Scales, Samples

and Theoretical Tolerances

By Jamie Lywood

07850 20 20 30

JL@empathy.co.uk

The Absolute Scale and how to interpret ERIC data.

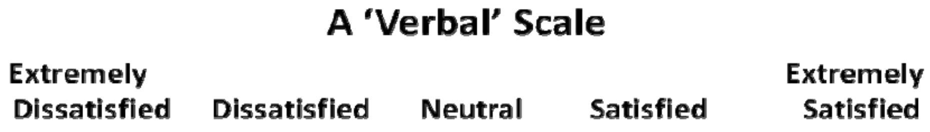
Over the years I have come across every conceivable measurement technique, numerous scales as well as the most horrendous and blatant misinterpretation of quite simple data. It is no surprise then that we are often asked about the H&Y Absolute Scale; where it came from? Is it valid? And, how can we interpret it? In this edition of Your Empathy Prophet™ we offer some explanation and examination.

The Absolute Scale

The Harding & Yorke scale is a 1 to 10 scale, using both Verbal and Numeric (aka Interval) criteria to distinguish various points on the scale. There has always been a lively debate on the various merits and disadvantages of 'VERBAL' (language based scales) versus NUMERIC (number based scales) as well as the actual number of points used.

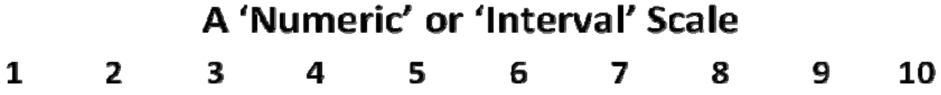
The general consensus is as follows:

- A numeric scale tends to be more accurate than a pure verbal scale as it allows less interpretation by the respondent between points - i.e. When asking the same customers to score a particular experience, scores of between 85% and 92% on a Verbal scale is measured at between 67% and 75% on the Numeric scale. Quite clearly they are not compatible as independent measures.



A 'Verbal' Scale

Extremely Dissatisfied	Dissatisfied	Neutral	Satisfied	Extremely Satisfied



A 'Numeric' or 'Interval' Scale

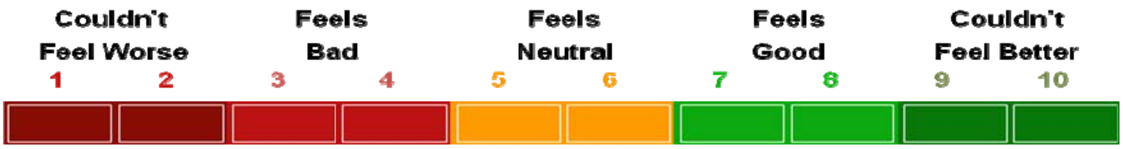
1	2	3	4	5	6	7	8	9	10

- A 10-point scale is considered most appropriate for high performing companies and those wanting to use their data to identify ways of improving their performance. This can only really be achieved through Numeric scales as the optimum number of points on a verbal scale is 7.

- When using a Numeric scale, the accuracy of the user is determined by the descriptions of the end points. This is essential and the best descriptions are those that denote absolutes – i.e. there is nothing beyond or below the descriptors ('couldn't be less satisfied' to 'couldn't be more satisfied').

- Verbal scales are more prone to misinterpretation and can be easily manipulated by their tone and sequence of verbs and descriptors used - although interpretations of these scales are often wrongly understood and articulated. Particularly poor scales are those found with 'strongly agree', 'agree' or 'very satisfied', 'satisfied' etc. They give an order without qualifying it in any way. Sure, we understand that 'very satisfied' is greater than 'satisfied', but we don't know by how much - and 'satisfied' to one person may be 'very satisfied' to another depending on their personal situation. It is not statistically acceptable to use means and standard deviations or to apply multivariate statistical techniques to establish the relationships between variables in Verbal data set although it is in Numeric ones.

Harding & Yorke have followed many of the arguments and conclude that a combined Verbal and Numeric scale from 1 to 10 with absolute determinants at each end satisfies every argument and gives the most precise and interpretable data available.



Furthermore, we take out skewed levels of expectation of the respondent by using our own employed researchers to analyse behaviours. As a result we are able to select, train and constantly monitor our researchers to ensure fairness, consistency and accuracy. We also ask researchers to collate and tag evidence (through soundbites, videoclips, scans or screen capture) so as to support the interpretation of our findings and help identify both examples of better practice and areas for improvement for our clients.

This 'evidential' approach has a huge catalytic effect on our client audiences and helps change and mould better practices from the outset.

Recently we have seen companies adopt advocacy measures. Any form of true advocacy is to be applauded, however some companies are already beginning to doubt the accuracy of the information collected and find little value in the approach. E.g. A company we have been working with, adopted an advocacy approach and yet couldn't understand why – if 77% of their customers said that they would recommend his company – were sales being driven downwards and no new customers referred to being 'referenced' when buying the company's products and services.

Fred Reichheld the founder of Net Promoter Score (NPS) recently warned companies (ECMW Conference 2008) using his measures not to link it to any kind of internal reward mechanism for at least 5 years, yet this is one of the primary reason for its adoption – that and the lesser cost of implementation.

Sample Size and Statistical Relevance

There are three factors that determine the size of the confidence interval for a given confidence level. These are: sample size, percentage and population size.

Sample Size

The larger your sample, the more sure you can be that their answers truly reflect the population. This indicates that for a given confidence level, the larger your sample size, the smaller your confidence interval. However, the relationship is not linear (i.e., doubling the sample size does not halve the confidence interval).

Percentage

Your accuracy also depends on the percentage of your sample that picks a particular answer. If 99% of your sample said "Yes" and 1% said "No" the chances of error are remote, irrespective of sample size. However, if the percentages are 51% and 49% the chances of error are much greater. It is easier to be sure of extreme answers than of middle-of-the-road ones.

When determining the sample size needed for a given level of accuracy you must use the worst case percentage (50%). You should also use this percentage if you want to determine a general level of accuracy for a sample you already have. To determine the confidence interval for a specific answer your sample has given, you can use the percentage picking that answer and get a smaller interval.

Population Size

How many people are there in the group your sample represents? This may be the number of people in a city you are studying, the number of people who buy new cars, etc. Often you may not know the exact population size. This is not a problem. The mathematics of probability proves the size of the population is irrelevant, unless the size of the sample exceeds a few percent of the total population you are examining.

This means that a sample of 500 people is equally useful in examining the opinions of a state of 15,000,000 as it would a city of 100,000. For this reason, The Survey System ignores the population size when it is "large" or unknown. Population size is only likely to be a factor when you work with a relatively small and known group of people (e.g., the members of an association).

The ‘confidence interval’ calculations assume you have a genuine random sample of the relevant population. If your sample is not truly random, you cannot rely on the intervals. Non-random samples usually result from some flaw in the sampling procedure. An example of such a flaw is to only call people during the day, and miss almost everyone who works. For most purposes, the non-working population cannot be assumed to accurately represent the entire (working and non-working) population.

Harding & Yorke Sample Sizes.

Typically H&Y consider that a sample of 40-50 interactions is sufficient to generate enough information for feeding back to a client with a good degree of confidence in the findings. Each individual interaction is measured on a minimum number of questions (47 for ERIC calls and several hundred for more in-depth studies). With H&Y projects and when samples within samples are to be considered then a minimum of 20 interactions is needed to deliver findings against the average (which still has to maintain a minimum of 40-50 interactions)

Number of interactions required for ‘Good’ degree of confidence

In-depth Analysis (Customer Empathy Audit)

	1 area	2 areas	3 areas	4 areas
Interactions	50	50	60	80
Possible different area cuts	1	2	3	4
No. of Questions per interaction	200	200	200	200
Total number of Questions	**10,000**	**10,000**	**12,000**	**16,000**

(Each area must belong to the same homogenous group and comparisons are made against the 'mean'.)

ERIC Analysis

	1 area	2 areas	3 areas	4 areas
Interactions	40	80	120	160
Possible different area cuts	1	2	3	4
No. of Questions per interaction	47	47	47	47
Total number of Questions	**1880**	**3760**	**5640**	**7520**

(Each area is treated as a separate unit and direct comparisons against each other and others from different companies / industries.)

The above scenarios give us the confidence that the information we feed back is accurate and can be relied upon for both choosing remedial action and rewarding performance.

The quantity of interactions required by H&Y can lead some to question the validity of the numbers and it is true that the situation is different from what you might automatically imagine would be the case.

For example, when considering the feelings generated by a Clients' customer base you can either generate a sample of the clients customers as above where, depending of course on the product or services offered, there will still be huge variations in the type of people being sampled or, as in the case of H&Y, you simply use a segmented base of the customer base to access, measure and assess the people belonging to the client. This can prove far more accurate and useful as it assesses an area of your business where you have absolute control. This makes remedial and change activities more robust and meaningful.

In conclusion, the way that H&Y access their clients is statistically more accurate than random sampling on the basis that the sample is a reasonably large sample of the clients own people – as opposed to a very small sample of their customers. Furthermore, people employed by a single unity are similar in the fact that they normally live in the same area, come from similar backgrounds, are trained and measured in the same way and hence adopt the same or typical corporate behaviours.

Add to this the fact that H&Y's researchers are all employed by H&Y, trained to the same level of accreditation, monitored consistently and are frequently tested through alignment measures and the degree of accuracy continues to improve.

Theoretical tolerances against a truly random population for around 50 interactions would be 0.2. However, taking into account all the factors above, the tolerances are reduced to around 0.025.

If you are still unsure try this exercise:

Prepare 5 columns of 10 numbers – these will represent your sample sizes.

Using Harding & Yorke's scale of 1 to 10 it is reasonable to assume that the sample will for the most part fall between a set of 4 parameters (i.e. between 4 and 8 on the scale).

Completely randomly populate the sample size with numbers between 4 and eight.

Sample Cut A	score	Sample Cut B	score	Sample Cut C	score	Sample Cut D	score	Sample Cut E	score
1	5	1	6	1	6	1	5	1	7
2	6	2	6	2	8	2	6	2	4
3	8	3	5	3	5	3	5	3	6
4	4	4	3	4	8	4	6	4	5
5	4	5	4	5	4	5	8	5	7
6	5	6	5	6	8	6	4	6	4
7	6	7	8	7	6	7	6	7	7
8	7	8	7	8	5	8	5	8	6
9	5	9	7	9	6	9	7	9	8
10	6	10	6	10	5	10	5	10	5
	56		57		61		57		59

Take the accumulative scores from the first set of 10 and divide by 10 to achieve the mean (56/10 = 5.6). Now add the second set of scores and divide by 20 to get the mean (113/20 = 5.65) etc.

Obviously the greater the number of Sample Cuts the less the impact on the whole.

	Cumulative	**Tolerance**
Sample of 10 (taken from A)	5.6	0.2
Sample of 20 (taken from A & B)	5.65	0.15
Sample of 30 (taken from A, B & C)	5.8	0
Sample of 40 (taken from A, B, C & D)	5.775	0.025
Sample of 50 (taken from A, B, C, D & E)	5.8	-

We can therefore assume (with relative safety) that the tolerance in this set is 0.025 and that to do more than 40 interactions would be pointless.

Over the next few pages we outline some ‘random sampling’ statistics and equations for further discussion.

Introduction

One of the questions sometimes asked of us is why we recommend a minimum number of calls of 40 for audits. The data that we derive from our analysis of interactions complements and supports the experiential evidence, and it should always be borne in mind that we only present the key threads that are apparent in all the interactions analysed.

Homogeneous Units

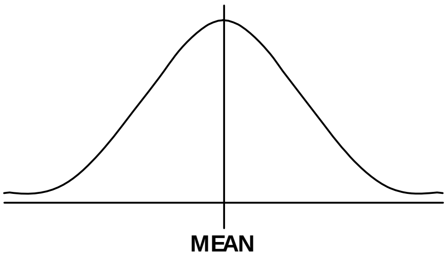
When commissioned to perform an Empathy Audit we always agree with our Client the cultural area that they wish to have audited. It is important to establish this as the number of calls is for homogeneous units. When we talk of homogeneous units – these may be from individual teams to whole divisions or companies. By homogeneity we mean that there are commonalities in the culture – threads that can be identified through research. It is these threads that we are looking for, and that we need a certain number of calls to have confidence that we can find them.

Statistical model

The simplest model for our work might be considered a Normal Distribution. It is the model we usually talk of, as it is one of the most widely known statistical concepts and closely matches our work.

A Normal distribution is where there is a continuous spectrum of possible data points, as in our Empathy scale, and where the data is symmetrically distributed about the mean [μ] with a specified variance [σ2]. The variance gives the shape of the Normal curve and relates to the spread of the data.

For the most part our top-level data is symmetric and the model can be applied.



Application to sample sizes

We assume that the possible targets within a homogeneous unit all have a possible Empathy rating, and that those Empathy ratings are Normally distributed with a mean [μ] that is the Empathy Rating that we seek to establish and report back to you.

We look for the first point at which we can establish that Empathy Rating with some degree of confidence, and at a sample size that is practical and commercially applicable. To do this we presume that our interactions will be drawn randomly from the whole population [bearing in mind that the probability of hitting a target closer to the mean is greater than hitting one at the extreme]. It is worth noting that an important mathematical theorem of statistics states that the mean of a random sample of size n drawn from a Normal population with mean [μ] and variance [$\sigma2$] has a Normal distribution with mean [μ] and variance [$\sigma2/n$].

To take a test of the mean of a defined Normal distribution we look for the limits of tolerance at a specified confidence level and varied sample sizes. The formula to use to do this is:

$z = (x - \mu0)\sqrt{n} / \sigma$ where:

- x is the sample mean
- z is the co-efficient derived from the level confidence level we specify [found in Normal statistics tables]
- $\mu0$ is the mean of the whole population
- σ is the square root of the variance of the whole population [the standard deviation]
- n is the sample size

We have to take representative means and variances from prior work on large target areas to give an indication of the figures for a nominal population. The mean we can take as 5.5, and the standard deviation is about 0.75, both of which tie-in to rational estimates.

We assume the mean of the sample is the same as the mean of the population:

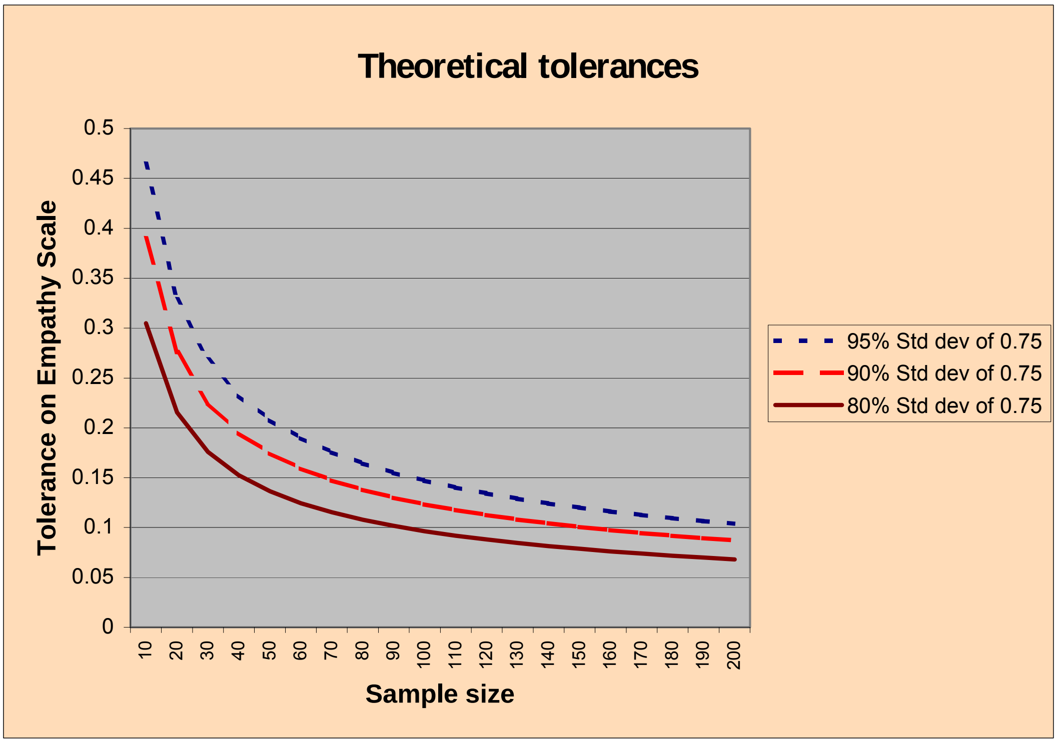
[$\mu = \mu0$]

What we are doing to find the tolerance is to find the bounds of this hypothesis at our specified confidence level.

So we calculate tolerances for various sample sizes from:

Tolerance $= \pm z.\sigma / \sqrt{n}$

The resulting data produces a graph as follows:



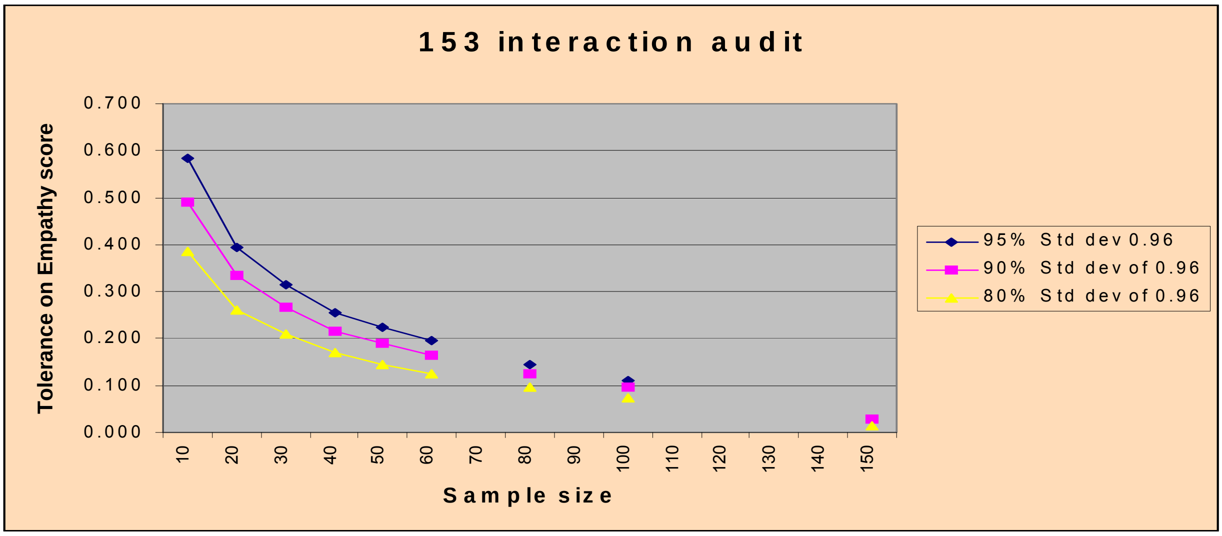
Between samples of 40 and 50 analyses we start to have solid confidence in the data produced, and recognise the commercial cost of increasing sample sizes above 50 calls to one homogeneous unit as potentially having diminishing returns. Of course we can increase sample size to cater for Client desires on accuracy and/or to enable cuts of data on sub-areas within the homogeneous unit.

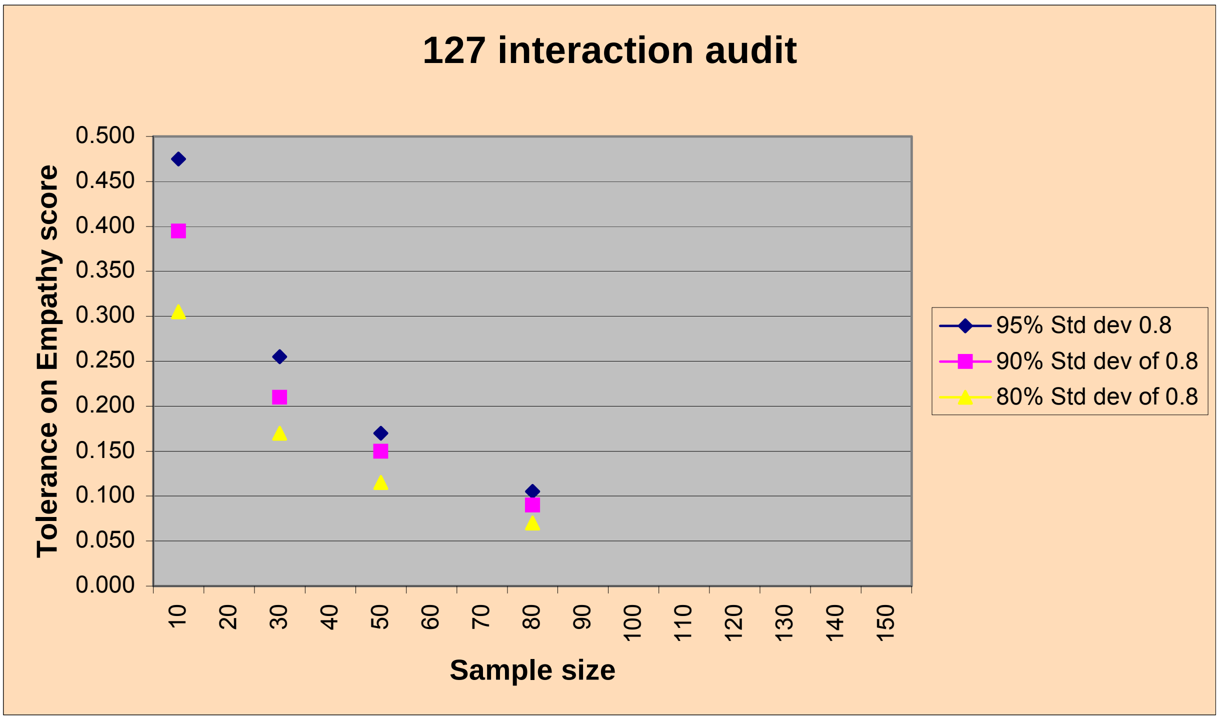
Practical Modelling

In order to develop our understanding [a desire that permeates our philosophy] we have explored the actual data from a number of audits to see whether they fit the Normal statistical model, and if so do they do so exclusively.

To carry out a test we used the data from an audit to represent a whole population and took repeated random samples of a specified size. Each sample was averaged. We repeated each sampling 10 000 times so that our distribution of averages was large enough for confidence. Having got the distribution of averages we identified tolerances for different levels of confidence for each sample size. The data from our larger audits have a close match – though there is a range of variances from about 0.65 to 1.

The graphed results show a similar curve to those from the statistical modelling with one small difference - that as the sample size approaches the population size the tolerance is much improved, as one would expect.



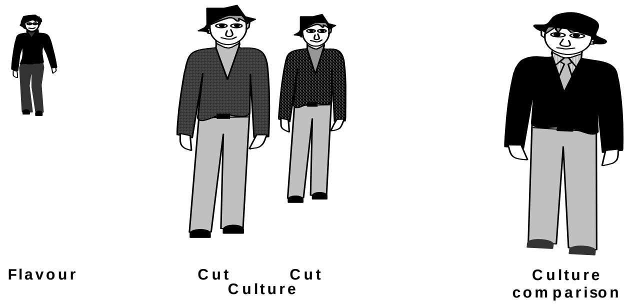


Conclusion

The Normal statistical model is a close fit to our actual data and vice versa. It provides us with sufficient confidence to specify the following:

Product Sample Rates

The sample rates below apply to all organisations where the number of targets exceeds the sample size. They can also be applied where the number of targets is down to a third of the sample size. In any situation where the number of targets is less than the nominal sample size we may suggest adaptations so as to give a confident view of the target area without compromising the research or overly impacting the relationship of our Client with their customers.



On Interaction/Analysis Audit Products:

- **10 interactions ≡ a flavour.** Conclusions comments/insights only. We cannot guarantee that the sample is representative; if it's part of a wider comparable sample we can feel more confident through association with the other interactions within the same broad culture.

- **20 interactions or more within a homogeneous group that together constitute a 'cultural' audit ≡ a cut.** Data cut primarily at section rating level, though can go to QID level. Differences must be of greater than 10% to feel confident in there being a difference between the different cuts of data. We are confident, given that the cultural threads are evident across all cuts, that the key inter-cut differences are valid.

- **40 interactions or more from a homogeneous group ≡ a culture or a cultural comparison.** The culture can confidently be analysed, and compared at a QID level to any other such sample within the same or other business, at the same time or at different points in time. Such a culture/cultural comparison group may be made up of a number of 'cuts' as long as the criteria for cuts is met. So one business unit with three departments wanting to have a culture and comparison between the three areas would need at least 60 interactions, given that the target areas could cope with 20 interactions.

On Literature and Correspondence Audits:

The material for these is usually from one of three sources:

- Boiler-plate correspondence
- Bespoke correspondence
- Published material

For auditing boiler-plate and bespoke material the same figures as above would apply, given that the material existed, and our preference would be to have material that is as it would be/has been sent out [same paper, envelope etc].

NB:
As with some web-sites and promotional material [Point Of Sale, publicity, annual reports etc] we would, at this point, only offer a flavour of how published material comes across. This is because there is no explicit interaction, though there is an element of potential relationship.

Specifically, we segment the Customer Experiences based on their expectation of the product or services they request or are targeted on by Clients (i.e. a Customer requests literature on Life Insurance receives a brochure on all types of insurance).

For Web based analysis we segment Customer Experiences based both on their expectation (as above) and their competencies (Novice, Intermediate or Advanced IT users).

On Survey Products:

- The same broad bands apply, though care should be taken in considering the number of targets, and the value of the targets to the survey.

With external surveys we usually need at least twice as many possible targets as we agree to hit [probably three times as many], as it is often difficult to contact and have time with people who do not directly work for our Client company.

Why is ERIC so important?

I was intrigued, and shocked, when I first discovered that no service-related metric in the world had been academically proven to correlate directly with profit. Like so many others I had been hoodwinked by claims of links with 'Shareholder Value', 'Loyalty', 'Revenue Growth' etc. that I believed these to be clear and unequivocal correlations with profit. How very wrong I was. Even now, when I first introduce the now known correlation between ERIC and Profitability (as measured by Return on Capital Employed), I get responses like "Yes, but they don't really - do they?" Yes they do and represent an increasingly powerful tool in a strategic armoury.

So what is a good score?

We can also make some fairly generalised assumptions of what is a good and bad score against the scale. For example a company scoring 6.5 or higher on the Empathy side will feel positively better than a company below this figure. These companies (scoring <6.5) tend to be unforgettable whereas companies scoring less than 4.5 are pretty unsustainable as far as the customer is concerned.

On the process side it is more difficult to differentiate those companies scoring higher than 7.5 whereas those scoring 7.5 clearly have a problem with how they process their customers.

These 'boundaries' will change over time, but individual company scores won't vary dramatically unless there are positive or negative influences on the overall culture of an organisation – this is unlike many Customer Satisfaction scores that are influenced by customer expectation at one specific moment in time.

Contact Jamie Lywood for references and further detail.

JL@empathy.co.uk

07850 20 20 30