MEASURING CUSTOMER SERVICE SUPPORT

Provided to the Manager of Customer Care & Support

Peta Roberts

August 2000

TABLE OF CONTENTS

1. BACKGROUND

Clients pay for services to support their use of software bought from Company X. These support services are accessible by customers through phone, fax and email for queries on Company X professional management range of software. Customer Care and Support (CC&S) runs the service which customers subscribe to on a yearly basis. Contracts for support services is separate to the original purchase of the software.

CC&S run a weekly phone survey of a sample of clients who have called in the previous 7 day period. The "courtesy call" program began in January 2000 and asks customers their opinion of the service they received when they last called.

2. EXECUTIVE SUMMARY

The nine variables being measured in the courtesy call survey do not capture all the variables that comprise customer satisfaction.

There is no significant difference between satisfaction levels of Platinum, Plus or Assist customers. Nor is there any significant difference between the State the customer calls from.

Women analysts provide a significantly better service to clients than male analysts.

The internal reliability of the scale used varies from week to week, and has only been reliable for about 50% of courtesy calls.

The data from the courtesy calls are not properly treated, and consequently has poor reliability. There is poor version control of questions in the scripts, with the consequence that interpretation of results is quite difficult.

3. RESULTS OF ANALYSIS

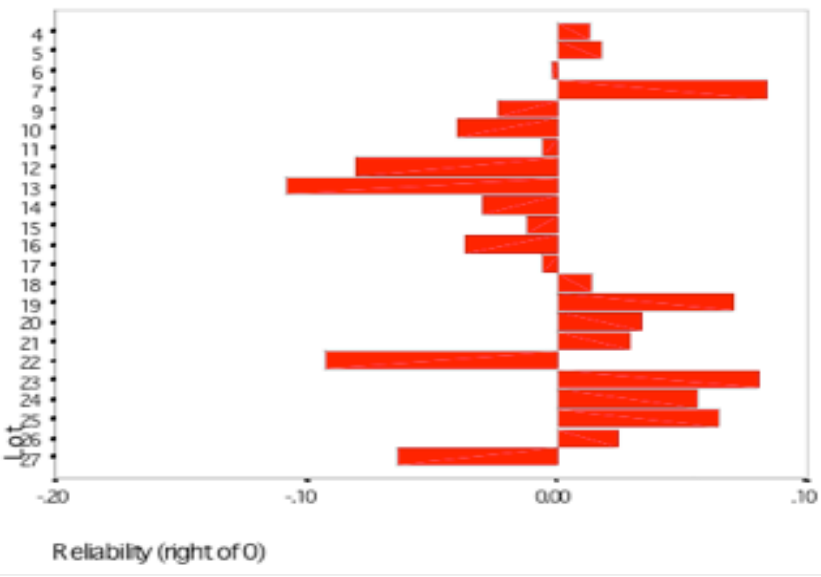
Scale Reliability

The first test to be done was testing the reliability of the scale used. This test shows whether the responses by each customer are internally valid. A perfect score of 1 shows that the survey is perfectly valid, and from1.0 down to 0.85 the results are acceptable as valid. Scores below 0.85 indicate the responses are not consistent.

Half of all lots tested as unreliable, while the other half was reliable. The survey started out as being reliable, but by the end of the first month, it became unreliable. This situation continued for over 2 months, but for the last 2 months, the survey has been reliable with 2 exceptions.

Graph 1 below graphically demonstrates which Lots are reliable and which aren't. (Note: the central 0 position is 0.8500 and everything to the right of it is above that value and everything to the left of 0 is below that reliability floor and hence is unreliable.

Graph 1 – Reliability of scale by each Lot



The implications of poor reliability is that is it difficult to draw any sort of conclusions about general trends in customer satisfaction from the data results.

So what determines reliability? Certainly the more calls that are made the less likely it is to end up with a skewed dataset, and extreme values are less likely to influence the whole of that week. However there is a low correlation between the number of survey calls made within a Lot and the final alpha score, so apart from a minimum number of surveys calls that need to be made to ensure statistical validity, there are other factors which determine the survey reliability.

Good reliability indicates individuals answer the questions using the scale in a consistent fashion. Explanations for the lack of reliability in half the survey results is problematic. Although the questions are tested out on a different sample each week, this should not lead to the significant drops in validity that occurs. It may be that the way the questions are asked by the DCT operator lead to some errors. This is possible given that the scripts are not always followed down to the written words. If different operators are used within the data collection on one Lot, then some changes may occur.

Another explanation is that the way the question is worded itself gives rise to inconsistencies, especially given that changes to the script question appear to been implemented without prior testing on how people understand the question.

This may also complicate the issue if the questions are not tested against each other. Although one scale is used throughout the survey, in some instances the scale does not appear to suit the question.

A final explanation may be that the customer is not reminded of the scale so that by the end of the questions, they may have forgotten the 5 scale points, and hence answer inconsistently. This emphasis the importance of constant close supervision of the carrying out of the data collection stage.

There were 5 of the 9 elements that, if deleted, could improve the reliability of the scale. Sat various Lots, these were TTTP, Accessibility, Rapport, Timeliness, Confidence. The wording of 3 of these 5 elements (Accessibility, Rapport, Timeliness) represents 3 of

4 questions which have been most changed throughout the survey. The fourth element (TTTP) is one of 2 questions which have been changed somewhat but whose meaning remains fairly continuous. The fifth element (Confidence) remained unchanged throughout the survey. However it had the lowest frequency for improving reliability if it were deleted (only 3 out of 24 lots) and had a the lowest impact on the alpha score of any these 5 elements.

Knowledge, understanding, communication and ownership were not indicated as questions which if dropped would improve the reliability of the scale. Interestingly, 2 of these 4 elements (knowledge, understanding) remained unchanged throughout the surveys. Communication changed slightly and ownership was discontinuous.

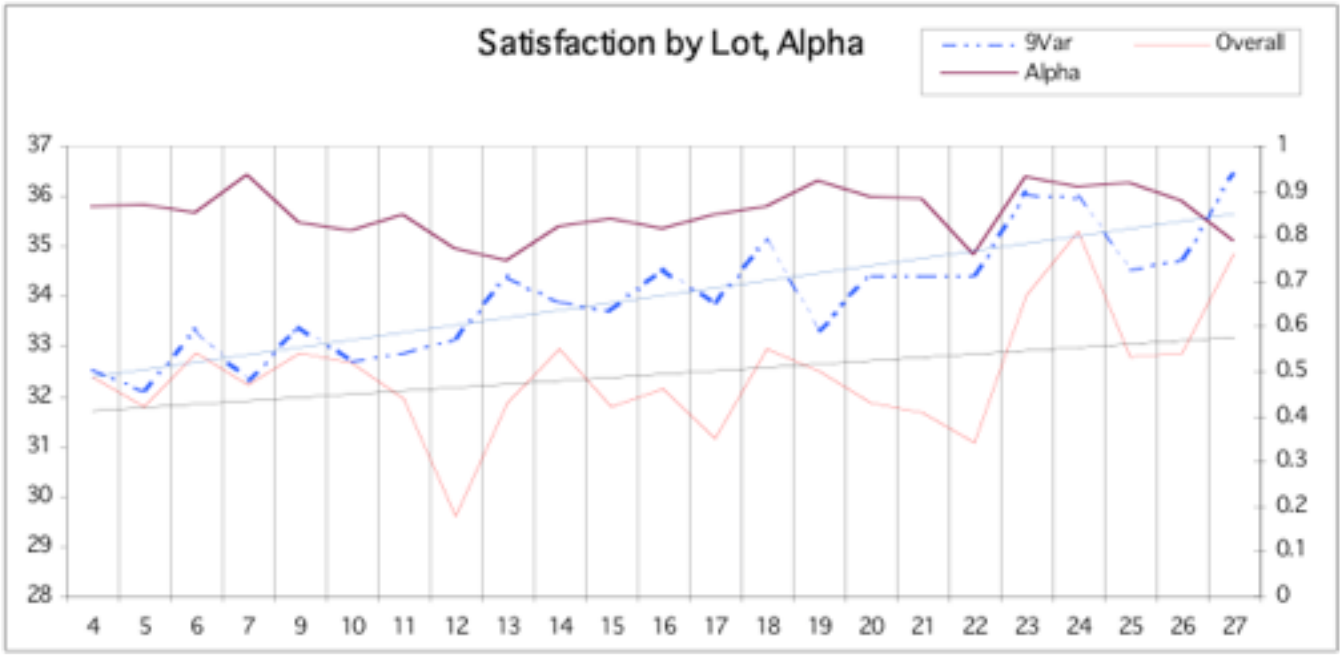
Components of Overall Satisfaction

Of the 10 questions asked, the final question asks the customer for their level of overall satisfaction. Theoretically, this final question represents the aggregate score for the previous 9 questions which dealt with individual components of customer satisfaction.

Hence the final overall satisfaction score was correlated against the sum of the scores for the previous 9 questions, and there was indeed a significant relationship at 0.6. Because it is not a perfect score of 1, there are obviously some elements of customer satisfaction that this survey is not measuring.

Satisfaction over Time

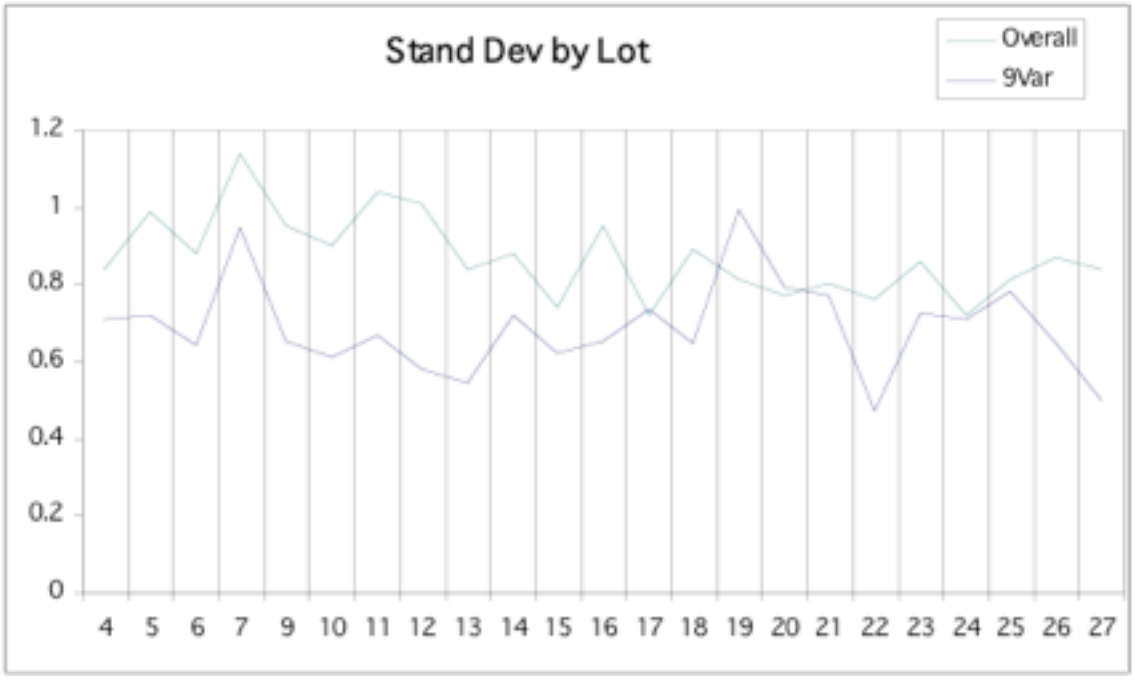
Analysis was done on whether or not satisfaction was improving over time. The following graph indicates satisfaction has been increasing over time. Satisfaction is measured using two scores – the score for the final 10^{th} question on overall satisfaction (Overall), and the combined sum of scores for the first 9 questions in the survey. A trend line indicates the slope of the increase. The graph also has the alpha score for the scale reliability.



The graph above shows that the slope is different for both scores, trending up more slowly on the overall satisfaction than for the sum of the 9 scores. Indeed, the graph shows that the trendlines are actually slowly moving apart.

So which score really measures satisfaction?

The following graph shows that the standard deviation of the two scores for satisfaction. Although they track each other, Lots 17,19 and 20 shows that there are large differences in variation, moving opposite to each other.



Drivers of Satisfaction

Correlation was used initially to discover any relationships between variables measured and overall satisfaction. The results were that:

- There is no significant correlation (0.01) between overall satisfaction and the number of elapsed days between the call being finalised and the courtesy call.
- There is no significant relationship between overall customer satisfaction and the where the customer is located (by State).
- There is no significant difference between the satisfaction levels of the three types of customers – Platinum, Plus, Assist.
- There is no significant difference between product type and overall satisfaction.

Staff Impact

There were 76 analysts named in the survey results. Given that some had left and some were still there, it was decided that results of any test for analysing satisfaction with individual analysts could produce unreliable results. There were a pool of 20 analysts with more than 50 calls to their name which could be used for future analysis.

Instead, a new field of gender was added, and customer's satisfaction was tested on whether the gender of the analyst made any difference to their satisfaction.

Indeed it was found that of all the elements tested, gender of the analyst had a significant impact on the reported overall satisfaction of the customer.

Best & Worst Aspects

Customer's views of the best and worst aspect of the service they received was analysed. Because the questions were open-ended, the answers have had to be interpreted and hand-coded into separate categories. This process itself nay give rise to some bias. Unfortunately the first version of the Courtesy Calls database was re-coded before the second revised edition was received.

Best aspect of last call

	Valid Percent	Frequency
Quick Answer	15.1	98
It gets solved	12.2	79
Friendly, patient staff	11.9	77
Named the staff who helped	10.3	67
Nothing was best	7.2	47
Staff knowledge	5.7	37
Once you get through, its resolved quickly	5.5	36
Is Improving	5.2	34
Good service, try their best	4.9	32
Being there	4.8	31
Good explanations	4.3	28
Got there eventually	3.4	22
Good follow through	3.1	20
Faxed resolution	1.5	10
Treated urgently,as requested	1.4	9
Prompt callback	1.2	8
Better than competitors	.5	3
Contact details of staff	.3	2
Quick via email	.6	4
I got a workaround	.8	5
Sub Total	100.0	649
Nil Response		1578
		2227

The response rate for both questions for this version of the dataset had a response rate of about 30% although this is misleading because not all customers were asked the question. .

Worst aspect of last call

	Valid Percent	Frequency
Poor response time	37.5	253
Nothing (positive)	23.9	161
Still unresolved, no help	5.9	40
Poor knowledge/undertanding	4.3	29
Not treated urgently	3.7	25
Wrong/conflicting advice	3.3	22
Limitations with software	3.1	21
Can't get through to right person	3.0	20
Poor attitude by staff	2.7	18
Poor follow through/lost query	2.5	17
Language barriers, accents	2.2	15
Poor communication of resolution	2.2	15
Named the staff member	1.6	11
Poor service	1.2	8
Hours of operation	.9	6
Unnecessary, if we were told	.7	5
Poor email management	.4	3
Faxes not clear	.4	3
Worse than competitors	.3	2
Poor access to same staff	.1	1
Sub Total	100.0	675
Nil response		1552
Total		2227

Almost 30% of the reasons forwarded for the bets aspect was the fact that the customer's question got answered. The major reason forwarded for the reason the interaction was the worst was the poor response time experienced by customers.

These extra comments may provide a guide to what other factors might be included in the survey which aren't currently and which might help explain the variables for customer satisfaction.

What data isn't compared

Accessibility

Most of the inquiries to CC&S come in via phone, although there are no figures on this. If it were possible with the sophisticated call centre equipment being used by CC&S to log against each call the time taken for the customer to get through, then that could be correlated against the reported satisfaction level of the client. The outcome of this analysis would be an indication of the average maximum time tolerated by customers on hold.

Also if the time on length of initial phone call by customer is recorded, this could also form part of the analysis.

But even this measure does not take into account drop outs, engaged signals, etc experienced by customers.

It also does not take into account the measures of reported call-back email turnaround time etc.

Changes

Service by CC&S changes on a daily basis according to a number of factors. Some of these factors may have greater impact than other changes and could be correlated on a timescale with overall satisfaction. These changes include:

- Processing changes
- Product upgrades
- Fluctuations in CC&S support staff numbers

Other Factors

Dates only are recorded - analysis of effect on satisfaction by day of week was not done.

Benchmarking – need evidence of how well Company X is doing in relation to its competitors

Lost client analysis

Although most of the inquiries to CC&S come in via phone, there are also enquiries by email and fax.

4. SOME NOTES ON THE SURVEY DATA RESULTS

How the Data was treated for analysis

To ensure that all original data remains unadulterated, a copy of the original electronic version of the dataset sent directly form Company X was made and analysis carried out on the copy. The data originally came as an Excel table. It was then manipulated, checked for obvious errors, and then transferred to a coded SPSS table.

There were 24 Lots for examination – Lots 4-28. This was the second dataset given to me by Company X because the first dataset had errors in scores for “Accessibility”.

Although it is master table, it had records for customers that were never called, although the call record was pulled out for sampling. Not all records pulled for sampling were actually sampled.

This dataset was examined for errors. Again there were errors in scores for Accessibility in Lot 28, so this Lot was deleted. Data for Lot 8 was deleted because the rules for the basis of the survey were not observed, and could have introduced bias into the dataset. Instead of calls from the previous week being used, data from 12 months previously was used. This was due to input error from the analyst who typed in 1999 instead of 2000 during the “get “process.

So from an original ‘master’ dataset of 2,492 records, the number dropped to 1,979 when those records where a courtesy call was not made was taken out of the sample. All of these cases were for the current year, and were for customers who were called. Some declined to participate, but this group is important to include because it makes analysing response rates much easier.

Before analysing the dataset in SPSS, some text fields were re-coded – these were State, Customer Contract Type, participation status.

As well, within SPSS Product codes was automatically recoded from text to numbers.

Best aspect and worst aspect were read and categorised. Eventually all open-ended comments for these two questions could be recoded into any of 20 categories. The date formats were standardised, and a new field was included to indicate the number of days separation when the call was finalised to when the survey courtesy call was made.

The data was treated in the analysis by Lot No. Each Lot comprised those closed calls pulled for surveying during that week.

Sample Size

The original program sample data was set to 120 records, but not all calls were included in the courtesy call program. The table below outlines for which Lot No (survey period) how many calls were extracted from the log records, how many survey calls were made, and how many customers participated. Currently the selection of logged calls is on the basis of at least 4 calls per Analyst per sample (week). Hence the data size now varies between each lot.

Response Rate

There were 2,492 records in the database covering from January-July. However this included records of all the calls extracted for use in the survey, although not all of these records were used in the survey. The actual number of calls made in the survey was 1,979. This comprised 1,903 who actually participated and 76 who declined to participate. This gave a response rate of 96.2% overall.

All of Lot 8 is unreliable because it was based on help inquires made in 1999 instead of 2000,and was excluded.

Survey Lot	Calls Logged	Calls Extracted	Survey Calls made	Number Participated	Response Rate
4	N/a	120	87	75	86%
5	N/a	120	86	83	97%
6	N/a	120	88	83	94%
7	N/a	120	92	91	99%
9	N/a	120	93	90	97%
10	N/a	120	95	90	95%
11	N/a	120	87	85	98%
12	N/a	120	72	64	89%
13	N/a	120	97	96	99%
14	N/a	120	95	90	95%
15	N/a	120	89	89	100%
16	N/a	123	97	95	98%
17	N/a	112	85	84	99%
18	N/a	121	93	93	100%
19	N/a	90	77	77	100%
20	N/a	96	91	91	100%
21	N/a	94	86	86	100%
22	N/a	88	77	75	97%
23	N/a	85	75	72	96%
24	N/a	98	87	83	95%

25	N/a	97	84	82	98%
26	N/a	89	76	74	98%
27	N/a	79	70	55	79%
TOTAL		2,492	1,979	1,903	96.2%

Type of Customer

The proportion of Plus and Platinum customers included in the survey is higher than in relation to the total number of Assist customers.

Contract Type	Total No. of Contract Types	Percentage of Contract Types	No. of Survey Calls**	Percentage of Total Survey Calls
Platinum	28	0.6%	33	2%
Plus	64	0.4%	65	3%
Assist	4,546	99%	1881	95%
Total	**4,638**	**100%**	**1,749**	**100%**

** May include repeat calls to same customers and does not count unique number of customers.

5. SUGGESTIONS FOR CHANGE TO THE COLLECTION OF SURVEY DATA

Survey Development

Over the weeks, the scripts had changed numerous times, but the data obtained during each of these weeks was still recorded under the element being measured. So even though the question asked about staff politeness changed drastically at one point, the feedback was still recorded under that element.

There has been poor appreciation that changes to the wording of questions may influence the answers that are received. DCT assured me that for 5 March the current script was used (version 9). This means that form about Lot 14 onwards, the survey questions has remained unchanged.

Only 6 of the 13 questions remained unchanged for the whole period and are illustrated below:

1. Staff Ability/Understanding	Continuous and valid
2. Staff Ability/Knowledge	Continuous and valid
3. Confidence (to re-use the service)	Continuous and valid
4. Best Aspect of the last interaction	Continuous and valid
5. Worst aspect of the lasts interaction	Continuous and valid
6. Any other comments	Continuous and valid
7. Time Taken (to answer query)	Mostly continuous and valid

8. Staff Ability/Resolution (later Communication)	Mostly continuous and valid
9. Accessibility	Discontinuous
10. Staff Ability/Politeness (later Rapport)	Discontinuous
11. Timeliness (to resolve query)	Discontinuous
12. Follow-up (later changed to Ownership)	Discontinuous
13. Overall rating	Discontinuous

See Appendix 1 for more details.

Recommendations for Changes to Script Changes and Data Collection

Between January and presumably March, the script used by the call centre staff had changed a number of times. The first version 2.0 had changed to version 9 by mid-March. It is important to constantly ensure that the questions being asked elicit responses that can be used, and is usefully measuring some aspect of customer satisfaction.

But there has been no version control whatsover over the scripts, which has flowed through to poor data management of the final scores obtained through the surveys. Despite a question being substantially re-written, no attempt has been made to record those scores as a separate data field. Erroneous conclusions may have been drawn from this contaminated data that will not represent the truth.

Attempts have been made to track down which changes were made to the various questions at what time, but neither DCT nor the analyst can tell when the changes were made. If they could it may be possible to re-create the original data set and then use the final results.

It is recommended that where questions be changed, however slightly, that the master data table be amended. A new column has to be created to now record the new scores for the new question, and the old scores for the previously asked questions must remain intact and closed off.

When new questions are added, they ought to be given a new separate field in the database so that the results are quarantined form similarly worded questions.

Recommendations for Data Changes

Date Fin = should be recorded as dd/mm/yy in the same format as the DCT data occurs, rather than as a string. This saves re-keying, and allows for the data then to be manipulated mathematically.

Names of analyst – there were 2 types of spelling of the same person – should be uniform spelling.

Record the week number that the customer had their last contact (from Mon-Fri). Also capturing the day of the week that the resolving call was made may be useful for analysing whether day of the week can affect the satisfaction levels.

Always check dates in the “get” last weeks’ data. I ditched all of Lot 8 because although the courtesy calls were made, these calls were based on calls made by the clients in 1999 due to a typing error by the analyst getting the data.

Recommendations for Sampling Changes

In the database containing the results of the courtesy calls, 20% of the records pulled for sampling were actually not sampled. If each week's sample is checked against this database to ensure that the same client is not called within the same three month period, then the data sample may be skewed if the checking process includes the 20% not yet sampled.

Validity of Data Collection

The results of the data analysis raise some doubts as to the whether the data collection method results in some bias. For example, it is up to the discretion of the surveyor to decide whether or not to ask the 3 final non-mandatory questions. Indeed the records show they were asked in a biased fashion. This bias may also be reflected in the way the previous questions were asked.

Both male and female staff undertook the surveys. Given that there is a gender bias in the satisfaction by customer depending on the gender of the initial analyst, then there nay also be a gender bias in the results of the courtesy calls. This has not been tested for.

Call volume data was nor available, therefore no analysis was done on whether sample size was adequate.

6. CONCLUSIONS

Although this brief initially seemed to be quite a straight forward task to analyse a dataset and draw conclusions for the business, it has proved to be quite a complicated task. It was assumed the dataset has the required data, but to turn data into knowledge about the customers of the business required a lot more than was in the dataset.

For example more information was required on how the data was collected before any of it could be used. More information would have been nice on the call volumes, changes to process, etc so more conclusions could have been drawn.

There are many complicating factors in drawing conclusions for the dataset, and so far only some factors have been identified as important factors in customer satisfaction.