BENCHMARKING: FROM THEORY TO PRACTICE

Jim Collins and Dan Mallet, Simmons

In 1995 Simmons introduced a major change in survey procedures, combining two separately fielded companion studies.

One study uses a Self-Administered Questionnaire (SAQ) to cover all media, products, and services. 20,000 respondents per year are selected by random digit telephone methods. Questionnaires are sent out and returned by mail.

The SAQ sample serves as the respondent database for delivery of all media, product, and service information in a single source, but does not itself serve to establish basic print audience levels. Instead, a separate Personal Interview (PI) study is used for this purpose. The Personal Interview study is devoted to print measurement only, using recent reading, and is based on a completely separate 20,000 respondents per year.

Benchmarking is the process used to impose Personal Interview audience levels on the Self-Administered respondent database.

At the time of the 1995 Readership Symposium in Berlin benchmarking had been applied to only one study release, and benchmarking was introduced essentially as a theoretical concept. Since then, Simmons has applied benchmarking in four additional semi-annual study releases.

Following a review of the benchmarking process, this paper will discuss our practical experience and results to date, in three areas.

First is basic workability. Does benchmarking do what it's explicitly supposed to, in the sense of matching audience levels where explicit control is exercised? Second, what happens where SAQ audiences are not explicitly controlled, such as fine demographic sub-groups and cells, and multi-title duplication? Finally, what effect does benchmarking have on the relationship between frequency and reading probabilities and in the use of frequency as a reader qualitative?

In answer to the first two questions, benchmarking has worked remarkably well. Mechanically, there have been few problems, despite our asking more in terms of explicit demographic controls than originally planned. In areas not explicitly controlled, the benchmarked SAQ is in very close agreement with the PI, and, if anything, the benchmarked SAQ appears to enjoy a small advantage in stability. The answer to the third question is complicated, though it is clear the SAQ frequency question does not have the same qualitative meaning as frequency questions in more traditional survey designs.

The Benchmarking Process

Print questioning in the SAQ includes a six month screen for all titles and frequency of reading for each title screened (out of 4 for magazines and Sunday newspapers, out of 5 for daily newspapers).

If the SAQ were the only study conducted, respondents would be assigned theoretical reading probabilities based on frequency of reading. Respondents answering 4 out of 4 would be assigned a reading probability of 1.00, 3 out of 4's would be assigned .75, and so on. Audience statistics such as average issue audience and two-issue reach (or turnover) would then be computed from these reading probabilities. This is standard treatment for self-administered, frequency-based readership studies.

Benchmarking extends the probability approach, with the express objective of assigning SAQ probabilities in a way that produces the same average issue audience and two-issue reach as the PI study.

The process is performed independently for each title. As a preliminary step, reading probabilities (i.e., read-to-screen ratios) at each frequency level are tabulated from the PI, along with the average issue audience and two-issue reach, as shown in Table 1.

Table 1: Example Personal Interview Data

		Average		Two
		Issue	Reading	Issue
	Population	Audience	Probability	Reach
	(000)	(000)		(000)
Non-Screens	165,704	-	-	-
Screen & Frequency:				
<1 of 4 Issues	3,537	444	.1255	832
1 of 4 Issues	8,507	2,543	.2989	4,326
2 of 4 Issues	4,991	2,485	4978	3,732
3 of 4 Issues	1,130	721	.6381	982
4 of 4 Issues	5,263	4,838	.9192	5,229
Total	189,132	11,030	.0583	15,101

Personal Interview reading probabilities are then assigned as initial reading probabilities to SAQ respondents with the same reported frequency level, as shown in Table 2.

Table 2: SAQ, Initial Reading Probabilities

			Average	Two
		Reading	Issue	Issue
	Population	Probability	Audience	Reach
	(000)		(000)	(000)
Non-Screens	158,387	-	-	-
Screen & Frequency:				
<1 of 4 Issues	7,508	.1255	942	1,767
1 of 4 Issues	9,161	.2989	2,739	4,658
2 of 4 Issues	5,586	.4978	2,781	4,177
3 of 4 Issues	2,282	.6381	1,456	1,983
4 of 4 Issues	6,208	.9192	5,706	6,168
Total	189,132	.0720	13,625	18,753
PI Target			11,030	15,101

This initial assignment of reading probabilities does not itself assure that SAQ audience values match those of the PI, because the SAQ and PI do not have the same distribution of respondents by frequency category. Probabilities are then adjusted, in two steps.

In the first step, SAQ probabilities are scaled or ratio adjusted to match the overall PI average issue audience. In this example, all probabilities would be multiplied by 11,030/13,625. As shown in Table 3, this matches the SAQ average issue audience to the PI, but does not yet force SAQ-PI agreement on two-issue reach.

Table 3: SAQ, Scaled to Average Issue Audience

			Average	Two
		Reading	Issue	Issue
	Population	Probability	Audience	Reach
	(000)		(000)	(000)
Non-Screens	158,387	-	-	-
Screen & Frequency:				
<1 of 4 Issues	7,508	.1016	763	1,448
1 of 4 Issues	9,161	.2420	2,217	3,898
2 of 4 Issues	5,586	.4030	2,252	3,595
3 of 4 Issues	2,282	.5166	1,179	1,749
4 of 4 Issues	6,208	.7442	4,620	5,802
Total Screens	30,745	.3588	11,030	16,492
Total	189,132	.0583	11,030	16,492
PI Target			11,030	15,101

Retaining the target PI average issue audience requires maintaining the current average reading probability for all SAQ screens (.3588 in this example). In the second benchmarking step, therefore, probabilities are scaled closer to or farther from the average SAQ reading probability. This matches two-issue reach to the PI without changing average issue audience. Table 4 shows final results in this example.

Table 4: SAQ, Fully Benchmarked

	Population	Reading Probability	Average Issue Audience	Two Issue Reach
	(000)		(000)	(000)
Non-Screens	158,387	-	-	-
Screen & Frequency:				
<1 of 4 Issues	7,508	.0078	58	116
1 of 4 Issues	9,161	.1994	1,827	3,289
2 of 4 Issues	5,586	.4192	2,341	3,701
3 of 4 Issues	2,282	.5741	1,310	1,868
4 of 4 Issues	6,208	.8849	5,493	6,126
Total Screens	30,745	.3588	11,030	15,101
Total	189,132	.0583	11,030	15,101
PI Target			11,030	15,101

The above applies to a single instance of benchmarking for one group of respondents (such as total adults in the example). In practice, the process is applied separately for Males and Females, and within each sex we want to match the SAQ to the PI on a number of different demographic characteristics.

To accomplish this, benchmarking is performed iteratively on a marginal basis, in a fashion much like the sample balancing weighting algorithm. PI read-to-screen ratios are used as initial reading probabilities for SAQ respondents, as described above.

At each demographic characteristic, probabilities in each class are benchmarked to the PI average issue audience and two-issue reach for the class, using the scaling and variance adjustments described. Successive demographics are dealt with in turn, and the entire process is repeated until the SAQ average issue and two-issue audiences closely match the PI on all demographic characteristics simultaneously.

A total of 13 demographic variables are included:

Age (6)
Education (4)
Occupation (5)
Individual Employment Income (8)
Race (3)
Hispanic Origin (2)
Marital Status (3)
Household Size (4)
Child Status (3)
Home Ownership (2)
Household Income (9)
Census Region (4)
County Size (4)

The numbers in parentheses indicate the number of classes used for the variable. Adjacent classes are collapsed when the number of SAQ screens in a class is less than 30.

The result of controlling so many demographics simultaneously is that virtually every respondent is assigned a unique reading probability.

Following the order of demographics shown above, in the first iteration the PI reading probabilities initially assigned to SAQ respondents would be adjusted to match PI average issue audience and two-issue reach separately in each of the six age categories. Coming into one class on Education, respondents in any one frequency group could have six different probabilities, depending on their age. At Occupation, respondents in any one frequency group could have 24 different probabilities, depending jointly on their Age and Education, and so on.

Workability

The benchmarking goals outlined above are quite ambitious -- matching average issue audience and two-issue reach for up to 57 sub-groups on 13 demographic characteristics, separately by sex.

The potential impediments to achieving these goals are combinations of PI audience targets and SAQ screen-ins that make it impossible to meet all targets simultaneously. There are several ways such conflicts might arise:

1. It may be impossible to match the PI average issue audience, either in total or in demographic sub-group (e.g., SAQ screens are lower than the PI average issue audience).

2. It may be possible to match average issue audience but not two-issue reach, again in total or in some demographic sub-group (e.g., SAQ screens are higher than PI average issue audience, but much lower than PI screens).

3. There may be a conflict between average issue audience in one demographic group and two-issue reach in another (hard to characterize simply).

Where such conflicts exist, average issue audience is given priority over two-issue reach.

Across the 230+ measured titles in the most recent survey period, there were no titles where a total sex average issue audience could not be achieved. There were about a dozen titles in each sex where the overall two-issue reach match was not achieved, usually because of a conflict with average issue audience in a demographic sub-group.

At the finer level, there were about 30 titles in each sex where the PI average issue audience target could not be achieved for at least one demographic sub-group, and about the same number where a two-issue audience could not be matched for at least one demographic sub-group.

The number of apparent conflicts at the sub-group level is largely driven by the number of demographics and classes considered. Conflicts arise when SAQ screens are materially lower than PI screens. While SAQ screens run a bit higher than the PI overall, there is sampling variability on both sides. Given the large number (and small size) of demographic classes considered, it is inevitable that there will be instances where SAQ screens are lower than the PI, especially for small magazines in small demographic groups.

On the whole, benchmarking has worked remarkably well. In a small number of cases there is a minor conflict that affects a few demographic sub-groups. For most titles, no conflict arises and close SAQ-PI agreement is achieved on all demographic groups controlled.

Non-Controlled Demographics

The list of explicit controls in benchmarking includes every demographic in the PI. Even so, explicit controls are not imposed on every class of every demographic characteristic, nor on demographics in combination.

The benchmarking approach rests on the assumption that matching the SAQ and PI on selected characteristics achieves reasonably close agreement on all characteristics, an assumption that has been confirmed by our experience so far.

As one example, Table 5 shows PI and SAQ Spring '97 Female audience ratings for Cosmopolitan, for nine demographic cells defined by age (18-34, 35-54, and 55+) and household income (<$30K, $30K-$60K, and $60K+). Age and income were each controlled marginally, so the SAQ and PI are in exact agreement on each row and column total. At issue is how the SAQ and PI compare in individual cells.

Table 5

Spring '97, Females 18+, Cosmopolitan Audience (as Rating %'s)

	Age 18-34		Age 35-54		Age 55+	
	PI	SAQ	PI	SAQ	PI	SAQ
HHI <$30K	22.8	22.7	11.5	11.3	3.5	3.6
HHI $30K-$60K	22.5	23.1	11.2	10.9	5.6	5.4
HHI $60K+	30.7	30.0	12.5	12.9	6.9	6.9

Looking at the PI, it comes as no surprise that the dominant dimension is age, with a strong younger skew. Cosmopolitan's audience is also upscale, especially within the youngest group. The SAQ mirrors this quite closely.

Table 6 adds corresponding audience data from Spring '96, a year earlier with completely different respondent sets for both the PI and SAQ.

Table 6

Spring '97 vs. Spring '96, Females 18+, Cosmopolitan Audience (as Rating %'s)

		Age 18-34		Age 35-54		Age 55+	
		PI	SAQ	PI	SAQ	PI	SAQ
HHI <$30K	S '97	22.8	22.7	11.5	11.3	3.5	3.6
	S '96	19.8	21.1	15.1	12.8	3.4	3.5
HHI $30K-$60K	S '97	22.5	23.1	11.2	10.9	5.6	5.4
	S '96	23.0	22.5	12.7	12.6	4.6	5.6
HHI $60K+	S '97	30.7	30.0	12.5	12.9	6.9	6.9
	S '96	29.9	28.0	14.6	16.6	8.6	5.1

There's a lot of information in this table, with horizontal PI-SAQ comparisons and vertical year-to-year comparisons in each cell, and it can take some time to read it.

What ultimately emerges from staring at many tables like this is:

- The basic "story" for the publication is the same, whichever of the four sets of numbers is examined;
- Year to year differences in the PI are generally larger than within year PI vs. SAQ differences;
- The SAQ is slightly more stable than the PI.

The small SAQ stability advantage is a consequence of the probability approach. All SAQ screens are counted towards average issue audience to a fractional degree, depending on reading probability. In the PI, in contrast, screens are counted towards audience on an all or nothing basis, depending on their recent reading answer.

The SAQ stability advantage is small, but shows up over and over again in all kinds of SAQ vs. PI comparisons. It has become something of a game in Simmons' offices, for example, to hide data sources and try to guess them from inspection of the numbers. While the SAQ and PI are always in reasonably close agreement, the SAQ is usually identified as the more reasonable, believable set.

That said, we continue to look for opportunities to improve the choice of demographic characteristics used in benchmarking. It is possible, for example, that benchmarking on fewer demographic characteristics and/or classes might actually improve SAQ-PI agreement, while taking more advantage of the stability benefit inherent in a probability approach. Selected demographic combinations might also be worth considering. With three years of survey data now available, we plan a comprehensive examination of these choices in 1998.

Audience Duplication

As already described, benchmarking is applied to each title separately, with no explicit attempt at controlling audience duplication between titles. Instead, duplication is the result of between title overlap in native SAQ screen and frequency answers, combined with benchmarked reading probabilities.

While there is some anecdotal evidence that the benchmarked SAQ somewhat understates duplication for some high duplication title pairs, this has not been borne out in the systematic examinations made to date.

After the first SAQ release (Fall '95), total adult audience duplication was computed for each of the 25,000+ magazine pairs. Average duplication in the SAQ across all title pairs was within 1% of the average in the PI. While this comparison does not directly address the target populations or title pairs that might be of particular interest to a specific user, it demonstrates that benchmarked duplications are at least reasonable.

More recently, in connection with EMRC-related committee efforts, SAQ and PI duplication were considered from a different perspective. The purpose was to assess duplication in the context of schedule reach/frequency, using a set of target populations and test schedules selected by the committee.

Target populations ranged in size from 30% of total adults (Women 25-54) to 1% (Adults 35+ with IEI $100K+). Test schedules included as few as five and as many as 27 magazines, with 100 to 3000 GRP.

Schedules were run against the SAQ and the PI for both Spring '96 and Fall '96. The reach/frequency model was MetherPlus, which is standard in Simmons' CHOICES system.

Across all schedules, the average SAQ vs. PI index on schedule reach was 100 -- reach for the average schedule reach in the SAQ was exactly the same as in the PI.

Individual schedules, of course, showed differences, averaging about 5% relative. Consistent with the discussion in the previous section, schedule results from Spring to Fall '96 were somewhat more stable in the SAQ than in the PI.
Future testing in this area will probably include schedule optimization, to determine whether the SAQ leads to different schedule choices than would the PI.

It's important to note that the observed SAQ-PI reach differences were smaller than the typical differences one might get from using different reach/frequency models on a single data set (i.e., SAQ-PI reach differences are within the bounds of modelling error).

Reading Probabilities and Frequency of Reading

In the standard recent reading questioning sequence (such as in the PI, for example), there is a integral connection between frequency answers and audience. This connection is not direct, in the sense that audience estimates are based on recent reading answers rather than frequency, but certain logical connections or boundaries are imposed. There cannot be more 4 out of 4's than readers, for example, and 4 out of 4's typically account for the lion's share of audience.

This sort of connection is not imposed on the benchmarked SAQ. Frequency answers come from SAQ respondents via a self-administered questioning instrument, audience comes from the completely separate sample of respondents via a personal interview.

Treated simply as separate studies, comparing the SAQ 4 out 4's to the PI average issue audience is not likely to lead to anything useful.

After benchmarking on a long list of demographics, the connection between frequency and reading probability is even less direct.

For any one title, there will be some demographic sub-groups where the SAQ screen-in rate is much higher than the PI average issue audience. Meeting the PI average issue audience in these sub-groups requires assigning relatively small probabilities to all SAQ respondents, even 4 out of 4's. At the same time, there will be demographic sub-groups where the SAQ screen-in rate is only slightly higher than the PI audience, which means all SAQ respondents must be assigned relatively large probabilities, even <1 and 1 out of 4's. As a result, the reading probability assigned to a particular high frequency respondent can often be lower than the reading probability assigned to a low frequency respondent in a different demographic segment, making frequency even less useful as a reader qualitative.

We continue to look for an acceptable frequency-like reader qualitative. Meantime, the SAQ frequency question must be viewed primarily as scaffolding for benchmarking, and really cannot be used as a reader qualitative in the traditional way.

Conclusions

As mentioned at the beginning, benchmarking has worked remarkably well. We are able to exercise benchmarking control over a far longer list of demographics than had originally been in mind, and the overall process lead to close SAQ-PI agreement even on characteristics not controlled, such as demographic combinations and title duplication. What was a theoretical idea only three years ago has now become a routine procedure.

Going forward, we seek to further document benchmarking performance and to make refinements in the handling of rare conflict situations and in fine-tuning the selection of demographic characteristics to be controlled.

Benchmarking is still a relatively new and evolving procedure. We will be sharing news of continued development at future symposiums.