

THE SEARCH FOR THE GOLD

Richard L. Lysaker
President
Audits & Surveys, Inc.

There is a substantial tradition in the U.S. supporting the development and usage of measures which can be used to validate audience estimates developed by syndicated services.

For over two decades the telephone coincidental method has been used extensively to evaluate audience estimates for both television and radio. It was used as recently as 1990 to evaluate the use of personal meters to measure television audiences. Electronic devices measuring cumulative current flow have been used to validate the reported levels of radio listening in automobiles. Magazine readership claims have been validated using the glue-spot technique. And, of course, the ARF Certitude Studies, using observed reading as a validator represent prior attempts at devising what has recently come to be called The Gold Standard.

The Gold Standard efforts, then, have occurred within the context of a long history of testing the validity of media audience estimates. The specific impetus behind the current work is based on the fact that in the U.S., there are two sets of magazine audience estimates which are used for the buying and selling of magazine advertising. These two sets of estimates come from two different methodologies: one based on the recency method and one based on the recognition method. The two methods get different results. The differences have diminished across time but some very large variations continue. Currently, for example, average issue audience estimates for two major publications vary by 10 million readers or more.

Prompted by these differences the Gold Standard Method was developed and experiments were designed and conducted to determine its validity.

The Gold Standard program has received financial support from a broad base including major publishers, magazine advertisers, advertising agencies and research companies as well as the ARF, Canada's PMB and Magazine Publishers of America.

The Gold Standard technique is based on the joint efforts of members of the ARF magazine validation committee, many of whom have been very active in all the past international readership symposia. Together we developed a methodology which, in our judgement, had a greater likelihood of being validated than any other.

At the most basic level the Gold Standard was designed to develop an accurate measuring stick against which the results of other methods can be compared and judged as correct or incorrect.

The first decision was to select a criterion for judging validity. Observed reading was chosen. Underclaiming would be measured by the degree to which the method captured observed reading. Overclaiming would be measured by the degree to which the method captured claimed reading which, by observation, could not have occurred.

In developing the Gold Standard Method, an attempt was made to address all known conditions which contribute to measurement error and limit them.

The nine elements of the method and the reasons for their inclusion are as follows:

1. The first time read yesterday model was selected to minimize 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.
3. The through-the-book method was selected since extensive research has shown that memory is facilitated more by recognition than by recall methods.
4. Full issues rather than stripped were used. This helps preclude missing readers who happen to have read only parts of the magazine.
5. The number of titles and issues included in the interview was limited to 12 to minimize 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 minimize confusion between readership of different issues of the same publication.
8. Readership is defined for the respondent to minimize variations in interpretation of the readership question. Readership is defined by having the interviewer say, "...When I say looking into the issue, this includes reading, looking into or paging through, or opening".
9. The first time question was not asked directly. Respondents were not asked "was yesterday the first time" etc. This question has a bias in favor of first time reading - as any "Yes - No" question has a bias in favor of the "Yes" response. Similarly respondents were not asked if the issue had been read before yesterday, as in that 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?"

Yesterday readers who have read the issue on one day are the only ones who contribute 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.

The readership claims a respondent could make include:

- Reading yesterday for the first time
- Reading yesterday not for the first time
- Not reading yesterday.

Based on observation, 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.
- Did not read yesterday or before

OVERCLAIMING/UNDERCLAIMING ANALYSIS GRID

Results of Observation

Readership Claim	Read Yesterday		Did Not Read Yesterday	
	First Time	Not First Time	Read Before	Not Read Before
Read Yesterday For First Time	Correct	Overclaim	Overclaim	Overclaim
Read Yesterday Not For First Time	Underclaim	Correct	OK	OK
Did Not Read Yesterday	Underclaim	OK	Correct	Correct

Under four conditions out of twelve, there is complete agreement between observations and readership. These combinations are labelled "Correct", as for example, when a first time reader claims first time reading; or a non-readers 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 - A first time reader would not be missed.

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.

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

Before validation testing began, the method was successfully tested for workability including: video taped sessions, Belson type interrogation of the video-taped respondents, and actual field trials.

The validation program consisted of a total of five studies, all of which have been completed.

Studies one and two have been reported previously. They represent some of 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 even for the largest US publications, overclaiming is possible for the vast majority of the population, while underclaiming, of course, can only occur among readers...usually a small percent of the population.

The other three studies were conducted late last year. Study 3 measures overclaiming and underclaiming of young issues for in-home reading.

The final two studies measure a different form of overclaiming: that which occurs when non first time reading is claimed as first time reading. In study four the first reading was observed in the home while in study five the first reading was observed in a public place.

Validation Study 1 was designed to measure overclaiming or underclaiming for young or prepublication 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: 25% of the interviews were validated.

The observed respondents were engaged in natural reading in a setting of their own choosing. This study like the other four was a double-blind experiment. Neither respondents nor interviewers knew the purpose of the readership interviews. Individuals observing the public place readership were not involved in the readership interviews.

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.

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

**RESULTS: OVERCLAIMING
 Prepublication Issues**

	# Observed Non-Reading	# Claimed FTRY	Overclaim Percent
Weeklies	501	2	.4%
Tri-Weeklies	499	-	-
Total	1,000	2	.2%

The level of underclaiming for pre-publication issues was also low, 7%. The capture rate was 93%. This level of underclaiming is as low as any we have seen before.

RESULTS: UNDERCLAIMING
Prepublication Issues

	# Observed Reading	# Claimed FTRY	Capture Rate	Under- Claiming
Weeklies	77	72	94%	6%
Tri-Weeklies	79	73	92%	8%
Total	156	145	93%	7%

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 times and places other than those observed. Thus maximum overclaiming of aged issues is four-tenths of one percent.

RESULTS: OVERCLAIMING (?)
Postpublication Issues

	# Observed* Non-Reading	# Claimed FTRY	Maximum Overclaim Percent
Weeklies	1130	4	.4%
Tri-Weeklies	1124	6	.5%
Total	2254	10	.4%

* Could have read yesterday away from observed public place

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.

RESULTS: UNDERCLAIMING
Postpublication Issues

	# Observed Reading	# Claimed Yesterday Reading	Capture Rate %	Under- Claiming %
Weeklies	26	18	69%	31%
Tri-Weeklies	32	29	91%	9%
Total	58	47	81%	19%

The eight cases of underclaiming for the weeklies were examined in complete detail but no patterns emerged to explain them.

This high level of underclaiming for aged issues when a yesterday reading technique is involved was totally 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 had always resulted in substantially higher capture rates.

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 was conducted using an expanded list of magazines.

It was clear that if the problem of underclaiming for aged issues was confirmed, the gold standard method was invalid. However, a decision was made that 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 program could proceed.

Validation Study 2 was designed to provide a fuller 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 measure of 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:	25% of the interviews were validated.

Operationally, Study 2 was identical to study 1.

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. One covered the

four magazines in the initial study and the second covered the other four magazines.

As was anticipated for the initial experiment, underclaiming levels for aged issues were low. 221 observations were made of test issue reading and 211 of these yield yesterday readership for these issues. Thus, underclaiming was 5% yielding a capture rate of 95%.

This high capture rate level occurred across all demographic groups and across various levels and types of reading exposure.

AGED ISSUES: UNDERCLAIMING

	# Observed Reading	# Claimed Yesterday Reading	Capture Rate %	Under-Claiming %
Monthlies/ Tri-Weeklies	98	96	98%	2%
Weeklies	123	115	93%	7%
Total	221	211	95%	5%

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, of course, is a maximum as respondents could have read these issues yesterday when we were not observing them.

AGED ISSUES: OVERCLAIMING (?)

	# Observed Reading	Claimed FTRY	Maximum Overclaim %
Monthlies/ Tri-Weeklies	1,102	1	0.1%
Weeklies	1,077	4	0.4%
Total	2,179	5	0.2%

The third study measures overclaiming and underclaiming of young issues which are read in-home. The method used has been named the "Spouse-Spy Test" and is the same as that reported for ARF Certitude Study # 2 as reported at the New Orleans Symposium. Housewives were recruited to unobtrusively observe their husbands for up to a few days after a subscription copy of a target magazine arrived in the mail, as it normally does. These observations resulted in classifying a respondent as a reader or a non-reader for the interviews which were conducted a day after

observation ended. Observations covered up to 3 days after magazine arrival for the weeklies vs 5 days for the monthly magazine.

This study like the others was a double blind experiment as the husband did not know he was being observed and the interviewer did not know the results of the observation.

The study was conducted among subscribers located in a random sample of zip codes located in two major metropolitan areas. Recruiting of housewives was conducted by telephone.

A total of 219 respondents were observed and interviewed. 105 were observed readers of one of the target issues while 114 were observed non-readers.

Of the 114 observed non-readers, one spontaneously indicated readership of the target issue at a doctor's office on the day before the interview. This left 113 observed non-readers, none of whom claimed first-time yesterday readership.

Of the 105 observed readers, 99 or 94% properly claimed first time yesterday readership. Underclaiming, then was 6%.

Thus, the Gold Standard Method passed the over and underclaiming tests covering subscribers' in-home reading or not reading of new issues.

STUDY 3 SUMMARY OF RESULTS

	<u>Overclaiming</u>	<u>Underclaiming</u>
Yes	-	6
No	<u>100</u>	<u>94</u>
	100%	100%

The final two validation studies measured the extent to which overclaiming could occur with the Gold Standard Method because readers who read a magazine on more than one day may claim first time reading for the most recent reading day.

In study # 4, the initially observed reading occurred in the home. Observations were made using the spouse-spy methodology used in study # 3. In study # 5 the initially observed reading occurred in a public place with the observation made in accordance with the methodology used in studies # 1 and 2.

In both studies four and five, the second observed reading of a specific issue occurred in a waiting room. For this part of the experiment, respondents known to be earlier readers of an issue were recruited to participate in a central location Coke and Pepsi taste test.

Respondents came to the central locations one at a time. When they arrived they were informed there would be a 15 to 20 minute

wait. The receptionist gave them a glass of water and a cracker telling them this would freshen their palate for the taste test to come ... reinforcing the belief that the ultimate purpose of their presence was to evaluate two soft drinks.

The respondents were then seated next to a table containing a copy of the magazine they had been observed to read previously. There was also another magazine available but it was selected as being unlikely to be of interest to the respondent.

Reading of the target magazine was not at all forced. In fact, one in three respondents did not pick up this magazine and read it. These respondents, of course, were lost to the experiment.

Next day interviews were completed from among the balance of the respondents who were all qualified as having been observed to read a target issue on two occasions. They, of course, could have read the issue on more than two occasions.

The timing of the second observation was spaced so that for some respondents the second observation occurred within 2 weeks of the first, while for others the elapsed time was as much as 5 weeks. This range of time was built into the study so that if any appreciable level of overclaiming occurred we could determine if it was affected by the amount of elapsed time between the first and most recent reading event.

The base for study # 4 was 115 respondents who were first observed to read the target issue in their own home and subsequently in a waiting room. For study # 5 there were 160 respondents who were first observed to read at a public place and subsequently in a waiting room

In both experiments the rate of overclaiming was 1%.

OVERCLAIMING LEVELS

	<u>Study 4</u>	<u>Study 5</u>
Initial Observation	Home	Public Place
# Observed		
Reading Twice	115	160
# Claimed FTRY	1	2
% Overclaiming	1	1

Therefore, the Gold Standard method passed all five validation tests. Surely other tests could be devised covering such situations as in-home reading by non-subscribers or second time reading occurring many months after the initial reading.

However, such experiments would be inordinately costly and are not in our judgement necessary as the existing completed experiments address the basic issues such additional experiments might cover.

In summary, we went for the Gold and were satisfied with the results.