A FRAMEWORK FOR QUALITY IMPROVEMENT OF QUEUEING SYSTEMS OF THE SERVICE INDUSTRY

M. F. Ramalhoto
Instituto Superior Tecnico, Technical University of Lisbon, Department of Mathematics
Av. Rovisco Pais, 1049-001 Lisbon, Portugal. Phone: +351-1-7143311; E-mail: d554@alfa.ist.utl.pt

ABSTRACT: It is developed methodology that provides a way of incorporating quality management concepts of customer’s satisfaction into the queueing system design. A new concept to obtain relevant data called the manager’s tetrahedron is proposed. A framework called total-quality-queue-management, based on the manager’s tetrahedron and on adequate queueing theory and decision making techniques, is outlined and discussed. This framework consists of four components: a stochastic demand model, a decision system, an outcome calculator and a scoring system.

1. Introduction

The percentage of service industries is far greater than the manufacturing industries in all countries. And manufacturers have also come to recognize that service (for instance, maintenance after sales) is now an essential element of competitiveness. In the so called “service factory” concept, the service is identified as the “fifth competitive priority” (as opposed to the traditional four competitive priorities-cost, quality, delivery and flexibility) in the manufacturing strategy.

In the service industry there are essentially two types of products to be considered: the product-service and the product-supply. The product-service can be defined as how has the service been provided and the product-supply is what has been provided. The product-service is provided, usually, through a service delivery process. The service delivery process is essentially described by a queueing model. One of the ultimate targets of most queueing system managers is to develop policies to provide consistently high product-service, for a wide range of customer’s types and arrival and service rates at “reasonable” costs. Section-2 discusses the product-service. Section-3 introduces a methodology that establishes the link between queueing systems and quality management. The Garvin’s quality dimensions-performance, flexibility, serviceability, reliability, courtesy and appearance, called here external quality dimensions are reviewed and interpreted in a queueing context. The quality dimensions of the service industry: timeliness, integrity, predictability and “customers satisfaction” are elected as the most relevant internal quality dimensions of the queueing system. A new concept called manager’s tetrahedron concept is defined. Section-4 deals with examples of recent research on queueing models that address flexibility, timeliness and customers perceptions of waiting and service, respectively. Section-5 outlines and discusses an empirical decision framework based on the manager’s tetrahedron concept and on the queueing model options selected to improve the quality of a queueing system of the service industry.

2. The Product-Service

In the quality improvement of the product-service we have to consider that: 1. Queueing and waiting in general are at the same time personal and emotional. On assessing the effectiveness of a product-service quantitative and qualitative aspects of human behavior towards waiting have to be addressed. If customers are pleasantly occupied while waiting: entertainment, social relevant information, job opportunities advertising, education, doing his/her own service, extra information about the queueing system itself (announcing delay estimates, guarantee a maximum possible delay or x% off, etc.), and if “social justice” in lines are preserved (single-line serpentine lines), or if the product-supply is a delight, then the customer’s perception of the length of the waiting time and its "reasonable" time limits may differ substantially. 2. Unlike the product-supply that, usually, can be sampled and tested for quality, the product-service can not easily. It is then important that the operator feels the pride of contributing to the higher standards of that queueing system. However, as it is recognized by several authors people who work in queueing systems, usually, are not aware that they too have a product to sell, and that this product is the service they are providing. 3. The product-service is delivered at the moment it is produced. In high competitive service industries customers will brand switch as a result of just one bad service encounter. To alleviate this time constrains it is important to have a communication mechanism in operation. The success depends heavily on
showing the customers that they have been heard by the system managers and their relevant opinions really make a difference. 4. Product-service quality always be a kind of balance between the expectations that the customer had and their perception of the product-service received. A higher quality product-service is one where the customer's perceptions meet or exceed their expectations. The primary area of difficulty is to identify appropriate means of communication. The so-called quality dimensions are used also to serve as a common language among the customers, operators and managers.

3 Quality Dimensions and the Manager’s Tetrahedron Concept

In general, there is different quality dimensions for the customer and for the manager. Therefore, I shall classify the quality dimensions into external (customer) and internal (manager). The quality dimensions: performance, flexibility, serviceability (responsiveness), reliability, courtesy (empathy) and appearance (tangibles) are called external quality dimensions. Performance is the primary operating characteristic of the queueing system. It can be “measured” by, for instance, the "absence or perceived absence of waiting time", "total sojourn time in the system not exceeding \( x \) units of time", “waiting versus competitive price”, etc. Flexibility is the queueing system’s built in ability to quickly respond to the changes of demand. It can be “measured” by, for instance, the duration of a traffic peak (how quickly it gets ride of it). Serviceability (Responsiveness) is the ability of the queueing system to respond to the individual needs of a particular customer. It can be measured by, for instance, the time to respond to those individual needs, to answer complains, etc. Reliability is the ability to always perform the product-service dependably, knowledgeably and accurately, and as expected by the customer. Courtesy (Empathy) is the caring, individualized attention provided to the customer, the effort to understand the customer’s needs, the ability to convey trust and confidence. Appearance (Tangibles) is the quality appearance of the physical environment and materials, facilities, equipment, personnel and communications used to produce the product-service. To quantify these last two quality dimensions is also required joint research work with other specialists, namely in human behavior, to set up the right questions to lead to the adequate way of quantifying them. These quality dimensions are of great value as system’s improving facilitators, but not necessarily in the ongoing business of monitoring and improving product-service quality and costs reduction. They reflect the views of the customer and not necessarily the real state of the system. They indicate the targets, under the point of view of the customers.

The manager needs to deliver what the customer expects, at reasonable prices, maintaining high productivity and profitability. For that, the adoption of the internal quality dimensions: timeliness, integrity, predictability, “customer satisfaction” might be quite useful. Timeliness is formed by the access time which is the time taken to gain attention from the system; the time queueing which is the time waiting for service (and can be influenced by the length of the queue and/or its integrity); the action time which is the time taken to provide the required product-service. Integrity deals with the completeness of service and must set out what elements are to be included in order for the customer to regard it as satisfactory. This quality dimension will set out precisely what features are essential to the product-service. Predictability refers to the consistency of the service and also the persistence or the frequency of the demand. Standards for predictability identify the proper processes and procedures that need to be followed. It may include standards for availability of people, materials and equipment and schedules of operation. Customer satisfaction is defined as the way to provide the targets of success, which may be based on relative market position for the provision of a specific queueing system.

Let me now introduce the manager’s tetrahedron concept. The tetrahedron’s vertices are, respectively, the customer, the market competitor, the operator (the heart of the production process), and the manager. All the links among these vertices are relevant for the quality improvement of the queueing system of the service industry. Ideally, the manager would be able to understand at any time the interactions among all of them. The external and internal quality dimensions represent steps towards a manager’s tetrahedron. The aim is to build up a fair partnership between the queueing system’s manager and the operators, customers, and in some cases even market competitors to facilitate easy and efficient communications, and commitment to the quality improvement and costs’ reduction of the queueing system, following a win-win strategy.

The internal quality dimension timeliness is linked to the theory of queues. Namely, the access time is linked to the theory of retrial queues, and the action time is the service time in the traditional queueing theory language. Parasuraman, Bury and Zeithami (1991), through external quality dimensions, have also
identified the service delivery process as the key to improving product-service quality and building customer loyalty. To improve the service delivery process, often means to improve the queueing model behind it. Most of the “product-service’ s failures/defects” are linked to “unacceptable access time” and “unacceptable queueing time”. The way of preventing those failures/defects relies mainly on adequate service delivery designs.

4 Queueing Theory in the Service Industry

From the queueing physics, born in 1915 with Danish telephone engineer Erlang, we can prove that in “crucial times” if nothing is done to spread out the arrival pattern, or to change the service rate the queueing system experiments very uneven traffic flows, and big failures/defects occur in the product-service. However not all the classical queueing results are sufficient or even adequate to the queueing problems posed by the service industry. Also very often, the arrival rate to the queueing system of the service industry is rather burst or subject to random fluctuation or periodically time-dependent. To design these queueing systems to specially meet the peak demands is not always the best action to take. Because, it can be costly, and the excess capacity can have negative psychological effects on the customer. Let me give some examples: (a) Ramalhoto and Syski (1996) have proposed and studied a queueing model that aims to provide managers with a way of dealing with some temporary peak situations, that is to say, to have high ranking in the flexibility quality dimension. The model is essentially a multi-server queue under the following additional decision rule-1: if the queue size exceeds \( b \) (the action line), introduce another server (or \( k \) servers \( k \geq 1 \)); when it falls below \( a \) (the prevention or alarm line), withdraw one server (or \( k \) servers, \( k \geq 1 \)), \( b > a \). Other rules could be considered as alternatives to rule-1 such as, for instance, rule-2: when the queue size exceeds \( b \), shorten the service time (by deferring some tasks to be worked out later, by dividing and scheduling when the service can be provided in multiple separate segments, or by reducing the quality of service, etc.). Or rule-3: identify classes of service needed by customers, each class requiring a different service time and being of different “value”, and treat the customers in separate queues when the total queue length exceeds \( b \). Which rule is preferable?. Section 5 addresses this question. (b) Usually, a customer whose call for access to the queueing system is unsuccessful at first attempt will repeat the call, once or several times, in quick successions thus giving rise to the phenomena of repeated attempts. That might cause serious difficulties if the queueing design of the system ignores this phenomena, which very often is the case in spite of the access time being part of the timeliness quality dimension. Ramalhoto and Gomez-Corral (1998) have provided heuristic numerical approximations to model this phenomena. (c) Queueing models with time-dependent arrival rates are obviously relevant for the design of queueing systems of the service industry. As well as queueing models that consider “affinity servers”, and the customers’ perceptions of waiting and service. As Larson (1998) has pointed out for the great majority of queueing systems customers the actual and/or perceived utility of participating in the system is a nonlinear function of queueing delay and multi-attributed. And the nonlinearity of the delay suggests, for instance, limits for the famous Little’s formula \( L=\lambda W \) (where \( L= \) mean queue length, \( \lambda=\)arrival rate, and \( W= \) mean waiting time).

Moreover, most of the probabilistic results needed to understand and control the stochastic behavior of the queueing systems of the service industry can not all be determined mathematically, and need an interdisciplinary approach. In the last section it is proposed a framework with the ability to jointly consider data management (from the internal and external quality dimensions selected), service delivery process design (from the set of the most adequate queueing models available) and decision making based on cost-benefit.

5. The Total-Quality-Queue-Management Framework

Usually, there are more than one queueing model able to respond to the needs of improving or redesigning of a particular service delivery process. Each queueing model option might lead to for instance different levels of delay, impact on “customer satisfaction”, and different costs. The aim, is to find a way that balances the customer waits and discomfort against operators idleness and costs.

Next, it is formulated a heuristic simulation-decision framework. That allows the evaluation of alternative queueing model options, and makes the necessary decisions by selecting those particular options that provide the best projected performance scores, in terms of a specified scoring criteria, based on the measures linked to the quality dimensions selected. It is called total-quality-queue-management framework.
The queueing model options are defined as “control parameters” in this framework. For instance, the queueing models corresponding to rule-1 to rule-3 in 4.(a), can be represented quantitatively by the following basic control parameters: $X_1$ - the regular size of the service staff. $X_2$ - the percentage by which the service times for each customer are to be reduced or expedited (as a function of queue length, or any other relevant quantity). $X_3$ - the amount by which the regular size service staff is augmented by other personnel (such as secretarial or clerical staff to meet periods of heavy demand). $X_4$ - the number of different classes of service needed by customers. $X_5$ - the percentage of the regular size service staff to allocate to each one of those different classes of service. Basically, the total-quality-queue-management framework consists of four components: a stochastic demand model, a decision system, an outcome calculator and a scoring system. The stochastic demand model represents our projection (and the uncertainties in our projection) of the rates of arrival and service requirements of the customers. The decision system searches systematically over the multidimensional space defined by the control parameters in order to find an optimal combination of values, $X_1^*, ..., X_5^*$, for these control parameters that will yield the “best” system performance, given the stochastic demand that has been specified for the particular problem. To enable the decision system to compute and evaluate the consequences of any specific set of control parameter values, it has to utilize the results of the outcome calculator and the scoring system. The outcome calculator and the scoring system have to be constructed as entirely separate and independent systems. The outcome calculator calculates (or projects) the specific outcome(s) that will result from any specific assumptions concerning the customers demand and any specific decision concerning the values of the control parameters. In particular, for any such combination of assumptions, the outcome calculator must be able to compute the pertinent outcome parameters (which are defined in terms of objective physical quantities such as the queue lengths, the customer waiting times, the service costs, and other pertinent descriptors of the outcomes), that may be needed to evaluate the real situation in terms of the selected quality dimensions. Clearly, the outcome calculator is concerned with the objective physical outcomes of the queueing system. The scoring system is concerned with the subjective “desirability” of the outcomes in terms of the customer’s expectations, perceptions of waiting and service, and current goals and objectives. The purpose of the score system is to assign to each outcome a ranking of the quality dimensions selected.

One of the interesting features of this framework is that we have two ranking schemes for the quality dimensions selected. The first is an inevitable consequence of the structure of the queueing system and its relevant physical laws (it reflects the voice of the real system), and the second reflects the customers’ perceptions and expectations of the queueing system (it reflects the voice of the customer). So, the comparison of both rankings might be very important to the queueing system’s learning process.

In conclusion in this paper the product-service is treated as “manufacturing in the field”. It is advocated that the product-service should be carefully planned, audited for quality control, regularly reviewed for performance improvement and customer’s reaction. And the framework proposed to achieve it relies on team work and reliable data through the manager’s tetrahedron, adequate queueing theory and decision making techniques.

REFERENCES


FRENCH RESUMÉ: Une méthodologie est développée pour permettre d’incorporer concepts d’amélioration de la qualité dans modèles mathématiques classique des file d’attentes. Un nouveau concept d’obtention des données clés est proposé sous le nom de “manager’s tetrahedron ”. Il y est proposé et debattu en quadre d’idée appelée “total-quality-queue-management” se basant sur le “manager’s tetrahedron” et sur la file d’attente “optimal” et les techniques de decision.