditional beliefs about questionnaire design and the postulated negative approach of informants to interviews as a task. *In particular I believe we have shown that 'conditioning to negative' and 'fatigue effects'* are artefacts and design effects, and not automatic human reactions.

I believe we have data which are of better quality, more acceptable and credible because more internally stable, and that this is due to constructing a more understandable and less confusing interview situation. There are still some problems with the technique but on the whole it is a great improvement.

However, I would not claim that the data are ultimately *true* any more than those form any other survey, because the only objectification of what surveys measure lies in the wording of the questions we ask. This can be extended as the only objectifications of the multi-dimensional behavioural complex we seek to measure lie in the total interview-complex, ie questions, prompt-aids and social situation.

REFERENCES

Allt, Brian (1975) 'NRS reading measures and the *Drive* study'. JICNARS

Allt, Brian (1983) 'The extended media list experiment'. JICNARS

Cornish, Pym & Brown, Michael (1980) 'Readership measurement reviewed'. JICNARS

Kalton, G & Schuman, H (1982) 'The effect of the survey question on survey responses: a review'. *Journal of the Royal Statistical Society* 145, Part 1

Tennstädt, F W R (1983) 'Effects of differing methods on the level of magazine readership figures' (Montreal Proceedings)

Whitley, Edward (1981) 'Some rotation effects in the British survey' (New Orleans Proceedings).