Analysis of Customer Patience in a Bank Call Center

Paul D. Feigin

July 24, 2006

1 Introduction

Recent interest in call center design and operations has been motivated by the ever-growing demand for providing continuous customer service in a wide variety of business settings. Call centers are therefore an important focus of the developing discipline of *Service Sciences*, which attempts to model and analyze service systems with the ultimate goal of improving the quality and efficiency of their operations.

A survey of the fundamentals required in order to understand the call center operations can be found in [3]. Queueing theory models form the basis of much of the analyses of call centers carried out, and include aspects of particular relevance such as abandonment and re-trial activity of customers. Concurrently with the theoretical modelling and analysis efforts, attention has also been directed to obtaining and pre-processing large amounts of detailed data from active call centers. These data are essential for the statistical analyses which are a vital complement to the theoretical modelling efforts. Examples of a variety of statistical analyses can be found in [2].

DataMOCCA ([4]) is a research-oriented data warehouse project aimed at providing an accessible repository of call-by-call records for several call centers. This repository can be used to study questions concerning customer, agent or system behaviors, as well as to provide a means of validating suggested models.

In this report we present an analysis of customer patience in a particular Call Center of a US Bank. In particular, we analyze the survival function of the *time until abandoning the queue*, which is the operational definition of patience. Obviously, this time is mostly *censored* as the vast majority of customers typically receive service before their *patience runs out*.

The queue we analyze is the one that customers enter in order to receive agent service, and is the one that the customer joins after s/he has completed interacting with a Voice Response Unit (VRU) and its follow-up. The VRU is the first station for virtually all customers who call the Center. It should be noted that about 80% of customers complete their service in the VRU, and their patience is not investigated in this report. (In particular, we do not know if these "VRU-only" customers leave the VRU satisfied, having completed their transactions, or whether some of them also abandon the system before satisfactorily completing their service.) Moreover, our analyses will also be restricted to one kind of service queue: that corresponding to Retail service.

We are particularly interested in discovering what may influence a customer's patience. For example, do particular messages about expected waiting time have an influence; and if so, what influence. Alternatively, is their any relationship between a customer's patience and his experience in the VRU stage of his call.

In the sequel we describe various aspects of the unique data resource, which was used for the analyses, and which is part of the DataMOCCA project men-

tioned above.

2 The data

The raw data are call-by-call records of all the calls arriving to the US Bank Call Center over the period March 26, 2001 through October 26, 2003. There are about 1.3 million calls each month that actually join the queue for agent service after the VRU and *post*-VRU activities.

Here, we will only describe the relevant aspects of the DataMOCCA repository, which is maintained as a set Microsoft Access databases. For more information refer to [4].

Each call is divided into one or more segments. A call segment record is constructed for each part of the call. This record provides detailed information on the interaction between the customer and system: including the party answering (VRU, message, menu or agent); the outcome of the interaction (abandoned in queue, terminated by agent, terminated by customer, transfer to another agent, etc.).

For our purposes, we consider only the first sub-call segments of each call. That is, suppose the customer was transferred from one agent to another, or after agent service s/he returned to the VRU and possibly thereafter sought further agent service. In such cases we only consider the waiting time until the first service is received when computing the censored *time to abandon* for the survival analysis.

The first sub-call consists typically of three segments: the VRU segment; a message and/or menu *post-VRU* segment; and an agent queue and service segment.

Our analysis is based on calls during the week from August 11 through August 15, 2003. During this period:

- 1,536,000 calls entered the Call Center system;
- of these 316,000 (20.5%) continued after the VRU interaction;
- of the latter 168,000 (53.1%) sought Retail service;
- of the latter, 162,000 (96.5%) arrived between 7:00am and 11:59pm (the opening hours for this service);
- and of the latter 158,000 (97.8%) went through the message/menu post-VRU station.

2.1 Defining Abandonment

In order to analyze customer patience we need to define exactly what we mean by a customer abandoning the queue before receiving service. We consider only those customers who went through the post-VRU stage.

A customer abandons if his service time was less than or equal to one second (unless the call was terminated during the first second of *service* by the agent). This definition takes into account the fact that those customers who abandon during the first second of *service*, have actually abandoned before the agent has been able to offer any service — in some cases, before he has even managed to *pick up the phone*. Note also that in some cases agents may terminate an incoming call as soon as it rings and in such cases we do not assign the caller to the abandon class.

The time at which a customer abandons is defined to be the waiting time in the queue for customers who abandon. (It does not include the extra second in case s/he abandoned after one second of *service*.)

2.2 The SmartQ system

As part of the post-VRU phase, there are sometimes announcements made by the SmartQ system which are designed to warn customers of heavier system load. When triggered, the system announces an expected waiting time (there are 6 options: <1min, 1min, 2min, 3min, 4min, >5min) and recommends that the customer return to the VRU. One of our interests is in studying the effect of this system on customer patience.

Of the 158,026 calls that entered the post-VRU after requesting Retail service, 16,670 (10.55%) received a SmartQ announcement. They were distributed as follows in Table 1.

SmartQ Announcement	Frequency	Percent
<1 min	13064	78.37%
=1 min	2059	12.35%
=2 min	1122	6.73%
=3 min	201	1.21%
=4 min	75	0.45%
>5 min	149	0.89%
Total	16670	100%

Table 1: Distribution of types of SmartQ announcements.

The above SmartQ announcement is accompanied by a menu suggesting that the customer return to the VRU. The following Table 2 shows that very few customers are actually influenced by this suggestion.

Out of curiosity, one may ask how well the SmartQ system reflects actual

	Continuation after post-VRU		Total
SmartQ announcement made	Return to VRU	Join agent queue	
No	0	141356	141356
Yes	223 (1.34%)	16447 (98.66%)	16670
Total	223	157803	158026

Table 2: The effect of SmartQ on returning to the VRU — August 11 through August 15, 2003.

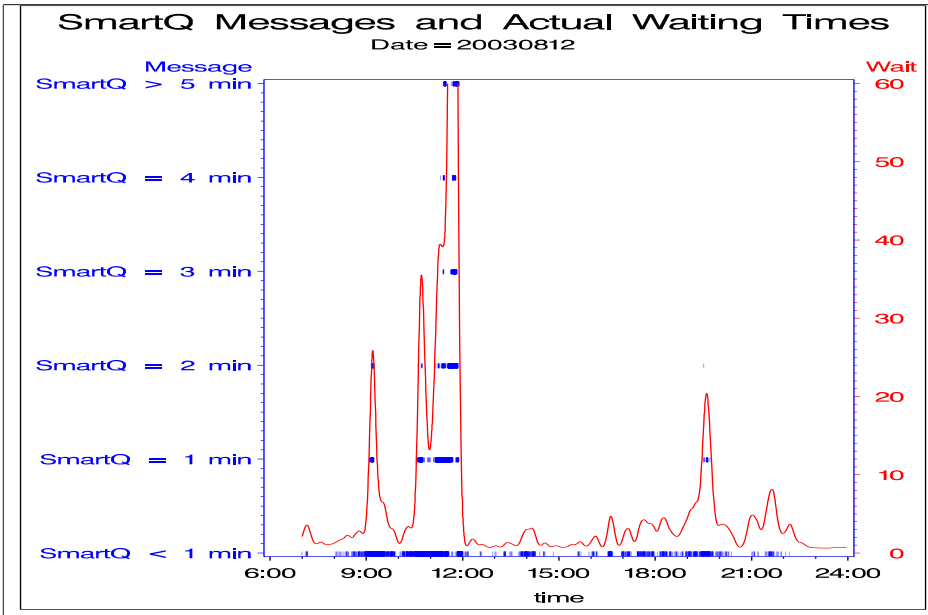

Figure 1: Comparison between density of various SmartQ messages (left vertical axis) and the smoothed waiting time (sec) over the workday for August 12, 2003.

system load on the Retail service agents. In Figure 1, we see a correspondence between the density of the SmartQ messages of varying degrees and the load on the system as reflected in the (smoothed) waiting times. This correspondence shows us that the SmartQ system does reflect the system load to some extent.

2.3 Invested Time before the Agent Service Queue

Among the factors that may affect the patience of a customer waiting for service is how much time he has already invested in the call — that is, the total time s/he has spent in the VRU and post-VRU phases of the call. Histograms and QQ-plots of this *invested time*, as well as of its components, are given in some figures below.

The QQ-plot (Figure 3) for the VRU time seems to suggest a quite good fit to the lognormal distribution. The histogram (Figure 2) does however suggest a bimodality with modes around 30 sec and 60 sec.

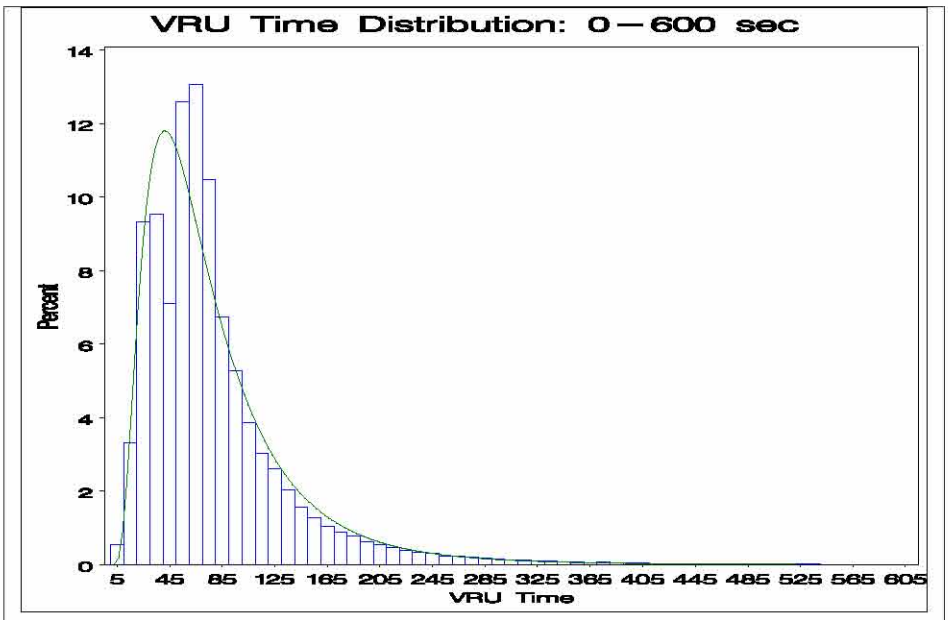

Figure 2: Histogram of VRU time - zoom-in to 0-600 sec. Lognormal fit overlaid.

The overall invested time before the customer seeks agent assistance has a histogram given in Figure 5, and an attempt at fitting a lognormal distribution suggests that the fit is not adequate — see Figure 6.

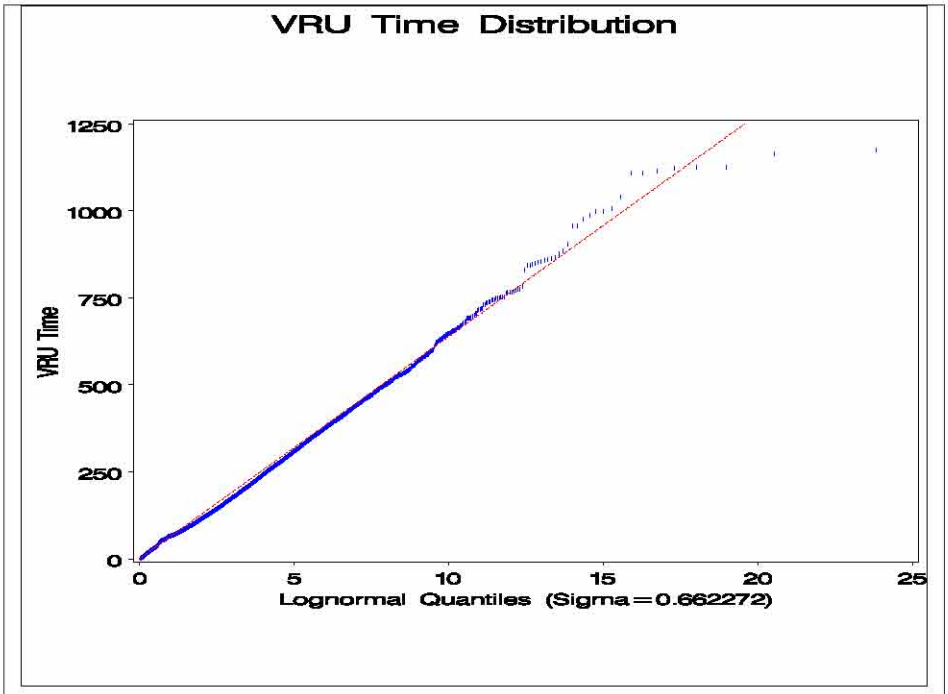

Figure 3: Lognormal QQ-plot of VRU time

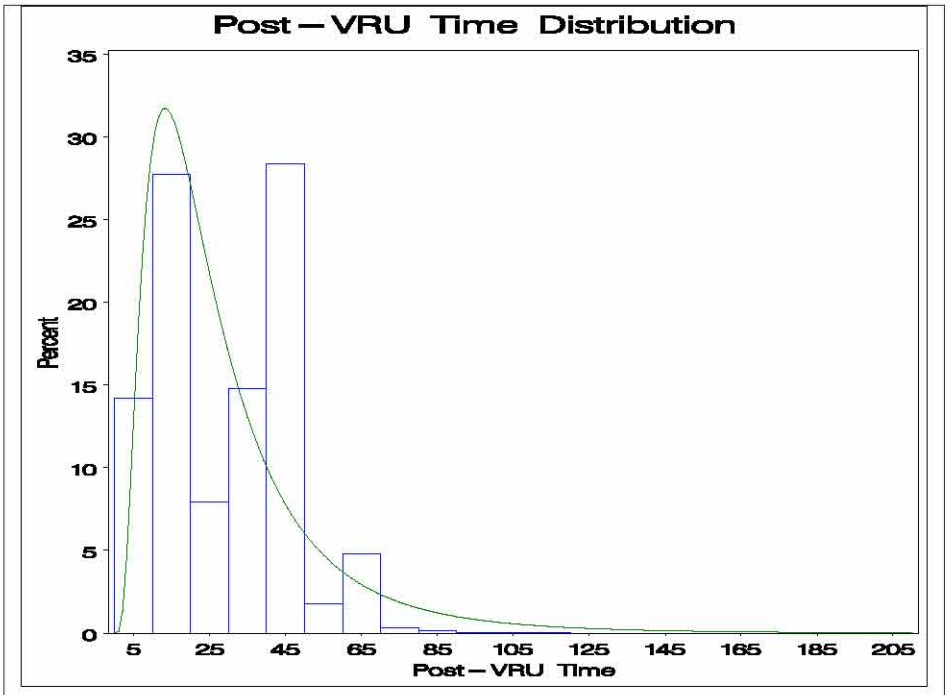

Figure 4: Histogram of Post-VRU time

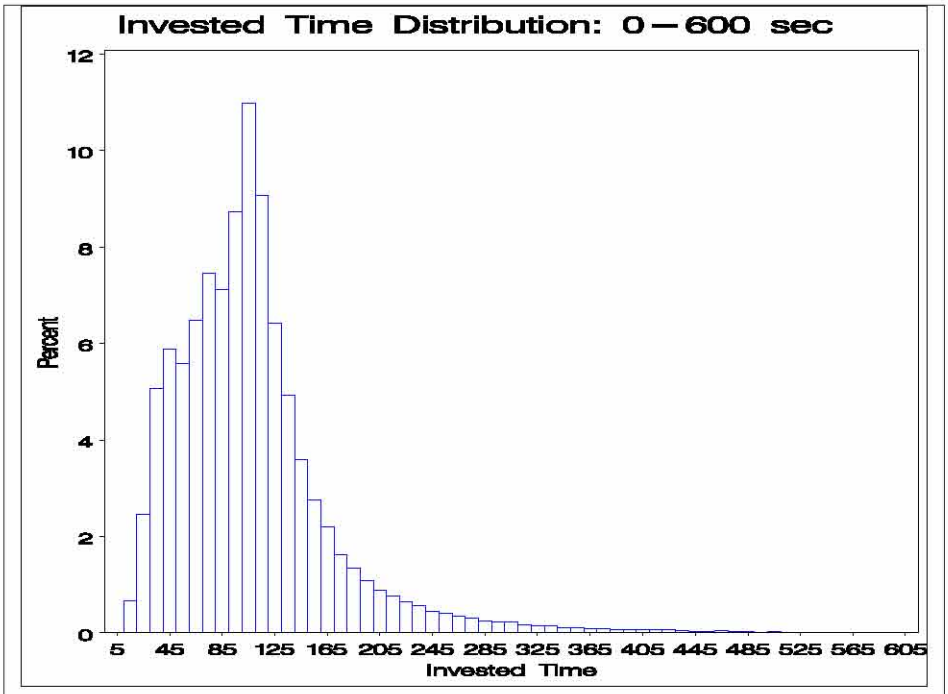

Figure 5: Histogram of total invested time- zoom-in on 0-600 sec

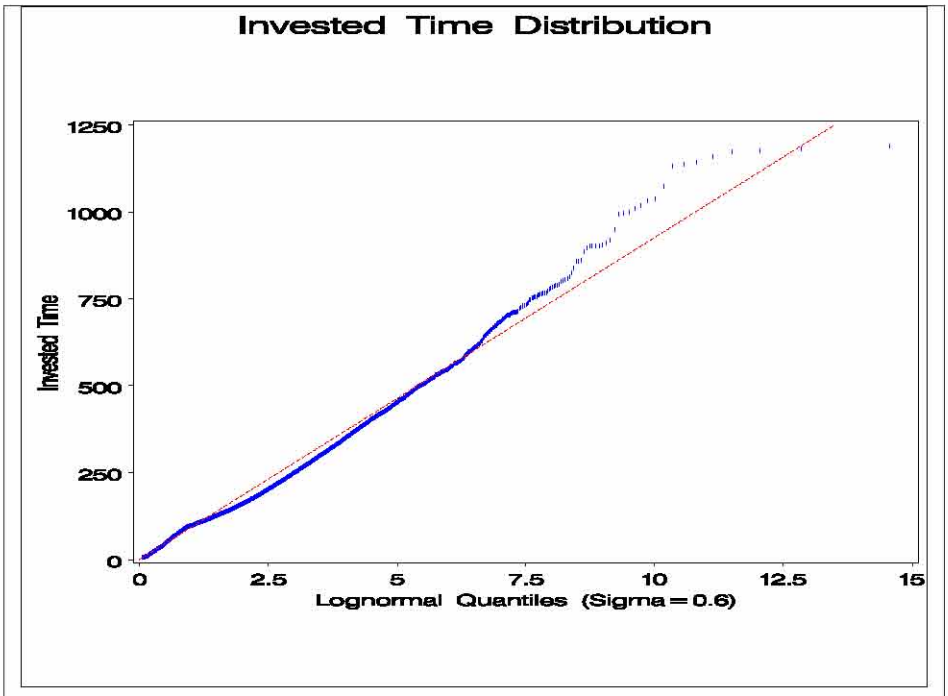

Figure 6: QQ-plot of total invested time

2.4 Proxy for Expected Waiting Time

A common assumption (see [1] and references therein) is that a customer's patience is related to his/her expectation or certainty concerning the true waiting time. One can argue that when a customer calls at a certain time on a certain day, s/he has some expectation of the wait in store, based on past experience with the call center. Thus one might consider a proxy for the expected waiting time to be the average waiting time for each period of the day for each day of the week.

In order to compute this proxy, we use the data for four weeks in August — from August 4 through August 29, 2003 — and compute the average waiting time in each hour period for each day of the week (Monday through Friday). Figure 7 plots these period averages, as well as the weekly average for each period.

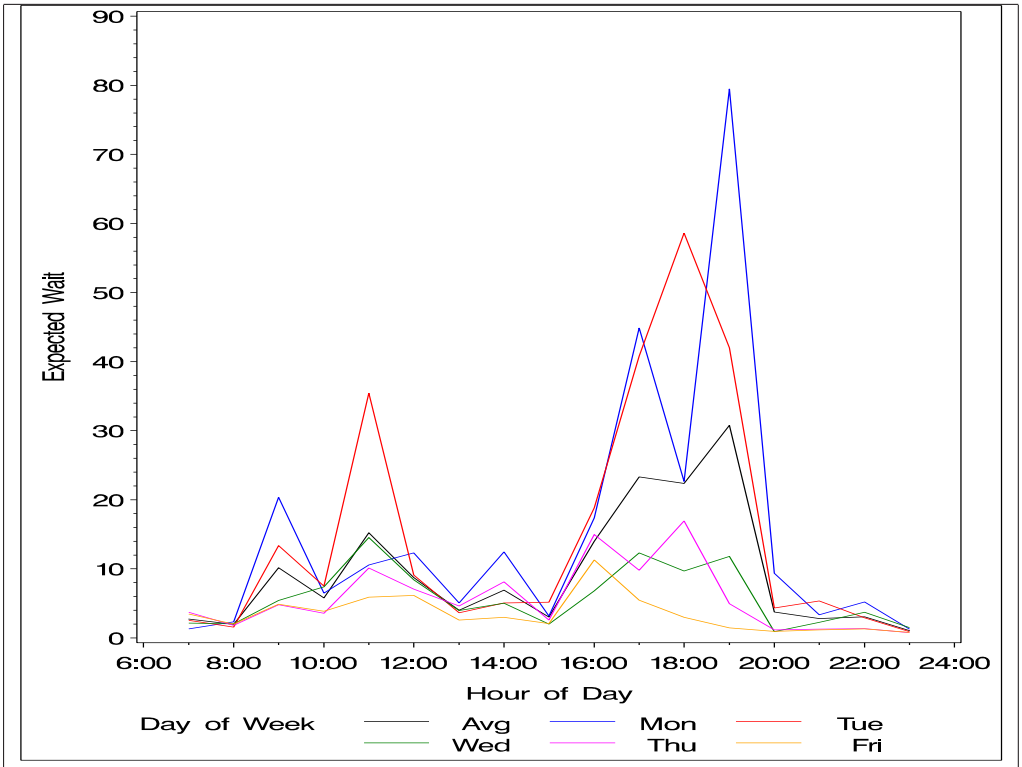

Figure 7: Hourly average waiting times for each day of the week - and weekly average

3 Nonparametric Survival Analyses of Patience

We use nonparametric survival analysis in order to estimate the patience distribution of callers, or of classes of callers. The time to abandon of the majority of callers is censored by the time to receipt of agent service. These analyses are

carried out using the Kaplan-Meier estimators and log-rank tests as provided by the SAS® Lifetest procedure.

3.1 Effect of SmartQ Announcements

Figure 8 shows the two survival curves for customers who did and did not hear the SmartQ announcement, among the 157803 callers who sought agent service after the post-VRU stage — recall that 223 customers returned to the VRU.

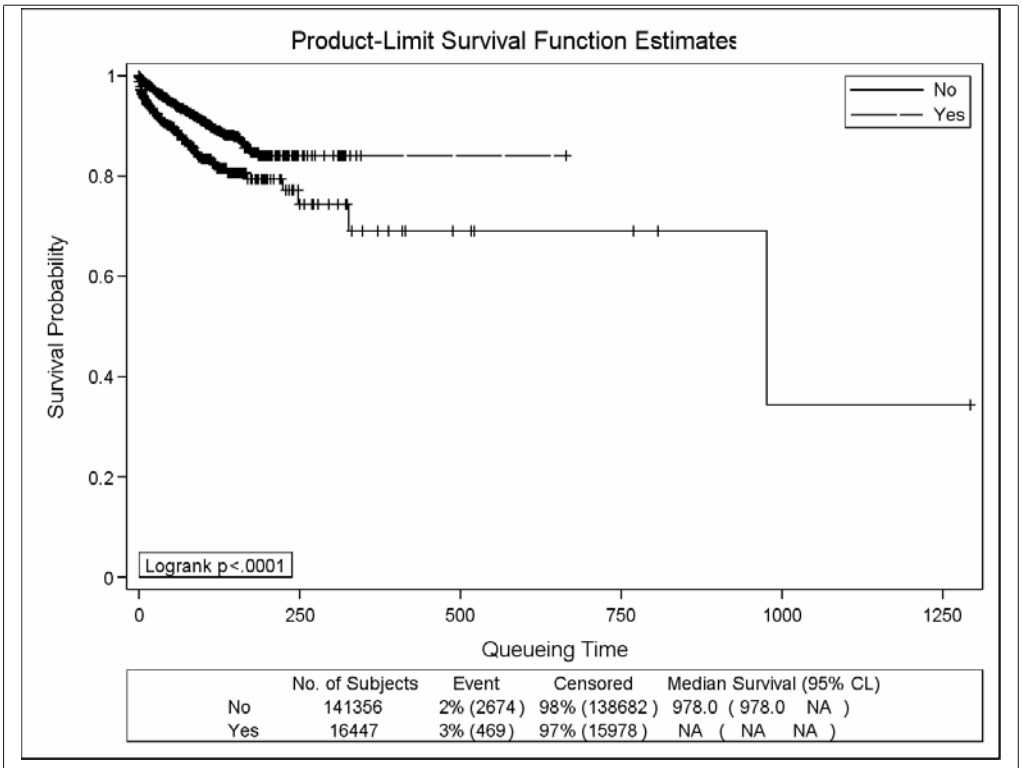

Figure 8: Survival curves for those who did (upper curve) and did not (lower curve) hear the SmartQ announcement

From this analysis we can see a clear increase in patience (=*survival* probability) among callers who heard one of the SmartQ announcements. The curves are statistically significant, but one needs to take care because even very small effects on patience could be detected as significant with such large sample sizes.

It is interesting to note that of the 16447 callers who heard a SmartQ announcement, 2.9% abandoned whereas only 1.9% abandoned from among the 141,356 callers who did not hear such an announcement. The apparent contradiction — the SmartQ announcements make callers more patient and yet more of them abandon — is due to the fact that these announcements occur when the system is under heavy load (see Figure **??**) and therefore even though patience may have been increased, so have the relevant required waiting times.

Of further interest is how the nature of the SmartQ announcement — the amount of expected waiting time — affects customer patience. By considering Figure 9 we see that the simple overall picture of Figure 8 hides some interesting phenomena.

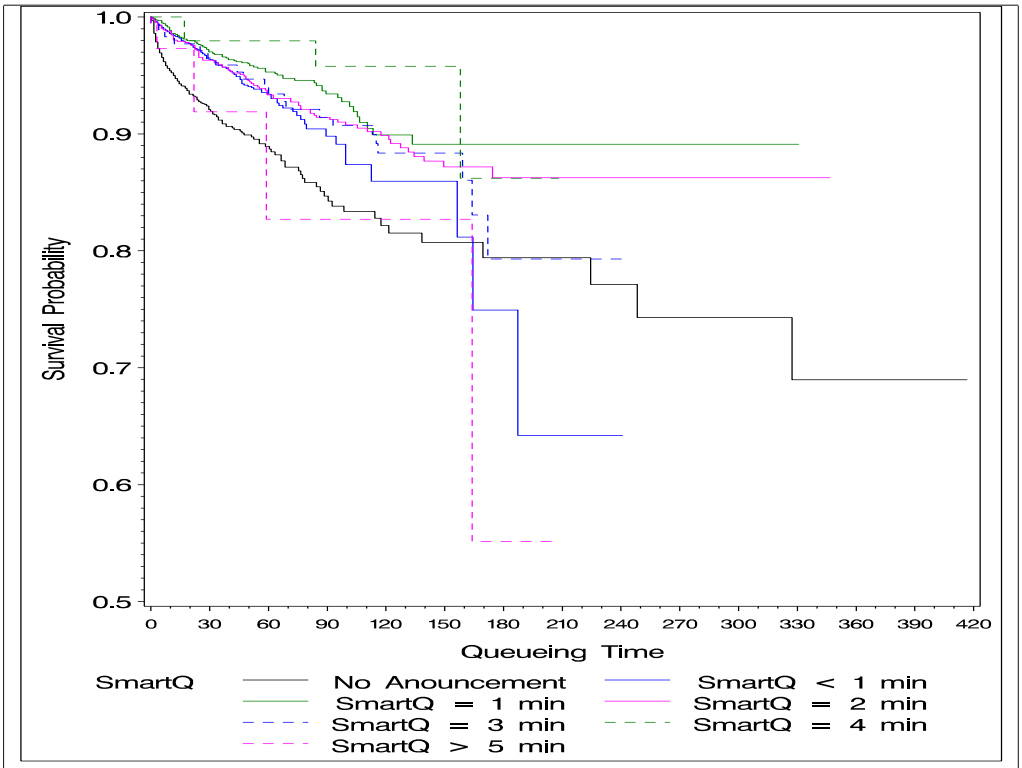

Figure 9: Survival curves for each message level of the SmartQ system — limited to queuing time less than 7 minutes

On the one hand there is a crossing of the survival curves for patience after no announcement compared to that with a SmartQ announcement of a less than a 1 minute wait. This crossing takes place after about 2.5 minutes. We can explain this phenomenon by as follows: telling a caller that there is a wait of up to one minute will cause her/him to abandon with higher probability after waiting 2 minutes than is the case for the caller that received no information about an expected waiting time. Such a caller has reason to believe that something is amiss and that s/he should try again later.

A word of caution when attempting to interpret the survival curve for the > 5 min announcement — it turns out that two thirds of the callers who received this announcement were only required to wait 0 or 1 second, with most of the others being served within one minute. Thus, their patience was not really taxed anywhere near the 5 minutes that were announced. The large drop in the survival curve at 164 seconds reflects a single abandonment among only 3 remaining callers. It seems that this announcement does not accurately reflect the status of the Retail queue, and may be due to some bug in the system.

3.2 Effect of Invested Time

Another factor that may affect a customer's patience when waiting for agent service is how much time s/he has invested in the call already. This time is the sum of VRU and post-VRU times. The median of this total invested time before entering the agents' service queue is 99 seconds. We compare the patience of customers who had shorter invested times (< 100 sec) and those who had longer invested times. We do this for customers who did not receive any SmartQ announcement, so as not to confound the fact that receiving such an announcement will in itself increase the total invested time.

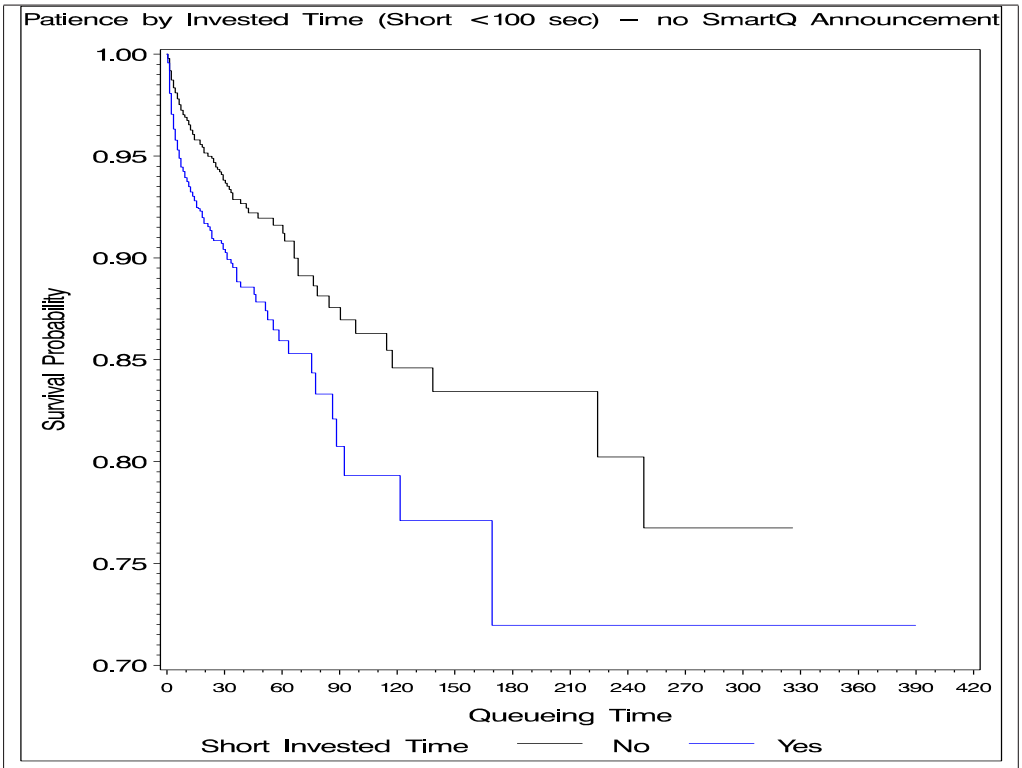

Figure 10: Survival curves for customers who invested less than 100 sec (blue curve) in the VRU and post-VRU stages compared to those who invested 100 sec or more (black curve) — limited to queuing time less than 7 minutes and customers who had no SmartQ announcement.

In Figure 10 we see a clear separation between the two survival curves. Those who have invested more time in the call already, are more patient when waiting for agent service. This difference is highly statistically significant ($p<0.0001$ for the log-rank statistic).

This fact has some quite obvious ramifications for operational management of a loaded system as far as maximizing the amount of customers who will receive service — minimizing abandonment. It suggests that when a customer enters the agent service queue he should be given a priority inversely related (not necessarily linearly) to the time he has invested in the VRU and post-VRU

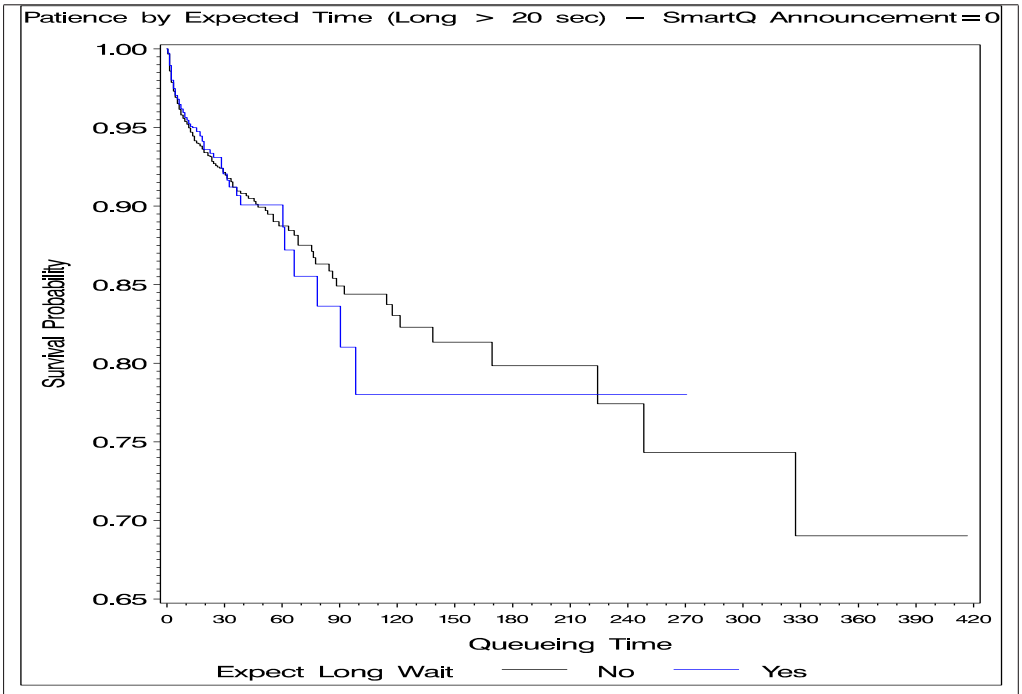

Figure 11: Survival curves for callers during hours of longer expected waiting times (>20 sec - blue line) compared to other hours (black line). Callers who had no SmartQ announcement.

stages. This strategy would mean that those who tend to be less patient would generally wait less, and those who are more patient, as determined by how much time they have already invested, would be allowed to wait longer. The fact that the assignment of such a priority is based on the customer's invested time in the system makes its implementation highly feasible and straightforward. It remains to investigate the effects of such an assignment of priorities using a theoretical queueing model.

3.3 Effect of Expected Waiting Time

For each day of the week, and for each hour, an expected waiting time (proxy) was computed based on averaging over four weeks in August 2003. In order to see if expected waiting time is a factor in determining the patience of customers, we dichotomized the expected waiting time to *long expected wait* (≥ 20 seconds) and *short expected wait* (< 20 seconds). Of the $5 \times 17 = 85$ hourly periods, 8 of them have average waiting times greater than 20 seconds.

In order to compare the patience of customers who call during these typically busier hours, we only consider customers who did not receive a SmartQ announcement. Note that there is very significant confounding between busier hours and hours with a high frequency of SmartQ announcements.

Figure 11 shows little evidence of a difference between the survival curves

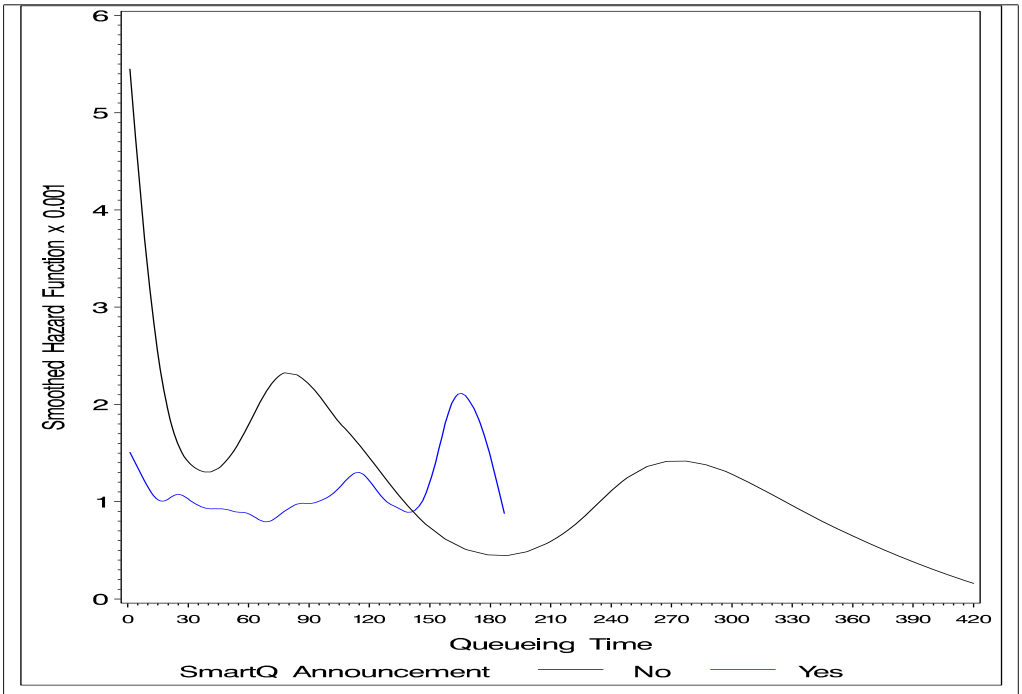

Figure 12: Estimated (smoothed) hazard function estimates for callers who heard a SmartQ announcement (blue line) compared to majority of callers who did not (black line).

for those who call during hours with longer expected waiting times compared to those who call during less busy hours. In fact, contrary to intuition, if there is any difference emerging in the plot (in particular, after 1 minute of waiting) it indicates less patience for those who are, theoretically, *expecting* a longer wait.

4 Hazard Function Comparison

Proportional hazard regression is a popular tool for the analysis of survival data. If its assumptions are valid, one can quantify the effect of various factors by using the hazard ratio.

In order to investigate its use for our data, we consider the estimated hazard functions for the customers who do and do not receive a SmartQ announcement. As well as evaluating the appropriateness of a proportional hazards model for the effect of the SmartQ announcement, looking at estimated hazard functions provides further insights into the nature of the effect.

From the estimated survival function ($S(\cdot)$), we estimated the cumulative hazard function ($\Lambda(\cdot)$) by computing $\Lambda(t) = -\log(S(t))$. Then the hazard function itself was estimated by differencing and dividing ($=\Delta\Lambda(t)/\Delta t$) and then smoothing. The resultant hazard function curves appear in Figure 12.

First of all, it is clear from Figure 12 that the proportional hazards model

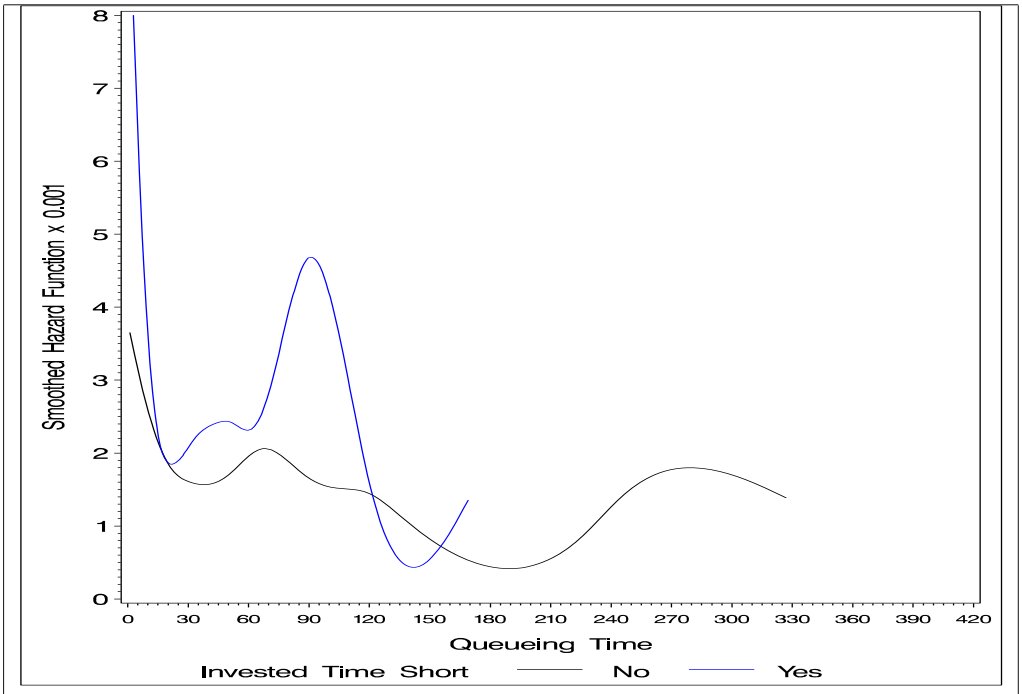

Figure 13: Estimated (smoothed) hazard function estimates for callers who invested a short time (< 100 seconds) before the agent service queue (blue line), compared to the callers who had invested more time (black line).

will *not* supply an adequate description of the effect of the SmartQ announcement, as the the two curves have quite different shapes. The effect of the announcement seems to be as follows:

- it reduces the initial relatively high abandonment rate by callers who otherwise, soon after entering the queue, tend to leave (possibly because they have no idea how long their wait will really be);
- it maintains the caller's patience beyond the 70–80 second mark, when again there is an increase in tendency to abandon for callers who heard no announcement; and
- there is a new peak of abandonment around 170–180 seconds which does not appear for callers who have held out this long among those who have not heard a SmartQ announcement.

A similar analysis was carried out for the estimated hazard function for callers who invested a short time (< 100 seconds) in the VRU and post-VRU phases compared to that for callers who invested a total of more than 100 seconds in those phases. Only callers who did not receive a SmartQ announcement are considered here.

In Figure 13, we can see that having invested a longer time before entering the queue causes callers to be initially much more patient (lower *hazard* for

early abandonment) and it also reduces the tendency to lose patience after 90 seconds.

5 Conclusions

The analyses described above are for a call center of a particular US bank. Some results may reflect particular properties of a bank call center, or even of the particular bank in question. For example, longer waiting times at certain times of the day may be the result of changing staffing policies, and may not be predictable by customers — see the apparent lack of effect of expected waiting time of patience discussed in Section 3.3.

Nevertheless, there are several interesting phenomena that are worthy of further theoretical and empirical study. Among them is the non-negligible effect of invested time on a caller's patience when waiting for agent service. Having invested more time in the pre-queueing stages, a caller tends to be more patient — a fact that could be used to manage priorities for customers entering the agent service queue.

Another phenomenon that is likely to have broader validity is the effect of announcements concerning anticipated waiting time. Such announcements seem to have the effect of prolonging a caller's patience, although there is some evidence to suggest that if the information given is not in accordance with the actual ongoing waiting experience, then the caller will abandon more readily than if s/he had not been given any information.

References

[1] Mor Armony, Nahum Shimkin, and Ward Whitt. The impact of delay announcements in many-server queues with abandonment. June 2006.

[2] Lawrence D. Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical analysis of a telephone call center: A queueing-science prespective. *JASA*, 100(469):36–55, March 2005.

[3] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5:79–141, 2003.

[4] Valery Trofimov, Paul Feigin, Avishai Mandelbaum, and Eva Ishay. DataMOCCA – Data MOdel for Call Center Analysis. Technical report, Technion, Israel, 2003.