# MULTIBASING℠
# DATA INTEGRATION WITHOUT REGRESSION TO THE MEAN

**Peter Walsh, Telmar Group/Harris Media Systems**

## Introduction

It is a post-modern world. Markets and media are fragmenting into lifestyle and psychographic niches with fewer distinguishing demographic characteristics. As this happens, the traditional practice of defining advertising target groups in terms of demographics becomes increasingly untenable.

Fragmentation also means of course that it is harder to achieve advertising goals. So, advertisers and their agencies must look more closely for synergies and efficiencies not only within campaigns employing a more diverse media mix, but also across different marketing programmes for the brand. And this demands a sophisticated level of integration in media planning.

Integrated planning requires 'currency' audience levels for separately measured media to be available for marketing target groups – i.e. users of products, services and brands, people displaying particular behaviours, lifestyles and attitudes, and so on. So, how can accurate ratings for TV and other media be obtained for product usage groups and so on when the audience measurement systems (apart from Print) capture little more than demographics?

Unless the planner is willing to work with single-source data in which audience levels typically differ significantly from the 'currencies', data integration is the only solution. The traditional methodology is fusion of respondent-level survey records.

However, the irony is that fusion relies mostly on demographics – the very limitation we are trying to overcome!

The realization that led to the idea of MultiBasing was that because demographics account for so little variance in so many product categories, respondent-level fusion *closes off* the opportunity to discover other statistical relationships in the data.

The idea was that it would be better to base a methodology on 'clusters' of respondents so as to retain the ability to identify relationships that arise more from people's *interests* than from their demographics. It is perfectly obvious that within any given demographic group, people who are interested in (say) gardening are more likely to buy gardening products *and* more likely to watch gardening programs on TV than other members of the same demographic group.

MultiBasing takes advantage of the fact that major readership surveys include questions about TV and other media in sufficient detail to enable those kinds of relationships to be identified and preserved when true audience levels are brought in from the respective 'currency' surveys. And with its capability of integrating multiple surveys, MultiBasing provides a platform for truly integrated media planning.

The methodology is described in this paper. After some terminology, the paper provides an outline of the multibase set-up and calculations involved in determining TV ratings for product usage groups. This is followed by a description of 'multi-dependent tree analysis' (MDTA), an analytical technique that is used to create the clusters.

Given that they gather a lot of data on other media, today's readership surveys are seen as forerunners of specially designed 'linkage studies'. The paper closes with some comments on how such surveys can become the foundation stones of the overall media research architecture of the future.

## 1. Terminology

### Linkage Study

*A survey that includes target group variables (e.g. product usage questions) and 'surrogate' media measures (e.g. questions on TV viewing)*. By including both types of variables, the linkage study makes it possible to detect non-demographically-based relationships between target groups and media vehicles. The linkage study will usually be a readership survey – e.g. MRI (the source of the data used here).

### Surrogates

*Surrogate media measures are 'shorthand' questions on the linkage study about TV viewing and other media exposure.* A simple frequency question such as "In a typical week, how often do you watch each of the following cable channels?" may be adequate for capturing relationships between cable channel audiences and marketing target groups, as would similar questions about networks, dayparts and programs.

**Linkage Cells**

*Clusters of survey respondents defined in terms of demographic and other variables that are common to the two or more surveys in a multibase*. Respondents in the same linkage cell have identical characteristics. The cells are isolated by MDTA in such a way that, across a wide variety of product usage variables and media measures, within-cell variance is minimized and between-cell variance is maximized.

**Multibase**

*A special database set up for MultiBasing*. It can be visualized as a large table with linkage cells as columns and media vehicles (audiences and duplications) as rows, filled by cross-tabulating the 'currency' audience database(s). Several media may be included (e.g. TV, Radio, Internet).

## 2. Multibase Set-up and Calculations

The example in this paper involves a very simple multibase structure that has been specially set up to illustrate the principles. Only 32 linkage cells were created so that it would be possible in the space available to provide a complete set of numbers that readers of this paper may use to replicate the process.

In practice however, depending on sample sizes and how many common variables are available, a set of linkage cells usually will number between 100 and 200.

Also, the 32 cells herein are simple combinations of gender, age-group and household income, whereas in practice, linkage cells are defined by whatever combinations of variables and codes may be found to simultaneously predict product and media usage (see the description of MDTA later).

**Notes About the Data**

- The author wishes to express appreciation to MRI for permission to use data from the MRI 2000 Doublebase.
- To provide a way of testing the calculations in MultiBasing, the MRI 2000 Doublebase was randomly split into halves, each of approximately 26,000 respondents. Split-half A represents the linkage study (MRI) and split-half B was made to simulate a TV ratings survey.
- Any two surveys of the same universe are likely to differ slightly in their population estimates and profiles. To reflect this, standard MRI respondent weights in split-half B were varied randomly up or down by 10% so that the population distribution across linkage cells would differ somewhat from that in split-half A. Also, the total 18+ population estimate was made to disagree by 3 million (compare Tables 1 and 2).
- The MRI surrogate TV audience measures were scaled down in split-half B in order to more closely resemble actual ratings. Where these appear or are discussed below they are referred to as "ratings". It would be easy for readers to be confused by this and to assume that the source is Nielsen Media Research or another TV ratings supplier, *but in all cases the 'ratings' are scaled-down surrogate TV measures from the MRI 2000 Doublebase*. The purpose is explained below.
- Splitting the MRI 2000 Doublebase sample made it possible to compare multibased estimates against empirical results. That is, TV ratings estimated by MultiBasing for product target groups can be compared against actual figures in split-half B which, unlike a regular TV database, contains the same target group variables. This comparison provides a direct check on the validity of the calculations.

## 2.1 Multibase Set-up

We begin by setting up a multibase to store ratings for all TV vehicles (i.e. dayparts and programs) that media planners want to work with. For each TV vehicle there is an audience rating and set of duplications for every linkage cell. Only one TV vehicle is shown below, FOX Sports. Audience duplications between vehicles are not shown.

**Table 1**

| **Cell** | **Demographic Linkage** (gender, age, h/hold income) | **TV Pop'n** (000) | **FOX Sports** rating |
|---|---|---|---|
| 1 | males, 18-34, <$30k | 9,384 | 6.6% |
| 2 | males, 18-34, $30-50k | 7,943 | 6.7% |
| 3 | males, 18-34, $50-75k | 7,508 | 7.2% |
| 4 | males, 18-34, $75k+ | 6,247 | 8.6% |
| 5 | males, 35-49, <$30k | 6,415 | 5.3% |
| 6 | males, 35-49, $30-50k | 6,626 | 6.5% |
| 7 | males, 35-49, $50-75k | 7,844 | 7.5% |
| 8 | males, 35-49, $75k+ | 9,977 | 8.2% |
| 9 | males, 50-64, <$30k | 3,908 | 5.1% |
| 10 | males, 50-64, $30-50k | 3,892 | 7.5% |
| 11 | males, 50-64, $50-75k | 4,230 | 7.3% |
| 12 | males, 50-64, $75k+ | 5,507 | 7.9% |
| 13 | males, 65+, <$30k | 7,850 | 5.8% |
| 14 | males, 65+, $30-50k | 3,299 | 7.9% |
| 15 | males, 65+, $50-75k | 1,373 | 7.3% |
| 16 | males, 65+, $75k+ | 1,080 | 9.3% |
| 17 | females, 18-34, <$30k | 12,636 | 2.9% |
| 18 | females, 18-34, $30-50k | 7,777 | 3.2% |
| 19 | females, 18-34, $50-75k | 6,918 | 4.1% |
| 20 | females, 18-34, $75k+ | 5,461 | 3.9% |
| 21 | females, 35-49, <$30k | 7,170 | 3.0% |
| 22 | females, 35-49, $30-50k | 6,958 | 3.2% |
| 23 | females, 35-49, $50-75k | 7,022 | 3.7% |
| 24 | females, 35-49, $75k+ | 9,735 | 3.7% |
| 25 | females, 50-64, <$30k | 6,234 | 2.9% |
| 26 | females, 50-64, $30-50k | 4,611 | 3.4% |
| 27 | females, 50-64, $50-75k | 4,282 | 3.4% |
| 28 | females, 50-64, $75k+ | 4,860 | 3.9% |
| 29 | females, 65+, <$30k | 12,391 | 2.5% |
| 30 | females, 65+, $30-50k | 4,069 | 4.0% |
| 31 | females, 65+, $50-75k | 1,386 | 3.5% |
| 32 | females, 65+, $75k+ | 921 | 4.2% |
| | Total | 195,515 | 5.1% |

## 2.2 Tabbing the Linkage Study

When a target group is selected, the following table is prepared by tabbing the linkage study (MRI). The target group in this example is golfers (i.e. played golf in the past 12 months). Perhaps the advertising campaign is for a brand of golf clubs. Also tabbed is the linkage study's surrogate audience measure among all people in each linkage cell (right-hand column).

The incidence of golfers shown below is simply the tabbed number of golfers divided by the cell population – e.g. 787,000 divided by 9,711,000 = 8.1%.

**Table 2**

| **Cell** | **Demographic Linkage** (gender, age, h/hold income) | **Linkage Study** (000) | Number of **Golfers** (000 | Incidence of **Golfers** | **FOX Sports** (surrogate) |
|---|---|---|---|---|---|
| 1 | males, 18-34, <$30k | 9,711 | 787 | 8.1% | 17.2% |
| 2 | males, 18-34, $30-50k | 8,997 | 1,196 | 13.3% | 17.7% |
| 3 | males, 18-34, $50-75k | 7,338 | 1,469 | 20.0% | 19.7% |
| 4 | males, 18-34, $75k+ | 6,864 | 1,798 | 26.2% | 19.4% |
| 5 | males, 35-49, <$30k | 5,487 | 306 | 5.6% | 15.0% |
| 6 | males, 35-49, $30-50k | 7,139 | 875 | 12.3% | 17.9% |
| 7 | males, 35-49, $50-75k | 7,887 | 1,672 | 21.2% | 16.5% |
| 8 | males, 35-49, $75k+ | 10,543 | 3,064 | 29.1% | 20.2% |
| 9 | males, 50-64, <$30k | 4,393 | 226 | 5.2% | 15.6% |
| 10 | males, 50-64, $30-50k | 4,251 | 479 | 11.3% | 17.1% |
| 11 | males, 50-64, $50-75k | 4,227 | 718 | 17.0% | 18.2% |
| 12 | males, 50-64, $75k+ | 5,783 | 1,431 | 24.7% | 19.7% |
| 13 | males, 65+, <$30k | 7,502 | 576 | 7.7% | 14.6% |
| 14 | males, 65+, $30-50k | 3,403 | 476 | 14.0% | 19.0% |
| 15 | males, 65+, $50-75k | 1,462 | 339 | 23.2% | 17.2% |
| 16 | males, 65+, $75k+ | 1,065 | 281 | 26.4% | 21.7% |
| 17 | females, 18-34, <$30k | 12,111 | 353 | 2.9% | 7.3% |
| 18 | females, 18-34, $30-50k | 7,580 | 475 | 6.3% | 8.9% |
| 19 | females, 18-34, $50-75k | 6,198 | 588 | 9.5% | 8.6% |
| 20 | females, 18-34, $75k+ | 6,061 | 720 | 11.9% | 12.3% |
| 21 | females, 35-49, <$30k | 6,830 | 266 | 3.9% | 7.2% |
| 22 | females, 35-49, $30-50k | 7,045 | 489 | 6.9% | 8.7% |
| 23 | females, 35-49, $50-75k | 8,538 | 565 | 6.6% | 8.4% |
| 24 | females, 35-49, $75k+ | 10,562 | 1,237 | 11.7% | 11.6% |
| 25 | females, 50-64, <$30k | 5,531 | 158 | 2.9% | 7.1% |
| 26 | females, 50-64, $30-50k | 4,528 | 294 | 6.5% | 9.1% |
| 27 | females, 50-64, $50-75k | 3,925 | 210 | 5.4% | 9.9% |
| 28 | females, 50-64, $75k+ | 5,121 | 406 | 7.9% | 8.3% |
| 29 | females, 65+, <$30k | 12,411 | 299 | 2.4% | 7.4% |
| 30 | females, 65+, $30-50k | 3,698 | 342 | 9.2% | 8.8% |
| 31 | females, 65+, $50-75k | 1,395 | 150 | 10.8% | 13.8% |
| 32 | females, 65+, $75k+ | 864 | 152 | 17.6% | 14.7% |
| | Total | 198,449 | 22,396 | 11.3% | 13.2% |

## 2.3 Target Group Estimation in the Multibase

The same cell-by-cell incidences of the target group as found in the linkage study are applied to the TV data in the multibase.

*Example:* 8.1% (incidence of golfers as found in the linkage study for cell 1) of 9,384,000 (population of cell 1 in the TV survey) = 760,000 golfers.

**Table 3**

| **Cell** | **Demographic Linkage** (gender, age, h/hold income) | **TV Pop'n** (000) | Incidence of **Golfers** | Est. number of **Golfers** (000) |
|---|---|---|---|---|
| 1 | males, 18-34, <$30k | 9,384 | 8.1% | 760 |
| 2 | males, 18-34, $30-50k | 7,943 | 13.3% | 1,056 |
| 3 | males, 18-34, $50-75k | 7,508 | 20.0% | 1,503 |
| 4 | males, 18-34, $75k+ | 6,247 | 26.2% | 1,637 |
| 5 | males, 35-49, <$30k | 6,415 | 5.6% | 358 |
| 6 | males, 35-49, $30-50k | 6,626 | 12.3% | 812 |
| 7 | males, 35-49, $50-75k | 7,844 | 21.2% | 1,663 |
| 8 | males, 35-49, $75k+ | 9,977 | 29.1% | 2,899 |
| 9 | males, 50-64, <$30k | 3,908 | 5.2% | 201 |
| 10 | males, 50-64, $30-50k | 3,892 | 11.3% | 439 |
| 11 | males, 50-64, $50-75k | 4,230 | 17.0% | 719 |
| 12 | males, 50-64, $75k+ | 5,507 | 24.7% | 1,363 |
| 13 | males, 65+, <$30k | 7,850 | 7.7% | 603 |
| 14 | males, 65+, $30-50k | 3,299 | 14.0% | 462 |
| 15 | males, 65+, $50-75k | 1,373 | 23.2% | 318 |
| 16 | males, 65+, $75k+ | 1,080 | 26.4% | 285 |
| 17 | females, 18-34, <$30k | 12,636 | 2.9% | 368 |
| 18 | females, 18-34, $30-50k | 7,777 | 6.3% | 487 |
| 19 | females, 18-34, $50-75k | 6,918 | 9.5% | 656 |
| 20 | females, 18-34, $75k+ | 5,461 | 11.9% | 649 |
| 21 | females, 35-49, <$30k | 7,170 | 3.9% | 279 |
| 22 | females, 35-49, $30-50k | 6,958 | 6.9% | 483 |
| 23 | females, 35-49, $50-75k | 7,022 | 6.6% | 465 |
| 24 | females, 35-49, $75k+ | 9,735 | 11.7% | 1,140 |
| 25 | females, 50-64, <$30k | 6,234 | 2.9% | 178 |
| 26 | females, 50-64, $30-50k | 4,611 | 6.5% | 300 |
| 27 | females, 50-64, $50-75k | 4,282 | 5.4% | 230 |
| 28 | females, 50-64, $75k+ | 4,860 | 7.9% | 386 |
| 29 | females, 65+, <$30k | 12,391 | 2.4% | 299 |
| 30 | females, 65+, $30-50k | 4,069 | 9.2% | 376 |
| 31 | females, 65+, $50-75k | 1,386 | 10.8% | 149 |
| 32 | females, 65+, $75k+ | 921 | 17.6% | 162 |
| | Total | 195,515 | 11.3% | 21,682 |

At this point in the calculations, a research supplier's restriction on the minimum sub-sample that may be used for reporting purposes (e.g. n = 125) is honoured. The estimated size of the target group (21,682,000 above) is also examined in unweighted form to ensure that it represents a sufficient number of respondents.

## 2.4 Initial Audience Estimation

Target populations are multiplied by the stored TV ratings to estimate the numbers of viewers. These values are summed across all linkage cells to obtain an *initial* estimate of each TV vehicle's audience within the target group.

*Example:* 760,000 golfers estimated in cell 1 from Table 3, times 6.6% (the rating of FOX Sports in cell 1 from Table 1) = 50,000 viewers. Summing down the right-hand column = 1,363,000 viewers.

**Table 4**

| Cell | **Demographic Linkage** (gender, age, h/hold income) | **Estimated Pop'n of Golfers** | **FOX Sports** rating | **Estimated Audience** (000) |
|---|---|---|---|---|
| 1 | males, 18-34, <$30k | 760 | 6.6% | 50 |
| 2 | males, 18-34, $30-50k | 1,056 | 6.7% | 71 |
| 3 | males, 18-34, $50-75k | 1,503 | 7.2% | 108 |
| 4 | males, 18-34, $75k+ | 1,637 | 8.6% | 140 |
| 5 | males, 35-49, <$30k | 358 | 5.3% | 19 |
| 6 | males, 35-49, $30-50k | 812 | 6.5% | 53 |
| 7 | males, 35-49, $50-75k | 1,663 | 7.5% | 125 |
| 8 | males, 35-49, $75k+ | 2,899 | 8.2% | 236 |
| 9 | males, 50-64, <$30k | 201 | 5.1% | 10 |
| 10 | males, 50-64, $30-50k | 439 | 7.5% | 33 |
| 11 | males, 50-64, $50-75k | 719 | 7.3% | 53 |
| 12 | males, 50-64, $75k+ | 1,363 | 7.9% | 107 |
| 13 | males, 65+, <$30k | 603 | 5.8% | 35 |
| 14 | males, 65+, $30-50k | 462 | 7.9% | 37 |
| 15 | males, 65+, $50-75k | 318 | 7.3% | 23 |
| 16 | males, 65+, $75k+ | 285 | 9.3% | 27 |
| 17 | females, 18-34, <$30k | 368 | 2.9% | 11 |
| 18 | females, 18-34, $30-50k | 487 | 3.2% | 16 |
| 19 | females, 18-34, $50-75k | 656 | 4.1% | 27 |
| 20 | females, 18-34, $75k+ | 649 | 3.9% | 26 |
| 21 | females, 35-49, <$30k | 279 | 3.0% | 8 |
| 22 | females, 35-49, $30-50k | 483 | 3.2% | 15 |
| 23 | females, 35-49, $50-75k | 465 | 3.7% | 17 |
| 24 | females, 35-49, $75k+ | 1,140 | 3.7% | 42 |
| 25 | females, 50-64, <$30k | 178 | 2.9% | 5 |
| 26 | females, 50-64, $30-50k | 300 | 3.4% | 10 |
| 27 | females, 50-64, $50-75k | 230 | 3.4% | 8 |
| 28 | females, 50-64, $75k+ | 386 | 3.9% | 15 |
| 29 | females, 65+, <$30k | 299 | 2.5% | 8 |
| 30 | females, 65+, $30-50k | 376 | 4.0% | 15 |
| 31 | females, 65+, $50-75k | 149 | 3.5% | 5 |
| 32 | females, 65+, $75k+ | 162 | 4.2% | 7 |
| | Total | 21,682 | | 1,363 |

The initial audience estimate can of course be expressed as a rating: Estimated audience of 1,363,000 divided by the estimated target group population of 21,682,000 = **6.3%** (i.e. initial estimate of FOX Sports rating among golfers).

If there is no special relationship between the target group and this media vehicle – i.e. no selectivity that is not already accounted for by the demographics used to create the linkage structure – then this initial estimate of the rating should be accurate.

This is very similar to the rating that would be found in a conventionally fused database that used the same 'critical matching variables'. However, the possibility exists that golfers are more likely than other people with similar demographic characteristics to watch FOX Sports. We can now determine whether this is so.

## 2.5 Cell-wise Estimation for Surrogate

Repeating the calculations in Table 4, a corresponding audience estimate is obtained for the surrogate in the linkage study.

*Example:* 787,000 golfers in cell 1 from Table 2, times 17.2% (the surrogate measure of FOX Sports in the linkage study) = 135,000 viewers. Summing down the right-hand column = 3,552,000 viewers.

**Table 5**

| **Cell** | **Demographic Linkage** (gender, age, h/hold income) | **Known Pop'n of Golfers** | **FOX Sports** (surrogate) | **Estimated Audience** (000) |
|---|---|---|---|---|
| 1 | males, 18-34, <$30k | 787 | 17.2% | 135 |
| 2 | males, 18-34, $30-50k | 1,196 | 17.7% | 211 |
| 3 | males, 18-34, $50-75k | 1,469 | 19.7% | 290 |
| 4 | males, 18-34, $75k+ | 1,798 | 19.4% | 349 |
| 5 | males, 35-49, <$30k | 306 | 15.0% | 46 |
| 6 | males, 35-49, $30-50k | 875 | 17.9% | 157 |
| 7 | males, 35-49, $50-75k | 1,672 | 16.5% | 275 |
| 8 | males, 35-49, $75k+ | 3,064 | 20.2% | 619 |
| 9 | males, 50-64, <$30k | 226 | 15.6% | 35 |
| 10 | males, 50-64, $30-50k | 479 | 17.1% | 82 |
| 11 | males, 50-64, $50-75k | 718 | 18.2% | 131 |
| 12 | males, 50-64, $75k+ | 1,431 | 19.7% | 281 |
| 13 | males, 65+, <$30k | 576 | 14.6% | 84 |
| 14 | males, 65+, $30-50k | 476 | 19.0% | 91 |
| 15 | males, 65+, $50-75k | 339 | 17.2% | 58 |
| 16 | males, 65+, $75k+ | 281 | 21.7% | 61 |
| 17 | females, 18-34, <$30k | 353 | 7.3% | 26 |
| 18 | females, 18-34, $30-50k | 475 | 8.9% | 42 |
| 19 | females, 18-34, $50-75k | 588 | 8.6% | 50 |
| 20 | females, 18-34, $75k+ | 720 | 12.3% | 88 |
| 21 | females, 35-49, <$30k | 266 | 7.2% | 19 |
| 22 | females, 35-49, $30-50k | 489 | 8.7% | 42 |
| 23 | females, 35-49, $50-75k | 565 | 8.4% | 48 |
| 24 | females, 35-49, $75k+ | 1,237 | 11.6% | 143 |
| 25 | females, 50-64, <$30k | 158 | 7.1% | 11 |
| 26 | females, 50-64, $30-50k | 294 | 9.1% | 27 |
| 27 | females, 50-64, $50-75k | 210 | 9.9% | 21 |
| 28 | females, 50-64, $75k+ | 406 | 8.3% | 34 |
| 29 | females, 65+, <$30k | 299 | 7.4% | 22 |
| 30 | females, 65+, $30-50k | 342 | 8.8% | 30 |
| 31 | females, 65+, $50-75k | 150 | 13.8% | 21 |
| 32 | females, 65+, $75k+ | 152 | 14.7% | 22 |
| | Total | 22,396 | | 3,552 |

Again this estimated audience can be expressed as a percentage of the known target population: 3,552,000 viewers divided by the target population of 22,396,000 = **15.9%** (i.e. estimated audience of FOX Sports among golfers, based on the surrogate measure in the linkage study).

## 2.6 Tabulation of Surrogate for Target Group

15.9% is the value *expected* to be found in the linkage study if there was no special relationship between the target group and the media vehicle left unaccounted for by the linkage structure.

However, the true incidence of the surrogate in the target group is a directly *observable* quantity in the linkage study – i.e. both measures are present in that survey. So now, tabulation of the linkage study gives us the *empirical* value for the surrogate among the target group: FOX Sports = 23.3%.

So far then, we have an expected rating based on the TV data and both expected and observed values from the 'surrogate' on the linkage study. The latter two values are used to make a final adjustment to the expected 'currency' rating, as below.

## 2.7 Residual Selectivity Index

The expected value for the FOX Sports surrogate was 15.9%, but tabbing the linkage study yielded an empirical result of 23.3%, which is 47% higher. This can be expressed as a 'residual selectivity index' of 147.

Finally then, the residual selectivity index is applied to the expected TV rating – i.e. 6.3% times 1.47 equals 9.2%. Without an actual question about golf on the TV survey, this is the best estimate we can obtain of the rating of FOX Sports among golfers.

Clearly, the residual selectivity index reflects the interest-based attraction that the media vehicle has for the target audience over and above any demographically-based selectivity. In short, golfers are more likely than most other people with the same demographic characteristics to watch FOX Sports.

It can be seen that the residual selectivity index compensates for regression to the mean. Regression to the mean occurs in conventional fusion because the demographics used to match respondents do not account for a great deal of the differences between people's product usage and media exposure. What this means in practice is for example that golfers are matched with non-golfers. As a result, within any demographic group such as high income males, fusion makes it appear that golfers are no more likely to watch golf on TV (or other sports for that matter) than non-golfers.

The extent to which MultiBasing overcomes this problem depends on the quality of the surrogates available in the linkage study. The more precise the surrogates, the more capable they are of capturing non-demographically-based relationships between target groups and media vehicles.

This is especially important for media vehicles that appeal to audiences with common interests (e.g. in a sport, hobby, occupation, lifestyle or other particular interests) rather than with a strongly differentiated demographic profile.

## 2.8 Test Result

Because the MRI 2000 Doublebase was split to provide a way of testing the calculations, we have a direct measure of the rating of FOX Sports among golfers. So, how does the estimate of 9.2% compare? The actual rating in split-half B was 9.3%.

The result obtained from conventional fusion would have been about 6.3%, as previously noted.

It is worth noting also that the rating of FOX Sports among males 35-64 with household incomes above $75,000 – a likely surrogate for golfers in the absence of a direct target group measure – was 8.1%. Why is this more accurate than the estimate that conventional fusion would have yielded?

Defining the target group as the demographic within which product usage (i.e. the incidence of playing golf) is especially high should generally tend to reflect more of a media vehicle's selectivity. However, it will fail to do so when no demographic group has a very much higher product usage incidence than average. And that, of course, is the very problem that planners so often face in today's consumer markets.

## 2.9 Run-time MultiBasing

At 'run-time', the procedure described above is applied to all vehicles that are selected as candidates. It is applied also to stored duplications *between* vehicles in order to provide complete data for reach/frequency calculation and optimization. MultiBasing involves a huge amount of calculation for an extended candidate media list.

It is not the purpose of this paper to discuss inter-media duplications. Suffice to say that surrogates on the linkage study are analysed against other media (e.g. magazines) in order to determine the variation from random in inter-media duplications at the vehicle level, and these variations from random are then applied to the 'currency'-based ratings.

## 3. Multi-Dependent Tree Analysis

### 3.1 Analytical Objective

The aim is to delineate linkage cells in such a way as to minimize within-cell variance and maximize between-cell variance across a wide range of both product usage variables and media vehicles. Before discussing the technique, the following may help to clarify the analytical objective.

In the foregoing example, the target group was golfers. Analysis of MRI would show that many more males play golf than females (as can be seen from Table 2). This means that the rating of any particular TV program among golfers should be closer to its rating for males than to that for females (where these ratings differ). That being so, for MultiBasing purposes we would need to store ratings separately for males and females. Think of this as the first branch in an hierarchical analytical 'tree'.

Analysis would show further that household income is another strong predictor of playing golf (it is an expensive game, after all). Analysis of the TV database would show that household income is also a predictor of certain TV vehicles' ratings. This means that in order to get good initial estimates of TV vehicle audiences among golfers, ratings must be stored for different income groups *among* males and females separately – i.e. a second branch in the analytical tree.

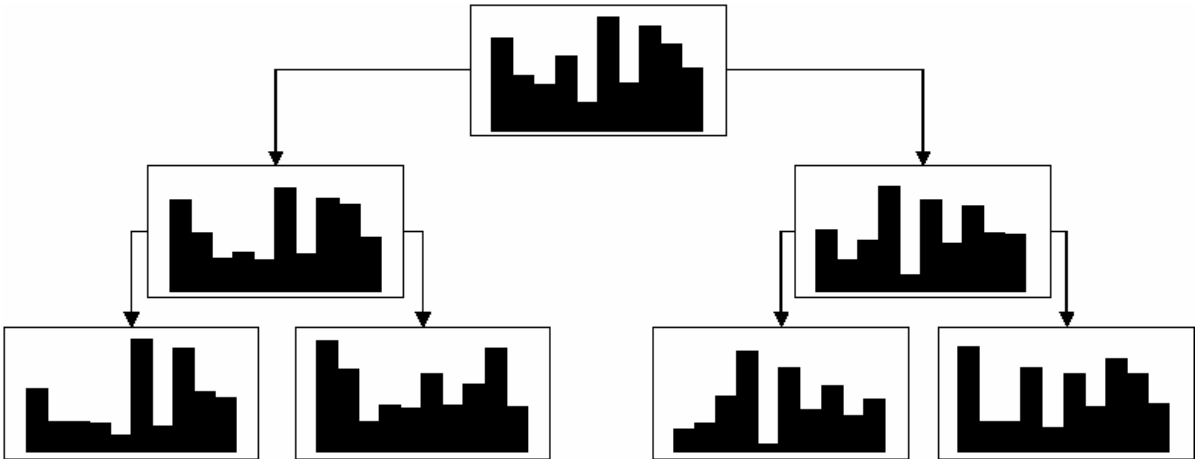
Further analysis would reveal that race is another predictor variable – i.e. high income white males are more likely to play golf than high income males in other racial groups. However, this doesn't necessarily imply that ratings must be stored separately for racial groups as well. This is because if a program's rating is *the same* among high income white males as among high income males of other races, then branching on race would not improve the accuracy of our initial estimate of that program's rating among golfers – i.e. it would not affect the outcome.

In general therefore, if a variable predicts target group membership but not TV ratings, then it is not essential to store TV ratings separately for linkage cells defined in terms of that variable. We can stop at the previous branch in the tree – i.e. high income males, in this example.

The same is true if a variable predicts TV ratings but not target group membership. Suppose that, among high income males, a program's rating is lower among those who are college graduates than those who are not. Suppose we set these up as separate linkage cells and store the different ratings. Again, the selected target group is golfers. In the MultiBasing process, the program's rating among golfers in this small part of the multibase will be a weighted average of the two groups. But given that the incidence of playing golf is the same among college graduates as it is for other high income males, the balance between the two will be the same as on the previous branch of the tree, and so the weighted average rating must also be the same as on the previous branch. Therefore we have not improved the linkage by branching on education among high income males.

It can be seen therefore that the primary analytical objective is to differentiate linkage cells in terms of common variables that simultaneously predict target group memberships *and* TV ratings (i.e. covariance). Obviously though, a general purpose linkage structure must be good for a wide range of products and media vehicles. So, for practical purposes, the analytical technique needs to be able to handle multiple dependent variables simultaneously. Being unaware of any existing software having that capability, we developed MDTA.

### 3.2 Description of MDTA



MDTA is based on so-called tree analysis (or AID), but with the capability of handling large numbers of independent (or 'predictor') variables and dependent variables simultaneously.

MDTA builds an hierarchical tree structure by branching at each node on the independent variable – and on the particular split of codes on that variable – that explains the greatest amount of remaining variance across all the dependent variables. Product usage and media audience variances are treated as separate quantities to be maximized simultaneously.

The process continues until no more cells are differentiated (i.e., no further variance is explained by the independent variables) or until cell-sizes fall below a pre-determined stopping rule.

The analysis is performed on the linkage study (MRI), and then the linkage cell coding is applied to the TV database prior to tabbing TV ratings into the multibase. Variables common to the two surveys must of course be harmonized beforehand.

The analysis does not need to be performed every time a new release of TV data becomes available. An advantage of MultiBasing is that the multibase can be updated simply by tabbing new data against previously defined linkage cells.

### 3.3 Specialized Linkage Structures

So far we have discussed the creation of a single set of (100 to 200) cells as a general purpose linkage structure. MDTA and the residual selectivity index ensure that results should be accurate for most product usage target groups.

However, under particular conditions a specialized linkage structure may be beneficial. For example, many readership surveys include some questions that are asked only of grocery shoppers. Such questions tend not to have as much influence in MDTA as ones asked of all respondents. Now, suppose that the survey does not include very precise surrogates for (say) daytime TV. Under these conditions, a specialized linkage structure for grocery product usage may yield more accurate initial ratings estimates which in turn would compensate for any 'softness' in the surrogate media measures.

Similarly, suppose that a linkage study asks a screening question about international travel followed by more detailed questions of those respondents who traveled overseas in the past year, but does not provide surrogate measures of travel-related TV programs. Again then, a specialized structure that is more specifically predictive of international travel behaviour would be beneficial.

Specialized linkage structures can be included in the same multibase as the general purpose structure. Separate 'partitions' are set up that users can select at run-time. Specialized linkage structures are MultiBasing's equivalent of 'custom' fusion (or 'fusion-on-the-fly') but with the added accuracy that is achieved by applying the residual selectivity index.

## 4. Overall Media Research Architecture

Data integration raises the issue of what common variables ideally should be measured on different media surveys to facilitate a high quality linkage structure. It is possible that linkages would be improved by adding a few carefully selected product usage, interests and lifestyle questions.

This can be determined by MDTA analysis of surveys such as MRI. It would undoubtedly be found that certain product usage and lifestyle variables are predictive of a raft of other consumer behaviour. For example, maybe the purchase price that was paid for a car would be more predictive than household income of such things as vacation travel, use of financial services, etc. If so, then the purchase price of the respondent's car would be a very useful variable to include on TV and other media surveys for the purpose of data integration.

Such an analysis is a project for another time.

As the need for integrated media planning grows, the quality of data integration will become an increasingly important issue. Arguably then, readership surveys that serve well as linkage studies will become central to the overall media research architecture of the future.

Linkage studies should include good quality surrogate measures for the media vehicles that planners want to work with, as well as the breadth of product usage and other variables.

However, a linkage study does not necessarily have to also provide the 'currency' measurement for Print, because with good surrogate measures on the linkage study, a separate Print survey can be multibased in the same way as TV and other media.

It is worth noting also that a linkage study does not have to be one huge questionnaire with hundreds of questions. It could consist of several independent data collections, so long as the linkage variables and surrogate media measures are common.

MultiBasing makes this kind of overall media research architecture possible because it works with separate databases rather than respondent-level fusion.

## 5. Conclusion

MultiBasing has some significant advantages over respondent-level fusion:

- MultiBasing can eliminate regression to the mean (depending on the quality of the surrogate media measures) so as to more fully reflect the true audience selectivity of media vehicles.
- MultiBasing preserves the integrity of the component databases – i.e. it does not create artificial new records.
- It can maintain 'currency' audience levels exactly without resorting to selective manipulation of respondent weights.
- Partitioning makes it possible for a multibase to include specialized linkage structures for particular product fields.
- Multiple databases can be integrated simultaneously – Magazines, Newspapers, TV, Radio, Internet, and even Outdoor. In principle it is possible to also include 'below-the-line' media such as in-store advertising and direct mail, if suitable audience research is available.
- Updating a multibase with newly released audience data is relatively straightforward – i.e. it only has to be tabbed in. This gives MultiBasing some timing and cost advantages.