FRIEND OR FOE - THE ROLE OF DATABASES IN MEDIA RESEARCH

Andy Brown, BMRB International Media Services
Paul Donato, Kantar Media Research

Introduction

The reason this paper is entitled "friend or foe" is a reflection of the way in which direct marketing databases have been regarded by users and practitioners in the market research industry. In the 80's and early 90's in Europe at least, the direct marketing or lifestyle database, that is to say information sources collected with the primary purpose of selling the names and addresses of respondents, were viewed in generally hostile terms. This was for two primary reasons:

i) The lack of any form of respondent anonymity was almost the antithesis of the traditional market premise.
ii) There was and still remains to a large extent today, a concern that the data gathering exercises carried out by the direct marketers, coupled with a greater level of consumer awareness of all things marketing, impacted negatively on the levels of co operation for market research.

The authors of this paper would argue that in recent years there has been a significant "warming" of the relationship between databases and market research. In essence the database has moved firmly in the direction of friend.

Before moving on to explain the ways in which databases are being employed in market and media research, it is worth perhaps outlining the merits and de-merits of each approach.

Benefits and disadvantages of Direct Marketing (DM) databases

Direct marketing databases tend to fall into two main categories:

i) Lifestyle databases - that is to say datasets that are generated with the intention of "renting" the names and addresses of the respondent.
ii) Customer lists - here we are describing usually a manufacturer or retailer’s list of customers and often the associated transactional data.

When comparing DM databases to Market Research (MR) the advantages of DM could be described as actionability, in the sense that the analyst can potentially communicate a commercial offer to the respondent directly and very cost efficiently. The DM datasets can often be extremely large in size offering high numbers of respondents in key demographic and product usage groups. Another key advantage to the owner of the customer database is the very fact that the information is owned by them alone. In some cases this may offer a competitive advantage.

One of the major disadvantages of data collected from DM sources is that in many cases the source is significantly unrepresentative, typically due to the method of data collection. Typically a response rate for a lifestyle questionnaire placed through the letterbox would be of the order of about 5-7%. An example of the data bias is that generated by the placement vehicle. In the UK ICD's (now Experian) Facts of Living study was placed via the UK's largest circulation TV listing magazine the Radio Times. One can imagine what the readership habits of the respondents looked like!

Increasingly the proprietors of the largest lifestyle databases are selling the product and media information in it's own right as their coverage of the population increases. Some databases claim to have up to 70% of the population in the UK, however much of this data is collected from different sources (and via different methodologies) over a number of years. This has led to another weakness of the DM industry in the UK in terms of the degree of professionalism shown by the industry practitioners with regard to claims about the databases. Finally the other feature of databases tends to be the very limited breadth of data held on each respondent. In many cases the dataset will contain extensive transactional data, but be limited to name, address and postcode in terms of descriptors.

Market research on the other hand typically benefits from the qualities of representativeness, and an implied greater accuracy through the anonymity of response. It is typically perceived as more professional, however lacks the actionability of the DM data. The other notable difference between DM and MR falls out of the issue of representativeness and accuracy, in that these two related MR benefits tend to come at a price in terms of sample design and often respondent incentives.

In the ideal world of information we would be able to combine the benefits of both DM and MR. That is to say large scale actionable data that is accurate, representative, and reasonably priced. How can these two sources be brought together to work in harmony while operating within the legal and ethical frameworks of the Market Research Code of Conduct and Data Protection law?

Outlined below are three different approaches as to how DM databases are being used in media research currently:

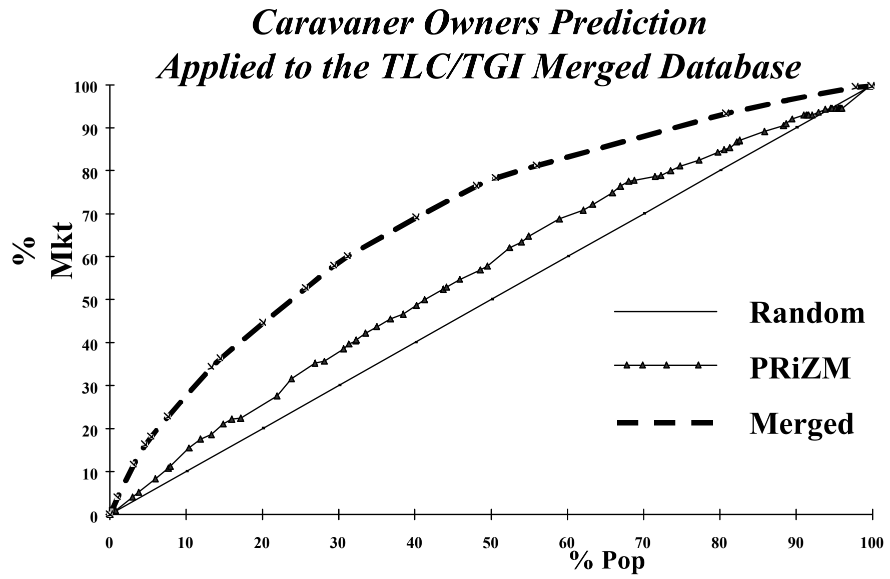
1. T-modelling –Linking readership studies to customer databases

In most major developed economies manufacturers, retailers and service providers are developing and exploiting customer databases. However as described above, there is generally good information on the commercial transactions of the individual, though when it comes to profiling, targeting and communicating with them there is a paucity of data. In many cases the database will hold name and address information and when for example the database member may have completed a registration form, occasionally basic demographics. Even then it can be difficult to know whether target groups on the database should be reached via, TV, print or direct mail. Equally as there is limited data on these individuals from the database there is little in the way of information for the planner to assess what creative approach to adopt. It was for these reasons that BMRB embarked on a development programme to review the ways in which data from the Target Group Index (TGI) survey could be linked with customer databases. TGI, which is based on a 25,000 sample of 15+ adults, is a widely used single source market research study in that it links product and media data collected from the same respondent.

The initial development work was carried out with jointly Claritas, who in the UK manage a direct marketing database branded as the Lifestyle Census. The Lifestyle Census is the aggregation of data from a large number of small lifestyle surveys collected via unsolicited letterbox questionnaires and product guarantee cards. Although the dataset is very large, in that around 13M households and 17M individuals are captured on the database, the range of variables available for analysis is limited to around 200. This is largely due to the fact that a lot of the questions are asked of only a subset of the database, while others are sponsored by advertisers who wish to have the exclusive right to mail them.

What we therefore wished to do was to take the representative wide-ranging data from the TGI, and to model it on to the Lifestyle Census. This would allow us to evaluate target audiences in a more detailed way, based on the modelled data from TGI.

The method BMRB employed was to take the individual data from three years of TGI creating a file of 75,000 respondents. Claritas in turn supplied BMRB with in effect the full Lifestyle Census file. A matching process at the individual level was carried out by BMRB. That is to say the respondents who were on the Lifestyle Census and had completed the TGI survey were identified. Having identified this group it was then possible to model via CHAID the variables collected on the TGI on to the Lifestyle Census. While this was in itself interesting, it was important that we were able to demonstrate higher levels of discrimination terms of segmentation than the existing methods of linking market research and databases. To this point, typically these methods had utilised a modelling of common variables of the two datasets, or more commonly the use of a geodemographic surrogate based on the postcode link. The table below shows a "gains" curve for two TGI variables compared with the Claritas owned Prizm segmentation system and modelled common variables.

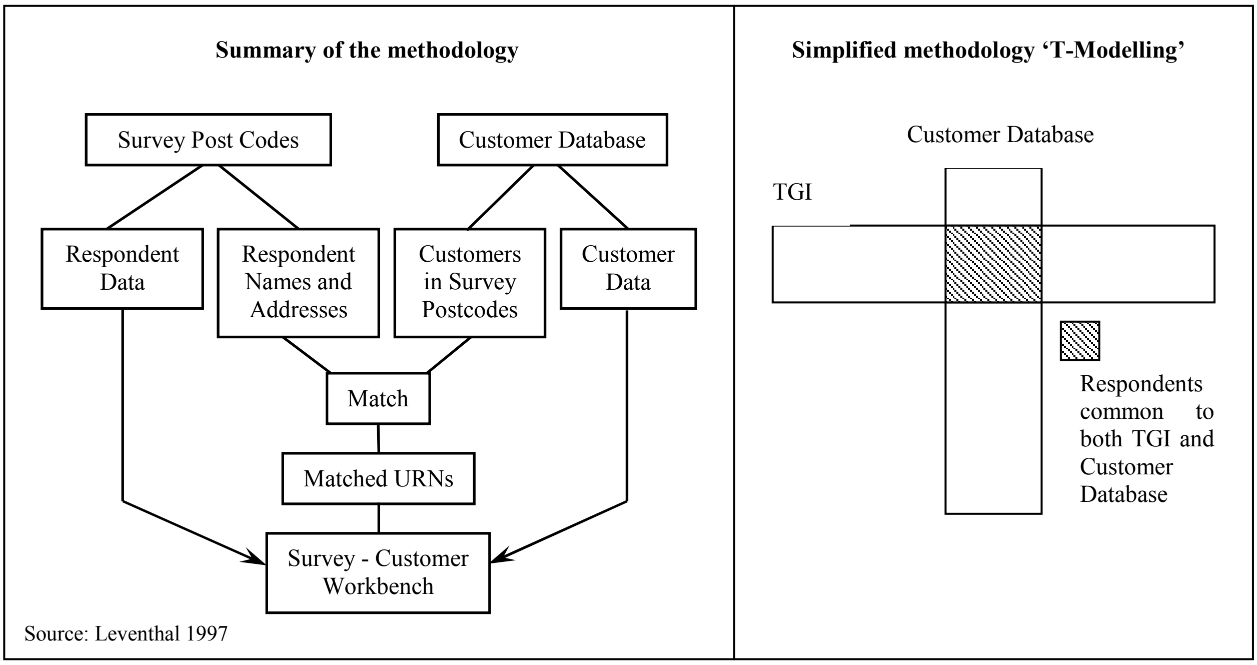


Base: TLC/TGI Merged Individuals (5907)

Source: Merged TLC/TGI 1995

Based on the success of this work we set out to evaluate applying the same technique to customer databases. The pilot client for this new approach was Barclaycard, a major credit card organisation in the UK. For many years Barclaycard were successful in marketing their services in the UK with a limited competitive set (mainly retail banking organisations) However during the mid 1990's there was a significant increase in the level of competition as the large US-owned card companies entered the market. The result was that it forced Barclaycard to really focus on development and retention of their existing customer base. Barclaycard hold records on 6.4 million cardholders. The details include basic demographics (age, sex and Mosaic code) as well as information on the different Barclaycard products the customers hold. They recognised the need to have a "fuller" picture of the customer in terms of their lifestyle and media consumption with a view as to what message to communicate to them and how.

In this case we merged 2 years of TGI a combined sample of 50,000 with the Barclaycard database. The initial stage of the process was to filter the information for the merge. To that end, the 38,000 postcodes were sent via an anonymised data file to Barclaycard for them to identify the "first stage" matches. That is to say all customers that fell into these postcodes were eligible for the second stage of the merge. Having extracted those people from the Barclaycard file they were then matched with TGI respondents by BMRB. Practically to facilitate the merge we create what is known as a "matchkey". That is to say a set of criteria to match individuals from the 2 datasets and to preserve the anonymity of the TGI respondents. In this case the matchkey was constructed using a combination of Postcode, house identifier, surname, sex and initial of the individual. Having created the matched group, the variables from TGI could then be mapped on to the database via Chaid analysis.



Summary of the methodology

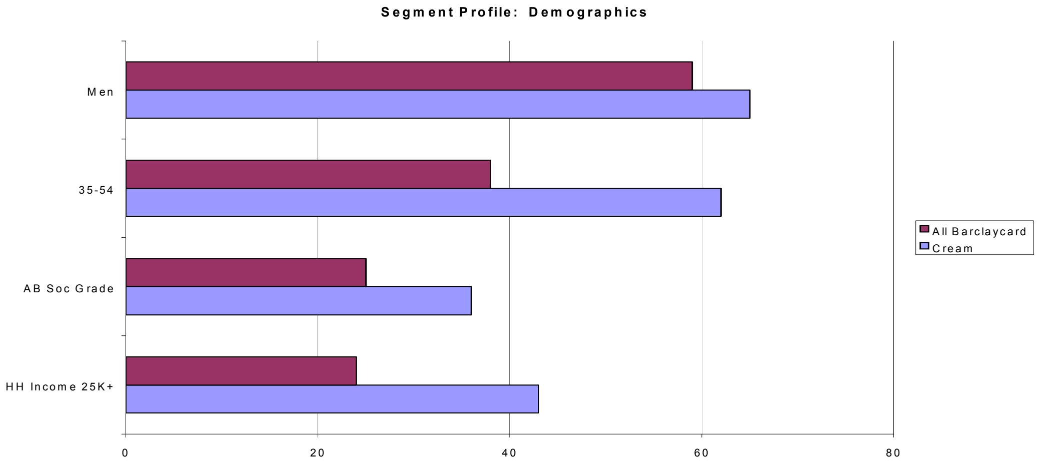
Simplified methodology 'T-Modelling'

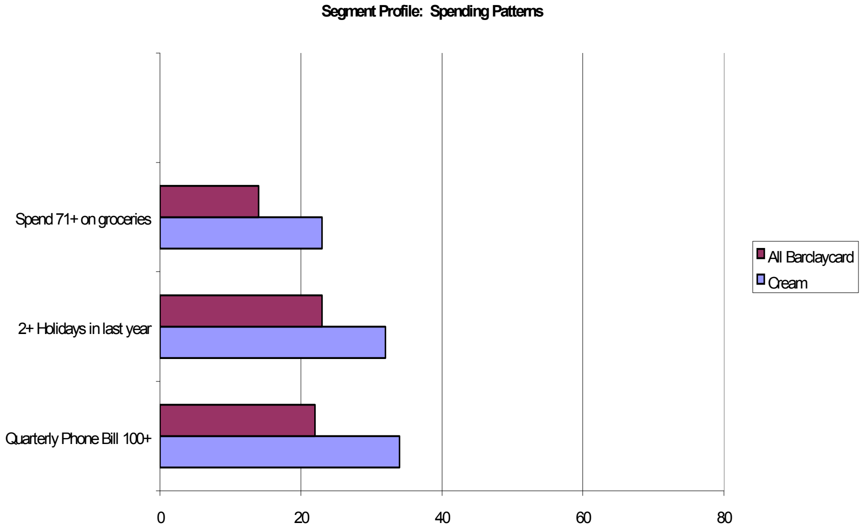
Survey Post Codes
Customer Database
Respondent Data
Respondent Names and Addresses
Customers in Survey Postcodes
Customer Data
Match
Matched URNs
Survey - Customer Workbench
Customer Database
TGI
Respondents common to both TGI and Customer Database

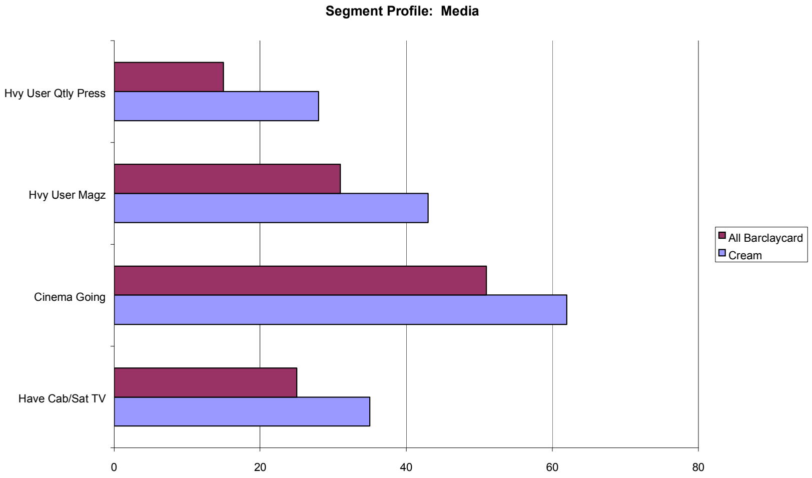
Source: Leventhal 1997

Due to the terms of the arrangement with Barclaycard we are restricted as to what we are able to present in terms of findings. Having previously created segments derived from transactional information within their customer database, and following the T-modelling exercise, it was now possible to link these segments with TGI variables. This means that the segments can be profiled in a more descriptive way both in terms of broader lifestyle behaviours but also importantly their readership and other media consumption patterns. For example Barclaycard had no way of determining whether a customer was also using another brand of credit card and which brand it might be.

While the first tables demonstrate the demographic and broad spending profile of the high value group described by Barclaycard as the "Cream" the last table shows that the segment can determined in terms of it's media profile. The table here merely assesses the weight of print consumption for newspapers and magazines, we could of course take this the next level of individual title.







Why is this new technique important? Well put simply this approach takes the marketer closer to a position of assessing a total communications plan. That is to say a below the line communication can be evaluated on a common platform to that used for the above the line media campaign. For example this means that there may be the opportunity to link a direct mailing with a promotion in a particular publication.

While this may be a positive development from an advertisers perspective, it should be recognised that there is a potential, that from a publishers perspective this may be seen as a negotiation device that may be difficult to respond to. That is because in effect we have created a unique proprietary media database for the advertiser.

While the T-modelling technique is still relatively new in the UK, it has been met with a generally positive reaction from the advertiser and media planning community. It is also probably true to say that much of the advertiser interest has been in the ability to better understand the customers on their database. The next step for BMRB is to further evaluate the quality and accuracy of the print media data. However if the outcome of the evaluation is positive, T-modelling offers a significant development in the field of media planning via an ethically sound linking of direct marketing databases and market research.

2. M2 –The use of direct marketing databases for measuring smaller magazines

A key advertising executive from the U.S. was speaking several years ago at the annual MPA meetings. Addressing over five hundred publishers, she argued the following:

“I can place a thirty second spot on Monday night football, and despite the fact that ABC doesn’t know the name or address of a single viewer, the following morning I get overnites on my desk that tell me whether my advertising reached the audiences that I paid for. That same week, I can run a two page spread in a magazine and despite the fact that the publisher may know the name and address of 50% of the audience, it will be a year before I get any numbers on the magazine and at that, it will be inactionable.

Magazines in the U.S. work in the most information rich environment compared to other media and countries. Indeed most major publishers have not ignored this resource but taken advantage of it as part of their business operations.

In this section, we will describe three ways in which databases will likely affect U.S. readership research over the next few years:

- Sampling
- The Research Instrument
- And, as in the UK, through predictive modelling.

Sampling

The subscriber list of a publication is a valuable marketing database in itself, being a collection of people who have demonstrated an affinity for the publication through the act of subscription. Although this list is an extremely valuable resource for potential research, these lists can also have a most dramatic impact on the business operations of most larger publishers. This is because in the case of some large publishers, revenue structures can be evenly distributed between advertising, circulation and database marketing/list rental and related product lines.

Indeed, database providers understand this aspect of databases. Magazines frequently overlay their circulation lists with commercial database providers in order to link database characteristics such as credit card ownership with their subscribers.

The commercial database companies are often bound by confidentiality agreements not to link subscriptions with household addresses for commercial sale to anyone other than the publication itself. The magazines generally retain this source of revenue for themselves for more direct sales through list brokerages.

However, the commercial database firms do use the information in profiling these households. They create punches like those listed in the brochure captioned above by aggregating information from warranty cards, magazine circulation lists and other related sources. For example a punch for tennis enthusiast may be created by the database provider for any household that appears on any tennis related list that it may acquire. These might include either: a tennis manufacturers warranty card list, a membership in the national tennis association or by virtue that they subscribe to a tennis magazine that has overlaid its list with the database.

This review is not intended to suggest that the role of databases in U.S. print research is ephemeral. Rather it is to suggest that there are several points of weak contact between commercial databases and readership research.

In the remainder of this section, we would like to describe work that has been done in the United States where commercial databases and circulation lists form are at the heart of the sampling, quality control and post research marketing mechanisms. In the UK, database marketers can be thought of as both friend and foe of readership research:

- a foe in the sense that research efforts are perceived as selling under the guise of research (sugging),
- and a friend in that syndicated research can realise many new predicative modelling applications and participate in a business sector that in more advanced economies can exceed total advertising in size.

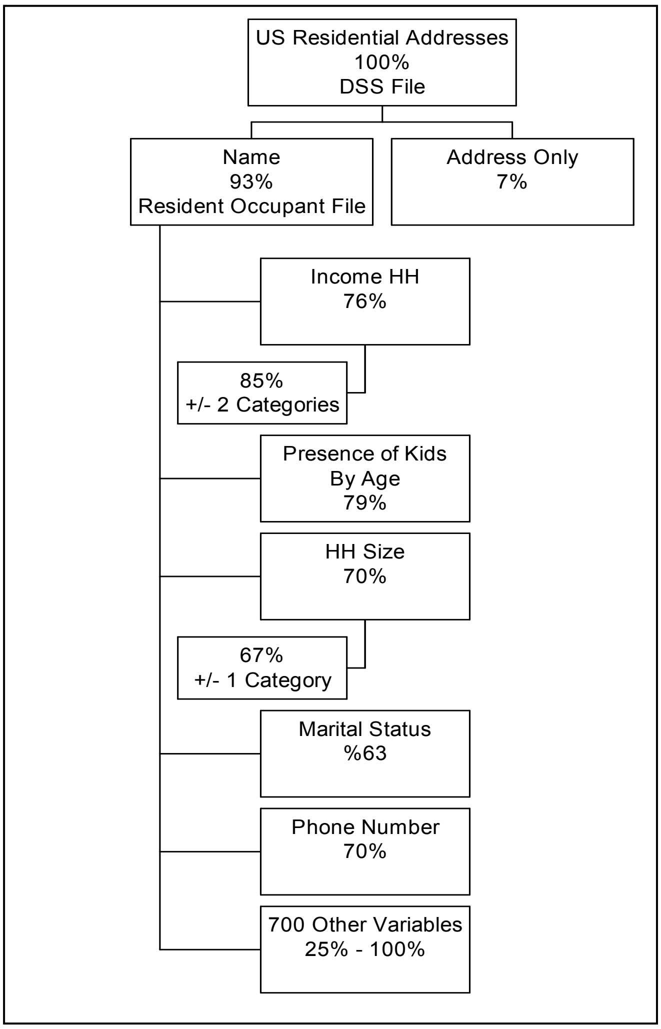
However, in the U.S., we are suggesting that databases will play much more than a significant yet ancillary role. We suggest that over the next few years, databases will form the core of the sampling and diagnostic systems used for readership research.

There is a very simple reason for this belief: increasing cost of personal interviews, declining rates among key reader populations, increased numbers of unmeasured magazines and most importantly, the role of multimedia planning software. However, the purpose of this section is not to prove that databases will be at the core of readership in the U.S. in the next few years, but rather to show how it will perform given our experiences to date.

To illustrate this performance we will draw on the experiences of the M^2 project in the United States. The sample for M^{2}s Total Audience Survey was drawn as a stratified sample (with an oversample of high-income households) from the database maintained by ACXIOM. The ACXIOM database is compiled continuously from multiple sources (telephone white pages, driver's license registration, electoral rolls, warranty cards, real estate records, consumer surveys, etc) and covers over 93% of all households in the USA. This file is then further enhanced by combining it with the DSS postal system file which will then cover 100% of all mailable addresses in the United States.

For each case in the ACXIOM database, there are as many as 700 variables (such as length of residence, type of dwelling, housing value, household income, automobile make, automobile model, presence of children, hobbies, interests, etc) available for use.

In the UK, the fact that these variables may be limited to a small proportion of the universe might be perceived as a significant loss in the quality of the sample. In the United States, such limits have been made the subject of a full diagnostic discipline with a simple means classifying the quality of the data. In the language of database marketing, the scope of such variable coverage is described as "population". A typical population profile of the total universe database is shown to the above.



The schema shows that currently, commercial databases in the United States cover:

- 100 % of the residential addresses
- 93% of which have names associated with them
- Of these 93%, up to 700 different variables are populated at different rates from a low of about 25% to a high of 100% for data of public record (e.g., mortgage status, assessed home value, …).

In 1998, M^2 ran a test of the predicted income levels that were listed on the database to the income levels that respondents provided for the same addresses. Our results agreed with a test performed by Leo Burnett in 1993 – that in 85% of the cases, the stated income levels were within two categories of the databased levels.

Household size shows a similar level of resolution, with 67% of the survey responses being within 1 person of the databased household size.

So the first element of databases, is that they provide one of the most significant opportunities for diagnostic evaluation of the sample. Databases provide estimators of sample demographics prior to the execution of the survey. With that, these data provide an excellent opportunity to evaluate the characteristics of the returned sample in contrast to the predesignated

The table below shows the response biases that were demonstrated when the databased characteristics of the predesignated sample (a mail syndicated total audience survey) were compared to the returned sample. The table shows that a mail total audience survey has surprisingly little response bias – especially in the upper income categories. If anything, professional/managerial categories are more likely to complete and return such surveys then lower socio-economic levels. The disparities of Age HOH and single person households show that the major response bias of mail surveys is a shortfall of younger respondents.

	TOTAL		
	Predesignated Sample	Returned Sample	Difference
DEMOGRAPHICS			
Residence 15+ years	21.70	23.80	2.1
Single family dwelling	77.00	79.50	2.5
Presence of children	25.10	27.00	1.9
Pro./mgr. Job	17.30	20.90	3.6
Own home	63.40	69.10	5.7
Married	45.50	50.60	5.1
Single	20.10	21.10	1.0
Age HOH 35 years+	62.00	67.20	5.2
Single adult HH	34.50	30.30	-4.2
Home value $250,000+	4.10	3.80	-0.3
HHI under $20,000	19.10	17.70	-1.4
HHI $20,000-$30,000	13.30	14.10	0.8
HHI $30,000-$40,000	12.70	13.80	1.1
HHI $40,000-$50,000	12.00	12.80	0.8
HHI $50,000-$75,000	20.50	20.30	-0.2
HHI $75,000+	20.80	20.00	-0.8

However, our understanding of sample impact on readership levels is even more significantly enhanced if we integrate circulation lists into the diagnostic process as it was done in the M^2 tests. We found that on average subscribers were about 10% more likely to return total audience surveys than non subscribers. However, this conclusion is affected dramatically by the magazines that are used in the comparison. Older skewing magazines found their subscribers yielding response rates that were 60% greater than average and young male subscribers returned surveys at a rate of some 20% less than average.

The point is that apart from all the other opportunities that they provide, commercial databases probably represent the best opportunity for diagnostic evaluation of a sample compared to any other potential element of sampling.

There are two other significant opportunities that databases represent for U.S. print research.

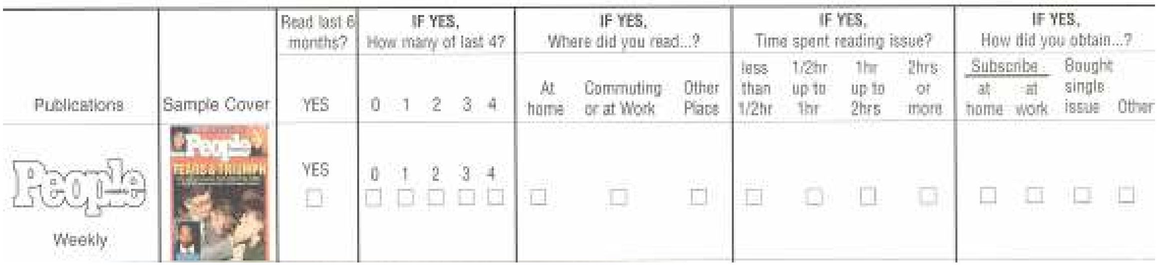
- The opportunity to conduct total audience research through the mail with higher respondent level incidence because of your ability to oversample readers by over sampling groups of magazine subscribers.
- The ability to create predictive models as were described for Barclay Bank. However, given the size of the direct marketing in the United States and the greater availability of database information, the potential impact of research platforms feeding the direct marketing industry is far more significant for the US.

The Research Instrument

Database sampling as described in the preceding section enables us to test total audience research methods by use of direct mail. Further it affords us the opportunity to oversample by readership groups in a way that makes it possible to measurement small previously unmeasurable titles.

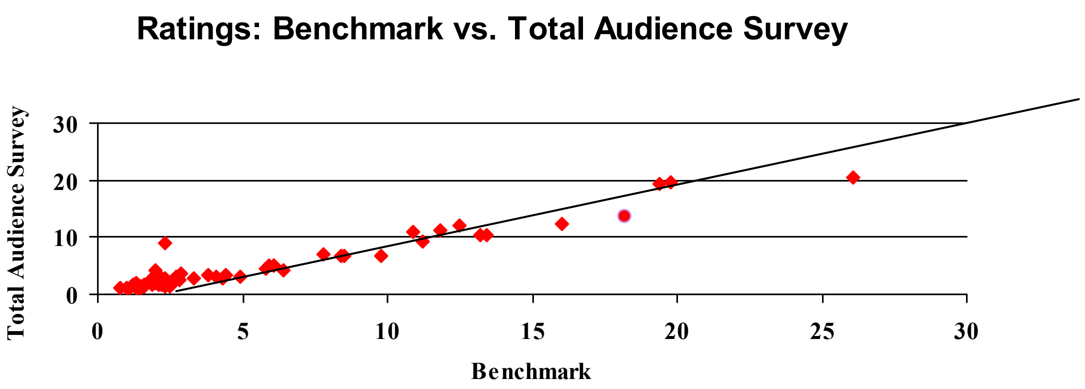
In the US we have been testing such an approach under the M^2 project. We have tested both audience measurement and accumulation measurement with very positive results. These results will be described briefly below in order to illustrate how database samples would enable a significant transition in how total audience is measured.

The M^2 survey instrument is a 16-page mail questionnaire. The first eight pages contain readership questions, while the last eight pages contain demographic and product usage questions. A total of 79 magazine titles were measured in its first execution. For each magazine title, we reproduced a four-colour picture of the magazine cover and we asked about reading in the last 6 months, the frequency of reading, the source of copy and the time spent reading.



		Read last 6 months?	IF YES, How many of last 4?	IF YES, Where did you read...?			IF YES, Time spent reading issue?				IF YES, How did you obtain...?			
											Subscribe		Bought	
Publications	Sample Cover	YES	0 1 2 3 4	At home	Commuting or at Work	Other Place	less than 1/2hr	1/2hr up to 1hr	1hr up to 2hrs	2hrs or more	at home	at work	single issue	Other
People Weekly		YES □	0 1 2 3 4 □ □ □ □ □	□	□	□	□	□	□	□	□	□	□	□

In the following chart, we show the audience ratings between the Total Audience Survey and a benchmark of recent reading. On average, the method delivered an audience that was approximately 10% lower than recent reading. A full analysis of the variation between frequency using covers and recent reading is available in an internal audit of the method. Suffice it to say that it was possible to measure many more titles as significantly lower cost.


Ratings: Benchmark vs. Total Audience Survey

However, with a proper use of database information, the value to US research might be much more than better data with more diagnostics for quality control. With proper use of information contained in commercial databases, it would be possible to oversample groups of people with a high likelihood of reading very small magazines.

Returning to the opening discussion of the US market, we said that very often, commercial database suppliers create lifestyle punches based on a compilation of warranty cards, magazine circulation overlays and other sources. These punches, illustrated a few pages earlier include such relevant categories as:

- Arts and Antique
- Auto Buff
- Bicycling
- Boating and Sailing
- Cable TV Viewing
- Camping and Hiking
- And so on, …

If we were going to measure titles which required twice the sample than we could expect through a traditional geodemographic oversample, we would buy these indicators and lay them into the database in the same way that we lay census tract income as a variable for oversampling.

Master Frame							
Name	Street	Apt	Town	Zip	Antiques	Boating	Auto
Janes	21 Elm		Irving	22991	1	1	0
Smith	23 Elm		Irving	22991	0	1	1
VVVVV	VVVVV		VVVVV	VVVVV	1	V	V
White	19 Fairway		Carmel		0	0	0
Agnew	18 Memory		Goshen			0	0
96,000,000					740,000	1,300,000	2,220,000

Today, print researchers can easily buy a file such as that listed above. The file will represent every mailable address in the United States. If we select one predesignated respondent from the list of antique enthusiasts their probability of selection is 1/ 740000. If we select one respondent from all respondents not on the antique list, their household probability is 1/(96000000 – 740000). Much of the information contained in the column marked "antiques" comes from magazine subscriptions.

The weights are proper and yield a probability sample. Multiple oversamples can be employed to select a sample that will provide sufficient reader respondent counts for several magazine fields – permitting the measurement of heretofore unmeasurable titles.

There is only one way to improve on this technique. That is to more properly manage the utilisation of circulation information as it contributes to the construction of these lifestyle indicators. As such a procedure insures more equal weights among all readers of a magazine.

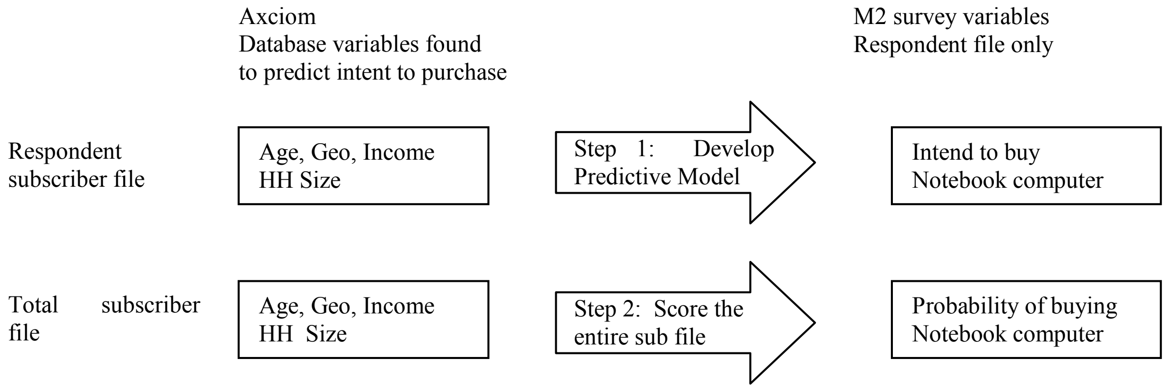
Post-Research Utilisation:
Predictive Modelling and Segmentation in the US

We have also developed several predictive models from the M2 database and loaded them back into the Axciom database from which we sampled. The nature of these models is similar to those described for the UK.

We have a large sample (of 18,000+ respondents) for whom we know (1) their answers to a set of very interesting questions (such as their likelihood to purchase personal computers or automobiles) and (2) their set of 700 variables on the ACXIOM database. We have build appropriate statistical models that predict consumer behaviour (such as the likelihood of purchase a notebook computer in the next 6 months) on the basis of their ACXIOM database variables. While the survey database of 18,000 cases is relatively small in terms of direct marketer, the predictive models applied to an entire magazine's subscriber list or the 96 million households in the ACXIOM database can yield significant numbers of qualified prospects to run large-scale direct marketing campaigns.

For the print industry, this provides a means of linking the research platform into part of the business mix of the publisher. The process is illustrated in the following diagram.

Predictive Models: Subscriber List


Axciom
Database variables found
to predict intent to purchase
M2 survey variables
Respondent file only
Respondent subscriber file
Age, Geo, Income HH Size
Step 1: Develop Predictive Model
Intend to buy Notebook computer
Total subscriber file
Age, Geo, Income HH Size
Step 2: Score the entire sub file
Probability of buying Notebook computer

A magazine supplies its complete subscriber list to the research company conducting the survey. There are certain database variables which are available for all households on the survey and the same variables are available for every address on the subscriber list.

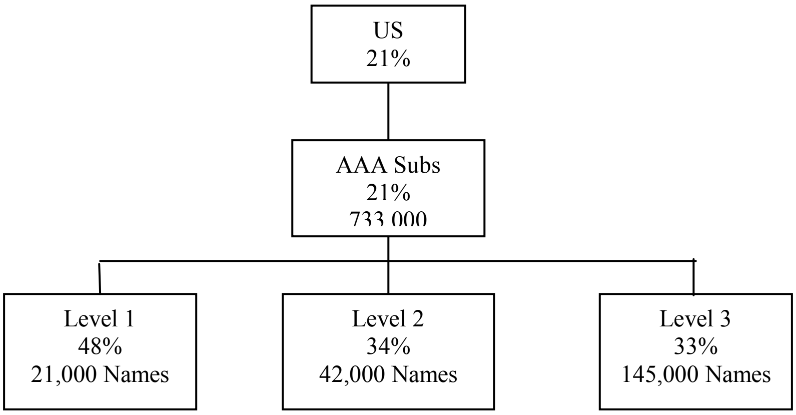
A model is developed from within the survey respondents wherein the database variables are used to predict the key dependent variable on the survey – in the case of our example, "intent to buy a notebook computer".

The same model is applied to the entire circulation list so that probabilities that each subscriber will buy a notebook can be applied to the circulation list.

At the end of the process, the publisher is in the position of selling a full run of the book advertising space but offering the advertiser, direct access to the segments illustrated below, who are the most likely to buy a notebook.

Segmenting the Subscriber List

Target: Subscribers who intend to buy Notebook computer in the next 12 months



In this example, a magazine with average penetration of subscribers who “intend to buy a notebook” can offer an advertiser the top segment of its readers who are most likely to buy a notebook.

The important thing about this application is that it brings research one step further into the business mix of the publishing industry. This is likely to benefit the future of print research because it is so amenable to accountability measures.

Currently, in the United States, the TDS or Transactional Database concept takes the integration of databases and print research one step further. TDS is a venture of MasterCard and Simmons in which retail segments are formed within readership audiences based on the credit card transaction files which have been maintained by MasterCard. Its premise is to produce retail based segments which would replace the geodemographic segments currently used as surrogates for retail behaviour.

TDS has shown that using geodemographic segments as a surrogate for retail behaviour is a very crude substitute. In a world where magazines are anxious to compete for the time based advertising that generally goes to television, radio and newspapers, creating awareness, image and demand through copy and precise temporal targeting through direct contact of high probability segments is a great opportunity to embellish the importance of research within the publishing industry.

Summary

Why is there this interest in databases and why is it relevant to today's media? As we move forward with both new media and traditional media, the database plays an increasingly important role. If one thinks of the world wide web and digital TV, both require either registration or subscription for pay per view, and furthermore they both link media exposure back to the individual in a way that traditionally mass media have not been able to. Thinking of the traditional medium of magazines. Here due to the sheer volume of title on the news-stand publishers have moved to direct distribution via subscription databases. The database is here to stay.

While there is more validation work to be done on the marriage of databases and market research, what is clear is that databases offer an increasingly cost effective method of providing cost efficient targeting and measurement solutions for magazines and other media.