Making customer satisfaction measures work - 5
Using relevant measures and metrics

Chuck Chakrapani

Once we resolve the issues discussed in previous articles in this series, we face problems related to the actual measurement of customer satisfaction. More specifically, we have two tasks. The first one is deciding what basic questions to ask. The second task relates to the use of a measurement scale. These two issues are discussed here.

Deciding on the basic CSM question components
Five basic questions form the basis of customer satisfaction measurement. These are incidence, frequency, importance, performance and an overall criterion measure. It is not absolutely essential to ask all types of questions, but they need to be given sufficient consideration before deciding to discard one or more of the types of measures.

Incidence
This variable relates to the relevance of a given service or the incidence of a given problem. Examples of incidence questions include:

- "Did you use an electronic search service in the past year?" and
- "Did you have any problem with your telephone service in the past three months?"

They act as filtering variables for further questions and as well are useful to define the relevant audience.

Frequency
Once we identify a customer who uses a service or is facing a problem, we then need to assess the frequency with which the service is used or the problem is encountered.

Once again, the purpose of this type of variable is to separate frequent from infrequent problems and to relate customer satisfaction to the frequency of the problem. Examples of frequency questions are:

- "How often did you use an electronic search service in the past year?" and
- "How often did you have a problem with your telephone service in the past three months?"

Frequency questions are usually asked only of those who reply in the affirmative to the incidence question.

Importance
A problem that is frequent (e.g., the elevator is always crowded) may not necessarily be considered important by customers who face it. Conversely, a service that is not used very frequently by the customer (e.g., emergency procedures in a hospital) may be vitally important.

Therefore, we need a measure of importance in addition to the frequency measure. Examples of importance questions are:

- "How important is it to you that your computer problems be fixed in the same day?"
- "Is it important to you that our service be available to you on-line?"

Performance
The previous three questions are a prelude to the central issue of performance. How satisfied is the customer with our performance? How does our performance compare to that of our competitors? In some cases, we may split this variable in two, one relating to how well we performed and the other relating to how satisfied the customer is with our performance. Examples of performance questions are:

- "How do you rate our ability to provide same day service?"
- "How satisfied are you with our same day service?"
- "How would you rate Company B [competitor] on this service?"

Overall criterion measure
It is also useful to have an overall criterion measure, such as overall satisfaction with the service provided. The overall criterion measure provides us with a means of computing the importance of the individual attributes (see the section on derived importance scores).

These five aspects of measurement - incidence, frequency, importance, performance, and the overall criterion measure - are basic to all CSM systems, although they may not be explicitly included in the questionnaire. For instance, if we interview only those who made more than 10 long distance calls last month based on telephone company records, we already know the incidence and frequency. Again, in some cases, we may avoid asking direct importance questions and decide to use derived importance scores instead. So it is not always necessary to ask explicit questions to cover these five aspects relating to customer satisfaction.

Choose the right metric
The next major step relates to the choice of the relevant metric to be used in measuring customer satisfaction. Attributes we choose to measure can be measured using a number of different scales. What type of scale should we use? Should we use a numeric scale (e.g., a 10-point scale), a verbal scale (e.g., Good, Average, Poor) or a binary scale (e.g. Satisfied, Not Satisfied)?

Measurement scales can influence the measurement
Satisfaction scores are influenced by the measurement instrument we use. For instance, if the average rating of overall satisfaction is 7.8 on a 10-point scale, what average rating can we expect the same attribute to have if it was measured on a 5-point scale? The current evidence is that different scales provide different ratings and they are not directly comparable. This suggests that we should be careful about the metric we choose to measure customer satisfaction. The second problem is that many metrics give inflated scores. Most customers do not use the lower points of a 10-point scale; when there is no mid-point on a scale, most neutral customers tend to move up rather than down (positive bias). So we need to look at measurement scales with two issues in mind:

a) *Weaknesses of measurement scales* . Most scales used in CSM have inherent weaknesses. We cannot avoid them, but we can choose a scale that introduces the least amount of bias for the problem under consideration.

b) *Inflated ratings.* We seem to get inflated ratings when we use standard numerical scales such as a 10-point scale. As we noted elsewhere, this is likely due to extraneous influences such as regression towards the midpoint of the scale, a respondent's tendency to truncate the scale, and the prevalence of a high proportion of low use customers.

These two issues are discussed in greater detail below.

a. Problems with measurement instruments
The first problem relates to the type of scales used in CSM studies. Some scales give more reliable results than others. The problems with scales used in marketing research include response bias, lack of clarity, and inability to differentiate among the objects rated. Scales are sometimes rejected because they lack 'face validity' (i.e., they don't appear 'reasonable' to the user). Sometimes a scale that works well in a face-to-face interview may not work in the same way in a mail or telephone survey. Other scales may be somewhat more difficult for respondents to use. Improving measurement instruments is an ongoing process. Therefore, the following comments should be taken as representing the current state of knowledge and not as definite solutions to problems of measurement. The observations below are based on the results of a number of studies by researchers such as Devlin, Brown and myself. As Susan Devlin reminds us, there is no 'best scale' that will provide the best results under all circumstances. It is best to experiment with a shortlist of different scales before settling on one that is best suited for your purposes.

Most scales that are used in service quality research fall into three major categories: verbal scales, numerical scales and comparison scales.

Verbal Scales
Verbal scales of evaluation use words rather than numbers to describe various scale points. Although typically

numbers are assigned to different scale points (e.g., 4=Very good; 3=Good; 2=Poor 1= Terrible), respondents using a verbal scale depend on words rather than on numbers to evaluate the service. Verbal scales can be binary scales, rating scales with a midpoint, or rating scales without a midpoint.

Binary scales
Binary scales have two alternatives such as:

- Acceptable - Unacceptable
- Good - Poor
- Satisfied - Dissatisfied.

Binary scales are excellent when the distinctions are clear in a customer's mind. They also force customers to examine their attitudes more closely and decide one way or the other. However, binary scales tend to make the customer's task difficult when he or she feels that the true answer lies between the two alternatives offered. For instance, the service offered may not be considered 'Unacceptable' but neither is it 'Acceptable' in the sense that the customer is happy with it. Therefore binary scales are suitable only when we believe that the respondents have clear-cut perceptions.

Rating scales without a mid-point
Examples of rating scales without mid-points are:

- Acceptable - Somewhat Acceptable - Somewhat Unacceptable - Unacceptable
- Excellent - Good - Poor - Very Poor
- Very satisfied - Satisfied - Dissatisfied - Very Dissatisfied

Scales such as these tend to suffer from a positive response bias. Studies show that those who have neither a positive nor a negative opinion about something tend to choose the lowest positive descriptor rather than the lowest negative descriptor. There is some evidence to show that this bias is not uniformly distributed across the population. This poses an additional problem. The differences in evaluation between two subgroups may reflect their bias towards a positive response rather than revealing genuine evaluative differences.

Rating scales with a mid-point
The scales described above may be changed to include midpoints. Here is an example of a verbal scale with a midpoint:

- Acceptable - Somewhat Acceptable - Neither Acceptable nor Unacceptable - Somewhat Unacceptable - Unacceptable
- Excellent - Good - Average - Poor - Very Poor
- Very Satisfied - Satisfied - Neither Satisfied nor Dissatisfied - Dissatisfied - Very Dissatisfied

As long as these types of scales are unambiguous, they seem to work in most situations. Ambiguity, however, can arise in the way a question is framed.

Subjective vs. objective scales
Rating scales can be phrased either from a subjective viewpoint or from an objective point of view. Subjective scales refer to a customer's personal feelings, as in the example below:

- Very Satisfied - Satisfied - Neither Satisfied nor Dissatisfied - Dissatisfied - Very Dissatisfied

An *objective scale* , on the other hand, attempts to elicit an 'objective' evaluation from a customer, for example,

- Excellent - Good - Average - Poor - Very Poor

Which is better - a subjective scale or an objective scale? Empirical research seems to support 'objective' scales. Why is this so? One possible hypothesis is that objective scales tend to suffer less from positive bias. An objective rating gives the customer a chance to dissociate his or her feelings - however mildly - from the evaluation itself. *'Poor' [service]* sounds less like a complaint than *[I am] Dissatisfied [with the service]* and, as a result, less subject to positive bias.

Numeric scales
Numeric scales, such as 10-point scales, appear to have inconsistent discriminating power. More often than not, they are poor discriminators. Another problem with these scales is what does an average score mean in words? While an average of 7.5 sounds high, what if averages for all institutions tested range from 6.5 to 8.5? What if the averages range from 7.5 to 9.5? There does not seem to be a consistent relationship between numbers on this scale and how a person would describe that number in words.

Numeric scales such as 10-point scales have inconsistent discriminating power.

Scales with a large number of points such as a 10-point scale tend to be particularly poor discriminators. In several studies of customer satisfaction, the analysis of results show that customers tend to mentally truncate the scale to the upper range, thus making all average ratings highly positive.

My analysis (unpublished) of a number of studies show that the failure of numeric scales to distinguish between competing products is much more pronounced for service quality/customer satisfaction than for product quality/product satisfaction measurement. It is not clear why this is so. However, one can hypothesize that the intangible nature of service (as opposed to a product) makes the rating seem 'subjective' and this, in turn, leads to a positive bias.

Comparison scales
Comparison scales compare actual performance with some other measure, such as one's expectations, as shown in the example below:

The competence of the staff was:

- Much better than what I expected
- Better than what I expected
- About what I expected
- Worse than what I expected
- Much worse than what I expected

Comparison can also be between one company and another which is perceived to be the industry standard. For example,

On responding to customer complaints

- Company A is much better than Company X
- Company A is better than Company X
- Company A is about the same as Company X
- Company A is worse than Company X
- Company A is much worse than Company X.

Comparison scales tend to distinguish companies and service rated better than the other two types of scales. This generally seems to be the case whether the comparison is between one's expectation and what is being delivered or between the performance of two companies.

What should we do if scores remain inflated after all this? We will discuss this issue in the next article.

Dr. Chuck Chakrapani of Standard Research Systems is a Toronto-based consultant, author and seminar leader. He works internationally. He is currently completing a book: How to Measure Service Quality and Customer Satisfaction *which will be published by the American Marketing Association early this year.*