TEST-RETEST RELIABILITY OF SCREEN AND READ

Peter Walsh, AC Nielsen-McNair

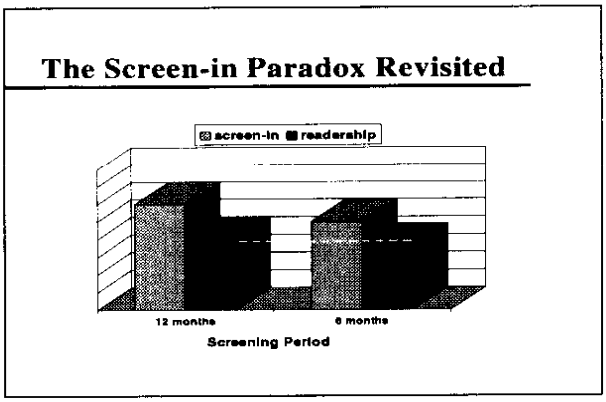
A theory was put forward at the Berlin symposium to explain the Screen-in Paradox. This paper describes an experiment in which we tested some implications of the theory with regard to the test-retest reliability of screening and readership questions. As predicted, an appreciable degree of unreliability was found, with significant impact upon readership estimates. The findings present a strong case for further experimental work being needed.

The Screen-in Paradox Revisited

The purpose of the screen is of course so that detailed readership questions need be asked only for titles the individual has read within the screening period. Without it we would encounter severe problems of respondent fatigue. There is, however, a downside. The screen prefaces subsequent measures with a question about reading in the past 12 months, or some other long period, over which we cannot realistically suppose people have clear recall. And so, some uncertainty is created from the outset.

The damage this does is evidenced by a phenomenon which has come to be known as the Screen-in Paradox. Its paradoxical nature is most readily illustrated as follows: Suppose we split a sample into halves and employ a six months screen in one half and twelve months in the other. Logically, in response to a subsequent recency question, the numbers claiming readership of a title within any period shorter than six months should be the same in both sample halves. But this is not what we would find.

Instead we would find more screen-ins over the 12 month period, naturally, but more readership claims as well - ie., more respondents claiming to have read monthly magazines in the past month, weeklies in the past week, and so on. So this is the paradox. Why are these numbers not the same?



This is not a problem with the Recency method exclusively; we would find much the same thing with Through-the-Book and, one supposes, any other readership methodology employing a screen.

This presents us with quite a challenge. At the very least it raises the issues of what screening period leads to the most accurate readership estimates and whether the same period should be used for all titles. It also calls into question the validity and reliability of readership claims in general, making it a particularly interesting phenomenon from a theoretical viewpoint.

The Theory

The Screen-in Paradox can be explained by a theory concerning response uncertainty *[Walsh, 1995]*. Response uncertainty is defined initially as the mental state characterized by two or more responses to a survey question being possible for the individual. By a response being possible we mean that it has some potential to be evoked. For instance, if asked which brand of toothpaste I bought most recently, either of the two brands I buy from time to time would be possible responses because I am not certain which one applies. Similarly, if asked for an opinion of the overall performance of the current Prime Minister, both approval and disapproval would be possible responses because my views on the subject are mixed.

Theory

Uncertainty is two or more responses being possible for the individual.

Every possible response has some potential to be evoked, and these potentials vary.

A particular response is certain if it is the only possible response for the individual.

Certainty on the other hand is the condition of only one particular response being possible. If asked how many children I have, the number 3 would be the only possible response because I am in no doubt about this being the fact of the matter, and it is inconceivable for me to be untruthful about it.

Potentials vary due to 'goodness of fit'. Because I buy Colgate more often than Macleans, the former response would have a better 'fit' for a question about my most recent toothpaste purchase. What is perceived as one's usual behaviour is a good clue with which to overcome uncertainty about any specific instance of that behaviour.

Likewise, approval of the present Prime Minister's performance would have a better fit than disapproval if I voted for his party at the last election. Goodness of fit potentially can be influenced by many things as well as past behaviour, such as moral and social values, psychological conflicts, selective memory, and so on.

While the potential for a response to actually be observed depends upon its goodness of fit, the choice that is made need not necessarily be the best fit. If the best fit was reliably determined in every choice situation then there would be no uncertainty. Therefore we can further define uncertainty as the condition produced when for whatever reasons - insufficient information or the complexity of the problem being likely candidates - mental processing fails to converge on the best fit (if there is one).

At least, it does not converge within the time available. In this sense, having a stranger on one's doorstep or telephone asking for immediate answers to questions one does not normally think about is very likely to create uncertainty, not just encounter it.

..... Theory

Potentials reflect variable 'goodness of fit' of alternative possible responses.

Because the 'best fit' is not a certain outcome, uncertain choices are indeterminate.

Nevertheless such choices tend to reflect the distribution of potentials.

An alternative way in which uncertain choices could be made is by their having a virtually random basis, as if a kind of mental Monte Carlo function. Philosophically this is not a very appealing idea, but it accounts equally as well for how things appear.

Whatever the mechanism, uncertain responses effectively are indeterminate. Nevertheless they tend to reflect goodness of fit. That is, if it was possible to repeatedly turn back the clock or ask the question in multiple 'parallel worlds', so that the individual had to respond again and again under exactly the same conditions, then the choice would not be the same every time, but the distribution of these responses would reflect the distribution of potentials for that individual.

..... Theory

The distribution of potentials is equivalent to a probability distribution.

The probability distribution can further be shaped by methodological factors.

Such influences would not be observed if people were certain about their responses.

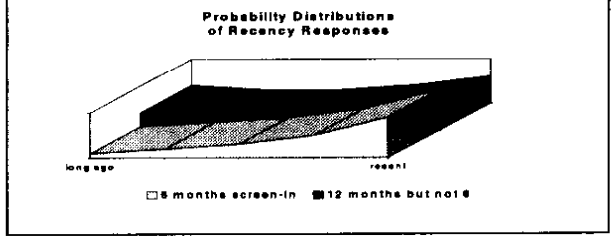
It is postulated that when there is uncertainty in daily life, choices are made in basically the same way. A great many of the choices we have to make are sufficiently uncertain for the best fit to be difficult if not impossible to discern. But somehow we manage to make them anyway. Importantly therefore, to at least some degree the theory preserves the symmetry between uncertain survey responses and real life attitudes and behaviour.

But goodness of fit is not the only factor. Due to indeterminacy, the distribution of potentials is equivalent to a probability distribution which may further be shaped by methodological influences such as the order in which brands appear on a showcard. Uncertainty empowers such influences; their effects would not be observed if respondents were sure about their answers to our questions. We can see this as the general mechanism by which, for instance, the structure of a recency or frequency scale can have a substantial impact upon the responses it produces. A wide variety of such phenomena in readership and elsewhere in market research can be understood in these terms.

Explaining the Paradox

We can now return briefly to our illustration of the Screen-in Paradox to summarize how the theory explains it: Some people would be uncertain about the screen and some would be uncertain about the read. Of course, they are different questions with different sets of response options. Consequently, even though there is a logical relationship between them, there are different distributions of potentials across the responses possible for the uncertain individual, and therefore a degree of independence between those distributions.

Partial Independence of Logically Related but Uncertain Responses



This creates the possibility - indeed the probability - that a respondent would screen out with the 6 month screen, and yet, if we asked the 12 month screen instead, they would screen in and then go on to claim readership more recently than 6 months. In aggregate, the recency distribution displayed by those people who would screen out with the 6 month screen but in with the 12 might not be very different from those who would screen in either way. In the '12 months but not 6' screener group there would probably be more reads within the period 7 to 12 months ago, but there would also be some in the more recent period.

Experimental Hypothesis 1

Experimental Hypothesis 1

Due to uncertainty, some respondents who screen out in a first interview should screen in for the same title in a second interview, even with no new reading in between.

Ie., we should observe a degree of test-retest unreliability in the screen.

If there is uncertainty about the screen, then the theory is that the respondent will have some probability of screening in and some of screening out. Therefore, if we screen publications on two separate occasions, then some respondents who screen out the first time should screen in on the second trial. This is the experimental hypothesis. Of course, since the two trials happen at different times there will be some reading in between, but we can allow for that.

The Experiment

This test-retest experiment screened newspapers and magazines on two separate occasions with the same respondents 7 days apart. Basically the approach was as follows: In the first interview, screen and read were asked for newspapers, while only the screen was asked for magazines. In the second interview, screen and read were asked for both newspapers and magazines.

Screen-ins from the first interview were carried forward automatically and read out as a reminder before screening the remaining titles again in the second interview. This can be seen from the wording of the second interview screen for magazines:

> *You told me last time that you have read .....* [READ BACK LIST]. *Now I'll read out the names of some other magazines. As I mention each one, please say whether or not you have read or looked into ANY issue at ANY TIME whatsoever over the past 3 months. It doesn't matter whose copy it was or where you may have read it. It could have been at home, at work, at a friend's home, or anywhere else at all, such as in a doctor's waiting room. The only thing that matters is that you spent at least 2 minutes reading or looking through it.*

Titles were organized into sets based on content similarity. In each set, the titles screened out the first time were screened again, but if there were no screen-ins for a set then it was skipped in the second interview with that respondent. This means that the number of respondents re-screened on any set in the second interview was much smaller than the total sample - a fact which should be kept in mind when considering the results.

As an experimental control, further sets were added to make the list of titles to be screened at least as long in the second interview as in the first. 61 newspapers and magazines were screened in the first interview and on average 70 were screened in the second. The average set consisted of 5 like titles which were read out slowly, with a longer pause between sets. Rotations were applied to sets and titles within sets, but newspapers were always asked before magazines.

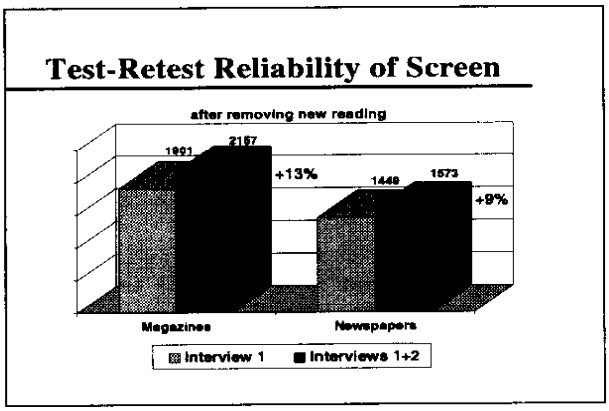
CATI was employed. This is not considered the ideal method of readership data collection, but we chose it as the most cost-efficient means of conducting repeat interviews for the purposes of this experiment. The sample was N=284. Telephone numbers and respondents within households (people aged 14+) were selected at random. The evening fieldwork was spread over the month of November 1996.

Test-Retest Reliability of Screen

Magazines

In the first interviews our 284 respondents generated a total of 1,901 magazine screens, an average of 7 out of 50 titles each. In the second interviews another 299 screens were obtained for the same magazines. 256 (86%) of the 299 additional screens were by respondents who, according to their responses to subsequent questioning (about first-time reading of the issue and frequency of reading the title), should have screened-in the first time.

These 256 additional screens represent an increase of 13%. This can be considered a conservative figure since, as stated earlier, only the respondents who originally screened-in to a set of titles were re-screened for that set in the second interview.



Possibly the increase could be due to confusion of title, which might be greater with telephone interviewing. But we analysed how many of each title's new screens came from respondents who originally screened-in for other titles with similar names compared with those for dissimilar names, and the results do not support that hypothesis - the slight difference was insignificant.

Nor did we find the additional screens to be significantly greater for monthlies than weeklies, or for small circulations than for large. There were some differences between the sets: the greatest additional screen-in rate was found for home and garden titles and the least was for young women's magazines, suggesting the possibility of an age bias. But no significant differences were observed between demographic groups, although of course the depth of this kind of analysis was limited by the modest sample.

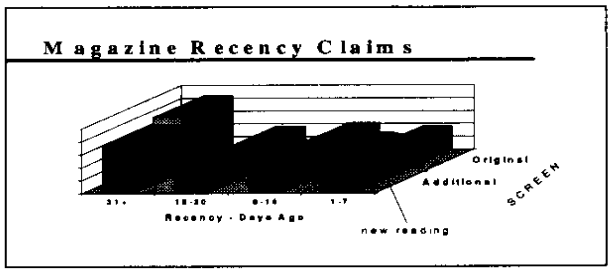
Newspapers and Inserted Magazines

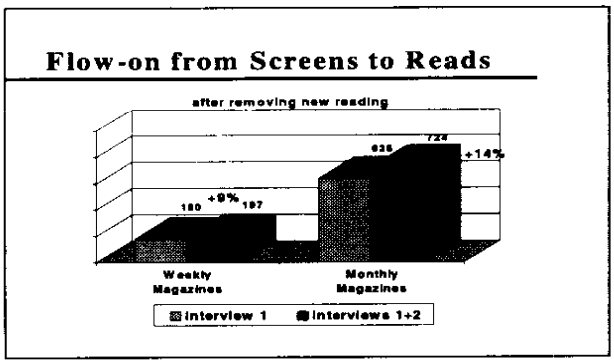
In the first interviews our 284 respondents generated 943 screens for daily and Sunday newspapers. In the second interviews another 127 screens were obtained for the same titles. The figures for newspaper-inserted weekly magazines were 506 and 78.

Combining newspapers and inserted magazines, 949 (65%) of the original 1,449 screens had read the title within the past 7 days. However, only 81 (40%) of the 205 additional screens had done so - ie., the majority of them had *not* read within the past 7 days, but instead had most recently read prior to the first interview and therefore should have screened-in at that time. They represent a 9% increase on the original screens.

Effects Upon Magazine Readership

After removing the additional screens which appeared to represent genuine new reading with respect to the screening period (ie., based on subsequent questioning), we compared the recency claims for weekly and monthly magazines to judge how much the additional screens would have contributed to readership estimates had they screened-in at the first interview.





Weekly Magazines

180 (26%) of 703 original screens for weekly magazines were within the past 7 days, compared with 17 (20%) of 85 additional screens. In this way, a 12% increase in screens yielded a 9% increase in reads.

Monthly Magazines

636 (53%) of 1,198 original screens for monthly magazines were within the past 30 days, compared with 88 (51%) of 171 additional screens. So, a 14% increase in screens yielded a 14% increase in reads. The increase in readership for monthly magazines was *pro rata* to the increase in screens.

Experimental Hypothesis 2

When there is uncertainty, the distribution of potentials is strongly influenced by the perceived frequency of the behaviour in question. Positive commonplace behaviour, such as reading, tends to be perceived as taking place more regularly or recently than is actually the case. Source-of-copy data for the vast majority of titles indicate that this leads to readership over-claiming, at least by people who enjoy a magazine or newspaper enough to buy it.

It follows from the theory that a lessening of uncertainty will reduce over-claiming. So we turn now to the primary hypothesis of the experiment. Assuming there is less uncertainty in a second interview, we should observe a decline in readership claims. This was measured for daily newspapers on the basis of how many days over the past week they were claimed to have been read.

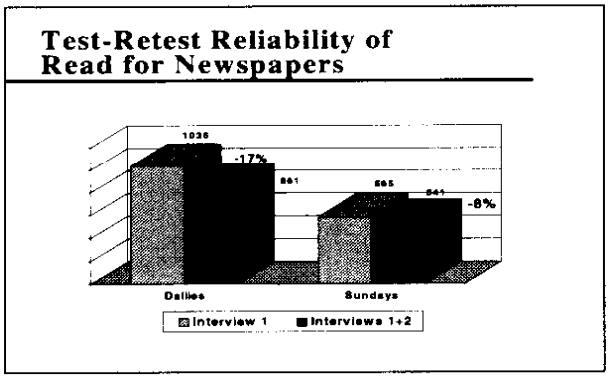
It is reasonable to assume there would be less uncertainty in repeat interviews because, now more familiar with what is involved, respondents should be more comfortable with the situation and task. Also, if the first interviews have the effect of sensitizing respondents to remembering their reading in the week between interviews, then there would be less uncertainty the second time. We recognized the risk that respondents might also be behaviourally sensitized - ie., to reading their daily newspapers - which would tend to increase the frequency subsequently claimed, but expected that the reduction in uncertainty should outweigh this.

Test-Retest Reliability of Read for Newspapers

In the first interviews, our 284 respondents generated 364 claims of having read daily newspapers over the past 7 days (ie., one or more issues). In the second this grew to 377. This larger number includes claims arising from the additional screens, 29 of which evidently represented new reading with respect to the screening period.

When day-by-day readership was asked however, the total number of reading events declined from 1,036 in the first interviews to 908 in the second. In other words, the average frequency fell from 2.85 to 2.41, or -15%.

If we consider only the respondents who screened in to a paper in the first interviews, then the number of daily newspaper reading events declined from 1,036 to 861, or -17%.



Thus the expected decline was observed in daily newspaper reading claims, demonstrating a considerable degree of unreliability in the original claims and supporting the hypothesis that reduced uncertainty leads to more accurate reporting.

The reference time-period for the day-by-day readership question was just the past 7 days. For Sunday newspapers and inserted magazines it was instead the past 3 months - ie., as in the recency question asked of all screens. Therefore we should not expect uncertainty to be moderated as much for Sunday papers and their inserted magazines as for dailies. Accordingly it was found among first interview screens that the number claiming in the second interviews to have read weekly newspaper titles over the past 7 days declined from 585 to 541, or -8%.

Conclusion

The theory, which originally was formulated to explain the Screen-in Paradox, predicts phenomena which readily can be observed by interviewing the same respondents on two occasions. It predicts the unreliability of screens and reads, in which regard what we have discovered from the experiment is cause for real concern.

While telephone interviewing was employed to minimize the cost, the same phenomena should be observed with other re-interview or longitudinal methodologies. With a readership diary panel for instance, the theory leads us to expect a drop-off in readerships for daily newspapers and weekly magazines after the first week or two, and for monthly magazines not much later.

The unreliability of responses to the screen flows through to the readership estimates. In the case of monthly magazines, a variation in screens (eg., due to a methodological change, or the behaviour of interviewers in the field) appears likely to produce a *pro rata* variation in reads. Clearly this is a methodological problem which needs to be addressed.

The primary purpose of the experiment was to test the effect of repeat interviews upon readership levels. The findings support the hypothesis of reduced over-claiming when there is less response uncertainty. With an appropriate experimental design, the same should be found for magazines as for newspapers. It is hoped that further experimentation along these lines will indicate how readership measurement can be made more reliable.