Richard Lysaker
Audits & Surveys Inc, USA

4.3 TOWARDS A GOLD STANDARD

As is well known, in the US there are two sets of average issue audience estimates used for the buying and selling of magazine advertising. One set of estimates, from MRI, is based on the Recency method while the other, from SMRB is based on the Through-the-Book method. While the differences in these estimates have decreased across time, there remain some marked differences.

As a result of these differences, there has been sentiment backed by some financial support that we should determine which set of audience estimates, if either, reflects actual readership behaviour.

The ARF established the Gold Standard Committee to pursue this goal. The committee consists of a knowledgeable and unusual group of participants. It includes representatives from extremely competitive publications as well as the heads of the competing research companies in the audience measurement business.

The committee's efforts were designed to develop an audience measurement system which can be used as a 'gold standard' – a measure against which all other estimates can be compared and judged as correct or incorrect. We felt this task to be sufficiently difficult not to be restricted to approaches which are economically feasible on a syndicated basis.

Together we have developed a methodology which, in our judgment, has a greater likelihood of being validated than any other.

Explicit in all of our decisions was a very specific criterion for measuring validity. The validity of the gold standard method would be tested by its ability to correctly capture reading which has occurred without incorrectly capturing claimed reading which has not occurred. Specifically, underclaiming would be measured by the extent to which the gold standard captured observed reading. Overclaiming would be measured by the extent to which the method captured reading which, by observation, could not have occurred.

In brief, the gold standard was designed to measure up against observed reading.

In developing the gold standard, we attempted to address all known major sources of measure ment error and minimise them.

Here are the nine elements of our method and the reasons for their inclusion.

(1) The first time read yesterday model was selected to minimise the effects of memory decay between the reading event and its measurement.

(2) A filter question was avoided as numerous studies have shown that it filters out readers. For example, one ARF Certitude Study showed that a filter question used in connection with the Recognition method would have filtered out 12% of the readers.

(3) The Through-the-Book method was selected since extensive research has shown that retrieval of past events is much more facilitated by recognition than by simple recall.

(4) Full issues were used with exposure to all items in the magazine. Obviously, this precludes missing readers who happen to have read only parts of the magazine.

(5) The number of titles and issues included in the interview has been limited to 12 to minimise respondent fatigue.

(6) The titles chosen include pairs of magazines which might well be confused because of similarity in content and appearance. This is the same idea as that behind the successful grouped titles work completed in the UK.

(7) Multiple issues of each title were included to minimise confusion between readership of different issues of the same publication.

(8) The procedure includes defining readership for the respondent. This should minimise variations in respondents' interpretation of the meaning of the readership question. Readership is defined for respondents, by having the interviewer say, "The next question about this particular issue deals with whether or not you have looked into it before now. When I say looking into the issue, this includes reading, looking into or paging through, or opening".

(9) We have avoided asking respondents "was yesterday the first time" etc. This question has a bias in *favour* of first time reading – as any 'Yes – No' question has a bias in favour of the 'Yes' response. Similarly, we have not asked if the issue had been read before yesterday, as in this case a 'Yes' answer is biased *against* first time reading. The actual question to determine first time reading is "Not counting today, on how many different days did you happen to look into this particular issue?" A yesterday reader who has read the issue on one day is the only one who contributes to average issue audience.

The grid which follows outlines all possible combinations of observed readership and readership claims for the proposed gold standard method. It also shows the effect of each combination on readership estimates.

This grid lists all possible readership claims. A respondent could claim:

- To have read yesterday for the first time or
- To have read yesterday not for the first time or
- Not to have read yesterday.

The grid also lists all possible observation results as a respondent:

- Could have read yesterday for the first time
- Could have read yesterday not for the first time
- Did not read yesterday but read before.

Under four conditions out of twelve, there is complete agreement between observations and readership. These combinations are labelled *'Correct'*, when for example a first time reader claims first time reading; or a non-reader claims to be a non-reader.

With this model three incorrect combinations of claims vs observed readership can occur without causing incorrect readership estimates. They are noted in the Grid as 'OK'. For example, if a respondent read yesterday but not for the first time and he failed to recall that reading – we would not be missing a first time reader.

According to the model, *underclaiming* happens when first time readers either fail to claim yesterday readership or claim first time yesterday readership was not the first time.

Overclaiming occurs when a respondent did not read for the first time yesterday but claims he did. This can occur under three conditions: in two circumstances readers, who read before, can claim first time readership, and non-readers can claim such readership.

Figure 1

Overclaiming/underclaiming analysis grid

	Results of observation			
	Read yesterday		Did not read yesterday	
Readership claim	First time	Not first time	Read before	Not read before
Read yesterday for first time	Correct	Correct	Overclaim	Overclaim
Read yesterday not for first time	Underclaim	Correct	OK	OK
Did not read yesterday	Underclaim	OK	Correct	Correct

OK = Disagreement between observation and claim does not effect readership estimate.

The proposed testing for the gold standard method involves testing it against all types of overclaiming and underclaiming as outlined in this grid (Figure 1).

Before validation testing began, the method was tested for workability including: video taped sessions, Belson type interrogation of the video-taped respondents, and actual field trials. These efforts indicated that we had a method which was ready to face the cold reality of trying to match results with observed reading behaviour. The total validation programme consists of five studies, two of which have been completed. Those completed include:

– Overclaiming and underclaiming of young issues in public places

– Overclaiming and underclaiming of older issues in public places.

Studies one and two represent what is quite possibly the severest tests of the gold standard method. They deal with public place reading which by consensus tends to be more casual than other types of reading. They also deal with overclaiming which, based on the ARF Certitude studies, probably poses a more serious problem for the recognition method than underclaiming. This is based on the fact that overclaiming is possible for the vast majority of the population, even for the largest US publications. The studies yet to be completed, for which about half of the necessary funding has been secured, include:

– Overclaiming and underclaiming of young issues for in-home reading

– Overclaiming: non first time reading claimed as first time reading for public place reading and in-home reading.

Validation Study 1 was designed to determine whether or not the gold standard method avoids or minimises overclaiming or underclaiming of young or pre-publication issues read in public places. It also includes a small number of observations to examine the reading of aged issues in public places.

A summary of the study design follows:

Sample: Barber and beauty shops, doctor and dentist offices. 24 establishments, 289 completed interviews.

Timing: Observations on Mondays and Tuesdays. Readership interviews on Tuesdays and Wednesday.

Magazines: *Newsweek* and *Time* *Family Circle* and *Woman's Day.*

Issue ages: One pre-publication issue, and two aged issues per title. Weeklies – 4 to 10 weeks old, Tri-weeklies – 5 to 11 weeks old.

Validation: Reading observers were observed. 25% of the interviews were validated.

The study was a double-blind experiment as neither respondents nor interviewers knew the purpose of the interviews. Individuals observing the public place readership were not involved in interviewing.

The data base for the entire experiment is 3,468 respondent issues. Each of 289 respondents were interviewed regarding 12 issues for which we had actually observed reading or non-reading to have occurred in public places.

The base for measuring underclaiming was 214 observed readings while for overclaiming the base was 3,254 observed non-readings.

The level of *overclaiming* for *prepublication* issues was extremely low. Based on 1,000 respondent issues where reading could not have occurred because prepublication issues were used, overclaiming in total was two-tenths of one percent.

These data reflect eliminating two respondents where reinterviews showed one respondent who was not observed to read, had read a subscription copy received in the mail the previous day, while the other respondent had read a neighbour's copy purchased at a news-stand.

This level of overclaiming for the gold standard method is substantially lower than any we have seen before. (Table 1).

Table 1

Prepublication issues: overclaiming

	Observed non-reading	Claimed first time yesterday reading	Percent overclaim
Tri-weeklies	499	-	-
Weeklies	501	2	0.4
Total	1,000	2	0.2

The level of *underclaiming* for *pre-publication* issues was also low. The percent underclaiming was 7% so the 'capture rate' was 93%. This level of underclaiming is as low as we have seen before (Table 2).

The level of *overclaiming* of aged issues was also extremely low. It should be noted that some of those classified as overclaimers could have read the aged issue yesterday at a time and place other than those observed. Thus maximum overclaiming of aged issues is four-tenths of one percent (Table 3).

Underclaiming of aged tri-weeklies was low, while that for weeklies was high. However, in both instances, the sample size was very small as Validation Study I was designed primarily to measure overclaiming and underclaiming of prepublication issues (Table 4).

The eight cases of underclaiming for the weeklies were examined in complete detail including thirteen demographic and reading variables as well as which interviewers observed the reading and which interviewers interviewed the respondent. No patterns emerged to explain these eight instances of underclaiming.

This level of underclaiming for aged issues when a yesterday reading technique is involved was unexpected. Prior published studies such as the ARF Certitude Studies and work sponsored by *Newsweek* (reported at the second International Readership Symposium held in Montreal) plus unpublished studies always had resulted in capture rates of 90% or better.

Because the sample size was so small for the cell covering underclaiming for aged weeklies and because all prior experiments with the yesterday reading method yielded different results, a full-scale test of underclaiming of aged issues with an expanded list of magazines was conducted.

Table 2

Pre-publication issues: underclaiming

	Observed reading	Claimed first time yesterday reading	Percent overclaim
Tri-weeklies	79	73	8
Weeklies	77	72	6
Total	156	145	7

Table 3

Aged issues: overclaiming (?)

	Observed non-reading	Claimed first time yesterday reading	Percent overclaim
Tri-weeklies	1,124	6	0.5
Weeklies	1,130	4	0.4
Total	2,254	10	0.4

Table 4

Aged issues: underclaiming

	Observed reading	Claimed yesterday reading	Percent correct
Tri-weeklies	32	29	91
Weeklies	26	18	69
Total	58	47	81

It was clear that if the problem of underclaiming for aged issues was confirmed, the gold standard method was invalid. However, if a full-scale look at the phenomenon detected underclaiming of under ten percent, we would conclude that the gold standard passed this test and the validation programme could proceed.

Validation Study 2 was designed to provide a full examination of the gold standard method in terms of underclaiming for *public place reading* of aged issues. Weeklies plus tri-weeklies and monthlies were included. This study also provided another solid measure of whether or not the method may generate significant overclaiming of aged issues which are read in public places.

A summary of the study design follows:

Sample: Barber and beauty shops, doctor and dentist offices.
10 establishments,
200 completed interviews.

Timing: Observations – Monday-Friday.
Readership Interviews – Tuesday Saturday.

Magazines: *Business Week, Newsweek, People, Time, Family Circle, Good Housekeeping, Reader's Digest, Woman's Day.*

Issue ages: Three aged issues per title.
Weeklies – 1 to 5 weeks old.
Tri-weeklies – 3 to 9 weeks old.
Monthlies – 1 to 3 months old.

Validation: Reading observers were observed.
25% of the interviews were validated.

This study, like the first, was a double-blind experiment.

The data base for the experiment is 2,400 respondent issues as each respondent was interviewed regarding twelve issues. Depending on which magazine was read, a respondent was queried using one of two forms of the questionnaire. Once covered the four magazines in the initial study and the second covered the other four magazines.

Study data base
200 respondents
12 issues each
2,400 respondent issues.

As was anticipated for the initial experiment, underclaiming levels for aged issues were low. 221 were made of test issue reading and 211 of these yield yesterday readership for these issues. Thus, underclaiming was 4.5% yielding a capture rate of 95.5% (Table 5 below).

The high capture rate level cuts across all demographic groups. These results suggest it may be highest among females, those 35–49 years of age and those with incomes of $50K plus (Table 6).

The high capture rate also cuts across various levels and types of reading exposure. These results suggest it may be highest among those who read over half of an issue and those who read as oppose to flip pages (Table 7).

Again, as in the first validation study, overclaiming levels for aged issues were extremely low. In total, it was only two-tenths of one percent. This is a maximum as respondents could have read these issues yesterday when we were not observing them (Table 8).

Table 5

Aged issued: underclaiming

	Observed reading	Claimed yesterday reading	Percent underclaim
Monthlies/ tri-weeklies	98	96	2.0
Weeklies	123	115	6.5
Total	221	211	4.5

Table 6

Age issue capture rates: demographics (total magazines)

	%
Male	92.9
Female	98.4
18–34	94.4
35–49	98.7
50 +	92.7
No college	95.1
Some college	96.2
Under $25K	91.8
$25 – $49K	95.7
$50 +	97.1
Total	95.5

Table 7

Aged issue capture rates: reading method (total magazine)

	%
Read over half	100
Read half	92.2
Read less than half	93.4
Read only	100
Read/flipped	94.3
Flipped	94.6
1– 6 minutes	94.7
7–14 minutes	96.4
15 + minutes	95.2
Total	95.5

Table 8

Aged issue: overclaiming (?)

	Observed non-reading	Claimed first time yesterday reading	Percent overclaim
Monthlies/ tri-weeklies	1,102	1	0.1
Weeklies	1,077	4	0.4
Total	2,179	5	0.2

The first tests of the gold standard involved public place reading and both represent among the most difficult variables any reading measurement system has to face. These tests have been passed.

However, there are three additional studies required to complete validation testing.

– A study similar to Study 1 which is based on in-home reading in primary households rather than public place reading.

– A public place study and an at home study which measures overclaiming due to non-first time readers claiming first-time reading.

Study 3 is a mirror image of Study 1 except that the observance of reading will be surreptitiously made by the respondent's spouse and will be limited to adults living in primary households.

Study 4 and 5 are studies of overclaiming when time has elapsed between multiple readings of a given issue among readers whose first reading occurs in a public place or at home.

These studies focus on overclaiming which can occur when a reader who reads an issue at one time and then, say, two or four weeks later reads the same issue. Overclaiming would occur if

the reader claimed yesterday readership after the second or Nth reading and claimed to have read the issue on only one day.

For readers whose first reading occurs in a public place, the observation will be made by interviewers, while for those whose first reading occurs at home, the observation will be made by the reader's spouse.

Both types of readers will be later observed in a waiting room where they have to 'participate in a soft drink taste test'. A total of 150 interviews will be completed among each group where a qualified respondent is one who has been observed reading the same issue on two separate occasions. In each case, half of the sample will consist of those whose reading occurred two weeks after the first observation. For the other half of the sample, the elapsed time between readings will be four weeks.

Initially another study to study overclaiming on aged issues was included in the validation programme. However, Validation Studies 1 and 2 involved questions about some 4,433 respondent issues where we measured the maximum overclaiming level for aged issues. Maximum overclaiming averaged three-tenths of one percent across the two experiments. Therefore, we judge that further testing in this area is not necessary.

Clearly we have made progress in developing a gold standard. But progress has been slow. To date we have secured about half of the funding required to complete the final three Validation studies. There is on the immediate horizon one development which should materially increase our ability to secure the remaining funding. And this development comes from outside the magazine advertising community. I am referring to the imminent development of operational single source measurement systems.

In brief, current single source systems will include direct measurement of sales, television advertising exposure, pricing, displays and couponing for large samples of consumers.

These systems will be used to measure the sales effectiveness of television advertising. The impact of promotion spending will also be analysed from the same data base. Furthermore, data from these systems may also be used as a basis for buying and selling television advertising.

There seems to be little doubt that single source systems will be among use shortly and will exert a powerful impact on the marketplace. This is extremely relevant because, as I am sure you've noticed, magazines are not a part of the systems being developed.

I believe magazines must become part of any widely used single source system in order to maintain their fair share of advertising revenue. How this might come about is well beyond the scope of this discussion. However, it is not at all unlikely that magazines will be measured in a single source system by a method different from both existing US measurement systems. Furthermore, magazines may not have a very large voice in selecting this measurement methodology as they will represent only a small influence on the total revenues of single source systems.

I submit, therefore, that a gold standard of magazine audience measurement is vital so the industry can make sure that any methods of measurement included in a single source system will develop equitable audience estimates that do not short-change magazines.