A Call Center Uses Simulation to Drive Strategic Change

Robert M. Saltzman
saltzman@sfsu.edu

Information Systems and Business Analysis Department
San Francisco State University
1600 Holloway Avenue
San Francisco, California 94132

Vijay Mehrotra
vijay@onwardinc.com

Onward, Inc.
888 Villa Street, Suite 300
Mountain View, California 94041

A large, customer-focused software company relied on simulation modeling of its call center operations in launching a new fee-based technical-support program. Prior to launching this rapid program, call center managers were concerned about the difficulty of meeting a proposed guarantee to paying customers that they would wait less than one minute on hold. Managers also wanted to know how the new program would affect the service provided to their existing base of regular, nonpaying customers. We quickly developed an animated simulation model that addressed these concerns and gave the managers a good understanding for the impact on system performance of changes in the number of customers purchasing the rapid program and in the number of agents. The one-minute guarantee would be fairly easy to achieve, even if the percentage of callers in the rapid program became quite high. Managers also gained confidence that, with appropriate staffing levels, they could successfully implement the new program, which they soon did.

Call centers are locations "where calls are placed, or received, in high volume for the purpose of sales, marketing, customer service, telemarketing, technical support, or other specialized business activity" [Dawson 1996, p. 35]. Inbound call

0092-2102/01/3103/0087/$05.00
1526–551X electronic ISSN
This paper was refereed.

SIMULATION—APPLICATIONS
INDUSTRIES—COMPUTER-ELECTRONIC

centers, which predominantly receive calls rather than initiate them, face the classical planning problems of forecasting and scheduling under uncertainty, with a number of industry-specific complications [Mehrotra 1997]. Well-known cases include L. L. Bean's phone-order business [Andrews and Parsons 1989], the IRS's toll-free taxpayer information system [Harris, Hoffman, and Saunders 1987], and AT and T's operational design studies [Brigandi et al. 1994].

In their planning, most call centers target a specific service level, defined as the percentage of callers who wait on hold for less than a particular period of time. For example, a sales call center may aim to have 80 percent of callers wait for less than 20 seconds. A related measure often used to assess call center performance is the average time customers wait on hold, or average speed to answer (ASA).

Eighty percent of calls waited five to 10 minutes.

Another key performance measure is the abandonment rate, defined as the percentage of callers who hang up while on hold before talking to an agent. It is both intuitive and well known throughout the call-center industry that customer abandonment rates and customer waiting times are highly correlated. High abandonment rates are bad for several reasons, most notably because a customer who hangs up is typically an unsatisfied customer and is much less likely to view the company favorably [Anton 1996]. Many analysts (for example, Andrews and Parsons [1993] and Grassmann [1988]) have directly modeled an economic value associated with customer waiting time and abandonment. Furthermore, such callers will with some probability call back and thereby help sustain the load on the system [Hoffman and Harris 1986]. Thus, call center managers must balance service level (a chief driver of customer satisfaction) with the number of agents deployed to answer the phone (comprising 60 to 80 percent of the cost of operating a call center).

A Specific Call Center Problem

We conducted our analysis for the managers of a technical support call center of a major software company. At the time, September 1995, it was the only company in its industry that provided free technical support over the phone to its software customers. Agents often provided a combination of software solutions and business advice to customers, and the customers perceived the service provided to be valuable. The software company's customer base and sales volume were growing rapidly.

Unfortunately, its call center had very poor service levels, with 80 percent of calls waiting five to 10 minutes, well above the firm's three-minute target. Predictably, abandonment rates were high, approaching 40 percent on some days.

To senior management, this situation was intolerable.

At first glance, the solution seemed obvious: simply add more agents on the phones, reducing waiting times and agent utilization. However, because this would increase the percentage of total revenues dedicated to free telephone technical support, it would adversely affect the bottom line.

Another obvious solution, supported in different parts of the organization, was to charge all customers for technical support—as the competition did—to cover the cost of more agents. However, there were two huge problems with this approach. First, the company had aggressively marketed the software as having free technical support. Second, to abruptly begin charging all customers for support would create a very negative impression, particularly in light of the currently poor level of service being provided.

To address what was widely perceived as a crisis, the company conducted many high-level meetings involving various parts of the organization, including the call center, product marketing, finance, and information systems. Within a few days, a proposed compromise solution emerged: the rapid program.

Company managers conceived of the rapid program as an optional service to be offered to customers who needed quick telephone support. For a fee, customers would receive priority in the telephone queues. The company would guarantee that their wait to speak to an agent would be less than one minute; if they waited longer, their calls would be free. The remainder of the customer population, which we refer to as regular customers, would continue to receive free technical support over the telephone.

The marketing group drafted a customer mailing describing the program as an added benefit. The information-systems group put a team together to examine how to modify the call-center agents' desktop systems to include billing capabilities. The finance group let out a sigh of relief.

The call center managers, however, had to determine very quickly whether to implement the rapid program. Specifically, they had to determine how many agents they would need to meet service goals for both rapid and regular customers. They were particularly concerned with the impact of priority queuing for rapid customers on the already-abysmal waiting times for the regular customers.

They commissioned us to conduct an analysis to help them understand the impact of the new rapid program on the call center's overall performance. The analysis had to be done in less than a week because the decision to launch the program was imminent. Our clients had very little knowledge of quantitative modeling, and they made it very clear to us that they were not interested in any particular methodology. Rather, they needed specific results that would help them make a sound decision on time.

The one-minute guarantee to rapid customers would be easy to achieve.

Solution Approach

Based on our understanding of the proposed rapid program, we thought that developing a small, animated simulation model would be the best way to tackle the problem. We had several reasons for this. First, the model would have to allow for two priority classes of customers. Second, a simulation model could represent one of the most important dynamic features of this system, call abandonment, while gathering output on a variety of performance measures of interest to management, including abandonment rates and service levels. Finally, the transient phase of the system accounts for a significant portion of the day and would have to be included in the overall measures of performance. Pilot simulation runs showed that, under most scenarios, the system did not reach steady state for at least an hour after the opening of business. Most analytical queuing models, on the other hand, assume the system is in steady state.

We were unaware of any analytic methods that could accommodate multiple priority classes, call abandonment, and both transient and steady-state phases. While the model results that we ultimately presented to management did not rely heavily on the transient phase, the animated simulation approach allowed us to demonstrate how the call center traffic within one time period (for example, 8:00 am to 8:30 am) would influence abandonment and waiting times for each class of customers in other intervals (for example, 8:30 am to 9:00 am). We thought that managers seeing system dynamics like this would be more likely to understand our model and thus be more likely to adopt its results. This advantage of animated simulation has been noted by many researchers, for example, Brigandi et al. [1994].

In hindsight, we realize it may be possible to approximate this call center reasonably well with an analytical queuing model. Green and Kolesar [1991], for example, explored how to estimate long-run average performance measures for a multiserver queuing system with nonstationary arrival patterns over a range of parameter values. Whitt [1999] demonstrated that a multiserver queuing model with certain assumptions about reneging can be developed to predict the distribution of customer waiting time. These results, however, do not apply to the transient phase nor do they address the issue of customer priority classes.

Moreover, at the time we were much more familiar with the latest advances in simulation modeling than with those in queuing theory. Consequently, we were confident we could conduct the necessary analysis in a week with simulation but were (and still are) unsure that it could be done with a queuing model.

Key Inputs and Data Sources

In modeling call center operations, data collection is a multifaceted challenge. One primary data source is the automatic call distributor (the ACD or phone switch). While the typical ACD stores a huge amount of raw data, our client had access to this information only through a handful of prepackaged reports, which provide aggregated results and little distributional information. Other model inputs were derived from (1) the results of management decisions, (2) raw data captured through other systems, and (3) business-planning assumptions.

For the rapid simulation model, we used five key types of input data.

The service-level targets we used in this model were provided by call center management. The target service level for rapid callers (80 percent of calls answered within one minute) was driven by marketing's desire to provide an attractive guarantee to entice customers to purchase the rapid-service option. The regular-caller target was far longer (eight minutes) and was viewed by management as an unfortunate consequence of the current budget constraints.

There was a great deal of uncertainty about what proportion (P_1) of callers would purchase the rapid-service option. Market research showed that many customers claimed to be willing to pay for faster service, but the many variables (pricing, seasonality in call volume and call content, new-product sales, and so forth) prevented it from generating anything more than a broad range of values.

Call center management forecast call volumes. Although we recognized their importance, we omitted time-of-day and day-of-week arrival patterns from our initial study for two reasons. First, we had to conduct the initial analysis quickly. Second, the client's primary emphasis was to understand the dynamics and trade-offs between key variables in the system.

A decision delayed can be worse than a bad decision.

We took the value of the average call-handling time (15 minutes) from the call center's ACD system; this value includes time spent talking to the caller and time spent afterwards on after-call work, or wrap-up time. Unfortunately, we could not obtain detailed distribution data from the ACD's reporting system, which provided only average handling times for different periods. However, we assumed that the call-handling-time distribution was exponential, an assumption we validated by plotting and analyzing call-by-call data from another database within the call center.

Determining a distribution to represent customer abandonment behavior was difficult. This is a standard problem in call center modeling, because observed abandonment behavior is an output of specific conditions (waiting times, customer tolerance) but must be represented as an input for planning models like ours.

We started with two basic assumptions. First, no matter how long the waiting time, no more than 40 percent of customers would ever abandon the queue. This was based historical data for time periods in which customer hold times were extremely long, when abandonment rates had peaked at about 40 percent. Management's feeling was that the rest of the customers simply had no choice but to wait to get technical support because of the importance of the company's software. Second, some customers (five percent) would abandon the queue as soon as they were put on hold. These can be viewed as customers whose problems are not very urgent or important. We made this assumption because historical data showed this level of abandonment even when customer waiting times were very short.

Using these data, we modeled the likelihood of a customer's abandoning the queue as a linear function of the time spent on hold. Specifically, for every two additional minutes spent on hold, our model assumed that the probability of abandoning the queue would increase by 3.5 percent, up to a maximum of 40 percent for a wait of 20 minutes or longer (Figure 1).

We would have liked to have had more detailed historical data about customer abandonment from which to generate a

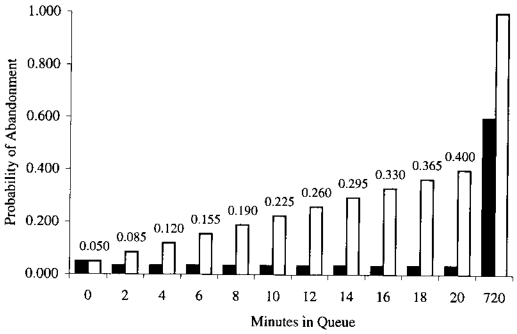


Figure 1: The hollow bars show the cumulative probability that a caller will abandon the queue after being on hold for various lengths of time. We assumed that five percent of callers would abandon as soon as they were put on hold. The solid bars show that for every two additional minutes spent on hold, the probability of abandoning increases by 3.5 percent, up to a maximum of 40 percent for a wait of 20 minutes or longer. This implies that 60 percent of callers would never abandon the queue, no matter how long their waiting time.

more sophisticated distributional model. However, we discussed our approach with the call center managers, and they found it to be a reasonably good representation of actual customer behavior.

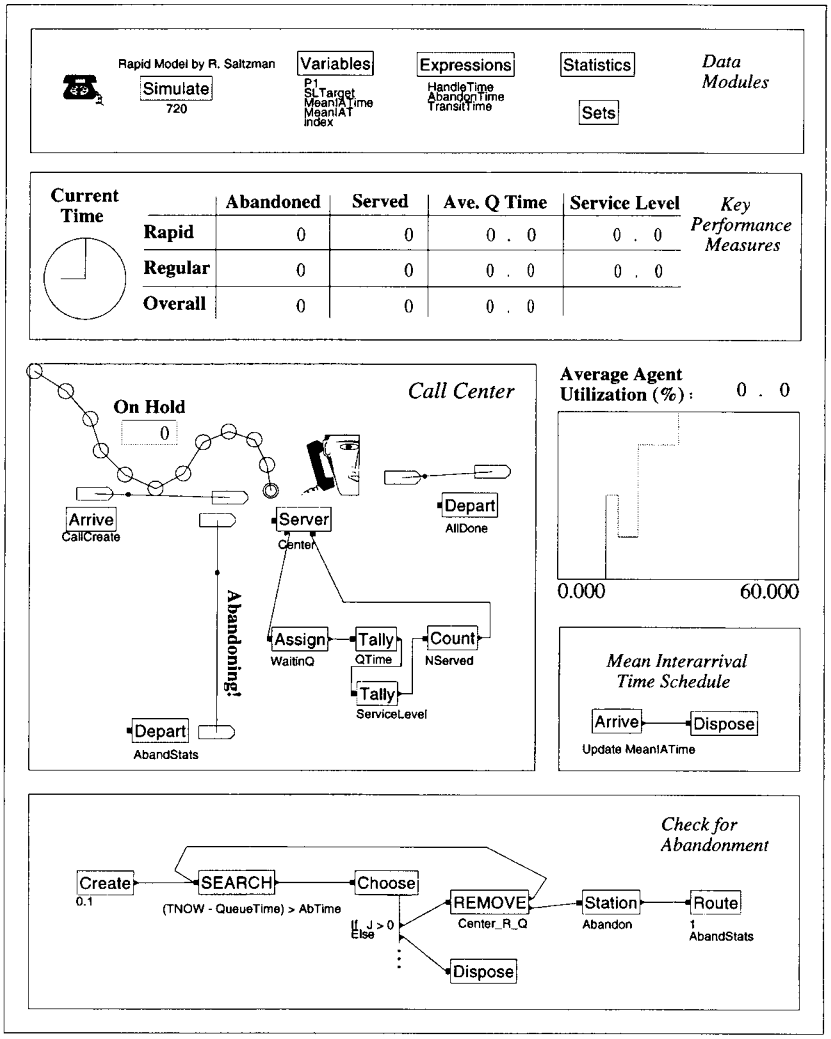
Model Description

Simulation modelers are generally advised to avoid building models that make a one-to-one correspondence with the elements of the real system under study [Pegden, Shannon, and Sadowski 1995, p. 32]. Our model, however, does represent a few key aspects of the call center essentially at full scale, that is, we used $S = 80$ to 90 agents and a large volume of callers typical of the actual system. We did this for two reasons: first, queuing systems are known to exhibit nonlinear behavior (for example, in terms of mean customer waiting time) as the number of servers increases, and second, the model would have greater credibility with management. In other respects, we kept the structure of the model simple because we had less than a week for design, programming, debugging, and analysis.

We built the simulation model using Arena, a module-based graphical simulation software package [Kelton, Sadowski, and Sadowski 1998]. Figure 2 shows the entire model as it appears on the screen in the Arena environment prior to execution. When executed, the model generates an animated picture of the call center's key operations.

We defined several essential run characteristics, such as the number of replications and the length of each replication, in the Simulate module. The Variables module contains model parameters, which were held constant for a given scenario, such as the percentage of rapid callers (P_1) defining the call mix, the service-level-target answer times (SLTarget) by call type, and the mean interarrival time (MeanIATime). In the Expressions module, we specified random variables, for example, the time to serve a call (HandleTime) and the amount of time a caller waits on hold before abandoning (AbandonTime). The Statistics module defines and writes out performance measures from each replication to data files that can be examined later using Arena's output analyzer. Finally, related objects, such as counters for the number of calls served by customer class, are grouped in the Sets module; this allows specific set elements to be referenced during execution by using the customer-class attribute as an index (1 for rapid customers, 2 for regular customers).

During execution, the model tracks a



	Abandoned	Served	Ave. Q Time	Service Level
Rapid	0	0	0 . 0	0 . 0
Regular	0	0	0 . 0	0 . 0
Overall	0	0	0 . 0	

Figure 2: We analyzed the call center using an Arena simulation model. The figure shows the entire model as it appears on the screen in the Arena environment. Five data modules at the top define essential run characteristics of the model, variables, expressions, and performance measures about which statistics will be gathered, such as the number of callers of each type who abandon the queue and who reach an agent. The 3 × 4 table continually updates the values of these performance measures. Calls are animated in the Call Center area where they can be seen waiting in line, occasionally abandoning, and being served. The modules at the bottom of the figure frequently check each call on hold to see if it is ready to abandon the queue.

number of system performance measures and continually updates their values in the 3 × 4 table: the number of callers who hung up without being served (Abandoned), the number of callers who reached an agent (Served), the average number of minutes served callers spent on hold (Ave. Q Time), and the percentage of served customers who spent less than the target time on hold (Service Level). The model also plots the average agent utilization, expressed as a percentage of the total number of servers.

The main flow and animation of calls occurs in the call center area of the model. Calls enter the Arrive module, with the time between arrivals being exponentially distributed, and they are immediately assigned random values for three key attributes, handle time, abandonment time and customer class. During execution, rapid calls appear on screen as small green telephones moving through the center while regular calls appear as blue telephones.

After arrival, a call is either put through to an available agent or put on hold to wait in queue until an agent is assigned to handle it. Rapid customers have priority over regular customers and, if they have to wait at all, appear at the head of the queue.

The call center's phone system had essentially unlimited capacity to keep calls on hold, so we specified no queue length limit in the model. The client's telecommunications department stressed that the call center had an exceptionally large trunk line capacity because of the firm's existing long-term contracts with its service provider.

During their wait in queue, some calls may abandon the system while the rest eventually reach the Server module. There, each call is allocated to one of the *S* agents, its waiting time on hold is recorded (Assign and Tally), a counter for the number served is increased by one (Count), and the call is delayed for the duration of its handle time. After service is completed, the agent is released and the call departs from the system.

Once per six simulated seconds, the model creates a logical entity to examine all of the calls on hold at that particular instant. If the Search module finds a call that has waited longer than its abandonment time, the Remove module pulls the call from the call center queue and moves it to the Depart module labeled Aband-Stats. Checking the queue for abandonment more often than 10 times per minute would probably be more accurate but lead to considerably longer run times (which we had to avoid).

The two modules in the mean-interarrival-time-schedule part of the model allow the arrival rate to vary by time of day. Periodically, the model creates a logical entity (Arrive) to update the value of the mean-interarrival-time variable MeanIATime based on the values contained in the MeanIAT vector. However, because we held MeanIATime constant throughout the day at 0.15 minutes to simplify the analysis, we did not take advantage of this feature.

Failing to offer its suffering customers an alternative for technical support would have been extremely damaging.

Verification and Validation

Because we had so little time, we verified and validated the model quickly and informally. To verify that the model was working as intended, we relied on our experience with building and testing simulation models [Mehrotra, Profozich, and Bapat 1997; Saltzman 1997], and on Arena's many helpful features. For example, Arena has a completely graphical user interface, many automated bookkeeping features that greatly reduce the likelihood of programming error, and interactive debugging capabilities that allow the user to stop execution and examine the values of any variable or caller attribute.

To validate that the model's operation and its output represented the real system reasonably well, we relied on the second author's experience as a call center consultant and intimate knowledge of the client's operations. Based on the model's animation and its average output over many scenarios, he made a preliminary judgement that the model was valid.

Subsequently, the model passed the most important test: the call center managers embraced it. They compared the queue lengths, waiting-time statistics, and abandonment rates in the base case to those in the ACD reports for specific time periods and found the simulated values largely consistent with what they saw in the call center. Once they were comfortable with the base case, they were eager to understand the impact of different adoption levels for the rapid program with various staffing configurations.

Experimentation and Results

The call center is a terminating system that begins each morning empty of calls and ends hours later when agents go home after serving their last calls. For simplicity, we took each replication of the model to be exactly 12 hours, so even though the calls in the system at the end of the day were not served to completion we counted them as served.

We defined scenarios as specific combinations of S, the number of agents, and P_1, the percentage of rapid callers. Testing six values of S and six values of P_1 led to a total of 36 scenarios. Since we had to run and analyze many scenarios in just a few days, only 10 replications were executed for each scenario. (We made later runs with 50 replications for several scenarios and obtained results that differed from those for 10 replications by less than five percent.)

Arena's output analyzer calculated summary statistics across the 10 independent replications [Kelton, Sadowski, and Sadowski 1998]. We report here only the means across these replications for each scenario. Although Arena generates confidence intervals for the mean, we did not report these to the client for fear of complicating the presentation of results (Table 1).

Under the circumstances we tested, rapid customers waited very little (less than half a minute on average among those who did not abandon) because of their priority over regular customers. Consequently, few rapid callers (3.3 to 4.2 percent) abandoned the system. The service level provided to rapid callers was above 95 percent in all but a few cases, achieved at the expense of a much lower level of service for regular callers.

Regular customers waited between 3.35 minutes in the best case ($P_1 = 0\%$, $S = 90$) to 18.5 minutes in the worst case ($P_1 = 50\%$, $S = 80$). When rapid callers made up 20 percent of the callers ($P_1 = 20\%$), the average waiting time for regular calls can be improved by 57.5 percent, from 10 to 4.25 minutes, by increasing the number of agents from 80 to 90. This would also reduce the abandonment rate of regular customers by 51 percent, from 21.5 to 10.5

		Number of agents, *S*					
P_1	Performance measure	80	82	84	86	88	90
50%	Ave. queue time (min.)—rapid	0.43	0.40	0.38	0.36	0.33	0.30
	Ave. queue time (min.)—regular	18.50	14.30	11.50	9.47	7.68	6.12
	Abandonment rate (%)—rapid	4.2	4.1	4.1	3.8	3.8	3.5
	Abandonment rate (%)—regular	32.0	28.4	24.6	21.0	17.6	14.5
	Service level (%)—rapid	89.5	91.2	92.3	92.9	94.0	95.1
	Service level (%)—regular	16.8	22.7	30.8	40.1	52.2	64.1
40%	Ave. queue time (min.)—rapid	0.33	0.32	0.31	0.29	0.27	0.25
	Ave. queue time (min.)—regular	13.10	11.30	9.55	7.99	6.62	5.32
	Abandonment rate (%)—rapid	4.1	4.0	3.9	3.8	3.7	3.5
	Abandonment rate (%)—regular	27.3	24.2	21.0	18.0	15.3	12.6
	Service level (%)—rapid	94.6	94.9	95.5	95.9	96.6	97.2
	Service level (%)—regular	21.0	28.5	38.0	48.3	60.7	72.1
30%	Ave. queue time (min.)—rapid	0.27	0.27	0.25	0.24	0.22	0.21
	Ave. queue time (min.)—regular	11.20	9.69	8.35	7.06	5.87	4.66
	Abandonment rate (%)—rapid	4.0	4.0	3.8	3.8	3.7	3.5
	Abandonment rate (%)—regular	23.8	21.0	18.4	16.0	13.7	11.2
	Service level (%)—rapid	96.9	97.1	97.9	98.1	98.5	98.4
	Service level (%)—regular	24.3	33.5	44.4	56.6	68.5	80.9
20%	Ave. queue time (min.)—rapid	0.23	0.23	0.21	0.21	0.20	0.18
	Ave. queue time (min.)—regular	10.00	8.78	7.61	6.50	5.30	4.25
	Abandonment rate (%)—rapid	3.9	4.0	3.9	3.7	3.8	3.3
	Abandonment rate (%)—regular	21.5	19.0	16.8	14.7	12.5	10.5
	Service level (%)—rapid	98.3	98.7	98.9	99.0	98.9	99.2
	Service level (%)—regular	28.9	40.1	51.1	62.9	76.0	86.3
10%	Ave. queue time (min.)—rapid	0.21	0.19	0.19	0.18	0.17	0.16
	Ave. queue time (min.)—regular	8.97	7.86	6.82	5.87	4.70	3.74
	Abandonment rate (%)—rapid	3.9	4.1	4.0	3.8	3.8	3.6
	Abandonment rate (%)—regular	19.3	17.2	15.3	13.4	11.4	9.5
	Service level (%)—rapid	99.1	99.3	99.4	99.3	99.6	99.5
	Service level (%)—regular	35.4	46.8	60.0	72.3	84.3	92.4
0%	Ave. queue time (min.)—regular	8.19	7.26	6.31	5.28	4.25	3.35
	Abandonment rate (%)—regular	17.8	16.0	14.2	12.4	10.5	8.8
	Service level (%)—regular	42.4	54.1	68.2	80.4	90.7	96.4

Table 1: We ran the model for 36 scenarios. Entries in the table represent average values across the 10 replications run per scenario. An abandonment rate for each class of callers was derived from the number of abandoned and served calls, expressed as a percentage, that is, Abandonment Rate = 100* Abandoned/(Served + Abandoned). Service level for rapid callers is the percentage of calls answered within one minute; for regular callers it is the percentage of calls answered within eight minutes.

percent.

Another key strategic question was what level of agent staffing would keep average waiting times reasonable for regular calls, that is, at or below eight minutes, while providing superior service for rapid callers (Figure 3). The model showed, for example, that if rapid callers made up 20 percent of the callers, 84 agents would be needed. In this scenario, rapid callers would have a very high service level of about 99 percent. Alternatively, if just 10 percent of callers were rapid customers, only 82 agents would be needed to achieve the eight-minute target average for regular customers.

Call center managers could use the graph to determine staffing levels if they changed the target average wait to another value, such as six minutes, or if more customers participated in the rapid program (Figure 3).

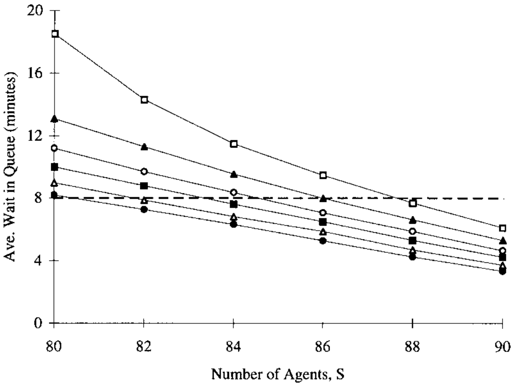


Figure 3: The lines represent (from top to bottom) 50, 40, 30, 20, 10, and zero percent rapid customer calls. For each percentage, we can see how the average wait in queue for regular callers decreases as the number of agents increases. The figure can be used to determine the number of agents required to keep average waiting times for regular calls at or below the eight-minute target, while providing superior service for rapid callers. For example, if rapid callers made up 20 percent of the callers, 84 agents would be needed.

Implementation

Once we had built an initial version of the model, we presented our preliminary results to a management team that had no previous experience with simulation. We explained the underlying concept of a simulation model as a laboratory for looking at different call center configurations and examining the impact of design changes on key performance measures.

Based on both the animation and the preliminary results, the managers were excited about the answers our model could provide. They asked us to quickly run some additional scenarios, which we analyzed and then reviewed with them a few days later.

The results of the simulation analysis gave the managers a good understanding for the impact on system performance of changes in the number of customers purchasing the rapid program (which they could not control) and in the number of agents (which they could control). In particular, the average queue time results made it clear that the one-minute guarantee to rapid customers would be fairly easy to achieve, even if the percentage of callers in the rapid plan became quite high, which helped drive the revenue projections for the program.

They were surprised at the dramatic impact adding agents would have in cutting the waiting time of regular customers. In our presentation, we gave the managers some general guidelines about how to make adjustments to staffing based on the proportion of customers who purchased the program. Most important, the results of our analysis gave managers confidence that they could successfully implement the rapid program.

We conducted our analysis and presented the results during one week in September 1995. Within a few days, the managers decided to introduce the rapid program. The firm did a lot of internal work very quickly to prepare for the start of the program. The company trained agents to handle rapid calls and changed call center desktop information systems (changing the user interface and modifying the back-end database and integrating it with the credit-checking and billing systems). Finally, it altered the call center's ACD logic to establish different call routing and priorities for rapid customers.

By November, the company had finished its preparations. Launched with a major direct-mail campaign, the program was eventually adopted by more than 10 percent of customers, and generated nearly $2 million of incremental revenues within nine months. Because of the success of this program, the company initiated a much more comprehensive fee-based technical-support service at the start of its next fiscal year.

Our simulation model made a major contribution to the launch of this program and the generation of this revenue stream. The results of our analysis played a key role in the company's decision to launch the program; in particular, they validated the feasibility of the "one minute or free" guarantee to rapid customers, which was an important part of marketing the program.

Our analysis had another more subtle effect on the rapid program. The decision to launch the rapid program was inherently difficult and contentious. Debate about the program's merits and risks could have caused weeks of delay. This delay would have been very expensive in terms of rapid-program revenues, because the call center's busiest months were December, January, and February. Our simulation analysis, however, provided a diverse group of decision makers (from the call center, finance, marketing) with a common empirical understanding of the potential impact of the rapid program on customer service, which in turn facilitated a much faster decision.

Lessons for Simulation Practitioners

This case study highlights several themes from which simulation practitioners can benefit (see Chapter 3 of Profozich [1998] for further discussion). This application of simulation is an excellent example of embedding a mathematical analysis in the midst of a larger decision-making process. In our simulation model, we did not try to determine which service program the call center should offer to customers because many business factors (including marketing, financial, and technical issues) influenced the definition of the rapid program. However, the results of our analysis enabled managers to evaluate the proposed solution empirically under different conditions.

The managers saw our simulation analysis as a vehicle for mitigating risk. They did not have to hold their breaths while they experimented on live customers with real revenues on the line. Instead they relied on the simulation for insights into what would happen when they changed the configuration of the call center. In many industries, managers increasingly see call centers as the customers' windows into their firms; poor strategic decisions can be very costly. By helping to prevent such errors, simulation can be very valuable.

Finally, the managers found our simulation analysis valuable because it helped them to launch the rapid program quickly. In today's competitive business climate, a decision delayed can be worse than a bad decision. For our client, failing to offer its suffering customers an alternative for technical support during the December-to-February busy season would have been extremely damaging. The simulation results helped the decision makers to quickly give the rapid program their blessing.

Retrospective Assessment

Looking back on our study, we can see ways in which we could have improved our modeling effort. First, we could (and probably should) have incorporated a nonstationary arrival pattern based on the historical call patterns. Although the model had the capability to generate a nonstationary arrival pattern, we did not use it.

With a few additional days for the study, we would have run more replications of the model. With more time and resources, we might have allowed for the specification of time-varying agent schedules (showing a variety of starting times, break times, and lunch times). Agent groups are another dimension we might have added, that is, some agents might be able to handle rapid calls only, others regular calls only, and the rest capable of handling both.

Including these additional dimensions in the representation of the call center would have caused the number of scenarios to explode. The business question of how the adoption rate of the rapid program would influence service levels for regular customers would very likely have been obscured by a host of lower-level issues. While decisions about schedules and skills are tactically important, they were not particularly relevant to deciding whether to launch the rapid program at all. (Mehrotra [1999] discusses different call center planning horizons.)

Conclusions

This project was successful in influencing a major business decision for several reasons. This project benefited from (1) a clear focus on specific business decisions and direct access to decision makers; (2) intimate understanding of key performance measures and data sources; and (3) quick delivery of the model and timeliness of the results. Also important was our experience in developing simulation models, a capability that most call centers do not have.

In conducting our analysis we faced two serious constraints: a limited amount of time and a powerful but general-purpose simulation package. Indeed, these constraints limit the use of simulation in the call center industry. The dynamic and frenetic nature of call center operations means that most decisions are made either in a reactive manner or with very limited lead time at best. Consequently, creating simulation models from scratch, even with user-friendly packages such as Arena, is usually not feasible.

Since we completed this project, several simulation packages designed specifically for call centers have appeared on the market, including the Call$im package [Systems Modeling Corporation 1997] that is built on top of Arena. They enable the rapid development of call center simulation models. Call$im, for example, is used by over 70 call centers around the world for planning and analysis studies.

Released initially in 1997, Call$im has a user interface built around call center terminology and includes such features as out-of-the-box integration with Visual Basic, animation, and customized output to spreadsheets. All of this shields users from the underlying programming, greatly simplifying the process of building call center models, importing data, conducting experiments, and analyzing results.

Would we have used Call$im for this analysis if it had been available then? Yes. Call$im's industry-specific objects—with such module names as Calls, Agents, and Schedules—contain detailed logic that mirrors the way call centers operate and default values that call center personnel find intuitive and reasonable. In addition, each module includes a variety of parameters that provide flexibility in model building and enable realistic representation of nearly all call centers. Automatically generated output statistics are also defined in call center terms, such as service level, abandonment rates, and agent utilization. A knowledgeable Call$im user could create and test the model shown in Figure 2 in less than one hour, using just six or seven modules for input data, model logic, and output creation. This package would have allowed us to add more model detail, to spend more time on analysis, and to test different scenarios for adoption of the rapid program.

Call$im is a far superior tool for creating call center models than Arena and its peers, but they are far better tools for building simulation models than general-purpose programming languages (which were once the state of the art). Arena's availability and flexibility were crucial to our successful analysis.

As call centers proliferate—there are an estimated 60,000 in the United States alone—and with simulation's potential to help managers configure and plan their operations, we expect the use of simulation will grow substantially in this industry.

References

Andrews, B. H. and Parsons, H. L. 1989, " L. L. Bean chooses a telephone agent scheduling system," *Interfaces*, Vol. 19, No. 6, pp. 1–9.

Andrews, B. H. and Parsons, H. L. 1993, "Establishing telephone-agent staffing levels through economic optimization," *Interfaces*, Vol. 23, No. 2, pp. 15–20.

Anton, J. 1996, *Customer Relationship Management*, Prentice Hall, Upper Saddle River, New Jersey.

Brigandi, A. J.; Dargon, D. R.; Sheehan, M. J.; and Spencer, T. 1994, "AT&T's call processing simulator (CAPS) operational design for inbound call centers," *Interfaces*, Vol. 24, No. 1, pp. 6–28.

Dawson, K. 1996, *The Call Center Handbook*, Flatiron Publishing, New York.

Grassmann, W. K. 1988, "Finding the right number of servers in real-world queuing systems," *Interfaces*, Vol. 18, No. 2, pp. 94–104.

Green, L. and Kolesar, P. 1991, "The pointwise stationary approximation for queues with nonstationary arrivals," *Management Science*, Vol. 37, No. 1, pp. 84–97.

Harris, C. M.; Hoffman, K. L.; and Saunders, P. B. 1987, "Modeling the IRS taxpayer information system," *Operations Research*, Vol. 35, No. 4, pp. 504–523.

Hoffman, K. L. and Harris, C. M. 1986, "Estimation of a caller retrial rate for a telephone information system," *European Journal of*

Operations Research, Vol. 27, No. 2, pp. 39–50.

Kelton, W. D.; Sadowski, R. P.; and Sadowski, D. A. 1998, *Simulation with Arena*, McGraw-Hill, New York.

Mehrotra, V. 1997, "Ringing up big business," *OR/MS Today*, Vol. 24, No. 4, pp. 18–24.

Mehrotra, V.; Profozich, D.; and Bapat, V. 1997, "Simulation: The best way to design your call center," *Telemarketing and Call Center Solutions*, Vol. 16, No. 5, pp. 28–29, 128–129.

Mehrotra, V. 1999, "The call center workforce management cycle," *Proceedings of the 1999 Call Center Campus*, Purdue University Center for Customer-Driven Quality, Vol. 27, pp. 1–21.

Pegden, C. D.; Shannon, R. E.; and Sadowski, R. P. 1995, *Introduction to Simulation Using SIMAN*, second edition, McGraw-Hill, New York.

Profozich, D. 1998, *Managing Change with Business Process Simulation*, Prentice Hall, Upper Saddle River, New Jersey.

Saltzman, R. M. 1997, "An animated simulation model for analyzing on-street parking issues," *Simulation*, Vol. 69, No. 2, pp. 79–90.

Systems Modeling Corporation 1997, *Call$im Template User's Guide*, Systems Modeling Corp., Sewickley, Pennsylvania. (www.sm.com)

Whitt, W. 1999, "Predicting queueing delays," *Management Science*, Vol. 45, No. 6, pp. 870–888.

An officer of the client firm wrote as follows: "At the time that the authors conducted the study, our technical support operations were in reasonably poor condition. While considering the decision that is described in this paper, we had a lot of concerns about its impact on all of our customers. Frankly, given the lack of data on hand about customer adoption of the new "Rapid" program, the major computer systems changes required to implement Rapid, and re-education of call center staff to handle these changes, we did not have a good understanding of what would happen if and when we launched this new initiative.

"The simulation analysis conducted by the Onward—SFSU team helped us get a much better understanding of the contingencies that we were likely to face. In particular, by quantifying the impact of Rapid customers and staffing/skilling decisions on service levels and costs, the simulation analysis gave us the confidence to launch the program. The results also gave us an excellent idea of how best to staff the different queues and how to adapt our staffing as we learned more about the customer population's acceptance of the Rapid program.

"I am proud to report that this program was a great success for our business, grossing nearly $2 million in a little bit less than nine months. The work done by the Onward—SFSU team played a big part in making it happen successfully."