DOES THE 'FIRST READING YESTERDAY' MODEL ADDRESS THE CENTRAL MEASUREMENT PROBLEM?

Peter Walsh

SYNOPSIS

The FRY model sets out to more accurately reflect the true relativities between the audiences of competing publications, by using only 'yesterday' readership for estimation purposes. Also, the task of recall is assisted by a series of preliminary orientation questions, as a way of counteracting certain biases.

FRY assumes that the central measurement problem is to do with the fallibilities of recency-recall. This is questionable.

This paper argues that what can be achieved with FRY is only a marginal improvement, because it does not deal effectively with the central measurement problem in readership research. It is concluded therefore that the harsh methodological trade-offs which FRY involves, are not justified by a sufficient increase in the validity of its estimates.

Some results of a recent daily newspaper survey, using a 'first reading within the past 24 hours' approach (FR24), are presented to argue an alternative view of the central measurement problem.

DOES THE 'FIRST READING YESTERDAY' MODEL ADDRESS THE CENTRAL MEASUREMENT PROBLEM?

On the surface, First Reading Yesterday would seem a logical way of dealing with the fallibilities of recency-recall. If survey respondents are able to report their reading behaviour with any accuracy at all, then surely, the recall of yesterday's reading should be the least error-prone. This is the basic proposition underlying the FRY model. It is, however, open to question.

This paper does not concern itself with two obvious weaknesses of FRY which have been pointed out by others. An example of the first is that a monthly magazine with an average issue readership (AIR) of 10% or less, will have fewer than 100 FRY claims in an otherwise respectable sample of N=30,000. Obviously, this is a quite insufficient base for detailed profiling, and represents a great sacrifice of data-collection efficiency.

A second objection is to do with reliability. For a monthly with an AIR of 5% and just 50 FRY claims (N=30,000), the statistical tolerance is 14% of the readership estimate. Were we instead dealing with 1,500 claims (ie., 5% of 30,000), then the relative error would be only 2.5% of the AIR. Thus, FRY is less reliable than what might be referred to as the 'whole-of-AIR' methods.

This implies that FRY estimates are inherently more volatile. Between-survey movements must be expected to often be contrary to circulation trends. Results within demographic groups, and hence reader profiles, must be even more unstable. It seems possible also that FRY claims could be more than usually prone to non-sampling biases (eg., method of contact, the interviewer's manner, prestige, order effects, promotional activity, etc.).

But our interest here is in the fundamental proposition that yesterday's reading is reported more accurately than reading over any longer period, *sufficiently so as to warrant the sacrifice of data-collection efficiency and reliability.*

Why have readership researchers in certain countries been willing to make that harsh trade-off? Clearly, FRY arose from Recency's lack of credibility in various places, due for the most part to its extremely high readers-per-copy (RPC) figures and sometimes peculiar differences between the RPCs of similar titles.

No doubt it seemed logical to those involved in developing FRY, that recency recall-periods should be truncated if they are too long for accurate claiming. But from other viewpoints, FRY is not so logical. For one thing, it is difficult to see how what is essentially still a recency-based method but now relies upon literally just a handful of claims for many titles, could be expected to improve the credibility of readership measurement.

VALIDITY

In the frustrating search for a more accurate method, it is easy to lose sight of something fundamental about readership models. In the case of Recency for example, "read or looked into within the past ..... (publication-period)" is an operational construct which may or may not be adequate for the purpose; it cannot and should not be regarded as a precise definition of readership.

That is, just as the Recency model conveniently 'assumes away' replicated and parallel readership, it can -- until established to the contrary -- equally as validly assume that biases produced by the fallibilities of recall are insignificant. Likewise with Through-The-Book (TTB), memory errors are assumed not to be so variable across titles as to significantly distort the picture.

Whether a readership construct is valid, and whether any bias is significant, ultimately are questions about the objectives of the research. The principal aim, clearly, is to reflect the *relative* readerships of competing publications with an accuracy sufficient for the sometimes finely balanced choices made between them. It is possible for this to be achieved even if the absolute values of the readership estimates are somewhat wide of the mark.

So, high RPC levels do not of themselves mean that a measurement system is not up to its task. Lower RPC figures, while no doubt more credible than those typically produced by Recency, are not necessarily any more valid.

The methodological trade-offs FRY involves, therefore would be justified only if its estimates were significantly more valid than those of alternative readership constructs.

Well, does a truncation of the recall-period result in a more accurate reflection of the *relative* sizes of the average issue audiences of competing publications? That is, are the dynamics of the predominant biases -- whatever they may be -- such that their effect is lessened significantly when the task of recall is made easier?

An alternative hypothesis is that the biases remain even when the task of recall is made as easy as possible. For example, if the prestige effect was the central problem, then it would seem not much less likely to influence 'yesterday' reading claims.

So, if the central measurement problem was due to prestige or any other effects *independent of* the fallibilities of recall, then this would create a strong argument for a radical rethinking of the readership construct.

FALLIBILITIES OF RECALL

There is plenty of evidence that, within clearly defined sets of competing titles, both Recency and Specific Issue methods (eg., the Cover and Index approach used in Australia) distort the true relativities even between titles of the same publication-period.

The credibility of Australia's current national readership survey is sorely afflicted by some controversial RPC relativities which defy logic and common sense. This isn't at all unusual on the world scene, of course.

Now, if memory failure (ie., forgetting) is significant -- ie., if the reading of one publication is under-claimed to an extent sufficient to affect advertising choices between it and one or more of its competitors -- then FRY's truncation of the recall-period could pay worthwhile dividends.

Evidently however, the most controversial symptoms of the problem have little or nothing to do with under-claiming. Rather, it is quite obvious that, although a certain amount of casual pass-on reading undoubtedly goes unreported, there is a far greater tendency towards *over-claiming*.

From what evidence is this obvious? With every justification and to their credit, proponents of FRY appear to have taken the view that a test of their readership construct's validity is whether the sales estimates it produces from source-of-copy questions, conform at least in relative terms with the audited circulations.

In a great many cases, the survey estimates are far too high. Some of this may be due to readers incorrectly claiming purchase, as distinct from over-claiming readership. However, this is likely to be a relatively small error. It does not explain why circulation estimates for some publications are so much higher than the actuals, while for others of the same type, having very similar 'sourcing profiles' (ie., readership structures, in terms of subscribers, casual buyers and pass-on readers), the extent of over-estimation is not nearly so severe.

Perhaps respondents are getting the titles mixed up. Presumably, this would be reduced by FRY. However, while confusion of title does appear to affect certain isolated sets of magazines, it does not explain what is clearly a much more general problem.

That can be said also of the prestige effect. In any case, as suggested earlier, this is largely independent of recall, and therefore it is arguable whether it would be moderated by FRY.

So, what now remains to make FRY's harsh trade-offs worthwhile?

TELESCOPING

It is only to be expected that people who last read a magazine (say) 5 or 6 weeks ago, would in many cases see it as having been more recent (eg., 3 or 4 weeks ago). Telescoping would thus seem capable of inflating readership estimates to a level approaching the actual cumulative audience of 2 issues.

How effectively might telescoping be reduced by FRY? It remains entirely possible, of course, that someone might mistake the day before yesterday's reading as having taken place yesterday -- especially if perceived as part of one's usual daily activity.

If telescoping exists at all -- and there is no question about that -- then for *any* reference-period whatsoever (eg., the past month, or yesterday), there is some probability, however slender, of *any* prior interval being telescoped into the reference period. So, even though it may be doubtful that many readers of a monthly magazine would see it as having a place in their *daily* routines, there is still the possibility that the day before yesterday is telescoped into yesterday often enough to have a serious effect.

1/30th of the unreplicated reading of a monthly takes place on the average day. Let's suppose that half the reading which took place the day before is telescoped into yesterday. The total FRY claims now represent 1/20th of the monthly's AIR, so that when grossed up, the estimate comes to 1.5 times the true readership. This is the same error as would be produced by the past 6 weeks being telescoped into the publication-period.

Therefore, not only is it probable that FRY fails to eliminate telescoping, but also, it is even possible that it *increases* its impact upon some of the estimates. The latter can occur if respondents are not thoroughly orientated towards thinking about what they actually did just yesterday. This in itself presents a problem of basically the same nature as that of readership measurement because, as now to be discussed, no set of questions requiring the recall of yesterday (or any other period) is guaranteed to invoke the temporal dimension of memory.

'PROTOSCOPING'

The term 'protoscoping' is coined to denote a typical response to questions requiring the recall of ordinary behaviour. It is similar in some respects to telescoping, but is distinguished from that by being independent of temporal recall.

That is, for example, claiming to have read a newspaper yesterday can in reality mean *not* that one did read it yesterday or even necessarily the day before, but rather, that reading the paper is regarded as 'usual' daily behaviour.

Were we able to independently examine the individual's behaviour over time, then instances of the claimed behaviour would definitely be found before very long. In a loose sense, the behaviour has been brought forward in time; thus, protoscoping is similar to telescoping in this particular respect. But the probability distribution of prior intervals falling within the reference period, would be somewhat 'flatter' than that of telescoping.

Protoscoping allows that readership estimates can be inflated to the cumulative audience of *more than* 2 issues. The potential for this to occur is related directly to the turnover of readers from one issue to the next, as is the case with telescoping.

Clearly, even moreso than with telescoping, FRY's ability to counteract protoscoping depends upon the extent to which the interview format succeeds in persuading respondents to put aside all notions about their 'usual' behaviour, and to focus instead only upon yesterday.

But the great difficulty here is that responses to any unrelated questions which might be asked about yesterday in order to create this effect (eg., "Please tell me how you spent the day.", etc.), are protoscoped as well. Also, it increases the probability of reading subsequently being claimed for yesterday rather than the true last occasion, because yesterday has now become the focus.

An alternative approach is to ask a series of sourcing questions of the type "Please think back to the last time when you yourself went into a newsagency. When was that?", the idea being to focus upon definite instances of behaviour without making yesterday 'the hero', so to speak.

The survey results presented below clearly indicate that this isn't very effective either.

EVIDENCE

Evidence that the central problem is to do with telescoping and, more significantly, *protoscoping*, can be found in the results of almost any readership survey which collects source-of-copy data. Presented here is an Australian newspaper study conducted by the author in 1989/90.

The national sample involved over 9,000 telephone interviews in which respondents were first asked a series of warm-up questions: "when last" they themselves went into a newsagency or shop and bought a newspaper; whether they have any papers delivered to their homes and, if so, when last; when last they read a paper brought home by another household member; when last they picked one up at work or anywhere they happened to visit, such as at friends' homes or business premises.

The sequence of title-specific questions thereafter was -- when last an issue "came into your hands"; if within the past 7 days, then day-by-day reading over the 5 weekdays up to and including yesterday; if today or yesterday, then when first started to read that issue; and finally, source-of-copy.

So, the readership construct was first-time reading within the past 24 hours exactly (FR24, perhaps) -- ie., from the same time last night, up until the time of interview. Interviewing was spread evenly across all nights of the week, and across the nine weeks span of the survey. Within-household selection was random.

No significant differences were found between the incidences of FR24 and FRY claims by the 'outs' when called back. Therefore, there was no evidence of a sampling bias having been introduced by our focus upon the past 24 hours rather than yesterday. For the reason which will become clear in a moment, we also double-checked that there were no socio-economic biases.

Source-of-copy was questioned in detail, with particular care being taken to distinguish between different household members as the actual purchaser of the copy claimed to have been read. Thus, there can be a substantial degree of confidence that the circulation estimates it produced reflect the survey's relative error-levels across titles.

National	Casual Sales* (000)	Survey Est.** (000)	Error Margin	RPC
quality morning broadsheet	80	148	+ 84%	3.8
financial morning tabloid	53	73	+ 37%	3.9
Market 1				
quality morning broadsheet	117	246	+110%	3.1
morning tabloid	250	414	+ 65%	3.4
afternoon tabloid	362	469	+ 30%	2.3
Market 2				
quality morning broadsheet	121	232	+ 92%	3.0
morning tabloid	357	443	+ 24%	2.6
afternoon broadsheet	151	159	+ 5%	2.1
Market 3				
popular morning broadsheet	113	182	+ 61%	2.5
afternoon tabloid	110	84	- 24%	1.7
Market 4				
popular morning broadsheet	70	113	+ 62%	2.5
afternoon tabloid	94	107	+ 14%	2.2
Market 5				
morning tabloid	105	133	+ 27%	2.6
afternoon tabloid	70	85	+ 21%	2.4

* actual and ** estimated circulation, both less actual home-delivery

Actual home-delivery circulations have been subtracted from the two left-hand columns, because nearly everyone who gets a paper delivered on any weekday, gets it *every* weekday. So, while these respondents might over-claim readership to a slight extent, this cannot produce a corresponding over-estimate of circulation.

Therefore, after subtracting that number of copies from columns 1 and 2, the percentages in the third column show the magnitude of over-estimation relative to the *opportunity* for it to occur.

For example, the quality morning broadsheet in Market 1 sells an average of 117,000 copies per issue, apart from home-deliveries. 117,000 is the number of copies *capable of being over-estimated*. The survey estimate of circulation, also reduced by the actual number of home-delivered copies, is 246,000 -- ie., 110% higher.

Overall average issue readerships are not shown above; the table instead compares the RPCs derived from them (ie., AIR divided by the total audited circulation).

The two papers with the highest RPC figures in the above table (ie., the national dailies), are exceptions in that they have relatively high rates of corporate subscription and pass-on at work. We should put them aside.

It can be seen that, in general, there is a strong relationship between the type of paper (ie., whether 'quality' or 'popular' morning broadsheet, morning tabloid or afternoon paper), the degree to which the potential for circulation over-estimation has been realized -- even by as much as 110% -- and the RPC figure.

With FRY in mind, it is worth reinforcing the point that these reading and purchasing claims were based upon recall of the past 24 hours only! Moreover, these claims were made after a series of warm-up questions designed to get people thinking about actual events, to improve recall of casual pass-on, and to make it clear that we were only interested in hearing about *actual* reading behaviour, as distinct from what they might regard as 'usual'.

Let's examine Market 2. The 92% over-estimation of casual sales of the quality morning broadsheet, doesn't imply that the *overall* AIR of that paper is 92% higher than it should be, because the AIR includes not only home-delivery readers, but also occasional readers who would be less inclined to over-claim readership than people who see themselves as fairly regular buyers.

But what we can conclude is that the true RPC of that paper falls somewhere between 3.0 and 1.6 (ie., 3.0 reduced by the extent to which casual sales have been over-estimated). Likewise, the RPC of the afternoon broadsheet in the same market, is 2.0 - 2.1.

With both being broadsheets and both regarded as 'quality' papers (although moreso the morning paper), what else is there about them -- apart from one being a morning paper and the other an afternoon -- to explain why the readership of only one of them apparently is over-estimated to such an extreme degree?

What distinguishes the morning paper, and indeed is a general characteristic of those whose sales have been over-estimated, is a higher turnover of readers from one issue to the next. This is shown below in columns 2 to 4. Column 2 is the average casual sales per issue expressed as a percentage of the market potential (ie., the population less home-delivery readers). Column 3 is the number of different buyers of 1 or more issues out of 5 (ie., the cumulative audience of buyers), percentaged the same way.

National	Error Margin	% of Potential*		C/S =
		Sales	Cume	
quality morning broadsheet	+ 84%	0.6	2.3	3.8
financial morning tabloid	+ 37%	0.4	0.9	2.3
Market 1				
quality morning broadsheet	+110%	2.5	10.2	4.0
morning tabloid	+ 65%	5.3	14.2	2.7
afternoon tabloid	+ 30%	7.6	17.3	2.3
Market 2				
quality morning broadsheet	+ 92%	3.8	14.2	3.7
morning tabloid	+ 24%	12.6	22.8	1.8
afternoon broadsheet	+ 5%	4.7	9.6	2.1
Market 3				
popular morning broadsheet	+ 61%	5.7	15.0	2.6
afternoon tabloid	- 24%	5.3	12.6	2.4
Market 4				
popular morning broadsheet	+ 62%	7.9	21.5	2.7
afternoon tabloid	+ 14%	8.9	16.2	1.8
Market 5				
morning tabloid	+ 27%	11.5	23.9	2.1
afternoon tabloid	+ 21%	6.0	13.7	2.3

* market population less home-delivery readers

3.8% of Market 2's population potential buy an average issue of the morning broadsheet, while 4.7% buy the afternoon broadsheet. But over 5 issues, 14.2% buy the morning paper at least once (3.7 times the number buying an average issue), while 9.6% (2.1 times the average issue) buy one or more issues of the afternoon paper. The method of calculation here is explained shortly.

One reason for this difference is that the morning broadsheet carries the bulk of classified advertising in Market 2. This is true of all the metropolitan morning broadsheets in Australia. From it, they obtain relatively more occasional readers.

THE EFFECT OF AUDIENCE TURNOVER

In the absence of any independent evidence indicating otherwise,
let's assume that the broadsheets in Market 2 -- Broadsheet A and
Broadsheet B -- both have an average of 2.0 readers per copy.
We are still excluding readership in home-delivery households,
where any tendency to over-claim is inconsequential as far as our
circulation estimates are concerned. Multiplying their casual
sales by 2.0 puts the AIR of Broadsheet A at 8% of the remaining
market potential, and 10% for Broadsheet B.

The nett audience to which these AIRs accumulate over 5 issues,
can be estimated by using the claimed incidence of readership
within each frequency group (ie., 0/5 to 5/5), to calculate how
many people will read at least 1 out of 5.

In the case of Broadsheet A's 1/5 frequency group, for example,
18% of whom claimed readership of what is in effect issue 6 (ie.,
any issue within the past 24 hours), it is estimated that 63% of
this frequency group (rather than all of them, as claimed) would
actually have read 1 or more out of the last 5. The error in the
cumulative estimates so derived, should be acceptably small.

Assuming that the beta-binomial distribution gives a reasonably
faithful representation of how this reading is spread throughout
the population, then the most recent issue read out of the last
5 of Broadsheets A and B, would be something like as follows:

	issues read most recently					
	issue 5	issue 4	issue 3	issue 2	issue 1	cume
Broadsheet A	8%	6%	5%	4%	3%	26%
Broadsheet B	10%	4%	2%	2%	1%	19%

If telescoping from issue 4 (the day before yesterday) into issue
5 (yesterday) was the only bias, and if it happened at the level
of (say) 50%, then A's measured AIR would be 11% versus B's 12%.
At 100%, they would be both 14%. However, the measured AIRs seem
too far apart -- viz., 20% to 13% -- for telescoping to be the
only bias. Across all papers, telescoping by itself does not
appear to offer a complete explanation of the AIR differences.

The concept of protoscoping, on the other hand, allows that some
proportion of *all* the cumulative audience of a certain number of
successive issues will claim readership of any particular issue,
virtually irrespective (within reason) of when in fact they most
recently read. They see themselves as readers and tend to claim
readership within what they know to be the publication-period.

CONCLUSIONS

It is evident from the results presented here, that our efforts to ward off the biases were at best only marginally effective. Compared with Australia's NRS, on which newspaper estimates are determined from day-by-day claiming over the past week, our experiment was a success of sorts. The RPC disparities between competing newspapers were not so great; the gaps narrowed in the expected directions.

However, it is clear that the questionnaire strategy -- similar enough to FRY in structure -- failed to keep *extreme and variable* levels of error out of the results. Because all these papers are competitive with one another, the results obtained -- like those of the NRS -- cannot be regarded as sufficiently valid.

With the benefit of hindsight, it is easy to see that many people would simply have 'protoscoped' their responses even to the warm-up questions, thus defeating the main purpose of these.

In other words, many survey respondents persist in reporting their perception of 'usual' behaviour. It is the path of least effort. Perhaps also, some are simply incapable of casting their minds back over the events of the past 24 hours.

So, while casual pass-on reading appears to have been more fully reported due to the orientation questions, the major source of error was not sufficiently moderated by a truncation of the recall-period from that used on the NRS. This implies that, in the magazine field, the methodological sacrifices FRY involves, are unlikely to be worthwhile. FRY does not deal effectively with readership's central measurement problem.

Obviously, these conclusions tend to give a pessimistic view of the prospects for improving the accuracy of readership claims by questionnaire strategies alone.

But, it is possible for a quite different readership construct to be devised, which provides the parameters needed to model the biases so as to derive accurate estimates.

An approach wherein response biases are statistically modeled as part of the estimation procedure, should not be regarded as a fundamental departure. As argued earlier, the validity of a readership construct hinges upon its ability to reflect the true relativities between publications with sufficient accuracy.

It is argued that this can be achieved only if we discover how to model the biases out of the basic data -- something which questionnaire design alone appears quite incapable of doing.

POSTSCRIPT

In October 1990, three of Australia's metropolitan afternoon papers were either closed or amalgamated with morning tabloids to create '24-hour' papers. Another had closed some eighteen months before. Arguably, these papers were denied the funds they needed to combat declining circulations, by the quite severe and mostly artificial RPC disadvantages created for them by the NRS.

There are now only two stand-alone afternoon newspapers left, and neither of them appears to be doing very well. So, it would seem too late to do much about the disadvantaged positions into which popular tabloids generally and afternoon newspapers in particular have been put by the biases in Australia's readership data.

Nevertheless, developing a workable solution to this intractable problem remains an important goal, because it affects the results for magazines even more severely than newspapers.