Donald Monk
Canadian Facts
Toronto, Canada

3.1 Validity: what is it?

PREAMBLE

Two months ago an envelope without any identification as to sender was left at the reception desk of my office. It contained a photocopy of a memorandum and, as its contents seem pertinent to this topic, a rough translation is given below.

PRAVDA CONFIDENTIAL

To: National Achievements Committee
3rd June 1982
Subject: Pravda Readership Claim

Although we have no readership data the committee wishes to claim that *Pravda* has more readers than any other publication in the world. My task was to assess whether such a claim would be disputed. To assess the situation I attended two conferences — in Stockholm and in New Orleans. Security was lax and there was no problem in attending without registration.

The state of readership estimates from other countries may be judged from some of the evidence presented:

i) Minor changes in question wording can alter results by 60%. (South Africa)
ii) The number of choices provided will alter results by 50% (West Germany)
iii) One US method gives results 88% higher than another method. (USA)
iv) Neither US method gives accurate results. (USA)
v) Everything works in Australia. (Australia)
vi) No one knows what a reading event is anyway. (UK)

My judgement is that our claim is unlikely to be disputed. Should that view be incorrect it will be easy to obtain several international experts to discredit any survey evidence used. They would seem to have a validity problem.

INTRODUCTION

The Pravda memorandum is clearly a travesty of the current state of readership research. But most of us have listened to scathing attacks on our methods by people in the advertising industry wishing to demonstrate their no-nonsense dynamism. So we cannot entirely dismiss this type of outsider perception of our industry. I suggest that the memorandum is regarded as a mild parody of a real situation. Readership research does have a validity problem. But so does practically every area of marketing and social research. What we need to assess is whether our current status is simply a reflection of the difficulty and complexity of the task, or whether we have neglected the issue of validity such that our measures are culpably in error. Therefore, as opening speaker on the topic of validity, I see my role as exploring some of the conceptual issues. But before examining the various types of validity, I will make some background observations.

For many years readership studies were conducted in individual countries as if in isolation from the rest of the world. Within each country doubts were occasionally raised about particular aspects of the measurement system: for example, replication and parallel reading in Britain, or the 'witch hunt' in South Africa. But until recently, there was little pressure for radical change, particularly on the part of publishers. Almost by definition any publisher who has remained in business is not suffering too much from whatever system of measurement is being used. Any change of method will have unpredictable consequences, producing some winners and some losers. The result has been greater demand for reliability of data than for validity.

The fundamental realities of the publishing business have not changed. But, for the researchers that serve that industry, the intellectual environment has altered, partly due to a greater degree of international communication enhanced by symposia such as this. The parochial orthodoxy has been challenged. Methods used in one country have been tested in other countries. People have become more open in admitting that not all aspects of their methods yield perfection. And no one now claims that their particular method is the absolute yardstick of truth. One result is that researchers have become more conscious of validity.

Although increased international communication has aided this process, some of the activity has been a mixed blessing. Some past conference papers seem to me to have more enthusiasm than methodological

3.1 Validity: what is it?

rigour. Simple comparability studies, that present aggregate data only, have been presented as validation studies. And the terms 'validity' and 'reliability' are often used as if they are interchangeable. I do not know whether this is a conceptual or semantic confusion. But I was finally persuaded to take on the task of defining some terms when I noticed that Wally Langschmidt's excellent book, which deals extensively with issues of validity, is called 'Reliability of Response in Readership Research'.

WHAT IS VALIDITY?

Respondent derived data may deviate from whatever is the truth at the time the data are obtained. Such deviations will be caused by random errors, by bias, or by a combination of the two.

The reliability of a measure is a function of the amount of random error in the system. This random error can have many causes quite apart from sampling considerations. Some examples would be: transient respondent factors (mood or fatigue); transient interviewer factors (mood or fatigue); situational factors (where interviewed, people present); mechanical errors (checking, coding).

A methodology that had no model bias but high random error would essentially be measuring the truth but with high variance. The average of repeated studies should approximate to the true situation.

A valid measure is one that has both good reliability and an absence of bias. But bias is the essential component of validity. Reliability is simply a necessary condition. One way of defining validity is as that quality of a method that measures the true situation with a minimum of random error. But perhaps we can get a better view of what is involved in having a valid method if we look at the necessary characteristics of an ideal readership measure. Such a measure will:

1.	Have a low level of random error.	**(High Reliability)**
2.	Be universally understood by respondents, who are psychologically capable of providing the information, and whose understanding is the same as our intention.	**(Construct Validity)**
3.	Be equally valid for all types of people and all types of publication.	**(General Validity)**
4.	Produce results that look sensible.	**(Face Validity)**
5.	Be internally consistent	**(Internal Validity)**
6.	Agree with outside criteria.	**(External Validity)**

This is a severe check list for any measure. It is particularly severe for a complex human activity that may be carried out at any time and at any place. In fact, the check list merely reinforces the obvious point that there is no possibility of our designing a test of total validity. Which means that we are left with making small skirmishes around the general target and making extrapolations that have no formal logical justification. These skirmishes are usually called partial validation tests, and typically examine one item of the check list. Such partial validation exercises are of two main types: **(a)** Analyses conducted on the data from established readership studies. **(b)** Special one-off experimental studies. I will review each of these two types of partial validation.

ARE THE DATA INTERNALLY CONSISTENT?

There are two types of validity check that can be conducted by secondary analysis of readership data. One is a type of face validity that queries whether the relationships fit in with common sense. The other is a type of verification procedure.

An example of face validity is checking the expectation that higher average issue readership estimates are associated with more frequent reading. For example, the use of three different survey methods might produce the following results. (See **Table 1**)

Method A has no face validity as it offends common sense. Method B has high face validity and the potential for general validity. The method C results are more typical. There is some face validity but there is some deviation due to bias or random error.

It should be noted that the method C results do not provide any indication of whether the discrepancies are due to the frequency scale, the classification of average issue readers, or a combination of the two. Historically, people have tended to assume that the results demonstrate a weakness in the average issue reading measure. But in most surveys, the frequency question is a much more difficult question in terms of recall period.

The second type of internal validation is when estimating procedures are verified. An example is the analysis of television viewing data reported by Appel *(1)* at the New Orleans Symposium to demonstrate

3.1 Validity: what is it?

TABLE 1

	Claimed number of issues in past 6 months						
	6	5	4	3	2	1	None
Reading probabilities							
Method A	0.40	0.50	0.30	0.60	0.30	0.40	—
Method B	1.00	0.83	0.66	0.50	0.33	0.17	—
Method C	0.89	0.76	0.60	0.47	0.38	0.31	—

telescoping effects. A second example comes from Canadian PMB data.

In the Canadian readership study a cross-check to circulation is used to derive an adjustment factor that is applied to primary readers. The cross-check is itself a form of external validation. But even if there was perfect agreement at the aggregate level, this would not necessarily indicate validity. A further test is to take sub-groups that have characteristics that one expects to interact with readership and compare estimates for those sub-groups. For example, different magazines have very different sex profiles, therefore, a pertinent test is to examine whether men and women provide similar primary household receipt estimates over different types of publication.

To avoid household composition bias, single-sex households were excluded. From the remaining data, the number of primary households per 1000 households was computed separately for male respondents and female respondents. Some results are shown in index form, based on the female estimate divided by the male estimate times 100. (See **Table 2**)

The results clearly indicate a sex bias in the propensity to claim primary household receipt for particular magazines.

This type of internal verification cannot demonstrate validity. It is similar to the position of being asked to prove the null hypotheses. A particular analysis can show that there is bias, but many analyses showing no effects cannot prove validity. They can, however, improve the researchers level of comfort.

ARE THE DATA EXTERNALLY CONSISTENT?

The main external criterion for comparison with readership estimates is circulation. At the face validity level, we expect similar types of publication to vary in circulation in approximately the ratio of the readership estimates. Also, if circulation doubles, we expect readership to increase, but not necessarily by one hundred per cent. We have no model that tells us how readership and circulation should co-vary, so at present we can do no more than apply a loose ordinal framework on our common sense expectations.

A more direct form of external validation is to derive estimates of circulation from the survey and compare them with known circulations. If the two figures agree it is assumed that, because the survey can measure circulation accurately, at least in aggregate, the readership measures are also accurate. There is, of course, no logical justification for this jump in reasoning, but as there is some relationship between circulation and readership an accurate estimate of circulation adds confidence to the readership data.

SURROGATE GOLDEN YARDSTICKS

Another approach to partial validation is the use of comparisons that cannot be totally justified, but which we believe to be nearer the truth. These can be segmented by type.

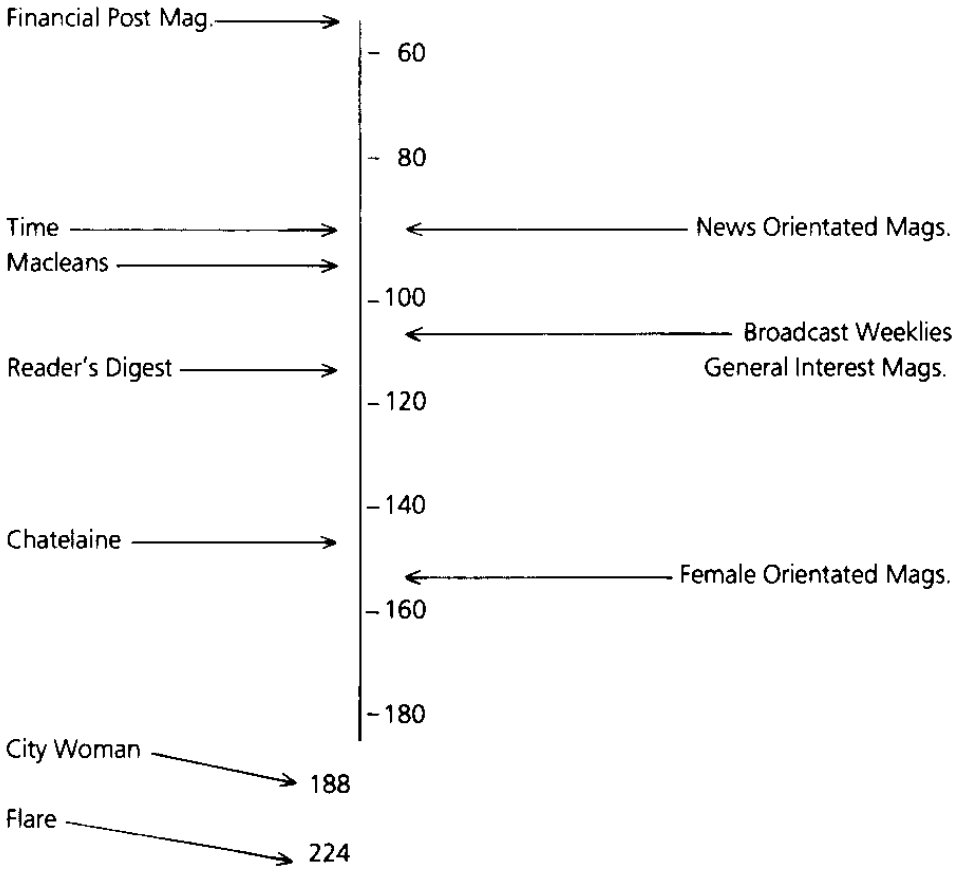
The Infinite Memory:

Ever since neuro-surgeons found that stimulation of brain cells sometimes restored to consciousness long-forgotten events, there has been a belief among certain people that all sensory experience is stored in the brain and the basic problem is to release it. Many quantitative researchers appear to hold this view, assuming that there is a form of words or type of visual aid that will stimulate astonishing feats of memory.

There is plenty of evidence from psychology and everyday life that the reconstruction of the past is an aid to memory. This works best when there is motivation on the part of the subject to recall the information, and the amount of information required is limited. Neither of these is typical of the average readership study.

But to some extent, they were present in the best known study of this type conducted twenty years ago by William Belson*(2)* in Britain. The merits of the intensive interview as a validating technique have been extensively debated over the years, but it has not elicited sufficient interest to be repeated. However, it has the advantage of

TABLE 2



setting out to assess all reading occasions rather than some limited sub-set. Also, it tends to give some explanatory information as to the causes of bias which are lacking in the more normative tests of validity.

Other methods in this category could be hypnotism, narcotics or brain scanning. They are untried methods for which we should keep an open mind. If one of them allows a person to give an accurate account of the nature and place of every meal in the past six months, most of us would be prepared to make the mental jump and assume that similar information on readership was equally valid.

Physical Checks:

The next group of tests involve some form of recording that people regard as reasonably objective; for example, observation or physical contrivances such as glue spots. Work in several countries, particularly South Africa and the US, have employed such methods. The methods used as yardsticks will have some random error but should be relatively free from bias. Therefore, they have value, but also the weakness of being restricted to particular types of reader or particular reading circumstances. They therefore lack general validity.

Chain Logic:

The third group is where a validation measure is selected that has good face validity and which can itself be partly validated using external data. The best example is the use of yesterday reading and yesterday first time reading.

Recall of yesterday's behaviour has reasonable face validity in that people intuitively feel that it is a time-period over which people can have accurate recall, and there is evidence from other research areas that good estimates can be obtained. Using measures of 'yesterday' reading and 'first time' reading, estimates of average issue reading are calculated. These estimates may be compared with other readership estimates

collected in the same survey or with readership estimates from other surveys. In neither case can comparisons be made at an individual respondent level so the comparisons are limited to aggregates. Cornish *(3)* used both types of comparison in his London Study and Joyce *(4)* has made extensive use of external comparisons to justify the higher readership figures obtained by Recent Reading in the US.

ADDITIONAL COMMENTS ON VALIDATION

A major problem with readership research is that the purpose is to provide precise measurements, yet we have no way of knowing what is the true situation. This means that, as researchers, we have a difficult balancing problem. We know that our methods probably provide reasonable estimates sufficient for the orderly transaction of advertising business. Yet we have difficulty proving our case and have poor defences against a well briefed interrogator. Because we do not know what is the true situation, our data are always inferences. The quality of those inferences is determined by the skill and effort we put into developing our measures and the evidence we can assemble to support what I will call their 'goodness'. This 'goodness' is based on an amalgam of partial validity checks, reliability checks, and verification that provide levels of comfort. This, in turn, means that we are unlikely to make a total breakthrough. Rather, our objective for validity should be to make progress a step at a time. But there probably need to be some guidelines as to how to make those steps, and the direction to take. I do not pretend to have the answers but I will comment on three possible areas.

Between title differences

As far as direction is concerned, I believe more emphasis needs to be given to between-title-differences. Although publishers may welcome a totally valid measure when it comes, in the interim their preference will be for a reliable method that has constant bias over all titles. Yet individual publications vary greatly in their distribution methods and the way they generate readership. For example, one simple way of classifying reading occurrences is between what I call: **(a)** Stationary people picking up moving publications (ie subscriber households), and **(b)** Stationary publications being picked up by moving people (ie waiting rooms).

Based on available evidence, it would seem that different readership methods vary in the efficiency with which they measure the different types of reading circumstances. Which means that they will be measuring different publications with different levels of efficiency. This implies that validity checks need to be applied across a range of reading circumstances rather than the one that is easiest to simulate within an experiment.

Explanatory and normative tests

If we are to design better methods, we require an understanding of both why particular types of bias occur and the limits of people's reporting ability. Too much of the work done in this area has been normative in nature, often in the form of simple comparability checks using aggregate data, and too little has been directed to the forming of hypotheses or knowledge. Although not an ardent supporter of the intensive interview, I believe we would have a better understanding of the factors that influence responses across a range of reading circumstances if a number of such studies had been conducted during the past twenty years.

Perhaps the most neglected area is a formal examination of the limits of people's ability to respond. Certain studies have been conducted under artificial experimental circumstances, but the range of possibilities is very large. For example, we know that a significant amount of occasional reading occurs at places such as dentists' and hairdressers'. The issue is whether people can accurately place such reading within a specific time band. But perhaps a starting point is whether people can accurately place their last visit to the dentist or hairdresser, irrespective of what they did there. As most people make appointments, there is an objective record of the date, and interviews can be arranged with people who visited three, four, five and six weeks ago. If people cannot give such information within the accuracy required for a readership model, it seems irrelevant what they did there. Some imagination assisted by some research funds would help in putting some constraints on methods. After all, if we do have a genuine interest in valid data, there is some argument for modelling data that we know can be obtained with reasonable accuracy, rather than attempting to collect any data that the modellers claim they require even though its validity is doubtful.

Reducing mythology

In readership research, we have too many examples of unreplicated evidence that enters the mythology of readership. Some of these studies may have been badly executed, simply unlucky, or compared with external data that were themselves in error. If so, this false knowledge acts as an inhibitor to progress. And even when some type of replication occurs and different results are obtained, the issue tends to be left in limbo. For example, the comparison between Recent Reading and Through The Book methods in Germany produced results at variance with similar data from several other countries. Yet I am unaware of any general view as to whether this was a breakthrough, an aberration, or something unique about the German people. I am not suggesting that all studies should be replicated. Some are specific to a particular publishing context and

inappropriate for export. Others are not worthy of replication or any other type of consideration. But unless some are carried out, we will continue to be handicapped with the mythology. Given the vested interest of many of the players, it is also advisable that such replication be conducted by reputable but neutral organizations.

CONCLUSIONS

We do have a validity problem, but not necessarily of the type suggested by my anonymous Russian correspondent. We face the same validity problems as everyone else attempting to measure a hybrid psychological and physical form of behaviour using self-reporting by people. As in many areas of applied technology, the results we obtain are probably better than we should expect on the basis of the theoretical evidence we have to support them. We need not feel depressed about the status of our measures, but we should feel challenged.

REFERENCES

1 Appel, V (1981). *Telescoping — the skeleton in the recent reading closet.* New Orleans Symposium.
2 Belson, W (1962). *Studies in Readership.*
3 Cornish, P (1982). *The London Experiment — an alternative recent reading method with partial validation.* ESOMAR, Stockholm.
4 Joyce, T **(a)** (1982) *The level of magazine reading.* New Orleans Symposium. **(b)** (1983) *Recent Reading.* ARF Conference.