Paul Sumner
Newsweek
London, UK

Jane Perry
Davidson Pearce Ltd.
London, UK

8.1 The modelling of readership behaviour

This session concerns itself with computer models of reading behaviour. The New Orleans symposium also contained a session which purported to devote itself to that subject. However, the treatment accorded the subject was inadequate and superficial. Too many of the papers were thinly disguised sales pitches, presented without context and with no serious attempt to explain how the models were constructed nor the implications of that construction for the user.

This topic is *not* peripheral. It concerns us all — as providers of research; as users of the data for editorial and advertising sales; as planners of advertising campaigns *and* as designers and sellers of models which analyse readership data.

The information collected from media research studies is subjected to more post-survey analysis than any other form of market research data. At present, for every £100,000 spent on collecting and publishing readership data, between £25,000 and £50,000 will be spent on computer analysis of that data. Straightforward cross-tabulation is a small (though increasing) proportion of this total analysis bill; the majority of the money provides schedule evaluations for publishers or advertising agencies.

For the publisher or advertising agency media research is rarely an end in itself — it is the beginning of a long and continuing process of buying and selling: of negotiating. If we adopt Brian Allt's terminology of a Press Negotiation Index we will have a clearer idea of the concerns that bring us to this symposium and the concerns which prompted this paper. If the choice of computer model can influence the media selection process (and it can) it is of vital importance that we should know what we are choosing — and why.

For the research companies the concern is just as clear. The data they so painstakingly provide can be used or mis-used, presented or mis-represented. But perhaps more important are economic considerations. The *volume* of post-survey analysis work has increased dramatically over the last decade; the *cost* of that analysis, as a proportion of total research costs, has declined. It is estimated that with each new generation of computers hardware and machine time costs have fallen by a factor of ten in real terms. Software costs have also fallen slightly but, being a people-intensive process, not significantly.

In comparison the costs of collecting the raw data that the models work on has escalated. To be fair, research costs have been well contained and have increased no more, and often less, than retail price indices. However, there will not be a real fall in the cost of interviewing and data collection — there has been a dramatic fall in the cost of computer analysis, a trend which will continue.

The implications are obvious. There exists a vast mountain of historical data — data which have scarcely been touched in terms of trend analysis, projection and forecasting. It is not inconceivable that these data can provide a solid platform for the construction of a readership model which would then need much less frequent updating; or updating from much more cheaply acquired sources such as circulation statements. Such a process may be desirable, even from the researcher's point of view, since it could release funds from what can easily become a sterile pre-occupation with head-counting to a better investigation of what goes on inside those heads.

This brings us, by a fairly circuitous route, to the main topic of this session, computer models of readership behaviour. What is currently available? What, given cheap computing power, might be available? What should be available? And finally, what is likely to be available?

Any one of these questions, fully addressed, could justify a lengthy paper on its own. In the brief time available we would like to offer a few thoughts and opinions for discussion and, we hope, to provide a context for the subsequent papers.

First, a definition. A model is a formal mathematical description of a process which permits the calculation, prediction or estimation of the value of a dependent variable (say the net reach of a magazine schedule) given the value of one or more independent variables (say last issue readership of individual magazines).

The values for the independent variables are obtained directly from research (last issue reach), provided through an intervening calculation process (probabilities derived from a frequency scale and last issue reach) or straightforward estimates (guesses — the effect of a strike, say). The modeller does not *need* to (but should) be concerned about the source of the input data. He/she is saying, in effect "If these are the values of the independent variables, this process X will produce this value Y of the dependent variable". The modeller is

8.1 The modelling of readership behaviour

concerned with devising, justifying and implementing process X. The models will work just as well with census data or guesses.

This being so it might be thought that the modeller can stand aloof from the media research debate, the lofty disinterested observer. Not so. Models which demand as input independent variables which can not be measured or estimated are at best intellectual curiosities. So the modeller must be aware of what current research or estimating techniques can provide. But that is not all he can do.

In designing a model every assumption must be made explicit. In this process the modeller clarifies the, often unstated, implicit assumptions. He can reveal flaws in the logic or point out additional or substitute measures that need to be collected in order to make the model work.

An excellent example of this process is the use of response functions to produce 'optimum' schedules. Building explicit models demonstrates that we do *not* know enough about the advertising process to provide the measures required for 'optimising' schedules. (Are three advertisements seen better than two? How much better? Computers need numbers. Is the same true for everyone?).

The modeller, the data provider and the end user are mutually dependent. The relationship should be symbiotic, not parasitic. The user states his needs from research, the researcher modifies the user's demands in the light of practicality and current knowledge, the modeller responds to both input availability and output requirements. That is what should happen, but too often communication is inadequate, resulting in a compartmentalised or blinkered approach.

In preparing this paper we asked friends and colleagues around the world to send us any articles or information that had appeared on new models of reading behaviour. The response was depressing and consisted largely of publicity brochures for computer bureaux. The underlying bases of the models were 'commercial secrets'.

While we understand that no computer bureau is about to give away its large (and it is large) investment in program development it is not too much to ask for a simple explanation of the model basis. Such an explanation would remove much confusion and apparent contradiction. (Non-additivity, for example, is not a mathematical error but a direct result of calculating C1's and C2's or probabilities for individual target markets rather than an average for the entire survey universe).

There are essentially only two different models currently available widely: individual simulation from derived probabilities, and Beta-matrix expansion from single-issue, two-issue and duplication data. (We exempt from this discussion the work of Dr. Morgenstern in France, which deserves wider dissemination).

The models do differ in their ability to introduce media and market weights and in formatting flexibility, but these differences are largely cosmetic. Approximations to the full binomial convolution and formulae models such as Metheringham, Sainsbury or Agostini are now only of historical interest since the full calculations can be cheaply performed.

The models are not particularly sophisticated in concept — in many ways less so than in the early sixties — and they are simple and cheap to operate. This simplicity is a mixed blessing. It trivialises the process of data analysis and disguises the need for an understanding of the basic calculation process.

We do not have the time to examine these current models in detail, or indeed access to the finer points of the individual implementations. We should remember that neither model provides an unambiguously 'right' answer.

In Europe in the late sixties and early seventies it was generally considered that individual simulation was clearly the best method. This arose because the readership model most used collected frequency data, making probability calculation simple, and because there was no good Beta-matrix model available. It is not now nearly so clear that individual simulation is superior. The model has two major faults as currently implemented. First is the attribution of probabilities. Probabilities are calculated and attributed to be consistent with the last issue claims. This provides internal consistency but does not accord with common sense. Why should a claim of (say) 4 out of 6 mean something in one target group and something else in a different group — for the same informant? The second problem is that of independence — the assumption that the probability of reading one publication is unaffected by the probability of reading any other publication. This again does not accord with common sense, and leads the model to over-estimate frequency and under-estimate reach.

Formulae models, whether based on the Beta function or some other two-parameter distribution, have their own problems. There are implicit if not explicit independence assumptions. Without adjustment (fiddle-factors) they can produce declining reach.

These objections do not and should not inhibit use of either model. They should require an understanding that models produce fallible forecasts, not absolute cast-iron predictions. We would like to think that that statement was obvious enough to be a platitude. It is not.

What is likely to happen next?

The main consideration in how schedule models will develop in the future is the development of computer hardware. As *processing time* becomes cheaper, and *programming* more expensive (people-intensive) there will be a concentration of analysis in the hands of one or two companies. As analysis costs fall, no company will be able to afford to develop new programs, or improve existing ones, unless it has a major share of a large market. The entry costs for schedule analysis models are already probably too high, in relation to the expected rate of return, to encourage any newcomers into the field.

A second factor which will encourage this trend is the development in international telecommunications links. These have also become progressively easier, faster, and cheaper in the last few years (it is very easy to forget how recently all this has taken place). This has several implications. First, program development costs can be spread over several markets. Data bases from surveys in many countries can be loaded on one central computer and accessed interactively from the individual home countries who receive the benefits of more advanced programs and a more powerful computer at no greater cost than dealing with a local bureau. An added advantage of this procedure is that demand for computer resources, and hence efficiency, can be spread more evenly due to time zone differences.

Second, the international satellite and cable links can themselves be used to transfer data, rather than merely to access it. This facility is rarely used at present, due mainly to cost and to a lack of urgency in making data available. Both these factors are likely to change in the near future, making the couriering of magnetic tapes as obsolete as the transfer of packs of punched cards. We will expect to see an increase in the development of networks of linked computers, transferring data and programs as well as messages, and probably acting as a back-up to each other in resources, in much the same way as the national grid provides electrical power in the UK.

What effect will these changes have on the models available? The major one will be that there will be less variety, and greater conformity, in the future. This is in line with a similar trend in the basic measurement of readership (although that may be debatable). However, the chances of any *improvement* in the actual models available are minimal. Left to their own devices, computer bureaux have a natural tendency to stick with what they have found to be acceptable and saleable, and to devote any resources available to improving presentation and formatting. We see no signs of any new kind of schedule estimation model in the near future. This would hold true, even if there were a major change in the measurement of readership, such as the introduction of panels. As Neil Shepherd-Smith has already pointed out a year ago (Admap, August 1982), the most likely use of such data would be in validating and sharpening existing models. It is unlikely that anyone would want to analyse the data directly for schedule evaluation on a routine basis.

We have not so far mentioned micros, which will undoubtedly cause major changes in all areas of data-handling over the next few years. The impact of desk-top micro computers on the models we have been discussing will not be significant. Exact duplication of the models for use on micros would be expensive to implement, and of no benefit to the current program owners. Further they would be very slow to execute — micros are not designed for matrix manipulation. Of course, a great deal of work has been done on approximation procedures which would speed up program execution, but at the expense of hard-won accuracy (or at least the elimination of one source of error). It is clear that many media surveys, particularly the smaller ones, are likely to become available in diskette form for micros over the next few years, but these will be primarily for cross-tabbing and cost-ranking purposes. We believe that schedule evaluation may be one of the last bastions of the main-frame computer.

We have outlined here the developments we think most likely in schedule analysis. Is this train of developments in the best interests of the user? We believe not. The computer bureaux, quite properly, are mainly concerned with the efficient operation of their businesses, and the optimum use of their resources of computers and programmers. They are not concerned with buying and selling advertising space, or with whether the models they sell are the most appropriate for the job. In most cases they, in the form of their employees and sales people, are purely middlemen in a basic (and pejorative) sense: they know neither what the models are used for, nor what their mathematical basis is.

The concerns of the media research business are, or should be, rather different. At present, there is a dismayingly low level of awareness of how the schedule models available work, and what they can appropriately be asked to do. We believe that the users of readership data, and the providers too, must in their own interests think about what is done, and what should be done. There is no more appropriate forum than a symposium such as this (in fact, there is no other forum at all, appropriate or otherwise). We believe that the bureaux themselves would welcome an agreed and user-originated definition of what is required, and that the

industry as a whole could rest happier if they believed that conscious decisions had been made. It is clearly as wrong that computer bureaux should be solely responsible for originating schedule models, as that research agencies should be solely responsible for devising readership questions.

It is right and appropriate that we should discuss now both what we want from computer models and how they can help us, before the decision is taken out of our hands. The dangers of allowing such decisions to go by default are major and real.