Paul H Chook
Ziff-Davis Publishing Company
New York, USA

2.2 The ARF comparability study

INTRODUCTION

During the past two years, I and several of my esteemed colleagues in the audience here have had a rather continuing involvement with the ARF Council which designed and supervised the Comparability Study.

While this involvement undoubtedly led to my being invited to prepare this paper, I am giving it not in any official ARF capacity but rather as an interested practitioner who has been both a producer and user of magazine audience research over the past 25 years. My purpose here, is to provide a factual report of what the ARF study set out to do and what it found, but at the same time I offer some observations about key issues which are drawn from other experience.

BACKGROUND

By way of background, throughout most of the 1970s, syndicated magazine research in the United States featured the work of two services - Simmons and TGI - each purporting to provide a measure of issue audience.

Simmons, in personal interviews, employed a version of the 'through the book' (TTB) method, using a skeletonised (12 item) copy of the issue as a memory aid. TGI, in self-administered interviews, employed a version of the 'recent reading' (RR) method, using the magazine name as the only memory aid.

When the two companies offering these studies were merged in the summer of 1978 to form the Simmons Market Research Bureau, SMRB announced that henceforth there would be a single research service. All interviews would be made in person and would cover close to 150 magazines but using a two-tiered system of measurement: weeklies, bi-weeklies and larger circulation monthlies would be measured by TTB; smaller circulation monthlies would be measured by RR.

This decision was apparently based in part upon SMRB's own historical analyses of Simmons and TGI audience numbers. Time does not permit reviewing these numbers here: in a nutshell, the audience levels produced by these services seesawed from one year to the next depending upon factors of survey execution.

To substantiate its position on the two-tiered system, SMRB offered at the conclusion of its new study to compare rates of audience change for those magazines measured by TTB in both years, with those for magazines shifted from the TTB to RR method. It was SMRB's expectation that the rates of audience change would be similar.

Research users and practitioners responded to this proposal with a number of concerns. Some felt that because of executional changes over the years, historical Simmons/TGI data comparisons could not be used to support a conclusion of comparability. Others pointed out that even if the historical data could be taken as supportive, they did not apply because of the changed conditions of readership measurement.

Changes in the recent reading method

Briefly, the SMRB version of RR was to be administered in a personal interview of modest length devoted essentially to the subject of magazine reading. The RR questioning would occur relatively early in the interview but would be done in two stages: respondents would first be asked to sort through a shuffled card deck containing reproductions of magazine logos to indicate which ones they think they'd looked into in the past six months. They would then be asked the recent reading question for each screened-in magazine.

Given all of these changes *personal* administration of the readership questioning, in a *shorter* interview, with magazine *logos* to aid recall in a *two-stage* questioning procedure, and with higher question and interview completion rates – one might reasonably expect to obtain *higher* readership numbers than those produced by the TGI version of RR.

Effects on TTB measurement

In the new mixed method interview the plan was to do the TTB screening using a masthead book for weeklies and mass monthlies first, then to do the RR screening and qualifying stages for smaller circulation monthlies, and then to finish off with the qualifying stage for the TTB screeners. In contrast, the traditional Simmons interview did not interrupt the flow of TTB for recent reading questioning.

Concerns were expressed that the shorter list of TTB titles, together with the changed interview flow could affect TTB audience levels.

Summarising industry concerns: the belief that historical Simmons/TGI data comparisons did not support the new plan; concerns about changes in administration of the TTB and RR methods; and the belief that a rate of change analysis would not be an adequate way of evaluating the new methodology; all led to the demand for a side-by-side *controlled experiment* in which the new and old methods could be simultaneously tested with all other factors held constant.

2.2 The ARF comparability study

THE SMRB STUDY (1978/79)

Meanwhile, SMRB moved ahead with their proposed study. Fieldwork began in the Fall 1978 and by Summer 1979 – well before the ARF study was completed – they reported the findings of their rate of change analysis and announced their decision to make adjustments to the data to be published in their Fall 1979 reports. Before turning to the ARF study, let us consider their rate of change findings and the effects of their adjustments.

There were 69 magazines common to the Simmons 1978 and SMRB 1979 studies: 36 weeklies and larger monthlies measured by TTB in both years, and 33 smaller monthlies shifted over to the RR method in the 1979 study.

For the weeklies and larger circulation monthlies, there was virtually no change in TTB reader per copy levels – a finding in accord with SMRB expectations. However, for the smaller circulation monthlies, reader per copy levels had increased on average by a margin of 90% – a finding not in accord with their expectations.

SMRB concluded that the new RR method led to overclaiming of readership, and decided to make a downward adjustment to the data. This was accomplished by eliminating those recent readers who reported reading fewer than two out of every four issues published. The remaining readers, designated QRR for qualified recent readers, came in at a reader per copy level 19% higher than when these magazines were measured by TTB. The adjustment procedure implicitly assumed that the entire disparity between TTB and RR levels was accounted for by those who claimed to be infrequent readers.

THE ARF STUDY

Let us now turn to the ARF study and see how its findings relate to the SMRB experience.

Design

The study grew out of appeals by the ANA, the AAAA, the MPA and by individual syndicated research users for an unequivocal test of methods comparability. At a cost approximating $450,000, it was funded on a tri-partite basis by over 80 advertisers, agencies, magazines and service organisations.

Briefly reviewing the design and operational features: it was done with a nationwide probability sample of people aged 18 and over during the period June through November 1979; it provided side-by-side comparisons of three audience methods during the same time frame; it was implemented by SMRB under ARF supervision and according to ARF specifications; and SMRB was selected to execute the field and data processing phases because of the organisation's experience and in order to maximise executional comparability with the regular SMRB study.

A total of 68 magazines was measured by three different audience methods. These included: 12 weeklies; 20 larger circulation monthlies; and 36 smaller circulation monthlies; all of which are measured in the regular SMRB study.

In addition, there were 56 dummy magazines – all smaller circulation monthlies – included to simulate the magazine list covered in the regular SMRB study.

A total of 4600 interviews was completed in three matched samples of 1000, 1800 and 1800 adults, respectively.

Sample 1 was an analogue of the traditional Simmons interview in which all magazines are covered by the TTB method. There were 68 magazines comparable to the normal Simmons list.

Samples 2 and 3 were analogues of the SMRB mixed methods interview. All told, 124 magazines were covered in these interviews, 34 by TTB and 90 by RR, closely comparable to SMRB practice.

The list of 68 test magazines was organised in two replicates of three groups each. The three groups A, B and C of the first replicate consisted of six weeklies, ten larger circulation monthlies and 18 smaller circulation monthlies; the three groups D, E and F of the second replicate were similarly constituted.

In Sample 1, all magazines in the six groups were measured by the TTB method. Hereafter, we will refer to this as TTB-T, for 'through the book traditional'. In Sample 2, magazines in the first replicate were again measured by TTB but in the context of a mixed methods interview. Hereafter, we will refer to this as TTB-M for 'through the book mixed'. Magazines in the second replicate were measured by RR.

Sample 3 provided the rotation by using RR to measure magazines in the first replicate and TTB-M to measure magazines in the second replicate.

Thus each of the 68 magazines was measured by each of the three methods. Further, when the two replicates are combined to provide average magazine comparisons, all persons in Samples 2 and 3 contribute to both the TTB-M and RR contrasts.

Concluding our discussion of the design, it is important to note three things: first, the study was designed strictly to assess *comparability* of three methods; it does not address the issue of *validity* – that is, which technique provides the *best* estimate of actual readership; second, the sample bases were established to permit reasonably dependable analyses of audience levels for *groupings* of magazines, but not necessarily for individual titles; third, calibration was not an initial goal of

2.2 The ARF comparability study

the study but rather an attempted remedy for non-comparability; this is important because calibration entails working with individual magazine ratings and would have benefited from larger sample bases.

Findings

Moving on to the findings, there are four matters to be addressed:

(**a**) do the methods in fact produce different results?
(**b**) if so, how different?
(**c**) what factors account for the differences?
(**d**) are the differences systematic enough to permit dependable calibration?

We compared audience levels for through the book mixed vs through the book traditional. On a base of all adults, there was virtually no difference between the two methods for weeklies.

The TTB-M method on average produced 8% higher scores than TTB-T for larger circulation monthlies and 12% higher scores than TTB-T for smaller circulation monthlies. Although not statistically significant, the patterns of difference for both magazine categories were consistent in most demographic groups.

It should be noted that in this tabulation and all that follow, monthly magazines are categorised as larger or smaller according to how they were measured by the 1979/80 SMRB, rather than according to the design layout which was predicated on the 1978/79 SMRB.

The differences between the mixed recent reading and through the book methods were dramatic – much more sizeable than contemplated at the survey start.

On average, RR produced 27% higher scores for weeklies, 80% higher scores for larger monthlies, and 96% higher scores for smaller circulation monthlies. All were statistically significant. The finding for smaller circulation monthlies was most cogent in that only these magazines were measured by RR in the regular SMRB study. The margin of difference was consistent with that found by SMRB in its rate of change analysis.

What factors account for these very substantial differences between the recent reading and through the book methods? The ARF study explored three types of factors: structural factors; reader factors; and, magazine factors.

By structural factors, we refer to the stages of questioning by which people are categorised as readers.

The first stage generates a screen-in rate; that is, the percentage of people who are classified as possible readers. TTB does this by having people inspect a masthead book; RR does it by means of a shuffle card deck.

The second stage generates a read-to-screen ratio; that is, the percentage of screened in readers who are classified as qualified readers. TTB qualifies by issue inspection; RR by a logo card cue.

A, the audience level is naturally the algebraic equivalent of the product of the screen-in rate and read-to-screen ratio.

Let us see to what extent screen-in rate differences and read-to-screen differences contribute to overall differences in audience estimates.

There were only minor differences in screen-in rates for the weeklies: 2% higher by the RR method. The differences were bigger both for larger and especially the smaller circulation monthlies - 12% and 19% respectively.

The differences in the read-to-screen ratios were much larger and also varied by magazine type. The RR read-to-screen ratio was 25% higher than the TTB read-to-screen ratio for weeklies and as much as 65% higher for the smaller circulation monthlies.

The differences at each stage combined to produce the overall RR vs TTB audience level differences in the following ways:

For all types of magazines, differences in the read-to-screen ratios were much larger than screen-in rate differences. Across all magazines, RR read-to-screen ratios exceeded TTB read-to-screen ratios by 51% – a margin five times as great as the screen-in rate difference of 10%.

We should note in passing that this finding contravenes what some US industry experts had strongly believed; namely, that interviewer and possibly interviewee concern about having to inspect a larger number of magazines substantially depressed screen-in rates for the TTB method and that this was the dominant reason why TTB audience levels were so much lower than RR levels. This clearly is not the case; while there is a definite difference at the screen-in stage for monthlies, the major factor is indisputably the read-to screen ratio.

Why should the read-to-screen ratio be lower for TTB than RR? Could it be *fatigue*? This seems unlikely, since once a person passes the screen and inspects the issue, there is no more burden to either interviewer or interviewee arising out of a positive rather than negative response to the readership question. Could it be *memory*? If so, it would suggest that the use of a 2 " × 4" logo card is a better stimulus to recall of recent reading than is a 12-item editorial item display to recall of issue reading. Possibly, but probably not likely. Could it be *issue age*? In my experience, and I am thinking particularly of the years I spent with the Politz organisation, this seemed to be a relatively unimportant issue. Our studies of reading days and of ad page exposures which enabled us to track the build-up of exposure day-by-day after issue appearance indicated that at the customary TTB issue ages, virtually 95% of the eventual issue audience had been accumulated.

2.2 The ARF comparability study

Moving on now to an analysis of reader factors, the one demographic characteristic that bore a clear-cut relationship to RR vs TTB differences was sex. The relationship is not apparent on an all-magazine basis, for which RR exceeded TTB audience levels by 67% for men and 65% for women. Nor was there any difference for the weeklies, most of which have dual audience appeal. There were, however, differences for larger and smaller circulation monthlies, with greater disparity for men among the former and for women among the latter.

This pattern suggested that the explanation might lie in editorial appeal, since the larger circulation monthlies have a heavier concentration of female-oriented magazines whereas the smaller circulation monthlies have a heavier concentration of male-oriented magazines.

This indeed appears to be the case. In a tabulation of minority sex reading - that is, male reading of female oriented magazines and female reading of male-oriented magazines – versus majority sex reading of these magazines, while RR exceeded TTB audience levels by an overall margin of 66%, the margin increased to 100% for minority sex readers and was even a bit higher for majority sex readers of gender-oriented magazines.

Another key discriminator was place of reading. Across all survey magazines, RR exceeded TTB audience levels by 41% for in-home readers and by 114% for out-of-home readers. The difference was directionally consistent for all magazine types, but more pronounced for the monthlies than for the weeklies.

Next we consider claimed frequency of issue reading.

RR audience levels were disproportionately higher among claimed infrequent than frequent readers. The differences were somewhat less pronounced for weeklies than for monthlies. For the smaller circulation monthlies in particular, the RR vs TTB difference among claimed readers of fewer than two out of four issues was three times as great as that among the two or three group (i.e. 271% vs 88%) and eight times as great as that among those claiming to read four out of four issues (ie, 271% vs 34%).

In the audience difference of each frequency of reader group according to the contribution of the screen-in rate and read-to-screen ratio, the RR screen-in ratios were higher for frequent and infrequent readers but similar to TTB screen-in rates for moderate readers. In contrast, the read-to-screen differences were systematically related to claimed frequency of reading - biggest among infrequent readers and smallest among frequent readers.

Interestingly, screen-in differences and read-to-screen differences contribute about equally to overall audience differences among frequent readers, whereas among infrequent readers, the read-to-screen differences are ten times the size of the screen-in differences.

Over 50 other factors were examined for association with RR vs TTB differences. Most were minimally related and some were confounded with other variables. None contributed appreciably to the variations previously identified for the factors of publication frequency, circulation size, claimed frequency of reading, place of reading and sex.

Calibration

Given these large and seemingly systematic differences between the RR and TTB-M methods, a considerable effort was made at calibration even though this was not a starting survey objective.

In concept, calibration is a two-directional question. It can be attempted either way: from recent reading to through the book or vice versa. The ARF effort focused on calibrating RR to TTB as a practical matter, since the genesis of the study was the compatibility of RR measurement with the Simmons organisation's heretofore exclusive use of TTB. Further, the calibration effort centred on the smaller circulation monthlies since these were the only ones measured by the RR method in SMRB's 1979 study.

As noted earlier, sample bases for the comparability study were established for analysing data for magazine groups. Since calibration involves working with individual magazines, we had perforce to employ data which were less precise than desirable for this purpose.

The ARF Council's consultant, Dr Ronald Gatty, developed two calibration approaches.

The first was a *regression* method which estimates a magazine's TTB-M level on the basis of two factors:

(a) its RR audience level.

(b) its RR male audience composition percentage.

The form of the regression equation was linear with all three variables expressed in natural logarithms. Separate equations were developed for males and females, using as input *data only for the smaller circulation monthlies* which were the subject of calibration.

The second approach was a ratio method which converts a magazine's RR screen-in level to an estimate of its TTB-M audience level. Separate factors were developed for each claimed frequency of reading group.

The basic form of the conversion factor, which we will call K, is:

$$K = \frac{\text{The number of TTB-M readers}}{\text{The number of RR screeners}}$$

Separate conversion factors were developed for men and women in each frequency of reading group. These average factors were then applied to the RR screen-in levels for individual magazines to develop estimates of their TTB reader levels.

2.2 The ARF comparability study

Note that the RR *reader* level itself plays no part in the adjustment procedure - only RR *screener* data are used.

Dr Gatty developed the factors using data for *all weeklies and monthlies* – and not just for the smaller circulation monthlies to be calibrated. This was done in order to bolster the sample base for each ratio.

As one would expect, the ratio of TTB readers to RR screeners increases with increased reading frequency. For both men and women, the factors range from a low of about 10% to a high of 70-odd%, but with some notable differences in the middle frequency groups.

The average implied factor for both sexes is virtually identical.

The fact that these ratios are based upon data for virtually all survey magazines is a bit troublesome in that the ratios for the magazines to be calibrated – the smaller monthlies – are expectedly different.

That is, since the smaller monthlies were previously shown to exhibit a greater disparity between TTB-M *readers* and RR *readers*, one would expect them also to exhibit a greater disparity between TTB *readers* and RR *screeners*.

This is in fact the case. For smaller monthlies, the ratio of TTB-M readers/RR screeners for men is 0.33 – about 10% lower than for the average magazine. For women, it is 0.27 – about 25% lower than for the average magazine.

In consequence, when average magazine factors are used for calibration, one would expect some inflation of estimated TTB-M levels – especially for women.

The ARF Council established four criteria for evaluating the adequacy of the calibration procedures. They were as follows:

(**a**) after calibration, the differences between audience levels produced by RR and TTB should be negligible on average.

(**b**) after calibration, the differences between audience levels produced by RR and TTB should be small for each magazine.

(**c**) after calibration, the demographic compositions produced by RR and TTB should be similar, magazine by magazine.

(**d**) after calibration, the audience duplication and turnover rates produced by RR and TTB should be similar, magazine by magazine.

It was not possible to make an evaluation based on the third and fourth criteria. For one thing, the data bases were not large enough to evaluate demographic compositions and duplication rates. For another, the comparability study did not measure audience turnover.

Consequently, the plan was to give a calibration procedure conditional acceptance if it met the first two criteria, leaving the evaluation on the third and fourth to subsequent testing with the SMRB data base.

As noted earlier, the evaluation was limited to the smaller circulation monthlies which SMRB had measured by the RR method.

Using the ARF data base, five sets of ratings were compared.

The first was RR unadjusted.

The next was a simple discount method which reduced each magazine's RR rating by the average RR to TTB-M difference. The simple discount reduces average levels but does not reduce the relative variation in magazine by magazine ratings. The simple discount thus serves as a stand-in for recent reading to assess the performance of alternative adjustments – QRR, regression, and the ratio method - in reducing relative variation.

The indexes next quoted for RR are somewhat different from those reported earlier because they are equally weighted rather than aggregative averages, and also because the analysis excluded three magazines with small TTB reader bases.

Relative to TTB, RR scores for men indexed at 197 and for women at 228. The simple discount method *ipso facto* reduced the indexes to 100 each. The SMRB QRR adjustment left differences of 22% and 26%. The regression method, derived from data for smaller circulation magazines only, *ipso facto* produced indexes very close to 100. The ratio method, derived from data for the broader list of magazines, produced indexes of 107 and 112. These differences of 7% and 12% are statistically significant at the 90% and 95% levels of confidence, respectively.

Regarding the absolute standard deviations of the individual magazine indexes: compared to the RR sigmas of 57 and 78, the simple discount sigmas of 29 and 34 reflect no real reduction: they are a necessary consequence of the reduction of RR scores by constant factors.

Against these benchmarks of 29 and 34, neither QRR nor the regression nor the ratio methods show any improvement in variance reduction. For women, the QRR and ratio methods actually resulted in larger relative variation than that present in the original RR set.

The regression and ratio methods were also evaluated using the SMRB survey data base. This analysis involved a comparison of 1978 TTB scores updated for circulation change with 1979 SMRB calibrated scores. The analysis covered 33 smaller circulation monthlies common to the two studies.

Consistent with the previous analysis, the regression method obtained close identity on average between RR and TTB. Also consistent with the previous analysis, the ratio method left differences of 8% and 18%, respectively, for men and women.

For the purpose of dealing with the second ARF

2.2 The ARF comparability study

criterion – that is, the margins of difference for individual magazines – we compare year-to-year shifts in rating levels with *signs ignored*.

Between 1977 and 1978, two normal years unaffected by changes in Simmons methodology or other extraneous factors, relative shifts for both male and female audiences were of the order of magnitude of 12%.

In contrast, as between 1978 Simmons and 1979 SMRB, the relative shifts were half again as big for regression calibrated audiences and twice as big for ratio calibrated audiences. In fact, the variations associated with the ratio method (which SMRB has now adopted) are as large as those ever experienced over the course of a year by the Simmons service.

Some examples of magazine to magazine variations found when the calibrations were applied to the SMRB data base: comparing two golfing publications, among men, both calibrations give *Golf Digest* a score about 20% lower than TTB while both give *Golf Magazine* a score about 50% higher than TTB. For three of the largest dual publications in the group, among men, the regression method produces ratings anywhere from 10% to 30% lower than TTB whereas the ratio method produces ratings from 10% to 20% higher.

Comparing five female interest magazines, *Mademoiselle* – the magazine with the largest index by the regression method – has the smallest by the ratio method, whereas *Apartment Life* – the magazine with the smallest index by the regression method gets a par index by the ratio method.

Again, for the three larger dual audience publications, the regression method produces adjusted scores lower than TTB for two out of three but the adjusted scores are higher for all three by the ratio method.

Against the full SMRB data base, there are large variations – appreciable enough to result in different buying decisions.

Based upon these and other analyses available in the published report, the following conclusions were drawn:
(a) in relation to criterion 1, that the differences be small on average, the regression method performed satisfactorily but ratio method did not.
(b) in relation to criterion 2, that the differences should be small magazine by magazine, neither calibration method performed better than the simple discount and both exhibited wide variations when applied to the SMRB data base.

Accordingly, it was concluded that the findings did not justify the acceptance of either calibration procedure as developed from the ARF study.

This, of course, does not imply that calibration is generally infeasible but simply that those developed were, in the Council's view, not sufficiently robust to assure that individual magazines would not be unfairly rewarded or penalised by them.

As a footnote, let me add that even if a statistically acceptable calibration procedure were at hand, there are still two further matters to consider. One, of course, is the problem of implementation: converting readers to non-readers or vice versa on the computer tape in a dependable non-prejudicial fashion. The other matter is to make periodic checks over time to certify that existing calibrations still work or to make appropriate modifications to them.

Closing comment

If there is one key fact that events of the past few years in the US make explicit, it is that there is in the abstract no such thing as a 'method'. The TTB method as nowadays promulgated by SMRB is not the same as the TTB method produced by the Politz organisation. The RR methods produced by SMRB and also by MRI are in no way comparable to that provided by TGI, and indeed despite their seeming similarities, there are some important variations in execution of the SMRB and MRI procedures.

All of this reminds us that we really dare not tamper incautiously with established procedures. *Methodological research should precede and be the basis for methodological change, and not the reverse.*

In the present instance, the ARF Comparability Study was of necessity undertaken as an after-the-fact evaluation of a *fait accompli*. Even so, the principal findings and their implications should help guide and shape future syndicated research efforts.

There is no doubt that, in the US and elsewhere, industry hunger for comparable data for long lists of magazines has led research suppliers to make compromises with methodology. That hunger has not abated but it is now accompanied by an explicit demand for dependable methodology as well.

I have often heard it said that advertising agencies do not require as much audience accuracy for planning print campaigns as do publishers for selling their respective magazines. This is undoubtedly true, and yet we have a growing consensus among agencies in the US that we can no longer afford compromised standards. In the interest of equity to the media, and out of responsibility to the advertisers who use them, standardised and tested methods must be employed. As recently as September of this past year, the AAAA made this position clear in their statement of 'Syndicated Magazine Research Specifications'.

Users really are willing to pay more for better quality. It is up to the suppliers to meet this demand.