

ESOMAR

**Worldwide Electronic and Broadcast
Audience Research Symposium**

SHANGHAI

4th-9th June 2006

Data Integration For The Multi-Media Planning Database

As described above, the TouchPoints survey itself provides a single source survey of multi-media consumption. In some situations like radio, the data is very similar to the media currency, in this case half-hour viewing. However, it does not have the same level of detail or granularity and for some media like posters, there is only a surrogate variable (albeit a powerful one), in this case time spent travelling by mode of transport. Therefore in order to provide a fully functional cross-media planning database it is necessary to integrate the various trading currency surveys.

One possibility is to use a mixture of ascription and calibration to extend the TouchPoints diary data to provide greater channel or title breakdowns and to force consistency with the currency data. This is indeed more appropriate for one or two of the media, but generally there are too many dimensions to control and there is a danger of losing the richness of the currency data. Therefore, in the main, we have opted for respondent level data fusion, using the TouchPoints data to provide valuable linking information between the currency surveys, with a final calibration stage to force consistency with the currency surveys. A key principle is that the TouchPoints media data is over-written by the fused currency data to avoid conflict in the integrated database.

As a final step in this process, Kantar Media Research's Target Group Index survey has also been fused in to provide a range of product usage and ownership planning groups.

A.1 The Hub Survey

The content of the hub survey has been carefully constructed to provide strong media linking variables so that important currency cross media interactions are not lost in the data integration process. For example, in addition to the time spent reading data provided by the diary, a questionnaire was used to record the familiar National Readership Survey style frequency and average issue readership information.

An important step in the integration process is the re-engineering of the TouchPoints sample to provide an appropriate hub or host survey onto which the media currencies are fused. As a minimum, the hub survey needs to be weighted to universe profiles derived from a large random probability sample. However, the TouchPoints sample is small (c. 5000) compared to the media currencies (e.g. National Readership Survey c. 36,000); a standard fusion would use only a small proportion of the currency survey sample and its effectiveness would be severely reduced.

The solution was to fuse the TouchPoints survey onto the BARB (UK Broadcaster's Audience Research Board) Establishment Survey which generates a sample of 50,000 adults in six months of fieldwork. This does not confer any precedence to television because the only interest is in the demographic and geographical data. This kind of fusion is equivalent to replicating TouchPoints respondents (each TouchPoints respondent is fused onto around 10 Establishment Survey respondents) but in a constrained way which converts the TouchPoints survey into a representative sample (just like weighting).

This kind of weighting/replication/fusion is legitimate if we always remember how the hub survey was created in subsequent currency fusions. If the fusion used all the demographic information available on TouchPoints as hooks **and** we found 100% success in matching these hooks, then none of the original TouchPoints information could have been lost. The TouchPoints media data will have been attached to the correct Establishment Survey respondents. Of course in practice the demographic matching is less than perfect but not significantly so. More importantly, because we are not trying to cross-correlate TouchPoints media data with a measurement exclusive to the Establishment Survey, considerations of regression to the mean do not apply.

We are not pretending that this increases the sample size of the TouchPoints Survey when analysed in its own right. Nor are we pretending that it will increase the individual media currency effective sample sizes. The effective sample sizes for measurement of media interactions would still be limited by the smaller of the two currencies and further reduced towards the original size of the TouchPoints survey. The point is that after fusion to the hub survey, when individual media are analysed the original currency sample size will be preserved.

The product of this re-engineering exercise is a large, high quality random sample with demographics and TouchPoints single source media data. This is now a suitable hub (or host) respondent level survey into which each of the media currencies can be fused or integrated.

A.2 The Data Integration Subjects

The integrated database provides a planning tool across nine media groups plus TGI. In some cases respondent level data fusion has been used to integrate the media currency. But not all of the currency databases provide a respondent level survey suitable for data fusion and in these cases alternative data integration processes were used. The data integration subjects are as follows:

Television	- Fusion from BARB
Magazines and Press	- Fusion from NRS
Radio	- Fusion from RAJAR
Regional Press	- Profile matching from JICREG
Posters	- Calibration from POSTAR
Cinema	- Calibration from CAA admissions
Internet	- TouchPoints
Direct Mail	- TouchPoints
SMS	- TouchPoints
Product Usage	- Fusion from TGI

An indication of the integration technique and the currency data source is given in the above list. Examples will be discussed below. In some cases the media does not have currency data, then the source is the TouchPoints survey itself.

A.3 Respondent Level Data Fusion

The currencies for television, radio, magazines and press have been integrated using respondent level data fusion.

We can (and have) fused the media currency surveys using only standard demographic linkages or hooks. This is a reasonable technique if nothing else is available because demographics are powerful discriminators of behaviour for all media, but it leaves open the following issues:

- (i) The relevance and discriminatory power of the demographic hooks can only be assessed in terms of each media separately. We really need to understand how the hooks relate to interactions between different media. **The TouchPoints survey provides this information.**
- (ii) There is a concern that demographic fusion won't be so good for cross-relating media consumption which may be more related to lifestyle and attitudes than other product usage type fusion subjects. There is evidence from previous evaluations of KMR's Target Group Ratings product that fusion does not perform so well for lifestyle led products like cosmetics, showing above average regression to the mean. **The TouchPoints survey provides media based linking variables. For example, if there is a respondent in the TouchPoints hub survey who is a heavy TV viewer but light radio listener, then the fusion process will match them separately with a heavy TV viewing BARB panel member and a light radio listening RAJAR diary respondent.**
- (iii) There is very limited opportunity for validation. **The TouchPoints survey provides this validation opportunity.**

A.4 Fusion Hooks – Media Imperatives

The hooks available for each media fusion comprise demographics, geographics and a set of media imperatives.

A media imperative is a summary of each respondent's consumption of the media to be fused. To be used as a hook, we must be able to construct the media imperative in both the hub survey and the currency survey to be fused, and be reasonably confident that they are measuring the same thing. For example, it is possible to calculate hours of viewing by time segment, by day of week, by channel group in both the TouchPoints hub and the BARB panel.

These patterns of viewing must be summarised to form a usable set of fusion hooks. In order to avoid subjectivity in this process, a principle component analysis was used to construct the media imperatives. A principle component is a linear combination (like a regression model) of hours of viewing by time segment, day of week and channel group which maximises the diversity between individuals. A relatively small number of principle components explain the majority of the systematic variation between individuals.

The principle components were constructed in the TouchPoints hub survey giving a functional model. Then given a particular respondent's hours of viewing by time segment, day of week and channel, the value of the principle component can be calculated for each TouchPoints respondent and each BARB panel member, using the same functional model. This process was controlled to allow for differences in overall levels and variation in viewing levels between the hub surveys and the BARB panel.

A separate set of media imperative hooks was constructed for each currency fusion. The radio imperatives are based upon hours of viewing by time segment, day of week and channel. The magazine/press imperatives are based upon claimed typical frequency of readership by publication group.

A.5 Fusion Hooks – Importance Weights

The principle of the data fusion process is to find a respondent in the media currency (donor) survey who has the same demographic and media imperative profile as a particular respondent in the TouchPoints hub (recipient) survey. This donor's media currency data is then assigned to the TouchPoints recipient and replaces their TouchPoints media data.

A large number of demographic and media imperative hooks (about a dozen of each) were used in the matching process. Inevitably it is not possible to find exact matches across all hooks. Where compromises have to be made it is necessary to give precedence to the more important hooks. Therefore we need to quantify their relative importance or discriminatory power through analysis of variance.

A key feature of this particular fusion exercise is that because we have the single source hub survey, we can evaluate the hooks in terms of the true object of the fusion, i.e. volumes and patterns of consumption across all media as measured by the TouchPoints half-hour diary. In this respect, the importance of a TV viewing based media imperative is tempered by its relative inability to explain variations in consumption of all the other media. Demographic hooks have a chance to gain their rightful place in the hierarchy.

Obviously there is a separate set of hooks for each media fusion and therefore a separate set of importance weights. For each fusion a multivariate analysis of variance techniques has been used to consolidate the patterns of consumption across all media to construct a single importance weight for each fusion hook. Since this is a media planning application, the input patterns of consumption were weighted by media to reflect share of advertising expenditure.

A.6 The Fusion Process

RSMB's data fusion algorithm uses a variation of Mahalanobis' Distances to quantify the similarity between recipients and potential donors. This allows for correlations and differences in scale between the hooks. There is also a cohesive piece of statistical theory which justifies the formula used to incorporate the importance weights into the distance measurement.

The analyses needed to calculate the importance weights are extensive and the computation of Mahalanobis' Distance is intensive. Some fusion practitioners have argued that this sophistication is unnecessary. This may be true if there are only a few demographic hooks but an equivalent to Mahalanobis' Distance is required when there are many, as in this case.

The routine for pairing donors with recipients makes a trade off between the closeness of their hook profiles (as measured by Mahalanobis' Distance) and donor frequency distribution (the number of times each potential donor is used). The greater the control of the donor frequency distribution, the more likely we are to preserve the media currencies in the integrated database.

At the end of the data fusion process, the integrated database is a large representative,

sample of the population. The information available for each respondent is:

- Demographic, geographic and geodemographic classifications
- A full Target Group Index product usage and ownership record
- A media usage record from each of the integrated media currencies which reflects the levels and patterns of that respondent's media consumption as measured by TouchPoints.

A.7 Reach and Frequency Planning

The primary application of the integrated database is multi-media schedule reach and frequency analysis. It is recognised that the database will also be used for non-commercial applications, but the reach and frequency application is discussed here for illustration of the thought processes involved.

For TV, in theory it would be possible to use the BARB panel's long term viewing records so that actual contacts with a TV schedule of commercial spots could be counted for each individual. However, all other media currencies have only short-term measurements of exposure and depend upon probability expansions to estimate longer-term contacts with a schedule. Further, each media currency has a different probability model (the National Readership Survey has several) for the expansion. This partly reflects the fact that the inputs to the models are different for each medium. For example, the NRS readership data is recency and frequency whilst the RAJAR radio data is from a one week quarter hour diary. A common denominator is required which embraces all media. The solution we adopted is personal probabilities. Not only do we believe this to be the best statistical solution, it is also probably the only practical way of getting systems into the market place.

If a person does or doesn't make contact with a specific media event (e.g. a TV spot, yesterday's issue of a newspaper, a radio station in a particular quarter hour) then their personal probability is 1 or 0. However, this is not indicative of that person's probability of making contact with the equivalent event on another day or in another week. As a basis for estimating that person's long-term contacts it is useless because it doesn't allow that they might change from 0 to 1 in the course of a schedule. What we need is their underlying probability of making contact with each media event, a number between 0 and 1. Then if that media event is repeated a number of times, we can use a Binomial expansion to estimate a particular person's probability of making 0, 1, 2, 3,contacts. These respondent level frequency distributions are then aggregated to form the full sample reach and frequency analysis. The bureaux who are providing the systems must be congratulated on their ability to perform such a computationally intensive task so efficiently.

The decision to use personal probabilities means that the database is transparent to the bureaux, guaranteeing a level of consistency in the market place. The following notes outline the methodology used to calculate personal probabilities for each media and the approach to integration for those media currencies which do not lend themselves to personal probabilities.

A.7.1 Television

Because long-run BARB meter panel data is available, personal probabilities were calculated directly from the data as a 12 week average by channel, day of week and time segment. Calibration is achieved by scaling probabilities to BARB published 4 week average ratings to tidy up distortions in the fusion process and to incorporate guest viewing and time shift data.

A.7.2 Magazines and Press

Because the NRS does not provide long-run readership data, a segmentation analysis was constructed to split the sample into homogeneous groups. There is an independent segmentation for each title. In addition to demographics, the vital segmentation factor is the claimed normal frequency – this preserves the duplications between titles. Within each homogenous group, every individual is assigned the same probability of reading – this is numerically equal to the Average Issue Readership of the group. Calibration was achieved by scaling probabilities to NRS published AIRs.

A.7.3 Radio

This falls between BARB and NRS because the RAJAR survey is based upon a seven day quarter hour diary. This segmentation is based upon demographics, claimed “ever listened” information and the radio media imperatives. Within each homogenous group and each station, day of week and time segment, every individual is assigned the same probability of listening – this is numerically equal to the average rating. Calibration was again achieved by scaling probabilities to RAJAR published ratings.

A.7.4 Regional Press

The JICREG data is not available as a respondent level dataset. For each of a number of demographic groups, the zero frequency and the Average Issue Readership are reported in aggregate for each title. Data integration was achieved directly by:

- (i) Identifying the hub survey respondents within each title’s circulation area, within each JICREG demographic group.
- (ii) Randomly selecting a proportion of these to be in the zero frequency group.
- (iii) Assigning the JICREG Average Issue Readership to every one of the remaining respondent in the demographic group.

The probability of reading is either zero or numerically equal to the AIR. These data are calibrated by definition.

A.7.5 Posters

The POSTAR respondent level travel survey is used to construct a functional model to predict reach and frequency for a particular schedule from site classification data. Functional models are inconsistent with the personal probability approach in the integrated database, so there is little value in conducting a respondent level fusion.

Instead, the TouchPoints hub survey provides time spent travelling by mode of transport for each respondent. The planning system specification only requires input of GRPs by panel size, by region. At this level, it can be assumed that time spent travelling is directly correlated with numbers of panels passed. The POSTAR site classification data provides total GRPs by region, demographic, mode of transport and size of panel. Therefore the TouchPoints time spent travelling data can be calibrated to provide total GRPs by mode of transport and size of panel for each respondent in the integrated database. These data are already calibrated to the currency by definition.

In this state the numbers are inconsistent with the personal probability approach. The problem is that we don’t know which nor how many panels are “available” to each individual. Our solution is to use the Poisson probability distribution to create the schedule frequency distribution. The Poisson rate parameter is derived by scaling the respondent level GRPs to aggregate schedule GRPs.

A.7.6 Cinema

The TouchPoints questionnaire provides recency and frequency information on

cinema visits. These can be combined to calculate the underlying rate at which each respondent visits cinema. These data are calibrated to CAA admissions data.

As with posters, we don't know which nor how many cinemas are "available" to each individual. Again the Poisson probability distribution is used to create the schedule frequency distribution. The Poisson rate parameter is derived by scaling the respondent level rate of cinema visits to aggregate schedule admissions (GRPs).

A.7.8 Internet, Direct Mail, SMS

Currency data for these media are all constructed from the TouchPoints survey itself. The relevant questions are:

- Internet recency and frequency, ISP and type of web-site.
- Daily counts of numbers of pieces of addressed and non-addressed mail, telemarketing calls and SMS text messages.