MODELLING CONSIDERATIONS AS AN AID TO READERSHIP RESEARCH DECISIONS

Peter Massons, Massons
Paul Sumner, Consultant

1.0 Introduction

This paper has arisen from issues we encountered in the design and writing of a new model of media 'reach and frequency' - the press model incorporated in the SESAME system of media survey analysis. The design of such models is a typical synergistic process involving the end user, the data supplier and the modeller.

The needs of the end user (advertising agency, publisher or brand manager) are paramount; they, after all, pay for the systems. They should set and prioritize the agenda for the information they need to improve their buying and selling roles.

The data suppliers must respond to this agenda with expert advice on the economic and technical feasibility of collecting certain volumes and types of data, couched in a language that the user can understand and act upon.

As experts, we may of course agree to differ, but to our end users we must have a clear explanation of the debated issues and the assumptions involved in making a particular choice of data collection methods.

The modeller, who is even more apt to use jargon than the data supplier, has an intermediary role to play. He must decide on the best practical way to convert the elementary (in the sense of components, rather than simplicity) data into the statistics required by the end user. Because this modelling process is often expressed in complex mathematical form, it has earned the unfortunate title of the 'Black Box' part of media systems.

This usage is doubly unfortunate in that it implies (i) that the average user cannot be expected to understand what takes place in the 'black box' and (ii) that there may be hidden assumptions disguising model inadequacies.

Neither of these implications is true.

The perspective of this paper is unashamedly that of modellers, but of modellers who have had detailed involvement in both survey design and data provision and data usage (as marketing directors of international publications).

2.0 Facts and Estimates

Every media survey is subject to both survey and sample error. This is too often forgotten. As modellers we assume that both error sources have been minimised.

But, suppose you have a counted measure of A.I.R. (average issue readership). Suppose that measure indicates (counts) that publication A has an A.I.R. of 2% based on a perfect random sample of 1,000 individuals.

The best we can state statistically is that 'we are 95% certain that the A.I.R. of A is between 1% and 3%'. (Consult your monograms or Statistics text books).

Many, if not most, of the debated refinements to A.I.R. measures deal with measure changes which are much smaller than this.

For a publisher of a genuine 2% reach magazine, 1% A.I.R. is bankruptcy, 3% is retirement to Monte Carlo. If the A.I.R. measure is good (3%) then the survey is excellent; if bad (1%) then the survey is faulty: Human nature.

If we are measuring 100 publications on a survey then it is reasonable to expect that 5 will record A.I.R.'s outside the confidence limits - higher or lower.

Data providers must emphasise these limitations of survey statistics. They must ensure that data manipulators use their common sense and historical trends to assess publication performance - not solely an individual measure.

The main proposal we make in this paper suggests an abandonment of time - sensitive, statistically variable measures and their replacement with pragmatic, longer term measures.

We want to emphasise that surveys do not provide facts. Only poor, good or better estimates.

3.0 Measuring A.I.R - 'Counting'

Average Issue Readership (A.I.R.) is currently the principle elementary statistic. In fact it is usually measured by asking a question about recent reading. When was the last time you read or looked at X. Answers within the usual publication interval count as average issue reading (A.I.R.)

Principal issues in readership research in the last decade or so, exemplified by these symposia, have concentrated on the validity and reliability to the A.I.R. measure. These issues include defining readership, parallel and replicated, reading, telescoping, rotation effects, media list length effects, etc.

You will all be able to provide your own extensions to this list. Although there is increasing consensus on many of these issues we are a long way short of agreement.

Currently A.I.R. is an operational measure. Readership according to P.E.S. *is* the estimates obtained by the efficient application of the P.E.S. readership technique. For EMS it *is* the estimate derived from the application of the EMS readership technique. They will not be the *same* even if the questionnaires are administered to the *same* sample.

Explanation of differences due to technique can provide valuable insights. However we see little evidence that a 'gold standard' of readership measurement is emerging.

Historic expectations produce an inevitable brake on change. Publishers pricing policies have evolved in a highly competitive market place over many years, using an established readership yardstick. Changes to the A.I.R measure will be resisted by those publishers who see themselves disadvantaged. At best we have an expensive education job and a suspicion, usually unjustified, of all readership figures.

The operational yard stick provided by an average issue measures is an excellent first step for media planning. We contend however, that it does not matter if that yardstick is 2'6" or one metre long. As long as historically accepted publication rankings and A.I.R. levels are maintained or only altered significantly (in the statistical sense) as a result of circulation or socio-demographic or editorial changes, then any competent derivation of average issue estimate is useable and acceptable.

(Note: this is not to suggest that attempts to improve the reliability and validity of A.I.R. measures are misplaced, only, as we shall see, that we can make more progress, more cost effectively, by a re-ordering of priorities).

4.0 The data-users needs

Average issue estimates, or cost-per-thousands derived from them, are only a starting point for data users. They establish a 'candidate list' of publications to be used for a particular campaign. Even here the publisher can argue, and attempt to demonstrate that 'qualitative' or duplication considerations can render an apparently expensive publication worthy of consideration.

The planners' task is to distribute advertising exposures against determined target groups to maximise 'response' - however defined - within their budgetary parameters.

The publisher needs to understand and be involved in this task to demonstrate their titles contribution to the chosen schedule.

We contend that the data-users' main pre-occupation is to examine (evaluate) combination of titles/insertions. That is, they are planning, buying and selling schedules of publications, not single insertions.

To perform this task they have to use models of readership exposure based on the elementary readership data.

5.0 Models

All discussions of readership behaviour are based on some 'conceptual' model of such behaviour. We restrict ourselves here - for time reasons - to a more limited usage, that of a 'mathematical model'

'A mathematical model is a formal mathematical description of a process which permits the calculation, prediction or estimation of the value of an independent variable (e.g. net reach) given the value of one or more dependent variables (e.g. A.I.R. or the probability of reading individual publications, duplication between pairs of publications)'.

Nearly all current 'reach & frequency' models (and all sophisticated ones) are based on the availability of, for each sampled informant and for each publication researched, the probability that the informant reads the publication.

5.1 Assigning probabilities - Frequency Scales.

Use of the average issue question alone does not permit estimation of the growth of publication reach from issue to issue. Still less does it allow the estimation of schedule performance. (The dichotomous 0%/100% claim predicts that all A.I.R. readers see all issues. Non A.I.R. readers see no issues).

To assign probabilities we need a frequency of reading claim (frequency claim) in some form. A fairly typical frequency of reading question might take the form. "In an average month these days how many issues of weekly publication X do you see - 4,3,2,1,<1, none."

Whatever form the frequency question takes (see further below) it segments the readership of each questioned publication into a number of discrete groups. We then assign probabilities as indicated by the calculation process (shown in Table 1).

Table 1 - Example of probability assignment

(Based on a national sample of 1000 informants)

	4	3	2	1	<1	none
Frequency Claim	200	150	50	50	150	400
A.I.R. claim	190	115	28	10	20	0
Implied Probability	0.95	0.76	0.56	0.20	0.13	0

Table 2 - Establishing 'Average Issue' reach

A.I.R - counted	A.I.R. - probability
190	200 x 0.95 = 190
115	150 x 0.76 = 115
28	50 x 0.56 = 28
10	50 x 0.20 = 10
20	150 x 0.13 = 20
463	463
= 46.3%	= 46.3%

Table 1 is read as follows:-

For example 150 informants claim to read 3 out of 4 issues of X. Of these 150, 115 have claimed to have read in the last issue period. We therefore 'interpret' a frequency claim of 3 out of 4 to imply a probability of (115/150 =) 0.76 for every such claimant for publication X of reading an average issue of publication X.

In table 2, based on the total sample, the A.I.R. count and the sum of probability - weighted frequency claims produces, as it must, **the same** estimate of the 'average' issue audience.

All sophisticated models of 'reach & frequency' accumulation use this method of probability attribution.

5.2 Consistency of probability attribution

We can adopt two different probability attribution strategies:-

(1) Calculate probabilities for the lifetime of the survey.
(2) Calculate probabilities at run time for the particular target market involved.

The second strategy preserves consistency between cross-tabulated counts of A.I.R. and probability-based estimates whatever subgroup of the survey is being analysed. It involves assigning different probabilities on each occasion to the **same** informant if they are considered as (say) a 2+ car owner or a £20,000+ income earner. We think this defies common sense and produces 'statistical artefacts' rather than reliable estimates. It arises because the tabulated A.I.R count is taken as the lead statistic. This strategy often also involves the use of ridiculously small sample sizes and a concomitant wide variety in the probabilities assigned, an unreliable base for comparing schedules.

The first strategy seems to us much better. Indeed we believe it can be extended to a radical conclusion - for the production of good, operational schedule 'reach and frequency' estimates it is unnecessary to ask the A.I.R question (on a regular basis).

5.3 A Testable Hypothesis.

When the authors of this paper were first involved, some 25 years ago, in the development of 'reach & frequency' models we were concerned to minimise the estimating task. Remember if you can, that the evaluation of a <u>single</u> schedule could cost as a much as £30 ($100) with our primitive computers - at least £300 ($500) in today's terms.

Probabilities were then, as now, the fundamental statistics for schedule evaluation. We can recall our conclusions from that period when we used routine statistical tests:

(1) There was no statistical reason to believe that: a claim of 3/4 (say) for 'Women's Weekly' implied a different probability than a claim of 3/4 for Woman's Own: or that a claim of 1/4 (say) for The Times implied a different probability than 1/4 for The Daily Telegraph.

(2) Frequency claims for publications of differing periodicity <u>did</u> imply different probabilities: even within a group e.g. weeklies, 'Women's Weekly' claims were different statistically from 'General Weekly' claims.

Essentially what we did was to assume that the frequency claim was the more reliable measure (and based on far more positive responses) than the individual A.I.R. scores. We used the totality of the A.I.R. evidence to attach a probability meaning to frequency claims within statistically determined groups (minimising the variance within groups and maximising the variance between groups).

This change of emphasis, from probability attribution based on individual A.I.R. counts to probability attribution based on grouped statistical analyses has important implications and leads to the following testable hypothesis:

"The answer to the frequency question **does** provide a good, operationally usable estimation of implied probability and by extension good, operationally useable estimates of schedule performance".

The kind of evidence needed to test this hypothesis is available, in great historical detail, for almost all national and international readership survey data bases.

6.0 Supporting Observations - Implications.

(1) If the historical evidence and statistical analysis supports the view that we **can** attribute probabilities from the frequency claim, then we can drop the A.I.R. question from routine readership surveys - providing opportunities for shorter interviews; more extensive informant data; multi-media data etc.

We will still need to ask the A.I.R. question for confirmation of the historical probability attribution and for assigning new publications to the appropriate group.

(2) Work on the psychology of memory suggests that is easier to report accurately 'usual' behaviour than to remember and locate accurately in time 'recent' behaviour. ('I do something most weeks' is easier to answer than 'I did something not more than seven days ago').

(3) We are not suggesting that all frequency scales are equally suitable (further observations below section 8).

(4) We are suggesting that there is a predictive value in the use of frequency scales which is not present in A.I.R. measures (Section 7)

(5) Use of frequency scale - attributed probabilities uses far more positive observations (i.e. not 'never read') than do A.I.R. measures. Typically between 2 and 6 times more 'positive' measures than A.I.R. and 4 to 20 times more positive measures than F.R.Y.

This will considerably reduce the number of 'rogue' estimates (outside 95% confidence limits) and the tedious explanation that such mavericks are not 'errors'. Such increased stability is desired by all of us.

7.0 IMU - Testologen Evidence

Lindberg (IMU-Testologen, Sweden) reported in the first symposium the consistent relationship found in Orvesto-Konsument, between single-issue audience estimates, derived from a frequency-scale alone, and circulation. Two of his examples are reproduced here (charts 1 and 2) showing this consistent relationship over a long period (1968-1981). He has many such striking examples.

These charts demonstrate far more consistency than similar mappings of A.I.R., against circulation, and an implied stability of readers-per-copy estimates, whose fluctuation has concerned us greatly in recent symposia.

Further, they demonstrate an element of 'prediction' in the frequency claim audience measure. 'Usual behaviour' seems to be interpretable as 'usual/near-future intended' behaviour. The frequency-scale audience measures often lead circulation changes.

Chart 1

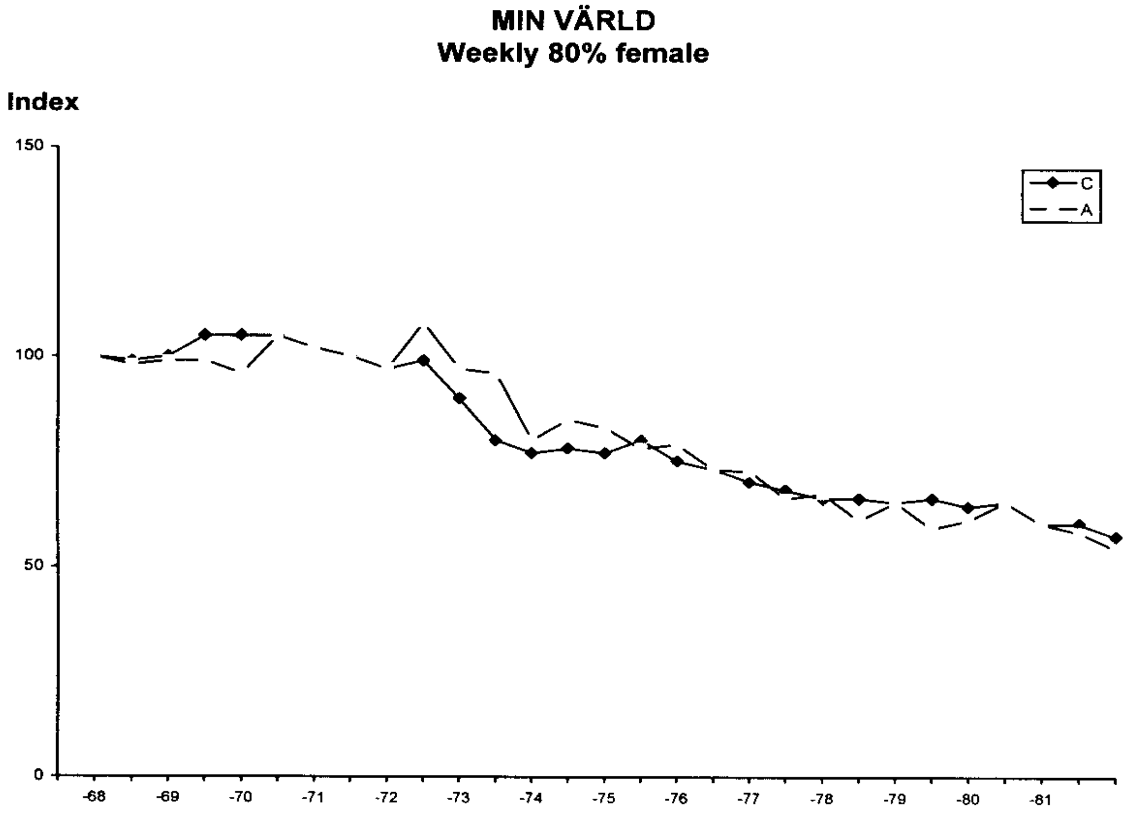
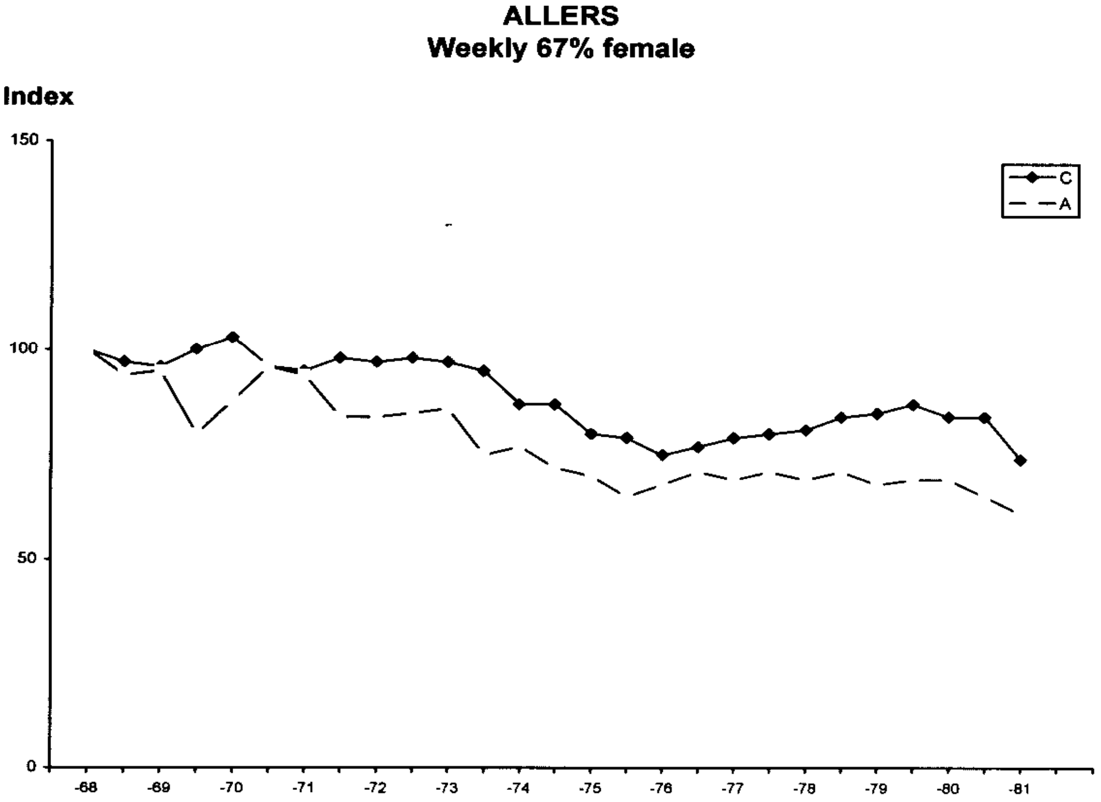


Chart 2



8.0 Some Frequency-Scale considerations

Since the first symposium there has been much discussion on the 'best' frequency scale to use. We establish a 'never reads these days' (zero probability) using a filter question.

Much of the ensuing debate becomes irrelevant if (as we believe) we can successfully drop the A.I.R. questions for routine surveys (e.g. A.I.R before or after frequency question, 'disclosed/undisclosed' recency scales etc.).

We contend that no frequency scale should have more than 7 positions (8 including never). There is abundant psychological evidence that humans can not meaningfully cope with more than seven categories.

We contend further that no scale should have fewer than 4 positions (5 including never). Most informants need this number of points to accurately describe their various reading strategies. Fewer than 4 points leads to a clumping of claims, and therefore implied probabilities, into too few groups. This in turn leads to model-derived frequency distribution estimates which are too 'peaky' and which distort schedule comparisons. (For a simplified demonstration of this effect see Appendix).

On the issue of verbal (semantic or numeric scales we prefer numeric scales as having far more consistent interpretation (if we allow '< 1' to be considered as a number). We note that nearly all verbal scales are now accompanied with a numerical guide (e.g. "Regularly' by which (mean at least 3/4 issues). We believe this can lead to conflict for the respondent as his interpretation of words may differ from the numbers.

The proposed statistical analyses will help to identify those scales which provide a stable and sensible segmentation of reading strategies.

9.0 Summary and Conclusion

The principal use of readership survey data is by media planners and brand managers in assessing the performance of schedules. Publishers too need to be involved in this task to demonstrate their publication's contribution to the schedule.

- Schedule performance statistics are derived from probabilities
- We can successfully derive implied frequency - claim probabilities from a frequency question alone-supported by detailed analysis of historical data.
- We can abandon the routine collection/counting of A.I.R. and replace it with a more stable frequency-claim-derived single issue measures.
- This process will lead to more stable/reliable/accurate frequency distribution estimates and therefore better/more accurate schedule comparisons.

Appendix

Effect of increasing frequency claim groups from 3 to 4

Suppose we measure publication X using the

α scale 'never, only occasionally (< 1 of 4), quite often (at least 1 of 4), almost always (at least 3/4)' **and the**

β scale 'never, only occasionally (<1 of 4), sometimes (at least 1/4), quite often (at least 2/4), always/almost always (at least 3/4).

We might find the following results:-

α	**0**	**1**	**2**	**3**
Frequency claimants	600	100	100	200
Probability	0	0.3	0.5	0.9

β	**0**	**1**	**2**	**3**	**4**
Frequency claimants	600	100	50	100	150
Probability	0	0.3	0.5	0.7	0.9

Both scales produce the same estimate of single issue reach - 26%

(α: 100 x 0.3 + 100 x 0.5 + 200 x 0.9 = 260
β: 100 x 0.3 + 50 x 0.5 + 100 x 0.7 + 150 x 0.9 = 260)

However if we use the standard binomial expansion to issues we obtain these rather different frequency distributions.

	sees 0	**sees 1**	**sees 2**
FD α	67.8%	12.6%	19.6%
FD β	67.4%	13.2%	19.4%

These differences may seem small, and they **are** for a single publication, but when multiplied out for many insertions in many publications of a typical schedule they can result in very different estimates of Frequency Distribution, with the shorter frequency scale producing more 'peaky' distributions and a distorted basis for schedule comparison.