Discovering Relationships between Service and Customer Satisfaction

Michael Buckley and Ram Chillarege

Center for Software Engineering
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598
mbuckley@watson.ibm.com (914) 784-7726

Abstract

Organizations spend significant resources tracking customer satisfaction and managing service delivery. Although a great deal of effort is expended in understanding what goes on within each of these areas, little or no effort has been applied to identifying and quantifying the relationships between the two. The objective of this research is to discover and establish potential relationships between service data and customer satisfaction. This understanding will enable more effective management, which will lead to improved quality, reduced cost and increased customer satisfaction.

This study uses three years of data from an IBM operating system to measure the correlation between 15 service variables and nine customer satisfaction attributes. The results show that:

- *There is a relationship between the service data and customer satisfaction. This is the first time the existence of such a relationship has been proven and quantified.*
- *The relative order of influence on customer satisfaction, of the four key service measures that are usually tracked, is defective fixes, followed by the number of problems, which in turn are followed by the number of defects and Days to Solution. The latter two were found to have little or no influence on customer satisfaction.*
- *There is a return on investment of at least ten to one, for each dollar spent on quality improvement efforts in development.*

Key Words: *Software Quality, Customer Satisfaction, Service Process, Correlation, Empirical Analysis.*

1. Introduction

There is an extensive body of literature on software metrics, and computer system failure analysis [1-12]. This ranges from research into particular aspects of software, such as code complexity [11], through schemes for in-process feedback [7, 10], to empirical analyses using actual failure data which is often gleaned from event logs [6]. However, these studies are generally restricted to a portion of the software life cycle, such as test, development, or field failures. There is little or no published research which covers the total software life cycle. That is there are no studies which show how variations in a parameter at one stage, say development, affect subsequent stages, such as customer satisfaction in the field.

The objective of this research was to identify and quantify the relationships between the data collected in the service process and customer satisfaction. These relationships, or indeed if there are any at all, are generally not understood at present. Thus, the situation is similar to that depicted in Figure 1, where there may be a good understanding of what goes on within each stage, but it appears as if there is an opaque curtain surrounding these stages.

There are a number of theories as to what the connections are, but these have generally not been proven. For example, if you ask if the number of problems on a product or the service call response time will have a greater impact on customer satisfaction, you are likely to get two conflicting answers. Thus, the intent of this research was to pull back the curtain shown in Figure 1, and to unveil the linkages between the stages, in order to facilitate prediction, understanding, and control.

There are good reasons for pursuing this objective. The first is that there are a number of parties that are keenly interested in determining what the connections are. These include the software development labs, the service organizations and the customer satisfaction survey group. The lack of understanding of the relationships means that they cannot evaluate how changes in one stage will affect the next, they cannot do trade-off analysis, and they cannot optimize quality improvement efforts.

Secondly, there are a large number of variables that are measured at each stage. However, measurements cost money and a focus on the wrong ones can divert attention and decrease efficiency. Thus, the critical variables need to be identified.

Another reason is that many companies need to reduce their service costs. This is especially true in the high volume shrink-wrapped consumer market, where the cost of a single call may be more than the profit on a product. However, they would like to decrease the service costs without adversely affecting customer satisfaction and thus they need to know what the relationships are.

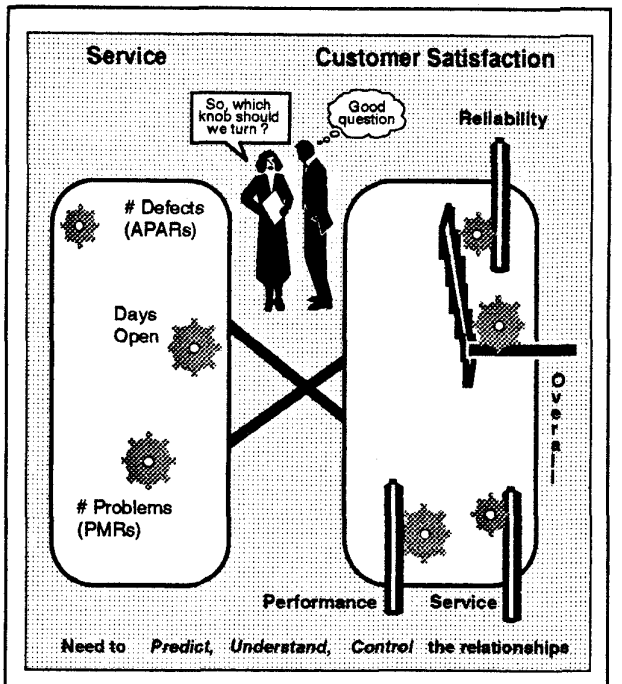

Figure 1: So, which knob should we turn ?

The last point to note is that there is a tendency towards sub-optimization. That is one understands and manages within each stage very effectively, but there is little or no effort (or method) to optimize across all the stages.

Identifying and quantifying "what drives what" will lead to improved quality, reduced cost, and increased customer satisfaction. This research contributes to these benefits by answering a number of key questions that development labs and service organizations encounter on a regular basis, namely:

- What service measures should one focus upon, to maximize improvements in customer satisfaction?
- What service measures can be "ignored" or discontinued ?
- What customer satisfaction can be expected given a particular call rate ?
- What return on investment can be expected from quality improvement efforts ?

The methodology that was used in this research is outlined in Section 2, which includes a description of the service and customer satisfaction processes and data. The results and main findings are presented in Section 3, and the key contributions are summarized in Section 4.

2. Methodology

The approach that was used, to determine if there was a relationship, was to use a number of years of real data collected on an IBM operating system product. The service and customer satisfaction data was extracted from various databases and the correlation coefficients between 15 service variables and nine customer satisfaction survey attributes were computed. The data sources and collection process are summarized in Figure 2.

The two sets of data were compared on both a per quarter basis and a per customer basis. In addition various offsets were used to allow for the time lag between service calls and customer satisfaction scores, and the sample size was varied as part of a sensitivity study. A cost - benefit study was also conducted to quantify the savings generated by investments in quality improvement. SAS was used for the statistical analysis which included scatter plots, time plots, and evaluation of the Pearson, Kendall and Spearman correlation coefficients.

The service data and the customer satisfaction survey data are described in the next two sub-sections. This is followed by a discussion of the analysis procedure.

2.1 Service Data

The service centers handle customer calls that are reported either electronically or over the phone. The calls cover a wide spectrum of problems from defects in the code, through non-defect related problems, to how-to and informational queries. These calls are handled by a three level service organization, which is summarized in the left hand column of Figure 2.

The first level, called Level 1, validates entitlement, does a quick data base search for known problems, and routes problems to the appropriate queues. Level 1 opens a Problem Management Record or PMR for each customer reported problem. If Level 1 is not able to solve the problem it is passed to Level 2.

The people at Level 2 are more specialized and have a greater depth of knowledge on particular products. They are generally able to solve the problem and close the PMR. If they are unable to do so and it appears that the problem is due to a *previously unknown* defect, then the problem is passed to Level 3, and an APAR is opened. APAR stands for Authorized Program Analysis Report and is IBM parlance for a defect in its code or documentation.

Once the root cause of the problem is identified and solved the APAR is closed and the symptom-solution database is updated so that subsequent callers may be directed to the appropriate fix. Fixes are distributed to the installed bases using PTFs or Program Temporary Fixes, until the next release of the product.

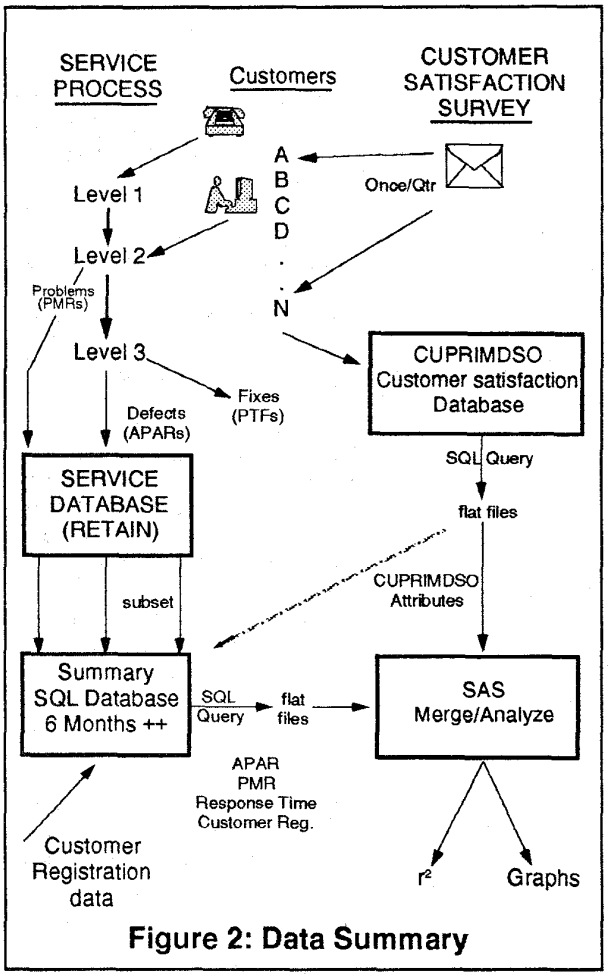

Figure 2: Data Summary

The IBM service process makes extensive use of a database and set of tools called RETAIN (Remote Technical Assistance Information Network). A subset of the RETAIN system is downloaded to a summary SQL database on a regular basis. The service data that was employed in this research was extracted from the summary SQL database, and covered over three years, from 1-Jan-1991 to 1-Mar-1994.

We focused on a subset of 15 service variables, since it would not be feasible to analyze the hundreds of service variables that are stored in the SQL database. The service variables that were evaluated are shown in Table 1. These variables were chosen because they span the major categories of data in the summary SQL database and more importantly because they represent variables that are closely managed and tracked by the service and development organizations. Thus, they are of particular interest.

VARIABLE	DESCRIPTION
# APARs Fixed	Total number of APARs fixed, all types
- # TVUA	Number of Valid Unique APARs
- # Route	Number of Sys-Route APARs
- # Invalid	Number of Invalid APARs
- # PE	Number of PTFs (fixes) in Error
# APARs Received	Number of APARs received
# APARs Open	Number of Open APARs
# PMRs	Total number of PMRs closed, all types
- # PS	Number of Preventive Service PMRs
- # IP	Number of Installation Planning APARs
- # DOPs	Number of Defect Orientated Problems
- # NDOPs	Number of Non-Defect Orientated Problems
# PMRs (Level 2)	Number of PMRs handled by Level 2
Days to Solution	Elapsed time from call open to solution given, for calls handled by Level 2
# Users	Measure of size of the installed base

Table 1. Service Variables analyzed in the research.

APARs are defects in code or documentation, as noted previously. They can be broken down into four types, which are Valid Unique APARs, Sys-Routes, Invalids, and PEs or PTFs in Error. A Valid Unique APAR is essentially a genuine defect, and it is created the *first* time the defect is discovered; Sys-Routes are essentially pointers from a Valid Unique APAR that are needed to propagate the APAR to more than one software component; Invalids are APARs that Level 3 rejects; and PEs are defects that are found in the fixes (PTFs) that were sent out to cure an existing defect. We analyzed the number of APARs received, and the number of Open APARs in addition to the above in case there was a backlog accumulating.

The problems (PMRs) can also be broken down into four types as noted in the table. Preventative Service and Installation planning PMRs are created when customers call up to get fixes or information related to upgrading or installing new software. The remainder of the PMRs are generally either DOPs (Defect Oriented Problems) or NDOPs (Non-Defect Oriented Problems). DOPs is IBM parlance for problems that are related to its own code or documentation. The first discovery of a defect generates an APAR and subsequent discoveries generate DOP type PMRs.

The remaining variables that we looked at are the number of PMRs handled by Level 2, and the Days to Solution for the problems handled by Level 2. This latter variable is an attribute of the Service Process, rather than of the product quality. The last variable, # Users, measures the size of the installed base, and was found to be approximately constant for the product analyzed in this research.

2.2 *Customer Satisfaction Survey Data*

The customer satisfaction data came from a mail survey that is conducted quarterly by IBM to assess customer satisfaction with various attributes of each product. The survey is mailed to one quarter of the installed base each quarter and thus the whole installed base is surveyed once per year. The assumption is that

the quarterly sample is representative of the entire customer base. This seems reasonable and there is no evidence to the contrary after many years of operating in this fashion. The typical response rate is about 35%.

The survey is done per product/platform and includes questions that apply to the entire product/platform, as well as questions on key attributes of the operating system and applications. The questions that cover the entire platform are designed to solicit detailed information on particular aspects of the product and supporting processes, such as, documentation, and the software distribution process. The attribute questions measure the key product attributes, namely: **C**apability, **U**sability, **P**erformance, **R**eliability, **I**nstallability, **M**aintainability, **D**ocumentation, **S**ervice, and **O**verall.

The attribute questions account for the majority of the questions on the survey, and thus, the survey is often referred to as the CUPRIMDSO survey, which is based on the first letter of each attribute. The survey typically includes 20 questions and is never more than four pages in length.

The customer is asked to provide a satisfaction rating for each attribute and application, using the scale show in Table 2. There is an example of an attribute question in Table 3. The customer has the opportunity to provide text comments when answering each question, in addition to providing a satisfaction rating.

Satisfaction Rating Scale		
1	=	Very Satisfied
2	=	Satisfied
3	=	Neither Satisfied nor Dissatisfied
4	=	Dissatisfied
5	=	Very Dissatisfied
N	=	Don't Know or No Opinion
_	=	Attribute needing most improvement

Table 2. Customer Satisfaction Rating Scale.

What is your satisfaction with the XXX application ?							
C)	Capability	1	2	3	4	5	N _
U)	Usability	1	2	3	4	5	N _
P)	Performance	1	2	3	4	5	N _
R)	Reliability	1	2	3	4	5	N _
I)	Installability	1	2	3	4	5	N _
M)	Maintainability	1	2	3	4	5	N _
D)	Documentation	1	2	3	4	5	N _
S)	Service	1	2	3	4	5	N _
O)	Overall	1	2	3	4	5	N _

Table 3. Example of an attribute question

The data from the survey is input to, and archived in, a CUPRIMDSO database. The data is then analyzed extensively by a group of dedicated specialists and statisticians, both for that particular quarter and in comparison to previous quarters. The work that is reported here deliberately avoids duplicating any of those analyses, because our objective was to add value by looking for relationships between the boxes (data sets), rather than by analyzing any particular box in more detail than is done currently.

The customer satisfaction data has a number of characteristics that the analyst should keep in mind. The first is that the data is subjective in nature. That is, it is qualitative rather than quantitative. This means that one would not expect to obtain as high a value for the correlation coefficients, as one would from purely quantitative data.

The second point to note is that customer satisfaction may be influenced by a wide variety of factors, including some that are outside the scope of the CUPRIMDO survey. For example, price, product availability, and easy of ordering. These influences are assessed using other surveys and tools, since no single instrument could assess all of the variables. The CUPRIMDSO instrument is aimed at product level characteristics and thus the results presented here are limited to those entities.

2.3 Analysis Procedure

The specific questions that this research attempted to answer were:

- Is there a relationship between the service measures and the customer satisfaction results ? Our initial hypothesis was that there should be a relationship in some instances.
- Are our expectations met in terms of:
  - The number of statistically significant results?
  - Specific relationships, such as number of defects (APARs) and the *Reliability* attribute ?
  - Are there any violations of our expectations ?
- What are the strongest relationships ?
- Which customer satisfaction survey attributes are the most and least influenced by the service variables ?
- Which service variables have the most influence ? In particular, should one concentrate on problems (PMRs), defective fixes (PEs), defects (APARs), or Days to Solution, if only limited resources are available ?

These questions were answered by creating a table similar to the one shown in Table 4. The flat files containing these tables were then read into SAS and the relevant correlation coefficients and graphics plots were produced. The generation of these tables was the most difficult and time consuming part of the research, because it necessitated considerable data gathering, sifting and filtering. The analyses and interpretation of the results is relatively straight forward once these tables are available, but the creation of these tables is fraught with difficulties.

	SERVICE VARIABLES			SURVEY ATTRIBUTES	
QTR	Number of APARs	Number of PMRs	Number to PEs...	% Very Sat. Maint.	%Very Sat. Overall..
1Q91	a	b	c	e	g .....
2Q91	b	l	a	h	j .....
3Q91	c	m	f	k	m .....
.	.	.	. ....	.	. .....
.	.	.	. ....	.	. .....
.	.	.	. ....	.	. .....
4Q93	x	y	z	r	s .....

Table 4: Example of table used for r^2 correlation analysis.

The first difficulty that one encounters is identifying and obtaining access to suitable data sets. Fortunately, IBM has a number of suitable data sets available that span many years, and the main task in the initial stages was to understand the processes that generated the data and how the various measures were computed and used.

The second issue is the fact that this is a large research space. There are a slew of entities and variables that can be changed and one must select a subset of those that are available. For example, the service variables that are analyzed, the level at which does the comparison - component, product, or platform - the definition of what constitutes a customer, etc, can all be varied. The high number of variables means that one must beware of "shopping for significance" among a large group of correlations, because a number of significant results will be found by chance alone [13].

The best way of avoiding this hazard is to state specific expectations before starting the analyses [12]. For example, it would be reasonable to expect that the *Reliability* attribute on the survey would be related to the total number of defects (APARs) and to the number of defect oriented problems (DOPs), whereas a relationship between the *Reliability* attribute and the Days to Solution is less likely. Thus, we enumerated a number of expected relationships before hand. In addition, we computed the number of relationships that would occur by chance alone. The results were checked against these two criteria to help validate the findings. We also examined each significant relationship to determine if it was plausible, and to determine if it was of practical, and not just statistical, significance.

The third problem is determining what unit should be used as the basis for comparison, that is, what should be used for the extreme left hand column of Table 4. For example, one could try to use time, or component names, or customer accounts, or release. Our investigation showed that the two best options were time, and customer account number.

Another issue is determining what measures to analyze and compare. Although, this would seem to be a trivial question it transpires that it can be quite complicated. For example, if one examined the percentage of customers that were **satisfied** with each of the CUPRIMDSO attributes, one would conclude that *Reliability* was the worst attribute. However, the opposite is true. This can seen by *also considering* the percentage of customers that are **very satisfied** with each attribute. It will be found that *Reliability* has by far the highest score for percentage **very satisfied** and this is the reason for *Reliability* having a low percentage **satisfied** score. Thus, one may need to examine more than one measure in parallel.

The last two issues to consider are what time periods, and what time offset, should be used during the comparisons. Although, we experimented with different time periods and time offsets, we generally used a time offset of three months. That is, the survey data for a particular quarter was compared to the service data from the previous quarter. The period used in the per quarter comparison was a quarter's worth of data (by definition), while we generally compared the previous twelve months of service data to three months of customer data in the customer level analysis.

The existence of a relationship and its strength was evaluated by computing the correlation coefficients between each pair of variables, and by examining the bubble and scatter plots for each pair of variables. SAS was used to generate the coefficients and plots. The research made use of three kinds of correlation coefficient, namely, Pearson, Kendall, and Spearman coefficients. It was found that all three were nearly always in agreement with respect to the degree and strength of the relationships between various pairs of variables. This agreement lends support to the conclusions. The use of more advanced statistical methods, such as canonical correlations and decision trees was considered, but did not seem to be appropriate given the small sample size and the limited scope of the study.

3. Results and Discussion

The analyses that were done can be grouped into four main areas. The first study involved a per quarter comparison of the data, and it proved to be useful and illuminating. However, the sample size was small and thus a second study was conducted on a per customer basis to increase the sample size. This was followed by a sensitivity study which was done to evaluate if changes in the sample size impacted the results and conclusions. Finally, a cost - benefit study was undertaken in order to quantify the savings from quality improvement efforts. The results will be presented in that order.

3.1 Correlation: by quarter

This series of analyses involved the comparison of the fifteen service variables shown in Table 1 with the nine CUPRIMDSO attributes, on a per quarter basis. Thus, the three years of data yielded a sample size (N)

of 12. The offset between the two sets of data was usually set to three months.

The results showed that there was a relationship between the service data and customer satisfaction in many instances. This was a particularly valuable finding since this relationship had never been proven previously, even though, many individuals believed that the two items were linked.

The scatter plot in Figure 3 and the correlation coefficients in Table 5 are representative of the results. These two exhibits show that there is a relationship between the two sets of data, although the r^2 values are relatively small. Small values are to be expected to some extent since the survey data is subjective and thus one is unlikely to find really high r^2 values, as noted previously.

N=12		Reliability vs APAR Total	Reliability vs # APARs Received	Reliability vs # TVUA	Reliability vs PMR Total	Reliability vs # DOPs
% Totally Sat						
Pearson	r^2	-.58	-.59	-.54	-.56	-.41
	p	.05	.04	.07	.06	.19
Spearman	r^2	-.60	-.58	-.54	-.61	-.41
	p	.04	.05	.07	.04	.19
Kendall	r^2	-.44	-.40	-.38	-.47	-.32
	p	.05	.07	.09	.03	.15
% Very Sat						
Pearson	r^2	-.41	-.36	-.27	-.62	-.62
	p	.18	.25	.40	.03	.03
Spearman	r^2	-.48	-.42	-.37	-.75	-.67
	p	.12	.17	.23	.00	.02
Kendall	r^2	-.35	-.31	-.23	-.63	-.47
	p	.11	.17	.30	.00	.03
% Satisfied						
Pearson	r^2	.20	.13	.04	.46	.54
	p	.54	.69	.90	.13	.07
Spearman	r^2	.21	.14	.13	.55	.55
	p	.52	.66	.68	.06	.07
Kendall	r^2	.20	.12	.08	.41	.38
	p	.37	.58	.73	.06	.09

Table 5: Expected Relationships: r^2 values for comparisons by quarter.

The values in Table 5 are those for the expected relationships that were established for the *Reliability* attribute before the correlation analyses were run. We see that some, but not all, of the expected relationships are significant. The strongest relationship is between the percent very satisfied with *Reliability* and the Total number of problems (PMRs). Table 5 also shows that there are many more significant results at the 5% and 10% significance levels than would occur by chance alone. This further supports the theory that the two sets of data are linked. The fact that the three types of correlation coefficient - Pearson, Kendall, Spearman - generally identify the same relationships as being significant, have similar r^2 values, and yield a similar number of significant results, increases our confidence in the conclusion.

We also note that none of the expectations in Table 5 are violated, with the exception of the percentage satisfied results. These indicated that there was a positive relationship between percentage satisfied with *Reliability* and the service variables. This is the ***reverse*** of what we would expect, that is, we would have expected a negative relationship in all cases. However, a careful examination of the raw data showed that the reason for the positive relationship was that the percentage very satisfied with *Reliability* is extremely large. Thus, the percentage satisfied must be small and hence we will observe a positive relationship.

The other comparisons exhibited similar characteristics to those in Figure 3 and Table 5. That is, there were many more significant results that would occur by chance alone; non of our expectations were violated; the r^2 values indicated linkages in many instances; and the three kinds of correlation coefficients provided similar results. Thus, we can conclude that the two sets of data are linked.

The additional comments that we can make from a detailed examination and comparison of all of the results are that:

- The three service variables with the "most influence" are the total number of defective fixes (PEs), the number of preventative service problems, and the total number of problems (PMRs). The "most influence" is defined as the variables that were involved in the greatest number of and the strongest relationships. The fact the number of PEs is the most influential variable is reasonable since one can well imagine that customers would become irate with defects in the fixes they receive, even though the absolute number of PEs is very small. This finding also validates an extensive "zero-PE" effort that has been underway in IBM.
- The number of sys-route APARs, the Days to Solution, and the number of Valid Unique APARs, were the three service variables with the least influence.
- The answer to our question as to which of the following had the most influence was that the order is:

  `PE > PMR Total >> APAR Total > Days to Solution`

  That is, defective fixes (PEs) and problems (PMRs) have roughly the same degree of influence and that influence far exceeds that of the number of defects (APARs) or the Days to Solution. In fact the latter had virtually no impact on the customer satisfaction. Thus, if limited resources are

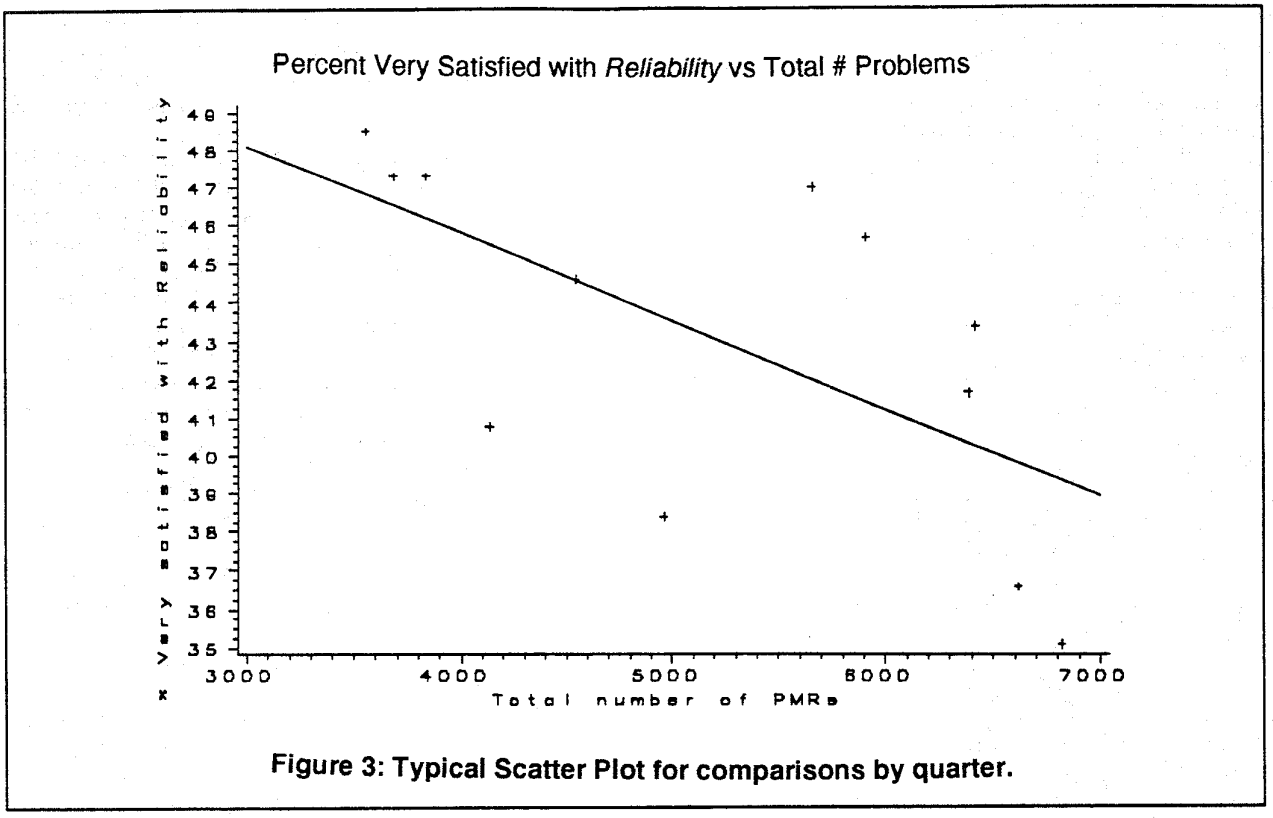

Figure 3: Typical Scatter Plot for comparisons by quarter.

available one should concentrate on the first two items.

- The different CUPRIMDSO attributes were influenced to different degrees by the service variables. In particular it was shown that, the two CUPRIMDSO attributes that were the most influenced were the *Overall* and the *Performance* attributes; while *Maintainability*, *Installability* and *Usability* were the three attributes that were the least influenced.

  These observations seem reasonable and we can provide plausible explanations for them with the exception of the strong relationship with *Performance* which we do not fully understand. We believe that the *Overall* attribute is strongly influenced because all of the service measures improved over the three years - often significantly - and so did the majority of the customer satisfaction attribute ratings. Thus, one would expect a relationship.

  The *Maintainability*, *Installability* and *Usability* attributes were generally the lowest scoring CUPRIMDSO attributes and they did not improve to the same degree as the other attributes over the three years. Thus, one would not expect a relationship with the service variables. We would also note that the service variables that we looked at are unlikely to influence these three attributes. For example, reductions in the number of APARs are unlikely to significantly impact usability.

There are a number of cautions that the reader should keep in mind. The first is that customer satisfaction data is subjective and that there is not necessarily a reverse relationship in all instances. That is, an increase in a parameter may have a positive effect on satisfaction, while a decrease in the same parameter may not have a negative impact, or visa versa. For example, the customer satisfaction analysts found that there is a negative cross correlation between the *Reliability* attribute and *Overall* satisfaction, but there is not a positive cross correlation between these two.

Thus, one must be careful to interpret the results correctly. For example, the

```
PE > PMR Total >> APAR Total > Days to Solution
```

result might lead one to conclude that they can ignore the latter two variables completely. However, this may not be the case as the lack of impact from the latter two variables may be a Maslowian triangle effect. Maslow hypothesized that people have a hierarchy of needs and that they stop worrying about lower level desires once these have been sated, and that they are then free to focus upon higher level desires [14]. For example, if a populace has abundant water, food and employment

they "ignore" these items and they can concentrate on cultural activities such as attending the opera. However, if you subsequently deprive then of food and water they will quickly revert to worrying about these entities, and forget about opera.

The same may be true with respect to the total number of APARs and the Days to Solution. Both of these parameters improved significantly over the three year period in the study and it may be that they reached a level where they were taken for granted, and where further improvements would not positively impact customer satisfaction. Thus, one is free to focus on the first two variables. However, if the latter two were to decline that may well have a negative impact on satisfaction. This is an example of where one might see a negative relationship without there being a corresponding positive relationship. Although, this Maslow theory seems plausible we did not attempt to prove it for the two sets of data at hand.

The second point to note is that the results presented above are for a particular product. While the authors believe that a similar general relationship will be found for many other products, the particular findings are likely to vary by product. Thus, while the number of PEs may be the most influential factor for the product considered here, response time may be the most influential service parameter for a home user software application. Thus, the analyses needs to be repeated for each product.

Other points to remember are that:

- The sample size was small, namely 12. Thus one should not read two much into some of the results, such as the absolute parameter values.
- Correlation does not imply causation [15]. This is true of any correlation study. The antidote is to examine the significant relationships to be sure that they are reasonable. This was done in this research and should be repeated for any new analyses.
- The results presented here are sufficient to gauge the relationship between the factors that were examined in both sets of data. However, one should not infer anything about the service variables or other factors that may influence customer satisfaction, that were not analyzed in this study.
- The last point to note is that the service attribute question is a recent addition to the survey and thus the sample size was very small, namely five. Hence, we did not attempt to draw any conclusions with respect to this attribute.

3.2 Correlation: by customer

This analysis employed a table similar to Table 4 to compare seven service variables to the nine CUPRIMDSO attributes, using the customer number as the basis for comparison. The service variables that were used were the first seven listed in Table 1. The full set of fifteen service variables was not used because the PMR data was not readily available on a per customer basis. The definition of what constituted a customer, and the time periods and time offsets that were used in the comparison were varied. The primary reason for doing this analysis was to increase N, and this analysis was successful in that respect.

However, the results invariably showed that there was no relationship between the service and survey data. This can be seen from the correlation coefficient values in Table 6. These values are typical of what was found for all of the variable pairs across all of the customer level comparisons. None of the expected relationships were significant. In fact there was only one significant results at the 5% level and in that case the r^2 value was small indicating that there was no relationship. Thus, the customer level comparison would lead one to believe that there was no relationship between the service data and customer satisfaction.

	Item	Pearson	Spearman	Kendall
Reliability	r^2	.04	-.13	-.12
vs	p	.77	.36	.34
APAR Total	N	55	55	55
Reliability	r^2	.05	-.009	-.009
vs	p	.72	.95	.94
TVUA	N	55	55	55
Service	r^2	.20	.21	.20
vs	p	.22	.19	.19
# PEs	N	40	40	40
Service	r^2	-.008	-.01	-.007
vs	p	.96	.93	.96
Days to Sol.	N	40	40	40

Table 6: Expected Relationships: r^2 values for customer level comparisons.

However, further investigation shows that this conclusion is incorrect. It was determined that the reason for the lack of correlation was that the same customer generally **does not** use the same identification number for service calls and for the customer satisfaction survey. Thus, it is misleading to compare the two sets of data using the customer numbers contained therein. A number of alternatives were explored to try and circumvent this deficiency but no solution was found.

3.3 Correlation: sensitivity to changes in N

A sensitivity study was conducted to determine if the results and conclusions were susceptible to changes in the sample size. The sample size used in the per quarter portion of the study was 12 and three variations on this were evaluated. These were an N of 13 by adding a quarter of data, and two different cases with an N of 11 that were obtained by dropping a quarter of data.

The sensitivity study showed that the results and conclusions were not affected by changes in N. We still

found evidence of a relationship; there were a similar number of significant results; the two most influenced, and the two least influenced survey attributes were the same; the expected relationships exhibited similar results; the overall order of influence for the service variables was similar; and finally we obtained similar, but not identical, r^2 and p values, for all values of N. The insensitivity to changes in N reinforces our faith in the conclusions.

3.4 *Cost - Benefit Analysis*

The research included a cost - benefit study that was conducted for two reasons. The first was to quantify how improvements in particular service variables would improve the customer satisfaction survey results. The second and more important reason was to try and quantify the "10 X" effort that had been conducted in IBM. 10X was a company wide stretch goal effort to improve product quality by a factor of 10.

The results of the cost - benefit study are summarized in Table 7. This table was created by starting with the Q4 1993 values for three key service variables and estimating how a ***hypothesized*** 50% reduction in these levels would impact a number of representative customer satisfaction scores. The changes in the other satisfaction attributes were also computed using the regression line coefficients but they are not shown in Table 7, due to space constraints.

	Number Defects (APARs) Worldwide	Number Problems (PMRs) USA	Days to Solution USA
Q4 1993 level	A	B	C
Reduce by	50%	50%	50%
New Level	.5A	.5B	.5C
Customer Sat. Improves by			
Attribute	Performance	Reliability	Performance
Measure	% Very Sat.	% Very Sat.	% Totally Sat.
Old	21.7	46.8	86.5
New	24.3	51	87.9
Change	2.6	4.2	1.4
% Change	12%	9%	1.7
Investment Cost	X	Y	Not Applicable
Annual Service Savings	12.7X	18.8Y	Not Applicable

Table 7: Cost - Benefit Study Results.

Table 7 also shows how much one would have to invest in the development process to reduce the levels by 50%, and how much that reduction would save in service costs. These two rows were built using a variety of data sources, including the service costs per APAR and PMR, the number of programmer years required to eliminate a certain number of APARs and PMRs for particular products, and the APAR and PMR rates for those products. We believe that the savings and investment numbers are accurate to within 90% of the true value.

The cost data was not computed for the Days to Solution parameter since changes in it cannot be directly translated to savings or to costs. For example, halving the Days to Solution will not create any savings unless the expended time per call changes. That is, if the expended time per call is 2 hours, then it does not matter if I expend that 2 hours over a day or over five days.

Table 7 shows that improvements in the service data has a positive influence on customer satisfaction. Thus, one need not be afraid that efforts to reduce service costs will negatively impact customer satisfaction. The table shows, for example, that a 50% reduction in PMR levels will improve the percent very satisfied with *Reliability* by 4.2%, which is a 9% relative change. Although, some of the changes may seem quite small, it should be noted that the customer satisfaction survey scores change slowly over time, and a change of a few percent may represent many years of improvement at normal rates.

The savings and investment data provide ample evidence that efforts to reduce service costs are worthwhile. The figures in Table 7, show a minimum 10 to 1 return on investment which is excellent. This 10:1 factor is in agreement with many cost of quality studies, which typically show that, the cost of a defect increase by a factor of ten as it moves from one stage to the next. The savings and cost data in Table 7 are averaged over a number of years in each case and thus one is unlikely to realize a 10:1 return on investment in year one. The investment and savings rates are generally non linear over time, and thus one will have to wait a number of years before the full benefits accrue.

4. Summary

This research established that there is a relationship between several service measures and Customer Satisfaction. The results are based upon an analysis of over three years of actual data for an IBM operating system product. Fifteen service variables and nine customer satisfaction attributes were analyzed. The implication of the results is that we can improve Customer Satisfaction by controlling the relevant service measures. Although, the existence of such a relationship was often questioned and some believed that it existed it had not been proven previously. The main findings are:

1. The four service variables that are mostly commonly tracked, from the fifteen that were analyzed, are the number of defective fixes (PEs), the number of problems (PMRs), the number of defects (APARs), and Days to Solution. This study found that the relative ranking for these four with respect to their influence on customer satisfaction is:

PE > PMR Total >> APAR Total > Days to Solution

That is, defective fixes (PEs) are the strongest driver of customer satisfaction and they are closely followed by the total number of problems (PMRs), while the number of APARs and Days to Solution have little or no influence on customer satisfaction. Thus, if resources are limited, the service focus should be on reducing defective fixes (PEs) and problems (PMRs), rather than on defects (APARs) or Days to Solution.

2. From a causal perspective, if we consider **all fifteen** service variables, the three that are the strongest drivers of customer satisfaction are (a) the number of defective fixes (PEs), (b) the number of Preventive Service problems, and (c) the total number of problems (PMRs).
3. From an effect viewpoint, the *Overall* and *Performance* attributes are the two customer satisfaction attributes that are the most influenced by the service data. The results show that there is little or no relationship between the service data and the *Maintainability*, *Installability*, and *Usability* attributes.
4. The cost - benefit study shows that for each dollar invested in quality improvement efforts one will save at least ten dollars in service costs. Hence, there should be a continued focus on improving the service measures, since this will reduce service costs **in addition to** increasing customer satisfaction.

These findings are specific to the product analyzed here. The authors believe that similar relationships will exist between these two data sets for other products, but that the specific details will vary by product. Therefore, the methodology presented here should be applied to data from other products in order to (a) validate the findings and (b) to determine what the specific links are for other products.

Acknowledgments

We are indebted to a host of people at IBM who provided invaluable insight and guidance, along with access to data and facilities. These include: Tom Byrnes, Bill Spencer, Al Beckmann, John Yang, Bill Bleier, Peter Santhanam, Ram Biyani, Jarir Chaar, and Kathy Bassin. We are especially grateful to Ken Fordyce, Art Nadas and Elliot Feit for contributing many valuable suggestions to the research.

References

[1] C. Jones, "Applied Software Measurement: assuring productivity and quality," *McGraw Hill*, 1991.

[2] W. S. Humphrey, "Managing the Software Process," *Addison-Wesley*, 1989.

[3] B. W. Boehm, "Software Engineering Economics," *Prentice Hall*, 1981.

[4] J. Gray, "Why Do Computers Stop and What Can Be Done About It," *Proc. 5th Symp. on Reliability in Distributed Software and Database Systems*, pp. 3-12, 1986.

[5] X. Castillo, "A Workload Dependent Software Reliability Prediction Model," *Proc. 12th Int. Symp. on Fault-Tolerant Computing*, pp. 279-286, June 1982.

[6] R. Chillarege and D. P. Siewiorek, "Special Issue on Experimental Evaluation of Computer System Reliability," *IEEE Trans. Reliability*, vol. 39, no. 4, 1990.

[7] R. Grady and D. Caswell, "Software Metrics: Establishing a Company-Wide Program," *Prentice Hall, 1987*, 1987.

[8] J. D. Musa, "Software Reliability: Measurement, Prediction, Application.," *McGraw Hill*, 1990.

[9] M. Sullivan and R. Chillarege, "Software Defects and their Impact on System Availability - a study of Field Failures in Operating Systems," *Proc. 21st Int. Symp. on Fault-Tolerant Computing (FTCS 21)*, pp. 2-9, 1991.

[10] R. Chillarege, I. Bhandari, J. Chaar, M. Halliday, D. Moebus, B. Ray and M.Y. Wong, "Orthogonal Defect Classification - A Concept for In-Process Measurement," *IEEE Trans. on Software Engineering*, vol. 18, no. 11, pp. 943-956, 1992.

[11] T. J. McCabe, "Complexity Measure," *IEEE Trans. on Software Engineering*, vol. 2, no. 4, 1976.

[12] M. Buckley, "Computer Event Monitoring and Analysis," *Ph.D. Thesis, Carnegie-Mellon University, Pittsburgh, PA*, 1992.

[13] S. D. Schlotzhauer and R. C. Littell, "SAS System for Elementary Statistical Analysis," *SAS Institute*, 1987.

[14] A. H. Maslow, "Motivation and Personality," *Harper and Row*, 1970.

[15] G. W. Snedecor and W. G. Cochran, "Statistical Methods," *Iowa State University Press*, 1989.