



**HOW TO BUY INFORMATION**  
**THE ABCS OF PURCHASING DATA ENHANCEMENT**

## **Contents**

	<b>Page</b>
<b>Introduction</b>	<b>1</b>
<b>Why Measure Data?</b>	<b>2</b>
<b>The ABCs of Data Enhancement</b>	<b>2</b>
<b>Understanding Data Quality</b>	<b>2</b>
- Overall Match Rate	
- Elemental Match Rates	
- Accuracy	
<b>Other Measurements</b>	<b>4</b>
- Data Interpretation and Representation	
- Data Quality	
- Delivery of the Data	
<b>Designing an Accurate Data Test</b>	<b>4</b>
- Testing the Test File	
<b>Case Study: HomeBank</b>	<b>6</b>
- Company Information	
- Current Data	
- Data Suppliers	
- Data Needs	
- Testing Procedures	
- Interpreting the Results	
- Making a Decision	
- Decision	
<b>Learning from HomeBank</b>	<b>12</b>
<b>Your Data Future</b>	<b>12</b>
<b>Where to go From Here</b>	<b>13</b>

## INTRODUCTION

Marketers in the UK often append, or enhance, their company's existing customer or prospect files with external marketing data, but there are no current standards for the process or simple guidelines for comparing data suppliers. The UK market has not yet set a level of expectation when it comes to purchasing data, leaving many companies to avoid the use of external data altogether.

This paper was written to establish standards for the UK marketplace and to help you understand the complex world of purchasing data so that you can be an informed customer. If you are responsible for your company's marketing efforts, you need to be able to evaluate different data providers in order to make the best marketing decisions for your business and maximise the power of the data available to you. *Caveat Emptor* ('Buyer Beware') should not be an accepted phrase within anyone's company where data is concerned.

It is a fact today that possession of the right information at the right time has a significant impact on your business. The lack of necessary information influences sales, marketing and production, even the future of your company.

Today's 'haves' are separated from the 'have-nots' by their ability to access the information they need when they need it. This extends beyond negotiations and business deals into the very fabric of your company. To do more than just survive you must learn more about your customers: What they buy, why they buy, where they live. The more you know, the more power you have to drive sales instead of sales driving you.

Before we begin, let us clarify what we mean by data enhancement. Data enhancement is the process of appending or overlaying external consumer marketing information to your existing customer or prospect files. Unlike purchasing a list, data enhancement adds valuable demographic, socio-economic and lifestyle data to the consumer or prospect files that you maintain in-house or have purchased.

**Note:** In this paper the words 'information' and 'data' are used interchangeably because everyone has a different view of what each means to them. Data and information are presented here as two names for the same tool. Our goal is not to define one or the other but to make sure that whatever it is called, you buy the best.

## Why Measure Data?

The first question that comes to mind is why do I need to measure the information I buy? One would imagine it is all good data. If not, how could the suppliers stay in business?

The answer is analogous to shopping for a car. All cars on the market today are acceptable and meet the basic requirements otherwise they would not be sold. Which car is best for you is another question.

The same holds true for data. No one buys a car without a test drive, you shouldn't have to buy information without one either. Before we design a data test drive, however, let's understand what we're testing.

## The ABCs of Enhancement Data

All measurements of human characteristics, traits, interests, attitudes, and purchase behaviour, contain some inherent level of inaccuracy. The reasons for this are attributable to anything from misspellings to statements of misinformation.

Consumers often misunderstand survey questions, fill in the wrong blank by accident, or even tick boxes to make a happy face on the form. There is no way to circumvent these inaccuracies or correct them in a systematic fashion. However, much data is quite accurate and can fully meet your needs. You must discover what you need from your data and which supplier provides this.

Be aware that there are a few data inaccuracies that may never be overcome; however, great strides are being made in the technology behind data compilation. Among these 'data holes' are the results of programming mismatches, but most are related to the actual contributor data sources. There is no such thing as perfect data, only levels of accuracy.

## Understanding Data Quality

When you are speaking with data suppliers, the term 'data quality' usually comes up. As defined in the business, data quality is described by the terms Overall Match Rate, Elemental Match Rates and Accuracy.

These are often the only factors that some companies consider when making a data purchase or formulating a test. To do this would be a mistake. While these measurements are very important there are other factors that may be just as, or more, important in the final application of the of the purchased data. Before we discuss these, however, let's define the terms.

**1. Overall Match Rate:** This refers to the number of records you receive from you data provider with respect to the number you submitted for enhancement. For example, if you sent in a list of 1,000 of your customer names and they returned data on 800 of these, you would have Overall Match Rate of 80%. This applies only to the total amount of records with data provided, not how much data was appended to each record.

*When comparing data providers many companies find match rates to be an extremely important variable, especially in modelling. Low match rates may mean that the company does not have a large enough representation of your customer base to give you the information you need. An overall match rate of 70-80% should be expected as a minimum from most large providers.*

*You need to define what the criteria are for the match rate quoted. Is it at the postcode level, the household or address level or the individual level? Make sure you get a clear definition from each provider of what constitutes a match. When comparing match rates, you want to make sure you are comparing apples to apples. Some providers may be able to physically match a name and address to a piece of data at the postcode level and classify that as a match. Other providers may only count it when a name and address is matched to data at a specific level such as household or individual level.*

**2. Elemental Match Rates:** This refers to the number of elements requested for each record versus the total number of elements appended to your file. Not all providers will be able to supply you with all of the elements you require. Conversely, some may have all of the elements present but return very few of them, i.e. they have a lot of blank fields.

*When comparing data providers be sure that you are able to receive all of the data elements you request. A company providing a 100% match rate that returns only half the elements you desire is probably not going to meet your needs.*

*It is also important to look at the average number of elements returned per record for the elements provided. A 100% Overall Match Rate with a 50% Elemental Match Rate implies that half of their database for this element contains blank fields.*

*Be aware that some companies measure Elemental Match Rates as the ratio of elements appended to matched records. In the 1,000 record example above, they would measure an ordered element with 600 matches for a single element as a 600/800 (80% overall match rate). This would compute as a 75% Elemental Match Rate. The real way to measure this number is 600/1000, or 60%. Don't be fooled by these inflated numbers.*

**3. Accuracy:** At first glance accuracy seems to be a simple thing to measure. Just pick 10 records at random then call those people and validate the information, right?

Unfortunately it's not that easy. To be statistically accurate your records should be diverse, chosen at random, and sufficient size (ten is usually **not** a sufficient size). In addition it is very important that the data is tested against a valid benchmark. See the data test section in the paper for a more detailed explanation.

**Other Measurements**

There are more aspects to data quality than just numbers. How easy the data is to understand, interpret, use and format is often as relevant as data accuracy. It would do you little good to get 100,000 records of 100% accurate data if you were unsure what you actually bought.

Listed below are a few of the other important parameters you should be aware of when designing a data test.

**1. Data Interpretation and Representation:** Is a data dictionary and file layout available for the data elements? Is it readable and understandable? Does it provide you with an accurate representation of the data?

**2. Data Quantity:** Was the match rate in each element high enough for you to use that element in modelling or list processing? What is the minimum number of returned elements that you will need for your modelling or lists? It is possible to buy data based upon Overall Match Rates and price and not get what you need to fulfil your job.

**3. Delivery of the Data:** Data delivery quality refers not only to how the data was delivered physically but also to the turnaround time and customer service involved. Less than one million records should have a five-day turnaround time. For one million records or more you should expect a maximum ten-day turnaround.

It is also important to measure qualities like customer service, responsiveness to questions, and how the company handles problem resolution.

**Designing an Accurate Data Test**

Purchasing information, especially large amounts of information, should be treated much like any other large purchasing decision. Competitive bids should be reviewed with respect to what they offer and their associated costs. All data compilers should offer you the ability to send a test file through their system before you actually purchase the information. This is your ‘test drive’ of the data and you want it to be as accurate and informative as possible.

To do so it is extremely important that a proper test is created. The following rules should be applied in order to create a fair, unbiased and statistically accurate sample of test records. Ignoring these guidelines will almost guarantee incorrect test results.

**Step One:** When creating your test file, each of the records must be chosen at random. Do not pick the first 1,000 records in your file or every tenth one. Use a random number generator or table instead to ensure a truly random sample. This may sound tedious but in reality, it can be easily programmed so that little effort is involved on your part.

**Step Two:** Each record chosen must have the same chance as any other of being part of the general population you are measuring. This means you cannot create a test file where most of the people live in Lancashire if you are measuring nationwide data. Likewise, if you are measuring data for Lancashire alone your test file must represent the whole county, not just a few towns. This sounds obvious but it often one of the failings of most tests.

**Step Three:** For any size file in which the records are truly chosen at random, a thousand records is usually a sufficient number. In order to allow for mistakes in the random choice of the records most companies use a higher number. You should expect that a test file up to 100,000 records should be no problem for all providers to test, and they should provide this test free or at a small cost (usually related to their processing costs). Listed below is a handy ‘rule-of-thumb’ table for test file sizes.

Number of Records in Database	Adequate Test File Size
0 – 100,000	2,000
100,000 – 250,000	10,000
250,000 – 1 Million	50,000
1 Million or more	100,000

**Step Four:** The level of acceptable error should vary by element and use. Do not run an accuracy test for each element with an exact match as your criteria unless this is absolutely necessary. Instead, choose a range of accuracy that is within acceptable parameters.

No data is going to be 100% accurate so design your tests to allow for some range of error. It is important that you measure the data the way it is intended to be used. For example, let us take the element ‘Individual Age’ and look at some test numbers below for a 1,000 record test file:

Provider	Number of Direct hits	Number of Hits +/-2 Years
One	500	575
Two	300	650
Three	250	800

Note that if you only counted exact matches in your tests you should probably choose Provider One. If you are comfortable with being off by plus or minus two years is acceptable then you would make the wrong purchase decision by choosing Provider One.

Choosing the proper testing ranges may be the most significant parameter you put in your tests. Pick your acceptable ranges carefully.

**Testing the Test File**

If you have followed the steps above you are now ready to test the provider’s data. Often it is wise to let an outside company run the test files and do the accuracy checking and analysis with you help, especially if you do not have in-house analysts to do the mathematics. However, this is not a necessity and is purely a business decision.

Contact each company and send the same test file through for enhancement to each company supplying data. Measure the turnaround time and customer service, as well as the documentation provided with the returned data and the level of knowledge of your company contact. This is especially important if you are looking for a company to establish a long-term contract with for data.

When the results arrive, send the enhanced data tapes to a phone-survey company for verification of the results (accuracy testing). One thousand completed surveys should be sufficient for any test.

Below is an example test for a fictional company to further illustrate the process.

**Example Case: HomeBank**

The following example for HomeBank should provide you with a good foundation for designing your own data test. Notice that throughout the case HomeBank follows the proper testing procedures and defines their data needs before testing.

**Company Information:** HomeBank currently has about 3 million records including prospect and current customers’ household information, wealth indicators, property information, automotive information and some lifestyle segmentations. They will use this data for modelling and to create scores for their prospect file.

**Data Suppliers:** HomeBank analysts have identified three companies that can provide all of the elements described above. To ensure no biases in the tests, they are known simply as Company A, Company B and Company C.

**Data Needs:** HomeBank has identified the elements they desire along with acceptable ranges. For relative importance a weighting system has been devised. A ‘3’ represents data that HomeBank absolutely must have for their models and marketing. A ‘2’ represents data that is important but is not a deal breaker. A ‘1’ represents data that would be nice to have but is not imperative. A ‘0’ is given for companies who do not provide any data for an element. Their testing framework is shown in Figure One.

**Figure One**

**HomeBank Testing Framework**

<b>Attribute</b>	<b>Weight</b>	<b>Oldest Data Accepted</b>	<b>Acceptable Accuracy Ranges</b>
Overall Match Rate	3	N/A	N/A
Elements Per Record	2	N/A	N/A
Data Usability	3	N/A	N/A
Metadata Provided	2	N/A	N/A
Data 'Freshness'	2	N/A	N/A
Delivery Time	1	N/A	N/A
Customer Service	2	N/A	N/A
 <u>Elemental Depth</u>			
HOH Gender	3	3 years	N/A
HOH DOB	3	3 years	+/- 3 years
HOH Occupation	3	2 years	N/A
Household Marital Status	2	2 years	N/A
Household Income	3	1 year	+/- £20,000
Number of Bedrooms	1	2 years	+/- 1 bedroom
 <u>Lifestyle Data</u>			
Have a Credit Card	1	3 years	N/A
Have Stocks & Shares	1	3 years	N/A
Interests	1	3 years	N/A
Sport	1	3 years	N/A

Key

HOH = Head or Heads of Household  
 DOB = Date of Birth  
 LS = Lifestyle Element  
 N/A = Not Applicable

## **Testing Procedures**

Following the testing procedures outlined above, HomeBank created a fair and valid test as follows:

One hundred thousand records were chosen at random from the 3 million records on file using a random number generator and following steps listed in this paper.

HomeBank sent this file to each of the three data suppliers for enhancement.

After the files were enhanced, HomeBank analysts judged each supplier on customer service parameters and called each with problems to judge responsiveness. The result of these separate tests were ranked and compiled into final test results.

To test accuracy, HomeBank used an outside survey company. Calls were made at random until 1,000 completed surveys were received.

The results from the tests are listed on the following pages and are ranked. The total scores represent the sum of the weighted scores multiplied by the rankings. The largest total score reflects the overall highest ranked data supplier. See Figures Two and Three on pages 9 and 10.

**Figure Two**

**Summary Table of Test Results and Weighted Scores**

**Testing Parameters:**

Test Record Size = 100M records  
 Elements to be appended = 12  
 Full file size to append = 3MM records

Attribute	Weight	Company A		Company B		Company C	
<b>Quality Attributes</b>							
Number of Matches		916,320		815,745		717,156	
Overall Match Rate	3	92%	3	82%	2	72%	1
Elements Per Record	2	7.70	3	6.94	2	6.69	1
Data Usability	3	3	3	1	1	2	2
Metadata Provided	2	3	3	1	1	2	2
Data 'Freshness'	2	2	2	3	3	1	1
Delivery Time	1	2	2	1 day	3	4 days	1
Customer Service	2	2	2	3	3	1	1
<b>Weighted Scores</b>		<b>40</b>		<b>30</b>		<b>20</b>	
<b>Elements Returned</b>							
HOH Gender	3	748,245	1	816,720	3	800,701	2
HOH DOB or Age	3	697,449	2	682,343	1	702,014	3
HOH Occupation	3	423,632	1	469,365	2	542,986	3
Household Marital Status	2	733,554	3	682,011	2	666,739	1
Household Income	3	726,582	2	626,567	1	816,720	3
Houseowner/Renter	2	579,838	2	499,826	1	642,835	3
Estimated Home Value	2	463,721	2	487,293	3	356,485	1
Number of Bedrooms	1	281,951	3	-	0	265,932	2
<b>Weighted Scores (non-LS)</b>		<b>35</b>		<b>33</b>		<b>45</b>	
Have Credit Cards	1	-	0	725,835	2	-	0
Have Stocks & Shares	1	803,855	3	-	0	-	0
Interests	1	781,943	3	-	0	-	0
Sports	1	815,309	3	674,375	2	-	0
<b>Weighted Scores (with LS)</b>		<b>9</b>		<b>4</b>		<b>0</b>	
<b>Total Elements Returned per Matched Record</b>		<b>7,056,079</b>		<b>5,664,335</b>		<b>7,794,412</b>	
<b>Aggregated Scores (w/o Accuracy Scores)</b>							
		<b>84</b>		<b>67</b>		<b>65</b>	

Figure Three

**Accuracy Test Results**

**1,000 total records verified by phone for each company**

Accuracy Scores	Range	Company A	Company B	Company C			
<i>Scores are listed as averages</i>							
HOH Gender	N/A	75%	2	74%	1	88%	3
HOH DOB or Age	+/- 3yrs.	55%	2	54%	1	74%	3
HIH Occupation	N/A	42%	1	45%	2	51%	3
Household Marital Status	N/A	89%	3	86%	1	87%	2
Household Income	+/- £10K	28%	1	29%	2	36%	3
Homeowner/Renter	N/A	35%	1	37%	2	38%	3
Estimated Home Value	+/- £20K	42%	1	46%	3	69%	3
Number of Bedrooms	+/- 1 bedr	66%	2	0%	0	68%	3
Have Credit Cards	N/A	0%	0	83%	3	0%	0
Have Stocks & Shares	N/A	85%	2	0%	0	0%	0
Interests	N/A	74%	3	0%	0	0%	0
Sports	N/A	62%	1	78%	3	0%	0
<b>Weighted Scores</b>			<b>36</b>		<b>36</b>		<b>55</b>
<b>Quoted price (£/000) at full volume with processing</b>		<b>£40.00</b>		<b>£33.00</b>		<b>£25.00</b>	
<b>TOTAL AGGREGATED SCORES</b>			<b>120</b>		<b>103</b>		<b>120</b>

**Interpreting the results**

First HomeBank reviewed all of the data returned to see if each company met all of their required elements.

Company B was unable to supply the element 'Number of Bedrooms'. This was a non-imperative element so no disqualification applied.

Company A was not able to provide one of the Lifestyle elements, Company B was unable to provide two, and Company C could provide none. All lifestyle elements were weighted with a 1 (non-imperative data) so no disqualification applied.

Looking at the total aggregate scores alone, Companies A and C are tied. However, while Company C had the highest accuracy score for their elements at 55, they also had the lowest overall match rate at 72% (a full 20% below Company A) and did not provide any of the Lifestyle elements.

While having a lower accuracy score at 36 than Company C, Company A had the highest Overall Match Rate (a very high number at 92%), the highest number of elements returned per matched record, the highest value in data usability and metadata provided, and rankings of 2 in data freshness, delivery time and customer service.

The question for HomeBank at this time is whether to choose Company A's very high quality scores and somewhat lower accuracy or Company C's lower quality scores and higher accuracy.

HomeBank decided they would have to look at the prices before making this decision. Quotes were reviewed for the full 3 million-record file enhancement. HomeBank wished to progress this file quarterly.

	<b>Company A</b>	<b>Company B</b>	<b>Company C</b>
<b>Price/M+Processing Quarterly Price</b>	£40 £120,000	£33 £99,000	£25 £75,000
<b>Yearly Price</b>	£480,000	£396,000	£300,000

**Making a Decision**

Company C was clearly the low cost winner. In a year's time the total cost for data enhancement would be £180,000 less than Company A's. To make a final decision the users of the data, the company analysts were asked for their views.

The analysts building the models, when queried, stated that the level of accuracy for the elements returned by Company A were within acceptable ranges and that the high match rate and number of elements returned were desirable.

Because they were using the data for modelling and not simply as a list file, it was more important for them to have enough data to make their models perform properly. They were concerned that the low match rate of Company C would be detrimental to the accuracy of their models. They stated they would prefer Company A's data to that from Company C if given the choice.

**Decision**

Company A was chosen as the winner of the contract because of their match rate, depth of elements returned and good scores in all other areas. Because the data would be used by analysts for modelling purposes, HomeBank decided it was worth paying the higher price for Company A's data in order to get the quantity and quality of data needed to make their models work properly. Even though HomeBank would have saved money by purchasing Company C's data, it would have been a wasted purchase since the data would not have met their needs.

**Learning from HomeBank**

It is important to review the fact that HomeBank decided to rank and weigh their data needs and incorporate these into the test. If weighting had not been used the results of the test may have been different.

It is not important that all tests be designed with weighting factors unless they make sense to your company and reflect your data needs. With respect to data testing there are no specific answers that can be provided on the choice of weighting factors. Common sense must prevail over empirical methods.

Note as well that despite Company C having close to the same score as Company A but with a lower cost, Company A's data was chosen. Test scores are good for comparisons and will help you to evaluate the results of the test but should not necessarily be used as the sole determinant of the outcome. Just as psychometric tests would be inappropriate as the only factor for hiring new employees, considering only the data test scores without other factors would be a mistake.

What is most important in a test is not that the numbers are added properly and the best number chosen, rather that the best data for your company is purchased.

### **Your Data Future**

Whether you are a current customer of a data supplier or have never purchased data before, the principles and procedures in this paper are the same. If you are currently buying data from a supplier and have not given it some extensive testing, perhaps now is a good time to do so. You might find that data from another company may more appropriately meet your needs. If you are new to data purchasing ensure you start off on the right foot.

Companies may also wish to run these tests every year or two, needs may change, weightings may change and companies may change. It is always better to be more informed than less informed, and unless you are enhancing a very small amount of data the costs for testing are prohibitive.

### **Where to go from here**

The next question that arises then is 'where do I go from here?' Fortunately this is the easy part. Contact suppliers that have data for sale that meets your needs and ask them how to run a test file.