EXPLAINING THE SCREEN-IN PARADOX

Peter Walsh, Kenning Asia-Pacific Pty Ltd

Abstract

With both the Recent Reading and Through-The-Book methods, average issue readership (AIR) appears to be influenced to a surprising extent by the screening question. If a modification to the method produces a change in the screen-in level, then a corresponding change in AIR is likely to be found even among more frequent readers - ie., respondents who should pass the screening anyway. It seems a paradox, but can be explained in terms of the ways respondents answer questions when uncertain. The processes shape responses to a wide range of other survey questions (eg., opinion polls). It is argued that researchers need to take a probabilistic view of response at the individual level, rather than a literal one. Implications for the design and interpretation of survey research are significant.

Introduction

"For MRI (Recent Reading), the study found a nearly one-to-one relationship. Relative year-to-year changes in screen-in rates are reflected in readership on a very nearly pro-rata basis." Mallett [1993]

"Except for weekly publications, further analysis ..... fails to support the theory (that, with the TTB method, readership levels should be unaffected by the screen-in level)..... *Also* (even for weekly publications) *where there was a compensatory change in read/screen ratio following a change in the screen-in rate, the change came principally from those claiming to be more frequent readers rather than from the less frequent readers as theory would suggest."* Appel [1993]

As current theory has it, the functions of the screening question are to capture all respondents who might qualify as readers, so reducing the number of titles for which the core readership questions need to be asked and providing a harmless outlet for prestige-influenced false claims.

Due to sampling error, interviewer effects and promotion, screen-in levels can change from one survey period to another, contrary to what is actually happening to cumulative readership (which theoretically is what is measured by the screen). But such effects cloud the issue here, so we will dispense with them.

A well-designed split-sample experiment will control measurement effects by ensuring among other things that the workload for respondent and interviewer is the same in both halves of the sample. It will yield a clean independent variable: either the number of titles screened, or the screening period. Let's consider the latter. When the screening period is manipulated in such an experiment - ie., results are compared for (say) 6 versus 12 month screening horizons - what will be found?

It will be found that the dependent variable - AIR - moves with the screen-in level. This is so with both the Recent Reading and Through-The-Book (TTB) methods, especially the former. But theoretically, any change in the screen-in under such conditions should be only among respondents we will refer to as 'peripherals'. With the Recent Reading method, peripherals are respondents whose last reading occasion was between the screening horizons (ie., 6 to 12 months ago). None of the change should come from those whose most recent occasion was within the publication period (eg., the past month).

In reality however, the change will be found among recent readers just as much as others. As Mallett showed, AIR based on Recent Reading moves up and down virtually pro-rata with the screen-in level. Extending the horizon (eg., from 6 months to 12) will tend to increase the screen-in level, and correspondingly AIR will go up. On the other hand, shortening the period or increasing the number of titles will tend to reduce the screen-in level for the average title, and AIR will go down.

With the TTB method, peripherals are respondents with a very low probability of having read the test issue. Their passing or failing the screening question should have minimal effect upon AIR. But, as Appel found, the 1980 ARF Comparability Study's controlled experiment with TTB screen-in generally did not reflect this expectation. AIR was affected even among regular readers. In fact, *most* of the absolute change in AIR came from such readers.

Interestingly, the 1983 Extended Media List (EML) experiment in the UK showed that the number of titles covered by a readership survey can be increased greatly without depressing AIR. But in conjunction with the increased number of titles, the EML moved to grouped title cards which in effect forced more titles past the screen. That is, if *any* title on a card passed the screen, then further questions were asked for *all* the titles on that card. As a result, AIR went up for most titles rather than down.

Can regular or recent readers so easily be influenced by the screening question? Or is it instead perhaps that some peripherals who pass a more inclusive screen then falsely claim frequent and recent reading? Either way, it seems a paradox. Or is there a logical explanation?

Attitudes, Opinions, and Information

Parallels in other applications of survey research point to a theoretical issue which is central to the field in general. And so we will digress for a while.

Implicitly, survey research generally rests on the premise that if respondents can answer a question at all, then the information which the question seeks to obtain must have been present in and accessible by the conscious mind; it is further assumed that the information is faithfully reported. Only a "don't know" is taken to mean that the information is either not present or inaccessible.

The reality however is that in survey interviews and other forms of dialogue we can very quickly formulate a response to something we might never have experienced, acquired any information or thought about. For instance, although someone might admit to having little or no specific knowledge of (say) their government's foreign policy, nevertheless they might not hesitate to give a response to an opinion poll concerning how the government is performing in that respect. Of course, what they are doing is converting a general attitude into a statement of specific opinion.

Let's define a general attitude (hereafter just "an attitude") as a disposition that we adopt to deal with situations calling for some kind of response when more specific guidance is unavailable - eg., when one doesn't have an opinion. Attitudes are extremely useful in daily life. Rightly or wrongly, attitudes determine how one behaves in relation to other individuals, social situations, politics, products, and all manner of things - including surveys. Indeed, because they are so useful, attitudes can withstand conflicting facts. Only with reluctance are they changed in the light of new information.

This is how it comes about that most people seem to have an 'opinion' when asked even the most novel opinion poll questions. Of course, it helps enormously to have not only an attitude to work with (eg., overall support for or opposition to the government) but also to be given some alternative answers from which to choose for the issue in question (eg., approve or disapprove).

This calls into question what an opinion is. Are we prepared to accept that it is something which can suddenly pop into existence when and only because an opinion poll question has been asked - ie., that in effect it is an artefact of the question? Or would we prefer to define an opinion as being like a belief which must exist prior to and independently of the question? If the latter, then 'opinion poll' is a misnomer. Such surveys more often measure attitudes than pre-existing opinions about specific issues.

Now, if a respondent happens to have a definite opinion about the subject, then that person's response to an opinion poll is pre-determined in much the same way as his response to being asked his age is pre-determined. That is, if he definitely knows his age, then this is independent of the question being asked and his response is fairly certain (although of course the possibility exists that he might lie about it).

However, few responses to opinion polls are pre-determined and in that sense certain. To give any response other than "don't know", the respondent probably needs to have an attitude towards the subject (eg., a general like or dislike of the government), but this is insufficient for us to be able to say that there is certainty about the response to a specific issue about that subject. That is, we can't assume that an 'opinion' expressed under such circumstances really existed as such before we measured it.

We should say therefore that up until the moment when such an 'opinion' is expressed, the outcome of the question is fundamentally uncertain.

The point is perhaps clearer when we think about the responses gathered by brand image repertory grids. Respondents are given (say) 20 image attributes from which to choose those that they feel apply to a particular brand. Typically the attributes have been mentioned in preliminary qualitative research, some by a lot of people but others by only a few. Now, of all the respondents who happen to associate any particular attribute with any particular brand, how many actually associated the two before the question was asked? Sometimes it might be many, but more generally, it is clear that many such associations are constructed only at the instant when the question is asked. Such responses seem right to the respondent; that is, they are in keeping with the attitude towards the brand.

The response to a survey question is pre-determined and certain if for instance the respondent has

- a definite pre-existing opinion about the specific issue addressed by an opinion poll - an opinion which might be expressed independently of the question being asked
- a definite pre-existing impression of a brand in terms of a particular imagery attribute
- clear recall of a specific reading event, upon which the response to a reading question is based.

On the other hand, the response is fundamentally uncertain if for instance the respondent -

- spontaneously translates an attitude towards the subject of an opinion poll into a statement of 'opinion' (eg., agree or disagree) in response to a specific proposition regarding that subject
- reports 'new' brand associations in response to a brand image question
  estimates how recently a publication 'probably' was read, or how many out of the last 4 issues probably were read, etc.

Strictly speaking it is impossible to tell whether a response is certain or uncertain in the above terms. We can be reasonably confident that responses to most demographic questions are certain. But it is undoubtedly true that, equally as typically, responses to readership questions are uncertain. Of course, some people will have the specific circumstances (eg., the recency) of a reading occasion clearly in mind, but more commonly respondents will fall back onto a cognitive device analagous to an attitude in order to answer the screening, readership and frequency questions.
To avoid terms such as 'cognitive device' and 'response model' in this paper, we will speak simply of "the way" a respondent answers a question.

Protoscoping

One such way has been termed 'protoscoping'. [Walsh, 1991] Protoscoping was proposed as a better explanation of a phenomenon observable in responses to Recency questions, which hitherto had been thought due to Telescoping. In brief, Protoscoping is where respondents answer time-related questions without reference to time-related facts; or answer issue-specific reading questions (eg., TTB) without definitely identifying the test issue as one which has or has not been read.

Protoscoping asserts that there is generally some probability that a respondent will claim (eg.) to have read a monthly publication within the past month when in reality it might have been much longer ago. Telescoping assumes that such an error is due to how time is stored in and retrieved from memory. Protoscoping instead asserts that the memory of elapsed time does not necessarily determine the response to a time-related question at all. At least some respondents (and the indications are that it is very many) instead formulate their response to such a question in a different way.

Further, it has been argued [Walsh, 1985] that because so many respondents answer readership questions in other ways, among certain groups of readers the measurement error can be very large. It can be so large that primary readership estimates (ie., estimates of buyers and subscribers from a source-of-copy question) can be up to several times the actual circulation.

But how is it possible that the way in which a respondent answers a readership question can be so much at odds with the facts? Here is an example: Let us suppose that, once a year or so, a person buys several successive issues of a particular motoring magazine. He does this when he is thinking about perhaps buying a new car. Then he stops reading it. When we ask him about it later, he 'protoscopes' the usual episode into the present. In effect he says "*When* I read this magazine, I usually read about 3 issues in a row." On this basis his actual response to the frequency question might be "3 out of 4".

In other words, what is associated with the title in this person's memory is a kind of formula enabling him to say what 'usually' happens *when the publication is being read at all.* It does not necessarily encompass the breaks *between* reading episodes; that would be rather irrelevant to remember.

Pre-determination and Certainty

If relevant 'facts' (ie., pieces of information *factual to the respondent*) are both present and accessible, then there will be a conflict between them and the way in which the respondent is otherwise inclined to answer the question. Somehow, out of this conflict, a response will be formed. But the 'facts' might play a relatively minor role, in much the same way as people's political attitudes can withstand a lot of contradictory information. So, although 'facts' may make themselves felt to a greater or lesser extent, the important point is that they do not necessarily determine the response.

On the other hand, if 'facts' are either not present, or inaccessible, or even just too much bother to retrieve, then the response will be entirely a function of whatever way the respondent chooses instead to answer. And so, the response is fundamentally uncertain. If the researcher was able to step back in time and

observe the respondent's actual reading behaviour, it would be impossible to reliably predict what the response to the survey question will be. It might be possible to determine what the response *should* be, but that is another matter.

To recap, responses to readership questions are pre-determined if and only if the respondent has a definite and accurate recall of a specific reading occasion. Typically this is not the case. When a respondent is uncertain, he will find some other way to answer the question. And so, we must regard his response as being uncertain in a quite fundamental way. It is uncertain not only because *he* is uncertain, but also because we cannot know which way he has answered the question. What we do know is that different ways of answering survey questions tend to produce different responses.

A deterministic counter-argument could be mounted here; it would go something like this: Even if the respondent answers 1the question in some other way, this in itself is the result of past experience and behaviour, and the response is therefore certain in the sense that it is fully determined by that history. However, this first ignores the fact that responses which can easily be manipulated experimentally are in a significant sense an artefact of the question. Second, if the researcher was able to step back in time to observe actual reading behaviour, *different* responses often would be expected. Third, because the respondent makes a choice, the way in which he will answer the question is not pre-determined.

A central argument being put forward here is that uncertain responses are artefacts of the interview in a way which is similar to a coin landing heads or tails is an artefact of someone having tossed it. This brings us to the next issue:

Uncertainty and Probability

For its own part, survey research is always based on a 'model' of some kind; explicit or implicit. There is no such thing as simply collecting facts, because even when facts are manifestly present, the questionnaire (ie., the choice of what information to collect and by means of what questions) together with the sampling, interviewing and analysis procedures, constitute a model in which there are many significant assumptions. Moreover an interviewer, with his own set of assumptions and attitudes, is interposed between the researcher and the respondent. Realistically then, can we ever say that the outcome of such a situation is pre-determined or certain?

Due to uncertainty, what we are about to measure with a survey question typically exists only as a probability distribution - a distribution at the respondent level - rather than as a certain piece of information. All responses which are possible for the respondent have some probability of being observed. This is similar to the situation when a coin is tossed. Until the moment it lands, heads and tails both have a 50% chance of being observed.

But unlike heads and tails, saying that all possible responses to a survey question have probabilities of being observed does not mean they have *equal* probabilities. Some responses will have much larger probabilities than others for the individual respondent. Nevertheless they all have *some* probability so long as they are possible for the respondent. The distribution is shaped by the way in which the respondent is about to answer the question, which is itself uncertain.

This, now, is the essence of the probabilistic view. For any individual respondent whose response is not pre-determined and therefore is uncertain, a response probability distribution applies.

The fact that such distributions exist makes it possible for the response to one question to be logically inconsistent with that to another elsewhere in an interview. This is often observed in opinion polls canvassing a range of issues. People are not always consistent in their 'opinions' when different questions bring different attitudes or contexts into play. Such contradictions would be less commonly observed if the 'opinions' were pre-existing, because cognitive dissonance would demand that most of them had been brought into harmony before the questions were asked.

This probabilistic view contrasts with what might be termed the 'literal' view of survey research which is that, assuming both the question and the response are perfectly understood, what we observe is a pre-existing 'fact' manifested in the literal meaning of the response. Necessarily therefore the literal view treats inconsistent responses as 'wrong' in a significant sense and problematic in analysis. But from the probabilistic view, logical inconsistency is only to be expected; simply on a mathematical basis (ie., the product of independent probabilities) there is a probability that it will arise.

Probability of the Screen-in Paradox

It is a simple matter to explain the screen-in paradox from the probabilistic viewpoint. First, there is some probability that even a regular or recent reader will answer the screening question in the negative on one survey, but would in the affirmative on another *even with all other things being equal*; that is, ignoring the various influences mentioned at the outset (eg., interviewer effects).

Likewise, there is some probability that a peripheral who is screened-in will go on to both qualify himself as a recent reader and answer the frequency question as if he was a regular reader. As above, this is simply a matter of probability due to the uncertainty typical of responses to readership questions.

Therefore, when screen-in levels are manipulated experimentally, or when they move up and down from one survey period to another, it is only to be expected that AIR will change as well.

Summary

The screen-in paradox can be understood from the probabilistic viewpoint. This holds that unless a particular response is pre-determined by overwhelmingly clear 'facts' being present and accessible in the respondent's mind *and* used to answer the survey question, it will be answered some other way. In this case, the response is fundamentally uncertain, and it is an artefact of the interview (ie., it does not exist as a 'fact' prior to the question being asked).

Prior to the question being asked in such a situation, more than one response has some probability of emerging at the respondent level. Therefore, there is some probability of two responses being given to separate questions which are logically inconsistent. In essence this is what is observed in the screen-in paradox, except that in this case the inconsistency is observed in the aggregate.

Some Implications

As the EML experiment showed, it is possible to increase the number of titles without experiencing a general decline in AIR. From this it follows that three factors (viz., the screen-in horizon, the number of titles screened, and how they are presented) can be manipulated to maintain a constant overall screen-in level so as to improve the comparability between surveys (for common titles). That is, if the average number of common titles screened-in per respondent is kept the same between two surveys, then this should make AIR more directly comparable.

Therefore the question of what the 'ideal' screening horizon is - this being an issue which exercises the minds of readership researchers - depends on the number of titles to be covered and how much freedom exists in the survey design to modify the way they are presented. But even using the 'ideal' horizon, and fine-tuning the methodology to introduce more stability to the average number of titles screened-in per respondent, AIR is still likely to move up and down due to various other influences upon the screen-in level. The movement can easily be greater than statistically expected, which as history shows has a negative impact upon the perceived reliability of the survey.

(The statistical expectation that a set of results should fall within a calculated range rests on the assumption that what has been measured is certain, and that only the sampling introduces any uncertainty. But we might ponder whether the concept of statistical significance really holds true if what we are measuring has a fundamental uncertainty all of its own.)

There are applications of survey research in which it is generally accepted that the information conveyed by responses should be treated probabilistically rather than literally. For instance, in studies of the potential for new products, "definitely will buy" is not taken as literally meaning that the respondent actually would do so if and when the new product comes onto the market. Rather, researchers look at the overall distribution. Probabilities (ie., weights) are assigned to the responses and the weighted sum is compared against normative data. Of course, there is an important difference between a purchase intention and a behavioural claim. One concerns the future and the other the past. But that distinction is not so significant if the way respondents answer questions about the past is not very different from how they answer questions about the future (ie., with uncertainty). So, just as we must assign weights to the responses to a purchase intention question, researchers also should assign them to reports of *past* behaviour if this is found to improve the accuracy and reliability of the estimates.

The suggestion that readership data should be calibrated (probably to circulation, although there are alternatives) is of course highly controversial. But the fact remains that with current methods, whatever screening horizon is chosen will have a material effect upon AIR. This makes it difficult to argue that current methods are based on valid measurement constructs.
This in turn negates the main argument against calibration. In the author's opinion, calibration is the direction which needs to be taken in order to arrive at more accurate and reliable AIR estimates. *As now*

can be seen, it might well be sufficient to calibrate just the screen-in level rather than the actual AIR. In practice this could involve establishing by some calibration technique a quota of past year readers for each title, so as to control the number of respondents going on to the readership questions.

Bibliography

Appel, Valentine; *How Changes in Screen-ins Affect Reads.* (Session Papers of the 6th Worldwide Readership Research Symposium, 1993).
Mallett, Daniel T.; *The Relationship of Screens and Reads and the Role of Screening in Readership Measurement.* (Session Papers of the 6th Worldwide Readership Research Symposium, 1993).
Walsh, Peter; *Magazine Sourcing.* (Session Papers of the 3rd International Readership Research Symposium, 1985).