THE FLAGSHIP RESEARCH JOURNAL OF INTERNATIONAL CONFERENCE
OF THE PRODUCTION AND OPERATIONS MANAGEMENT SOCIETY
Volume 2• Number 1 • January - June 2009

Joscm

Scenario Analysis within a Call Center Using Simulation

Marco Aurélio Carino Bouzada, PhD
Universidade Estácio de Sá (MADE)
marco.bouzada@estacio.br

ABSTRACT: This paper works on and presents the results of several analyses − scenario and sensitivity − made with the help of Simulation and focused on dimensioning questions of handling capacity in a large Brazilian call center. The objective is to measure the sensitivity of the call center's performance to potential modifications of critical variables. The bibliography related to the application of such tool in call centers is reviewed, and the way by which the problem is treated nowadays is described in detail. The methodology used to achieve this article objective involved a simulation model in the Arena Contact Center software, which worked as base case upon where the scenario and sensitivity analyses could be performed. This paper comes to the conclusion that Simulation is a tool perfectly adequate to its purpose as long as it could be able to show, for the studied call center, mainly that: (i) it is possible to reduce the operator contingent; (ii) fair variations on the demand pattern can impact too much the performance indicators; and (iii) it is possible to improve the service level if an aggregated handling format is adopted.

Key Words: Call Center, Dimensioning, Simulation.

1. INTRODUCTION

Call centers are operational centers installed with the purpose of utilizing both Communication and Information Technologies, in order to automate a great volume of different activities and telephone services, not only of calls received (incoming calls), but also those generated by the center as well. The inbound centers, where the calls come from the clients, are characterized by a system constituted by several call attendants that receive calls from other people, usually clients, even if only potential clients that wish to get information on this or that subject, buy a product, be able to answer to a certain research, update the data already known, register occurrences or post a complaint, amongst other requests (GROSSMAN et al., 2001; HAWKINS et al., 2001).

According to Mehrotra, Profozich and Bapat (1997), managers and designers of call centers have a job much more difficult today than it used to be in the past. With more products and services being especially created, disposed and ready to use, sold and assisted by technicians out in the market, the call centers have spent much efforts to provide different service levels for different types of clients, which present different needs. Today, the telephone systems allow a great flexibility and can determine how the calls might be routed and put on line. Nevertheless, at the same time, this makes the planning and the analyses even more difficult, due to the fact that they allow a link between multiple call centers, prioritization of certain calls, existence of different abilities between operators and customization of calls routing.

Today managers must be able to know exactly what is going on at call centers in order to know how calls, routes, priorities, operators and their capabilities, outsourcing, peak periods and other factors which influence the level of service, the disconnect averages and the utilization rates (BOUZADA, 2006).

Labor costs represent almost 70% of the total industry, justifying the need for an efficient management and the importance of a quantitative approach for the dimensioning of a service handling capacity, which consists on a trade-off between this cost and the determination of the right service level; in other words, in having the right number of qualified people and resources at the right moment, in order to work with the forecast working cargo, maintaining the quality patterns and the required service level. This way, and according to the call centers' day to day more and more complex operations, the use of better accurate models on the dimensioning of the size of the personnel team, of the industry that works with great financial volumes, has been considered

as more important than ever (HALL; ANTON, 1998; ALAM, 2002; BOUZADA, 2006).

As studied by Mehrotra and Fama (2003), and Hall and Anton (1998), the call centers are interesting objects for the simulation studies, because: (i) they cope with more than one type of call, where each type represents a line; (ii) the calls received in each line arrive by chance – as time goes by; (iii) in a few cases, agents make calls proactively (especially in telemarketing or charging calls), or as a return for a call received; (iv) the duration of each call is random, as well as the work that the agent executes after the call (collecting of data, documentation, research...); (v) the progress on the systems which route the calls for the agents, groups or locals, make the logics behind the call center even more sophisticated; (vi) agents can be trained to answer only one type of call, several types of calls or all types of calls with different priorities and preferences specified for the routing logics; and (vii) the great amount of money invested in call centers, on both forms, capital and work, is capable to justify the use of this so powerful tool.

In accordance to Hall and Anton (1998), the call centers can use the Simulation to test (and eventually justify its implementation), if certain changes can improve the system before its implementation. The best call centers use this tool effectively and efficiently, in order to project the system, to manage the operation and to plan for the future, in face of possible scenarios.

Due to this fact, this paper describes the dimensioning problem of the handling capacity of a large Brazilian company of call centers, looking for an immediate proposal presentation of several scenario analyses, changing a few important parameters of the system. Such analyses are rendered viable by the use of Simulation along the work, whose objective is to measure the sensitivity of the call center's performance to potential modifications of critical variables.

2. LITERATURE REVIEW

According to Anton (2005), the main expense in typical call center is due to human resources (64% of the total costs), far beyond those related to technology (16%), the second more expensive issue.

Therefore, and in order to reduce staff requirements, one of the key duties consists of managing the call centers queues, which occur when there is no agent available to handle a client, which waits on a virtual line from which he will leave only when an operator is set to attend him or when he disconnects the call. As observed by Brown et al. (2002), in the case of call centers, the virtual queue is invisible among the clients and among the clients and the operators.

In the call centers scenario, Araujo, Araujo and Adissi (2004) say that the queues discipline, when well managed, is a strong ally for the call centers production planning and controlling area, which have as a goal to achieve the expected results with scarce resources, turning this area more and more important for these companies. Additionally, a significant reduction to the clients waiting time can be obtained.

A few call center characteristics make it difficult to apply analytical formulas from the Queue Theory for its modeling, including: generic distribution for the handling time, time-varying arrival rates, temporary overflows and abandonment (CHASSIOTI; WORTHINGTON, 2004).

Chokshi (1999), Klungle and Maluchnik (1997), Hall and Anton (1998), Mehrotra and Fama (2003), Avramidis and L'Ecuyer (2005), Klungle (1999) and Bapat and Pruitte Jr. (1998) go beyond a few recent factors that contributed to the increase on the demand for the use of the Simulation tool in the call centers sector: (i) the increasing importance of the call centers for a good number of corporations, due to the fast increase of information, communication and technological gadgets, increasing the need to use scientific methodologies on decision makings and tools for its strategic management instead of using the intuition, only; (ii) the increasing complexity of call traffic along with rules more and more viewed on the skill-based routing; (iii) the uncertainty more and more predominant at the decision problems usually found on the operational management of call centers phone desks; (iv) fast changes on the operations and improvement of the re-engineering activities resulting from the increase of joint-ventures and acquisitions, business volatility, outsourcing options and the utilization of different channels in order to reach the consumer (telephone, e-mail, chat...); and (v) the availability and accessible price of the computers, together with a range of simulation applications in call centers, available in an everyday market less and less complex, intuitive and easier to be assimilated and used.

The Simulation, according to Mehrotra (1997), explicitly shapes the interaction between calls, routes and agents, as well as the random individual incoming

calls and the also random duration of the handling service. Through the use of Simulation, managers and analysts translate the call centers gross data (call forecast, distribution of the handling times, schedule hours and the agents abilities, call route vectors, etc.), in handling information on the service levels, clients abandonment, use of agents, costs and other important performance measures of a call center.

According to Chokshi (1999) and Klungle and Maluchnik (1997), the use of Simulation to help management decisions in a call center allows the following benefits: (i) to visualize future processes and be used as a communication tool; (ii) to validate the processes premises before its implementation; (iii) to analyze the impact of the changes (scenario studies) in detail; (iv) to foresee the aggregated needs of resources and to schedule the working team; (v) to measure the performance indicators; and (vi) to estimate impacts on costs and economies.

The first usage of the Simulation in a call center, as said by Hall and Anton (1998), is the evaluation when one may verify "where the call center is". The key-question is "how efficient is the operation nowadays?" The goal of this evaluation is to establish a point of departure (and reference) for the change.

Mehrotra, Profozich and Bapat (1997) and Yonamine (2006) speak about the second usefulness of Simulation in a call center: the Simulation allows the fast and accurate understanding of how the operational performance of the call center will work when facing specific scenarios (based on modifications caused by external or management initiatives such as, for instance, the adoption of a new technology, a new business strategy or the increase of the amount of work), before any change is effectively made, but not interfering on the operation of the call center' phone desks, as well as not impacting its budget. This way, a few questions may be answered amongst others: (i) Which is the impact of a call overflow? (ii) Which are the trade-offs in the act-prioritizing special clients? (iii) Will the service improve if dispatchers provide basic pieces of information to clients? (iv) Which are the potential gains associated to the adoption of a predictor dial?

In accordance to these authors and to Gulati and Malcolm (2001), Bapat and Pruitte Jr. (1998) and PARAGON (2005), a simulation model can be used (and has been used more frequently than ever) – besides normally allowing graphics and animations – to contemplate a few other critical aspects of the modern receptive centers of all sizes and types, such as: (i) a specific service level; (ii) flexibility on the distribution of time between incoming calls and of handling time; (iii) consolidation of the central offices; (iv) skill-based routing; (v) multiple types of calls; (vi) simultaneous lines; (vii) call disconnect patterns; (viii) call returns; (ix) overflow and filling of capacity; (x) waiting lines prioritization; (xi) call transference and teleconferences; (xii) operators preferences, proficiency, time learning and schedule. The outputs model can emerge in shape of waiting time, call disconnecting average amount, (both with the possibility of differentiation on the call types) and level of the operators utilization (with possibility of the operator types differentiation). And, due to the applicability of this approach to the real and complex characteristics of call centers, the Simulation can make its dimensioning and management more reliable.

Mehrotra and Fama (2003) and Klungle (1999) envisioned future tendencies capable of impacting the simulation of call centers, such as: (i) the operational complexity, which will continue to increase – more waiting lines, more variation on the operators scale and combination diversity among skills and route rules – forcing the analysts to create richer models; (ii) emerging of more Simulation softwares specialized in call centers, whose importance tends to follow the role that the Simulation will assume in the process of remodeling the central offices that are necessary to the dealing with the new complexities; and (iii) a higher understanding by the executives that the call centers are main components of the clients' value chain, discharging a wish to understand the inherent risks of any operational configuration and the consequent improvement of the quality of the collected data and accuracy of the parameters (such as distribution of time between incoming calls, handling time, waiting time, average of disconnecting and others), holding more robust results.

Gulati and Malcom (2001) used Simulation to compare the performance for three different calls programming approaches (heuristic, daily batches optimization and dynamic optimization), revealing opportunities for improving the outbound call center process within the studied bank. The model outputs provided a way to check the system performance compared to the management goals and showed that the non-heuristic approaches achieved better results, but not during the whole day.

Miller and Bapat (1999) described how Simulation was used to project the ROI related to the acquisi-

tion and utilization of a new call routing technology for 25 call centers. Demanding US$ 17 million on investments and an operation cost of US$ 8 million per year, it was needed to check if the technology would cause enough benefits (cost reduction, operators productivity increase and possibility to handle more calls) in order to ensure its implementation on a national extent.

Lam and Lau (2004) wrote about a restructuring effort of a Hong Kong company which supplies service for computer and office equipment. As long as there were many opportunities available to improve the process, Simulation was used to explore the different options and evaluate the results of the existing call centers restructuring. The simulated results analysis confirmed that the great improve opportunity consisted of the joint of the current resources in a sole call center.

Saltzman and Mehrotra (2001) presented a study where Simulation was used by a software company which intended to visualize its call center operating before the launching of a new paid support service program. They wanted to verify if the goal – the paying customers waiting less than 1 minute before being handled – would be achieved. The management also wondered which would be the new program impact to the service offered to the non-paying customers regular basis.

3. THE CASE

3.1 The company

Contax emerged by the end of the year 2000 as a natural extension of Telemar's business, in a branch of the economy which did not invest much in technology and qualification of the customer's service, in order to help its clients on their operational management, aggregating value on the relationship with final customers (CONTAX, 2006).

Presently in Brazil, Contax is the largest growing company in this industry, with a growth of almost 60% in 2005, when it invoiced R$ 1.129 millions. It is known as the largest enterprise of this branch based on the number of attending service positions, and the second largest in terms of sales and work force, inside the national territory (OUTSOURCING, 2005).

The Contax capital is 100% national and today it operates with more than 22.000 positions at the customer service, almost 50.000 employees and more than 40 clients, with Telemar being the largest one (representing approximately 60% of the sales). The main products related to this client are: (i) Velox; (ii) 103; (iii) technical support and repairs; (iv) Oi; and (v) 102, which receives calls of customers that need information from the telephone directory.

The dimensioning process of handling capacity

The dimensioning consists of the analysis that may customize physical, technical and personnel structures of a call center towards the objectives of the customer service operation that begins with the forecast of the demand within the days.

The 102 product was chosen to illustrate the dimensioning problem, since its demand is the most foreseeable and, therefore, being possible to measure the quality of a dimensioning process independently i.e., departing from the premise that the input – demand forecast – presents a good quality. Besides that, there are only two types of clients of the 102 product: plus (a paid service) and basic (a service free of charge).

The service level for this product is related to the waiting time of the final client at the telephone line, from the moment the incoming call arrives to when it is answered. In other words, it is the time which the client remains waiting in line, listening to the background song and waiting for the operator. More precisely, the level of the service consists of the percentage of calls – amongst the completed ones, only – that wait no more than 10 seconds to be answered.

As only the calls answered count in the computation of the service level, the disconnections are not accounted (and, therefore, not punished), for effects of the service level. Nevertheless, they are measured through another indicator (abandonment rate) and Contax pays fines when this rate exceeds 2% in a month. As this may happen, to avoid the disconnect is seen as a priority, to the detriment of the service level, as long as this is maintained above a minimum value. The service level does not involve formal requirements of the contract (as the abandonment rate), but does influence the commercial relationship and dignity; i.e., it is interesting to not prioritize only the abandonment and, as a consequence, not worry with the maintenance of the service level in decent values.

The dimensioning routine – isolated for each product (basic and plus – due to the priority of the last over the first) – begins by computation of the daily needs of operators, departing from the forecast amount

of calls, the average handling time (AHT) and the average time that operators are busy per day. After that, the need of operators (converted to the 6-hours-operators pattern) is compared to todays' resources availability, discounting the losses concerning vacation and absence. The result of this comparison is the balance or the deficit of the labor for each day of the planned month. The output of this first step is the amount of operators that need to be hired or dismissed in the referred month so that the required numbers can be achieved.

From the moment the contract decision is taken, or the dismissal is decided and implemented, the planning team can look forward to a more detailed analysis – daily dimensioning. This must be done for one day only, and this pattern format must be repeated to the other days of the period, since the scheduled hours of each employee must be the same on everyday of the month.

In conclusion, a volume of calls and an average handling time (necessary numbers for the dimensioning) shall be chosen to be used as a pattern for the dimensioning of all days of the month. The chosen day for the pattern is, usually, the fifth day of higher movement. This way, the dimensioning will guarantee the desired service level for this day and all the not-so-busy days, but not for the four days of higher demand, when there will be a loss in the service level. Nevertheless, this does not represent a problem, because the agreement related to the 102 product involves a monthly level service and not a daily level.

Over the day chosen as a pattern for the dimensioning of the month is applied a curve that shall reflect the daily demanding profile, i.e., which daily volume percentage will happen in the first half hour of the day, in the second half hour of the day, …, and in the last half hour of the day. This curve is shown based on the calls report received at each period of half hour for each day of the week. Concerning the 102 product, the curves of each day of the week are very similar (mainly from Monday to Wednesday, with a little increase of the volume in the afternoon of Thursdays and Fridays), and on Saturdays and Sundays they happen to be a little different.

The result of this process is a forecast call demand (volume and AHT for each half hour). Using the concepts of the Queue Theory and with the help of the Excel Supplement Erlang formulas, called Turbo Tab, it is computed the necessary amount of operators that will be handling the demand of each period with a minimum pre-established service level (normally 85% of the calls being answered before 10 seconds – for plus clients – and 75% – for basic clients).

The last month contingent of operators is then considered. Due to the amount of operators that are initiating their work at each day period and the daily work load of each one of them (4 or 6 hours), a sheet computes how many operators will be available for each period of half hour. This information is then compared to the operators need for each period of 30 minutes, previously calculated.

Over the actual operators scale, the planning team will work on the changing of the operators' availability for each period of the day, in order to achieve the desired service level. The objective is to assure a certain amount of people in each scheduled hour, throughout a trial-and-error process, during which it will be necessary to analyze several factors, such as daily working hours load, working laws aspects, union agreements and available physical space. In the case of the 102 product, the balanced scale (alternating times with the operational contingent over or under the requirements) can be used, since what really matters for commercial purposes is the daily average level service.

During the staffing process, the planning team makes experiments by modifying the quantity of operators that begin to work at each period of time. These changes consequently alter the quantity of operators available in each period of half hour. The sheet containing the Erlang formulas uses this information then to estimate the service level for each period of half hour and for the day, which depends also on the forecast demand.

At this interactive process, the principal motivation of the analyst is to maximize the day's average service level. The level of the service in each hour band, itself, does not present a great concern to the analyst who, nevertheless, tries to avoid great deficits of operators assigned in relation to the demanded within the hour bands of the day.

The concern about daily deficits does exist because, in hours with a higher deficiency of operators it is possible to register a great incidence of abandonments. And this could be very bad for two reasons: fines for excess of call abandonments and the possibility that the client left without an answer returns the call later on and waits until getting an answer, therefore deteriorating the service level.

This dimensioning effort main goal is to provide a better adjustment between the demanded and offered capacity during the day.

On the last part of the dimensioning and staffing processes, the analyst tries to estimate how the operation level service will be (percentage of calls answered in less than 10 seconds), on all days of the month (until here the computation was based on the fifth day of larger movement, only). The intraday distribution of operators elaborated during the past steps is repeated on all days of the month and, along with the daily call demanding forecast as well as with the demand intraday behavior profile, is able, therefore, to estimate – through the Erlang Methodology – the service levels to be obtained for each day and hours, within the month in question.

Methodology used to perform the scenario and sensitivity analyses

For the real world of call centers, the Queue Theory is the best analytical methodology to be used, but there are experimental methods – as Simulation, for instance – that should be even more adequate for an industry with an operational day to day as complex as modern call centers, as suggested by section 2 of this paper.

The employment of the Simulation allows us to contemplate the highlighted characteristics of the same section, including the abandonment behavior (it is possible to consider that a percentage of clients that disconnected their calls, will return and try a new contact in a given amount of time, which can be modeled by a statistical distribution) and a flexibility on the definition of the handling time distribution.

The idea consists in simulate by computer and in a few seconds, the call center's operation during periods of 30 minutes. This way, it is not necessary to experience in practice some of the dimensioning alternatives in order to know the consequences; the experimentation is made in a virtual environment. Nevertheless, it is possible to visualize the operation itself (with the calls arriving, being sent to the queues and handled afterwards) and what would be going on, in detailed forms (practically as being *in loco*), in order to understand why a certain period of the day presented a service level so low, for instance (instead of only accepting the number supplied by the analytical formulas).

For the dimensioning and staffing of the operators to handle the plus clients of the 102 product, in August of 2006, it was used the premise (originated on the demand forecast) that 586 calls would come to the phone desk with an AHT of 29 seconds in the first half hour of the day (from 00:00 a.m. to 00:30 a.m.). The staffing team requested then 12 operators for this period.

In the software Arena Contact Center it was built a model to simulate how the system would behave in this period, with the same demand premises (volume and AHT) and with the same operational capacity (12 agents).

As the calls come to the phone desk without any kind of control, this process can be considered a random one, the conceptual basis suggesting therefore that the call arrivals rate might be shaped through a Poisson process. The concepted simulation model implemented this process with a mean of, approximately, 0.33 calls arriving per second (or 586 in a 30 minute interval).

In relation to the handling time, the Erlang distribution uses to better shape this process and, therefore, it was used with a mean of 29 seconds. It requires nevertheless an additional parameter (k) related to the variance of the data around the mean. The standard deviation of the distribution is equal to its mean divided by the square root of k. To be able to consider a moderate variance of the data around the mean, the model takes the Erlang distribution with $k = 4$, resulting on a variation coefficient of 50%.

In order to allow a correct interpretation of the clients' abandonment behavior, it was necessary to perform a research close to the Contax basis, which includes the disconnected calls of the 102 product. The research showed that the waiting time of the calls disconnected historically present a mean of about 2.5 minutes, following a distribution not too far from an exponential one. It was also necessary to model the return behavior of the disconnected calls. To make it possible, it was used the premise that 80% of the disconnected calls are recalled between 1 and 9 minutes after the disconnectment happens (uniform distribution).

The simulation of the call center operation during 30 minutes was replicated 100 times in the software in a period of 142 seconds, and the first results indicate that, in average, 595 calls were generated in each replication. This number is a little higher than that demand premise of 586 calls, due to the fact that, in the simulation, a few of the disconnected calls were

replicated and put on the queue again. From the generated calls, 579 calls in average were effectively handled by the operators in each replication, generating an AHT of 29.35 seconds.

From these calls, 541 were handled before 10 seconds, resulting in a service level of 93.31%. From the 595 calls generated in each replication, 14.5 (in average) were disconnected by the clients, generating an abandonment rate equal to 2.44%. Amongst the disconnected 14.5 calls, 11.5 (79,41%) returned to the queue a few minutes after the disconnectment. In average, the operators were busy 78.75% of the time, during this period.

From this basic scenario, several scenario and sensitivity analyses were studied for a better understanding of the system operational behavior facing potential changes of its main parameters. This methodological proceeding is similar to the ones used and described by Miller and Bapat (1999), Gulati and Malcom (2001), Saltzman and Mehrotra (2001), Lam and Lau (2004) and Yonamine (2006), whose works were pointed during section 2 of this paper.

Analysis of scenarios, sensitivity and results

All foregoing results depart from the premise that the handling time follows an Erlang distribution (with a variation coefficient equals to 50%). Nevertheless, it is possible to have the handling time presenting a different variance and that this parameter being able to cause an impact on the most important results.

The sensitivity analysis about the variance of the handling time tries to measure this impact. The same simulation was repeated a few times using the software, always with the same mean on the handling time (29 seconds), but with different values for *k* (and, consequently, for the variation coefficient), from where the relevant outputs were collected, and the main performance indicators (service level and abandonment rate) could be obtained, which are shown on the following Table 1.

As *k* increases, the variance of the handling times diminishes. Due to the homogeneity of these times, the system becomes more stable, presenting as the most noticeable consequence an increase of the service level. But the abandonment rate is not as clear in its tendency, although it may look like falling as the variance of the handling time diminishes (*k* increases). The variation on these outputs is not too large, but is far from being irrelevant, revealing a significant potential impact of this parameter on the most important results. Thus, the correct consideration of the handling time variance – and not only of its mean – reveals to be extremely necessary in order to obtain accurate results.

Table 1 – Service level and abandonment rate for different values for the parameter *k* of the Erlang distribution, from 00:00 a.m. to 00:30 a.m., Aug/06, 12 operators

k	Variation Coefficient	Calls				Service level	Abandonment rate
		created	handled	before 10 sec	abandoned		
1	100%	597	580	528	16,17	91,03%	2,71%
2	71%	602	585	539	16,09	92,12%	2,67%
3	58%	598	584	539	13,27	92,31%	2,22%
4	50%	595	579	541	14,52	93,31%	2,44%
6	41%	599	583	545	15,46	93,44%	2,58%
9	33%	597	583	546	13,92	93,77%	2,33%

Source: Table elaborated from the results obtained by the software.

The worst performance (service level = 91.03% and abandonment rate = 2.71%) occurred in a situation in which the variance was the largest possible (k = 1; variation coefficient = 100%). This is an Erlang distribution case in which it coincides with the exponential distribution, the same format used to model the handling time in the analytical methodology employed by Contax to estimate the indicators.

This evidence arouses a curiosity related to the verification of the results of a simulation which considers another type of distribution – since, in accordance with what was indicated on section 2, the behavior of the handling time in call centers can present different formats – usually used, as well, to model this variable: the lognormal. Using the same original simulation model, but modifying this variable distribution to fit the high-

lighted format (with the same mean – 29 seconds, as well as keeping the same variation coefficient of 50%), 100 replications were run in the Arena Contact Center software.

In average, 597 calls were generated, 583 handled (544 before 10 seconds) and 13.95 abandoned. The resulting service level and abandonment rate were 93.40% and 2.34%, respectively. These indicators are too close to those obtained with the Erlang distribution where k = 4 (93.31% e 2.44%, respectively), giving a certain reliability to these values and suggesting that any of the two formats commonly used to model the handling time can be used indistinctively.

Simulation also allows the scenario analysis (What-if?). At the given example being studied here, since the employment of 12 operators in the highlighted period generated a service level (93.31%) comfortably higher than the minimum goal (85%), what would happen with this indicator after a reduction of 1 operator? Would it be possible to keep it above the goal?

Within this scenario, an average of 576 calls was handled, from which 491 before 10 seconds. The service level obtained in this scenario with 11 operators was of 85.23% then. This value is still above the 85% established for the plus clients. In other words, the 12th agent was not missed (in terms of achieving the service level goal), even in spite of his absence having lowered the service level in more than 8 percentual points.

But it would not hurt to also know the impact caused by this reduction in the abandonment rate. And, from the 604 calls generated in each replication, 26.2 – in average – were abandoned by the clients, resulting in rate equal to 4.34%, this revealing a great impact on this performance indicator. Nevertheless, if this indicator was not to be considered as so important, the use of Simulation could cause the savings of one agent for this time band, what can effectively happen in some scenario in which the abandonment rate can be seen in a lower level.

The operators reduction impact can also be visualized in case of higher deficits of handling availability through a more complete analysis of sensitivity. The same simulation was replicated a few times in the software, always with the same parameters, but varying at the number of operators. The relevant outputs were collected, and from them the main performance indicators (service level and abandonment rate) could be obtained, which are shown on the following Table 2.

Table 2 – Service level and abandonment rate for different amounts of operators,

from 00:00 a.m. to 00:30 a.m., Aug/06

Operators	Calls				Service level	Abandonment rate
	created	handled	before 10 sec	abandoned		
9	719	540	160	168,02	29,64%	23,38%
10	639	567	354	67,68	62,46%	10,60%
11	604	576	491	26,21	85,23%	4,34%
12	**595**	**579**	**541**	**14,52**	**93,31%**	**2,44%**

Source: Table elaborated from the results obtained by the software

As expected, as the handling availability decreases, the service worsens, as well as its performance indicators, mainly the abandonment rate, which quadruplicates after a 2 agents reduction. With this contingent, the service level is much lower than the goal, but cannot yet be considered unacceptable, which occurs in the 9 operators scenario, when it becomes 3 times lower than the original value. With this handling availability, the amount of abandoned calls exceeds those ones handled before 10 seconds, making the abandonment rate almost as high as the service level!

That reveals a great impact of the reduction of the amount of operators on the system performance (indicating the need of the dimensioning activity to be developed with much care); and suggests that hiring 10 agents for this time band would be the indispensable minimum, characterizing a scenario for which the performance indicators would be bad, but not catastrophic.

The most refined decision on how many agents to hire effectively for this time band should take in account the potential costs involved in the inclusion/ exclusion of 1 or more operators on/from the map of scheduled times. This way, Contax could question whether it would be willing to spend one additional monthly labor cost in order to improve the service level and the abandonment rate for the highlighted time band by the amounts shown in the analysis.

As described in section 3.2, the dimensioning of the operational capacity is made separately for each product – basic and plus. Nevertheless, the Simulation allows – in a very similar way to what described Saltzman and Mehrotra (2001) – a visualization of what would happen with the operation in a scenario in which different clients – basic and plus – could be handled under an aggregated form (by the basic and plus agents), but maintaining the priority for the plus clients and, this way, dismantling the queue discipline normally used by analytical models – FIFO or "First In First Out".

For the sizing and staffing of operators to handle basic clients of the 102 product during August 2006, it was used the premise that 399 calls of these clients would come to the phone desks, with an AHT of 34 seconds on the first half hour of the day (from 00:00 a.m. to 00:30 a.m.), for which was indicated a staff of 7 operators.

The idea now consists on simulating – in order to observe the system behavior – the scenario where these calls would be aggregated to the calls of the plus clients during this same time band (whose premises were mentioned before, in this same section), forming a sole queue. The 12 plus operators and the 7 basic ones would be able to handle both types of calls, but with different proficiencies and priorities, that characterize the skill-based routing, a mechanism feasible only under experimental approaches, as explained before in section 2.

But basic clients do not behave as plus clients do, and their different characteristics must be contemplated by the model. Their handling time, for example, is, in average, a little higher. For model purposes, the same Erlang distribution was used for this time, with a variation coefficient of 50%.

Generally, the basic client is more patient before disconnecting the call and this waiting time was also modeled as an exponential distribution, but with a mean of 3.5 minutes (instead of 2.5). It was also considered that a smaller amount of the clients that disconnect the call (70% instead of 80%) try to repeat it within a space of time also lower: between 1 and 6 minutes (uniform distribution).

Concerning the priorities, the plus calls are preferential upon basic ones and should be handled, as long as it is possible, by plus operators – theoretically more capable – should the plus operators be busy at the moment, the plus calls would be handled by the basic operators. The basic calls, in turn, would be preferably handled by basic agents, theoretically less capable, in order to let the best agents free for more important calls.

When the calls are not handled by their preferential agents, their handling time is changed. The model takes in consideration the fact that a plus call being handled by a basic agent (less capable) lasts 10% more to be completed; on the other hand, if a basic call is handled by a plus agent (more capable), this lasts 5% less to be completed.

The simulation was replicated 100 times in the Arena Contact Center software. In average, 413 basic calls and 591 plus calls were generated in each replication, 19.8 and 9.5 of them being abandoned, respectively. The resulting abandonment rates were 4.80% for the basic clients and 1.60% for the plus clients. Amongst the disconnected basic calls, 14.2 (71.46%) returned to the queue a few minutes later, and amongst the plus calls, 7.6 (80.44%) did the same. 393 basic calls and 581 plus calls – in average – were effectively handled in each replication. From these, 303 basic (77.25%) and 575 plus (98.96%) were handled before 10 seconds (service level).

Comparing the performance indicators of the plus clients with prior values obtained in the scenario with segmented handling (abandonment rate = 2.44% and service level = 93.31%), it is easy to conclude that the aggregated operation became fairly better for these clients. Such results were expected, since 7 basic agents started handling plus calls. It is true that the 19 agents also handled basic calls, but only when there was no plus call waiting.

This preference allowed a considerable improvement to plus clients handling and, certainly, reduced a little the quality of the service for basic clients. But it is interesting to notice that, at this scenario, the service level for neglected clients (77.25%) remained still above the goal to be achieved (75%). The abandonment rate (4.80%) became a little high, but it is not considered unacceptable for basic clients.

Simple as it seems, the handling aggregated format improved the system performance for plus clients without neglecting too much the quality of the service for basic ones. This happened because the agents could handle their non preferential calls while idle, allowing the increment of the operation quality as a whole. This kind of analyses and conclusions would not be able to be performed/obtained through analytical methodologies, being only feasible by means of an experimental approach, such as Simulation.

The AHT for basic clients was 32.65 seconds, a little lower than the value of 34 seconds, used on the AHT premise, because a few of them were handled by faster agents (plus). For plus clients, the AHT was of 30.37 seconds, a little higher than the 29 seconds premise, due to the fact that a few of them were handled by slower agents. The basic calls awaited, in average, 6.24 seconds before being handled, while the plus calls awaited only 1.34 seconds. Such disparity is due to the handling preference for the latter calls.

The basic operator utilization rate was 89.59%, very similar to the plus one (88.84%), mainly because both types of agents were qualified to handle both types of calls. Both types of agents were most part of the time (51-52%) handling the preferential and more numerous plus clients, spending the remaining time answering basic clients (37-38%) or idle.

The decision on how many operators shall be scheduled and on which basis they should be segmented or aggregated in each time band, is entirely up to the Contax planning management. Nevertheless, it may be possible that the company is interested on the analysis of the impact of other variables – that are not under their control (parameters) – on the service level and abandonment rate for the central desk, which would be possible using Simulation, according to section 2.

The scenario analysis can be used to find out what would happen with these performance indicators if – for instance – the calls volume during a given time band was 10% higher than the forecast.

During the simulation of this scenario (but with the same handling staff), 637 calls were handled, in average, from which 540 – in average – before 10 seconds. The service level for this scenario was consequently of 84.68%. This value is barely lower than the 85% established goal for plus clients and fairly lower than the 93.31%, that would be obtained if the demand had behaved in accordance to the forecast.

From the 672 calls generated in each replication, 32.2 were disconnected, in average, by the clients, implying an abandonment rate equal to 4.80%, which reveals a great impact of the added demand to this performance indicator.

Thus, a simulation of this scenario showed that the original operators contingent (12) – facing an unexpected 10% demand increase – would be able to practically assure the service level goal of 85%, but would also interfere too badly with the abandonment rate.

And what would happen with the service level and the abandonment rate if the demand was underestimated (also in 10%), although not in relation to its volume, but to the AHT?

In this scenario, 575 calls were, in average, handled; from these, 482 (in average) before 10 seconds. The resulting service level was of 83.85%, a value a little lower than the goal (85%) and fairly lower than the 93.31%, that would have been obtained in case the demand had behaved according to the forecast plan.

From the 606 calls generated in each replication, 29.8 in average were disconnected by the clients. The consequence is an abandonment rate equal to 4.92%, revealing a great impact of the AHT on this indicator.

Like the scenario that reflected an increase on the calls amount, the simulation of this situation showed that the original contingent of 12 agents – before an all of a sudden 10% increase on the AHT – would be able to assure a service level almost equal to the 85% goal (in this case, a little more distant), as well as to increment (a little more) the abandonment rate.

Based on the scenario analysis with a demand more intense than the forecast, it is possible to conclude that a not too large variation (10%) in relation to the forecast values, may impact directly the performance indicators, especially the abandonment rate. This conclusion requires much care for the calls amount and AHT forecast. Another important revelation is related to the probably surprising fact that the impact can be even higher when the increase occurs with the AHT, when compared to a same magnitude difference on the calls amount. This may lead to an accuracy on the AHT forecast being even more important than the accuracy on the calls volume forecast.

Therefore, it could be interesting to investigate the impact of higher variations on the AHT (for more and for less) on the system performance, through a more complete sensitivity analysis.

The simulation of the basic model was replicated in the software a few times, but with different values for the mean of the handling time (from amounts 30% lower to 30% higher), from where the relevant results were collected. These results were then organized on Table 3, which follows and also computes and presents the main performance indicators, i.e., service level and abandonment rate for each scenario.

Table 3 – Service level and abandonment rate for different values for AHT,

from 00:00 a.m. to 00:30 a.m., Aug/06, 12 operators

AHT (sec)	Δ	Calls				Service level	Abandonment rate
		created	handled	before 10 sec	abandoned		
20,5	-30%	583	582	582	0,84	99,92%	0,14%
23,4	-20%	586	583	580	2,68	99,50%	0,46%
26,4	-10%	589	582	571	6,34	98,13%	1,08%
29,3	**-**	**595**	**579**	**541**	**14,52**	**93,31%**	**2,44%**
32,2	+10%	606	575	482	29,82	83,85%	4,92%
35,2	+20%	640	567	356	68,39	62,80%	10,69%
38,1	+30%	683	548	223	128,01	40,77%	18,74%

Source: Table elaborated with results obtained by software

Therefore, higher AHT values than the forecasted ones rapidly deteriorate the system performance in terms of service level and abandonment rate (especially this last one), revealing an enormous potential impact of that variable on the most important results. This way, Contax should devote its best efforts to avoid the AHT increase, making their operators conscious about the destructive consequences of an increase on this value.

Analyzing the upper part of Table 3, it is possible to conclude that performance indicators improve considerably after a reduction of only 10% on the AHT: the service level increases in almost 5 percentual points, exceeding 98%, and the abandonment rate falls to less than its half (1.08%). This suggests that it might be worth while to invest on agents training, in order to try to reduce a little the handling time. Unfortunately, it is also true that higher reductions on this time do not hold advantages so remarkable (especially for the service level, already found close to its optimum value) for the system. In other words, the cost involved in diminishing the AHT in 10% may be possibly compensated by the benefits resulting from this reduction, but it is difficult to believe that the same would occur with more drastic reductions on this variable.

Similarly to this, there should be other questions addressing the same kind of issues: (i) what would happen with the abandonment rate if the client became more impatient and began to disconnect calls after, for instance, 1.5 minutes (instead of 2.5) in average, without being handled? (ii) what would be the impact of this change on the service level?

At this new scenario proposed and simulated, 580 calls, in average, were handled, from which 551, in average, before 10 seconds. The service level for this scenario was, consequently, of 95.09%. This value is a little higher than the 93.31% that would have been obtained with the previous clients' abandonment behavior, as well as fairly higher than the 85% goal. Even increasing in 40% "the clients' impatience", i.e., diminishing the average waiting time before disconnectment from 2.5 minutes to 1.5 minutes, the impact on the service level was small. Maybe one should expect a higher increase on this indicator, due to the fact that if there are more clients quitting the queue, it would be more usual that the remaining calls could wait less before being handled.

What happens is that, according to this model, 80% of the disconnected calls return to the queue a few minutes later, overloading the system once again and not allowing the service level to increase so much. This type of analysis and conclusion would be practically impossible through analytical approaches, which do not consider the abandonment behavior.

If the management is interested only on the service level, it might not become so interesting to make great efforts in order to forecast with great accuracy the clients' average waiting time before disconnecting the call. This is due to the fact that a not so small change of 40% on this average time is incapable of

impacting intensively the service level. If, however, there is an interest on monitoring the abandonment rate as well, it is fundamental to analyze the impact of the new scenario on this indicator.

From the 600 calls generated in each replication, 19.3, in average, were disconnected, generating an abandonment rate of 3.21%, fairly higher than the previous 2.44%.

In order to verify in a more complete way the performance indicators sensitivity in relation to the average waiting time before disconnectment (AWTBD), the same simulation was repeated a few times by the software, with different values for this variable. The needed outputs were collected to compute the main performance indicators (service level and abandonment rate), which are presented on the following Table 4.

Table 4 – Service level and abandonment rate for different values for the AWTBD, from 00:00 a.m. to 00:30, Aug/06, 12 operators

Average waiting time (minutes)	Calls				Service level	Abandonment rate
	created	handled	before 10 sec	abandoned		
0,5	624	578	562	46,65	97,26%	7,47%
1,5	600	580	551	19,25	95,09%	3,21%
2,5	**595**	**579**	**541**	**14,52**	**93,31%**	**2,44%**
3,5	593	582	540	10,02	92,75%	1,69%
4,5	591	582	534	9,10	91,91%	1,54%

Source: Table elaborated from the results obtained by the software

As one may observe, the abandonment rate is fairly sensitive to the AWTBD, especially at lower levels for this variable. Therefore, its correct consideration seems to be important on the obtainment of accurate results, even though the fact that the service level reveals a small sensitivity to changes at the same variable.

The planning manager can also be interested on knowing what would happen if the contracting company became more demanding in relation to the service level and came to redefine its concept, changing it to correspond to the percentual amount of clients that waited less than 5 seconds (instead of 10) before being handled.

In the simulation of this scenario, 579 calls in average were handled, from which 489, in average, before 5 seconds. The service level for this scenario was, consequently, 84.44%. This value is fairly lower than the 93.31% obtained with the original definition of service level, and, what is more important, a little lower than the 85% goal established for plus clients.

This comparison reveals that a redefinition on the concept of the service level can impact by a not so small way this performance indicator, something reasonably expected. It is even possible that, exactly like it happened with the illustrated example, the current operators configuration becomes not sufficient anymore to offer a service level in accordance with the pre-established objective. In this case, it may be important to find out how many additional agents would be needed to allow this performance indicator to go back to levels higher than the goal.

In order to find this out, it is necessary to verify if the addition of 1 agent only is good enough for the objective to be achieved. In the simulation of this scenario with 13 operators, 580 calls were handled, in average, 534 of which (in average) before 5 seconds. The resulting service level was 91.99%, a value higher than the goal (85%) and than the 84.44% obtained with 12 agents, but a little lower than the original 93.31%.

This shows that the hiring of an additional agent is capable of making the service level go back to the expected goal, but its benefits do not compensate the negative impact on this performance indicator caused by the redefinition of its concept.

The main results achieved by means of these scenarios and sensitivity analyses are summarized and commented on the following Table 5.

Table 5 – Main results obtained by scenarios and sensitivity analyses

Scenario		Impact on performance indicators	
		Service level	Abandonment rate
Handling time variance increases		Small increase	Erratic behavior (no bias)
Handling time distribution: Lognormal instead of Erlang		No changes	No changes
Amount of operators decreases		Huge decrease	Gigantic increase
Handling aggregated format	Plus clientes	Fair increase	Fair decrease
	Basic clients	Above the goal	Acceptable
Calls volume increases		Fair decrease	Big increase
AHT increases		Fair decrease	Big increase
AHT decreases		Fair increase	Big decrease
Clients become more impatient		Small increase	Fair increase
Clients become less impatient		Small decrease	Fair decrease
Service level concept becomes more demanding		Fair decrease	No changes
Service level concept becomes more demanding + 1 more operator		Small decrease	No changes

Source: Table elaborated by the author

4. CONCLUSIONS

During this research, several simulation models were built, completing different real call centers features and for different possible alternative scenarios, in order to compute performance indicators and suggest solutions concerning operation sizing. In general, it was clear that the Simulation allows an easy evaluation of the impact of changes on the original characteristics of the operation on the performance indicators. It also allows one performing several sensitivity analyses related to a few operational parameters.

Specifically, the scenario and sensitivity analyses developed during this research made us visualize how Simulation can give support to decisions related to the process of dimensioning a call center, since the results mainly revealed that: (i) it is possible to reduce the operator contingent in some of the time bands of the day, without much interference on achieving the service level goal; (ii) not too intense variations on the received call volumes and on the mean and variability of the handling time can impact a great deal on the performance indicators, especially the abandonment rate, pointing to the need to forecast these values with much accuracy; (iii) it is possible to improve considerably the handling performance for preferential clients, without much interference on the quality of the service for basic clients, if an aggregated handling format, with priorities, is adopted; (iv) in case the clients become more impatient and disconnect the calls in a faster way, the impact on the service level would be small, but very significant on the abandonment rate.

4.1 Suggestions and recommendations

In order to shape more accurately the situations to be simulated, it would be interesting if future researches could make an effort on the direction of finding out (based on the calls historical map) the correct statistical distribution and the variability for the time between incoming calls and for the handling time. Several Simulation studies assume the basic premise that these times follow an exponential distribution and develop researches related only to these variables means for each time band. But the results obtained can be sensitive to the distribution format and to the variability of these times.

Following this same way of thinking, an empirical research could collect information related to the impact caused on the handling time when calls are not answered by the type of operator used to do so, in a consolidated handling system for different types of calls. The right consideration on this impact tends to generate more accurate results for indicators on call centers with aggregated handling.

Another issue that certainly reveals a great potential of operational improvement to be analyzed in future researches is related to the multi-product operator

(CAUDURO et al., 2002). There is a feeling about being economically advantageous to use the same operator to handle two or more different operations at the same time in order to reduce his idle time. This could happen in case of relatively similar operations on which the same operator could work and which present complementary demand behaviors along the day, the week, or the month. The achieving of this supposed economical advantage in terms of cost-benefit could be checked through a well detailed simulation model, as suggested by Bouzada (2006, p. 244-245).

REFERENCES

Anton, J. (2005), "Best-in-Class Call Center Performance: Industry Benchmark Report", *Purdue University*.

Alam, M. (2002), "Using Call Centers to Deliver Public Services", *House of Commons Paper*, London: The Stationery Office Books.

Araujo, M., Araujo, F. and Adissi, P. (2004), "Modelo para segmentação da demanda de um call center em múltiplas prioridades: estudo da implantação em um call center de telecomunicações", *Revista Produção On Line*, Vol. 4, N. 3, p. 1-20.

Avramidis, A. and L'ecuyer, P. (2005), "Modeling and Simulation of Call Centers", *Winter Simulation Conference*, p. 144-152.

Bapat, V. and Pruitte Jr, E. (1998), "Using simulation in call centers", *Winter Simulation Conference*, p. 1395-1399.

Bouzada, M. (2006), *O uso de ferramentas quantitativas em call centers: o caso Contax*, Thesis (Ph. D. in Business Administration), Rio de Janeiro: UFRJ/COPPEAD.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyin, S. and Zhao, L. (2002), "Statistical analysis of a telephone call center: a queueing-science perspective" (working paper 03-12), *Wharton Financial Institutions Center*.

Cauduro, F., Gramkow, F., Carvalho, M. and Ruas, R. (2002), "O Processo de Mudança e Aprendizagem no Call Center de uma Empresa de Telecomunicações", *EnANPAD*, p. 1-13.

Chassioti, E. and Worthington, D. (2004), "A new model for call centre queue management", *Journal of the Operational Research Society*, Vol. 55, p. 1352-1357.

Chokshi, R. (1999), "Decision support for call center management using simulation", *Winter Simulation Conference*, p. 1634-1639.

Contax (2006), *Contax Contact Center*, <www.contax.net.br>.

Grossman, T., Samuelson, D., Oh, S. and Rohleder, T. (2001), *Encyclopedia of Operations Research and Management Science*, Boston: Kluwer Academic Publishers, p. 73-76.

Gulati, S. and Malcolm, S. (2001), "Call center scheduling technology evaluation using simulation", *Winter Simulation Conference,* p. 1841-1846.

Hall, B. and Anton, J. (1998), "Optimizing your call center through simulation", *Call Center Solutions Magazine*, p. 1-10.

Hawkins, L., Meier, T., Nainis, W. and James, H. (2001), *Planning Guidance Document For US Call Centers*, Maryland: Information Technology Support Center.

Klungle, R. (1999), "Simulation of a claims call center: a success and a failure", *Winter Simulation Conference,* p. 1648-1653.

Klungle, R. and Maluchnik, J. (1997), "The role of simulation in call center management", *MSUG Conference,* p. 1-10.

Lam, K. and Lau, R. (2004), "A simulation approach to restructuring call centers", *Business Process Management Journal*, Vol. 10, N. 4, p. 481-494.

Mehrotra, V. (1997), "Ringing Up Big Business", *OR/MS Today*, Vol. 24, N. 4, p.18-24.

Mehrotra, V. and Fama, J. (2003), "Call Center Simulation Modeling: Methods, Challenges and Opportunities", *Winter Simulation Conference,* p. 135-143.

Mehrotra, V., Profozich, D. and Bapat, V. (1997), "Simulation: the best way to design your call center", *Telemarketing & Call Center Solutions*, p. 1-5.

Miller, K. and Bapat, V. (1999), "Case study: simulation of the call center environment for comparing competing call routing technologies for business case ROI projection", *Winter Simulation Conference,* p. 1694-1700.

Outsourcing (2005), *Ranking*, <www.callcenter.inf.br>.

Paragon (2005), *Simulação de Call Center com Arena Contact Center,* <www.paragon.com.br>.

Saltzman, R. and Mehrotra, V. (2001), "A Call Center Uses Simulation to Drive Strategic Change, *Interfaces*, Vol. 31, N. 3, p. 87-101.

Yonamine, J. (2006), *O Setor de Call Centers e Métodos Quantitativos: uma Aplicação da Simulação*, Dissertation (M. Sc. in Business Administration), Rio de Janeiro: UFRJ/COPPEAD.

AUTHOR'S BIOGRAPHY

Marco Aurélio Carino Bouzada is a Production Engineer (UFRJ, 1998), M. Sc. in Business Administration (COPPEAD/UFRJ, 2001) and Ph. D. in Business Administration (COPPEAD/UFRJ, 2006). Currently a professor belonging to the permanent staff of the Estacio de Sá University M. Sc. in Business Administration Program, to the Escola Superior de Propaganda e Marketing Graduation in Business Administration Program and to the COPPEAD/UFRJ Finance Specialization Program, with experience on topics like Statistics, Quantitative Methods, Operational Research and Business Games.