Industrial Management & Data Systems

Optimal design of a multi-server queueing system with delay information
Miao Yu Jun Gong Jiafu TANG

Optimal design of a multi-server queueing system with delay information

Miao Yu, Jun Gong and Jiafu Tang
College of Information Science and Engineering, Northeastern University, Shenyang, China

Received 21 May 2015
Revised 16 July 2015
4 August 2015
Accepted 10 August 2015

Abstract

Purpose – The purpose of this paper is to provide a framework for the optimal design of queueing systems of call centers with delay information. The main decisions in the design of such systems are the number of servers, the appropriate control to announce delay anticipated.

Design/methodology/approach – This paper models a multi-server queueing system as an M/M/S + M queue with customer reactions. Based on customer psychology in waiting experiences, a number of different service-level definitions are structured and the explicit computation of their performance measures is performed. This paper characterizes the level of satisfaction with delay information to modulate customer reactions. Optimality is defined as the number of agents that maximize revenues net of staffing costs.

Findings – Numerical studies show that the solutions to optimal design of staffing levels and delay information exhibit interesting differences, especially U-shaped curve for optimal staffing level. Experiments show how call center managers can determine economically optimal anticipated delay and number of servers so that they could control the trade-off between revenue loss and customer satisfaction.

Originality/value – Many results that pertain to announcing delay information, customer reactions, and links to satisfaction with delay information have not been established in previous studies, however, this paper analytically characterizes these performance measures for staffing call centers.

Keywords Service systems, Balking, Predicting and announcing delays, Reneging, Satisfaction with delay information

Paper type Research paper

1. Introduction

For the design and management of call centers and many other service systems, it is common to use delay information system informing customers about their anticipated waiting time. Delay information system captures customer psychology associated with the uncertain waiting. Managers have several objectives in providing such delay information, such as modulating demand by announcing times of high congestion, enhancing satisfaction with inevitable waiting, in all, stirring the system in order to maximize her revenues. The purpose of this paper is to study this feature, its impact on the performance of the system, which directly address the delay information satisfaction, customer abandonment, and some important service levels in order to maximize the firm revenue.

Delay announcements affect customers' behavior in terms of abandonment (balking and reneging), in turn, and have significant impact on system performance. When the system announces a customer about her anticipated delay, she will decide right away either to hang up immediately according to her estimation that her delay is too long or to start waiting in the queue. So delay information modulates customer reactions

This project is supported by the National Nature Science Foundation under Grant No. 71271052 and Fundamental Research Funds for the Central Universities (No. N120804003).

Industrial Management & Data Systems
Vol. 116 No. 1, 2016
pp. 147-169

0263-5577
DOI 10.1108/IMDS-05-2015-0201

(Taylor, 1994). At the same time, for customers who enter the queue, delay information, would further have the effect of increasing patience as a result of reducing the uncertainty. Furthermore, when providing delay information is longer than the virtual delay in waiting experiences, customer may not choose the call center because of no trust any more. Therefore, making delay information is especially important in call center.

In call center settings, the service encounter is unlike face-to-face service encounters at other service sites, such as restaurants, hotels, and banks. This service system provides a service in a kind of invisible queue, therefore, the best means of providing and controlling customers' satisfaction may be providing products and service efficiently and quickly in call centers. Call center managers usually deal with satisfaction issue by traffic management. Currently, a large number of service industry observers have emphasized the importance of incorporating satisfaction metrics into these firms to balance customer behavior and cost control. So call centers need to balance the traditional efficiency and quality needs with the emphasis on customer relationship management. That is instigating similar changes in the functioning of call centers.

To the best of our knowledge, that is the first generalization of delay information queuing model to incorporate customer satisfaction. In particular, previous researches usually ignored the role of satisfaction with the delay announcement economically. That is the true in practice, customers may feel dissatisfied even entering service thereby never choosing the firm because of no trust with delay information. In addition, measuring customer satisfaction can be done through other process-related metrics by accounting for some details, such as short abandonments and quick answers. However, these metrics have not been detailed in such a system with delay information.

As call centers have matured as the main service delivery channel of some service industries, it role has become a revenue generator for the firm. For revenue management, there is the main challenging problem related to capacity sizing. On the one hand, revenue is generated by serving a customer, and high-quality service because of higher staffing level determines more revenue. On the other hand, staffing cost plays a crucial role in the determination. Finding the optimal trade-off becomes equivalent to maximize total revenue. However, optimality of staffing has mostly been viewed as a cost minimization issue (Mandelbaum *et al.*, 2002; Baron and Milner, 2009). Few papers have focussed on call center profit when they make a capacity sizing decision.

In predicting delays for arriving customers, this paper builds on previous analytical studies on the work of Jouini *et al.* (2011). Furthermore, some extensions make the performance prediction with delay information by the queueing model much more accurate. The model of single-class call centers will not be directly applicable to many current call centers, nevertheless, we think these results and analysis could provide useful and new insight.

The main contributions of this paper can be summarized as follows. First, we propose to characterize customer satisfaction with delay information to modulate customer reaction, and develop an approach to explicitly derive the expression. Second, we provide a comprehensive list of metrics that including customers satisfaction and abandonments. In such a system with delay information, what is new here is we propose new metrics and develop an approach to explicitly derive their expressions. Third, that abandonment form in this system means the loss of revenue, which allows us to capture the characteristic that revenue are a direct function of capacity sizing. We focus on these call centers offering service and bringing revenue, in which of these features have already been analyzed in the call center and literatures (similar to the case in Aksin and Harker (2003)).

The rest of the paper is organized as follows: In Section 2, we discuss a brief review of literature. In Section 3, we formulate impatient behavior and delay information satisfaction in a queuing system, by which we give some metrics that include abandonments and customer satisfaction. Then, we show how to explicitly computer these service levels in a convenient way. In Section 4, we use some extensive numerical experiments to show the role of satisfaction with delay information in the performance measure call center, then conduct a numerical analysis in which we draw comparisons between the service levels. In Section 5, We conduct a numerical analysis on how to make optimal announcement choice and staffing level under various customer reaction and system parameters. In Section 6, we provide a framework for the optimal design of call centers with the right metric to avoid some unwanted behavior, and we also show the negative effect on optimal revenues after using some specific service level constraint. Finally, in Section 7, we generalize the key insight derived from the analysis, and conclude with the limitations of our results and future research directions.

2. Literature review

The literature related to the subject of this paper spans mainly three areas. The first area is concerned with predicting delays for arrival customers with the psychology of waiting. The second area pertains to abandonment phenomena. The third area is related to customer satisfaction and staffing. None of this literature considers the model with a combination of all of these factors.

The relevant literature on delay information is large and growing. In broad terms, there are three mains areas of research on delay information. The first area studies the effect of delay information on system dynamics. One of the first representative papers are Hassin (1986), then a large number of studies on the impact of delay information subsequently focussed on the system performance in the invisible queue; e.g., see Whitt (1999), Guo and Zipkin (2007), Allon *et al.* (2012), Aksin *et al.* (2013), Jouini *et al.* (2014), and references therein. Research has found that informing customers of delays is beneficial regardless of the model used, but the optimal amount of precision in the announcements varies from model to model. Furthermore, the importance of modeling customer responses in the following literature is emphasized. Armony *et al.* (2009) study customer responses to delay information by requiring an equilibrium analysis. Jouini *et al.* (2011) study a model where customers react by hanging up immediately upon hearing the delay announcement if the announced waiting time is too long and might subsequently renege because of impatience. As shown in the recent work by Yu *et al.* (2015), they posit that delay announcements impact customer behavior in a complicated way. The second area studies alternative ways of estimating customer delay in service systems; e.g. Nakibly (2002), and Ibrahim and Whitt (2009a, b, 2011). The third area is mainly focussed on customer psychology in waiting situations, which subsequently leads to customer response: e.g., see Hui and Tse (1996) and Munichor and Rafaeli (2007). This paper falls into the first main area of research.

Moreover, another direction in modeling abandonments without delay information has reference significance (Gans *et al.*, 2003; Mandelbaum and Zeltyn, 2009). An important feature of call center is customer abandonment, since it is a real phenomenon that delayed customers do not accept waiting in a call center. So far, there are some empirical evidences regarding abandonments that can be found in Brown *et al.* (2005) and Feigin (2005), they, respectively summarize the abandonments analysis of a real record in a banking call center. On theoretical aspect, lots of theoretical models can be proposed, such as the simplest Erlang A model. Garnett *et al.* (2002) propose an asymptotic analysis

of a Markovian model with abandonments in the heavy-traffic regime. They mainly characterized the relationships between staffing, the offered load, and system performance measures such as the probability of waiting time and the probability of abandonment. This can be viewed as an extension of Halfin and Whitt (1981) by adding abandonments.

A satisfied customer will have a higher preference to again choose service from the same firm than a dissatisfied customer. In the field of customer relationship management, Anderson and Sullivan (1993) have shown that customer satisfaction is a good predictor for the likelihood of repeated purchases and revenue growth. Some researches state that customer satisfaction increases the firm's profitability (Au *et al.*, 2002; Nie, 2000). For the field of call center, Dean (2002) points out some issue of service quality could affect customer loyalty in call center, and he investigates real consumers in an insurance company and bank using call center in order to validate his perspectives. For the significance of staffing, most call centers determine the minimum number of agents for satisfying customer waiting requirements (Kim and Ha, 2010). In the same time, there are some up-to-date researches with reference value for the optimal design for customer satisfaction and staffing in the system of internet chat services, e.g. Luo and Zhang (2013), Tezcan and Zhang (2014), and Zhang *et al.* (2011). Hence, the researches on customer satisfaction and staffing often go in tandem.

These literatures motivate some assumptions of our models. So far, these results that pertain to delay information satisfaction level, customer abandonment behavior and link to a staff dimensioning problem with revenue optimality have not been established. This distinction will be discussed further below.

3. Model overview

This section starts out with a formulation of the underlying queueing system, where anticipated delays are announced to customers upon arrival. Specifically, we take into account the queuing process with abandonment, and this model mainly characterizes satisfaction with delay information and other different metrics. The resulting model where the policy for each queue is first-come-first-served (FCFS), and abandonment is not allowed once a customer starts service.

3.1 The queueing model with delay information

We consider a firm which provides a service with a call center and delay sensitive customers, and the waiting time of a customer in this system with delay information could be anticipated and announced to the customer upon her arrival as delay information, as shown in Figure 1. Further, we assume that customers are differently aware of such delay information and it exactness because of different patience levels. Customers arrive according to a Poisson process with rate λ. There are s homogeneous servers, which is the decision variable in this model. The service times are independent and exponentially distributed with mean $1/\mu$. For these customers, assume that the initial random patience time of customers T are random and independent and identically distributed and under a given continuous distribution. Given that these patience time are exponentially distributed with parameter γ. System load ρ is written as $\rho = \lambda/s\mu$.

On arriving, a new customer could get service immediately if the number of customers in the system is less than s. If all of the agents are busy, the customers have a probability α_0 of balking before any delay information is provided. This features models a portion of extremely impatient customers who call with the idea to hang up at

once while they need to wait of service. The remaining customers may decide to balk due to the delay information or to accept the announced delay. Let d_n denote the announced delay and $p^B(n)$ denote the probability of balking when the random patience threshold exceeds the delay d_n. We assume that the balking behavior of customers is independent:

$$p^B(n) = P(T < d_n) = 1 - e^{-\gamma d_n}. \tag{1}$$

We denote the distribution of a customer's virtual delay after hearing the information by D_n, where n is the number of waiting customers ahead of her or him and virtual waiting is defined as before. Consider a customer who waits in the queue after hearing delay information, her updated patience threshold is still an exponentially distributed with rate γ'. Let $g_n(t)$ and $G_n(t)$ denote the probability density function of D_n and the cumulative distribution function of D_n. Because the random variable D_n is the downcrossing time from the state $s+n+1$ until absorption in the state s, D_n can be characterized by a hypoexponential distribution. Then, we obtain:

$$g_n(t) = \sum_{i=0}^{n} \left(\prod_{j=0, j \neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) (s\mu + i\gamma') e^{-(s\mu + i\gamma')t}, \tag{2}$$

$$G_n(t) = 1 - \sum_{i=0}^{n} \left(\prod_{j=0, j \neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) e^{-(s\mu + i\gamma')t}, \; t \geqslant 0. \tag{3}$$

In this paper, we announce to the customer the delay d_n as discussed and reviewed in (19), where d_n corresponds to a given coverage probability β, that is to say that the virtual delay of a new customer cannot exceed the anticipated delay with a probability β. It is given by:

$$d_n = G_n^{-1}(\beta). \tag{4}$$

Specially, there is a version of the fact that we announce the expected delay to customers, where instead of announcing the delay information ensuring β coverage. So we announce to the customer d_n the expected value of the random variable,

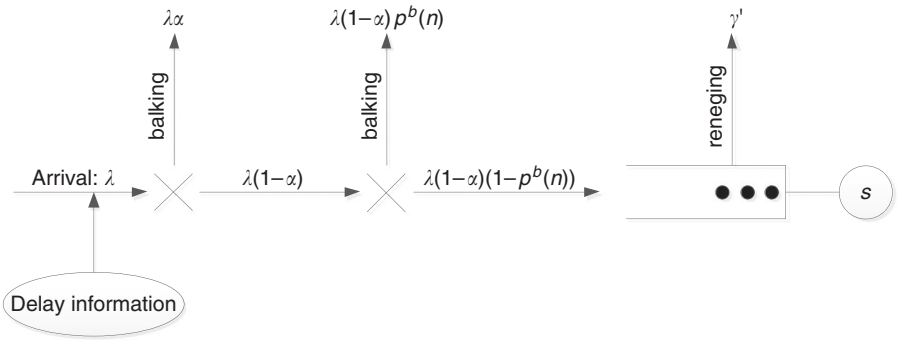


Figure 1.
A call center with delay information and customer reactions

denoted by $E(D_n)$. This substitutes D_n characterized by a hypoexponential distribution, we have:

$$E(D_n) = \sum_{i=0}^{n} \frac{1}{s\mu + i\gamma'} \tag{5}$$

In what follows, let t_k denote the initial patience threshold of the kth customer, in such a system, customers will update their patience threshold according to d_n and the initial random patience t_k upon hearing the delay information, as the value $\theta t_k + (1-\theta)d_n$, where $\theta > 0$. Here, θ represents the weight coefficient which determines the updated patience. Thus, we obtain the relationship between the initial patience threshold T and the updated patience threshold T'. This is the assumption above mentioned that the updated patience threshold T' is assumed to be exponentially distributed with rate γ'. In previous research, the quality of the approximation of the exponential distribution has been validated. Hence, in this paper, we still assess the assumption and make use of a numerical method to computer the rate γ'.

3.2 Characterization of customer behavior

We derive some related steady state probabilities by making use of the Poisson arrivals see time averages property (Wolff, 1982). Birth-and-death rates are both state-dependent as Figure 2. $p(i)$ denotes steady state probability that the number i of customers are present in the system at a random instant ($i > 0$). Let $L(t)$ denote the system state of representing the number of customers in call center at $t \geqslant 0$, which $\{L(t), t \geqslant 0\}$ is a Markov birth-and-death process. When the system reaches some steady state, that is new customer enters the call center with λ. If $i < s$, arrivals all can enter the system, so the birth rates are λ, and departures are the completion of service. Otherwise, if $i \geqslant s$, an arrival will immediately balk upon the announced delay, so the birth rates are $\lambda(1-\alpha_0)(1-p^B(n))$ according to the analysis above, in addition, departures are the completions of service or abandonment, then the death rates are $s\mu + (i-s)\gamma'$.

In the stationary regime, the stationary probability of i customers in the system, denoted by $p(i)$ for $i \geqslant 0$, is then given by:

$$p(i) = \frac{\lambda^i}{i!\mu^i} p(0) \quad \text{for } 1 \leqslant i \leqslant s, \tag{6}$$

$$p(i) = \frac{\lambda^i}{s!\mu^s} \left(\prod_{j=1}^{i-s} \frac{1-p^B(j-1)}{s\mu + j\gamma'} \right) p(0) \quad \text{for } i > s, \tag{7}$$

with:

$$p(0) = \left(\sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \sum_{i=s+1}^{\infty} \frac{\lambda^i}{s!\mu^s} \left(\prod_{j=1}^{i-s} \frac{1-p^B(j-1)}{s\mu + j\gamma'} \right) \right)^{-1}. \tag{8}$$

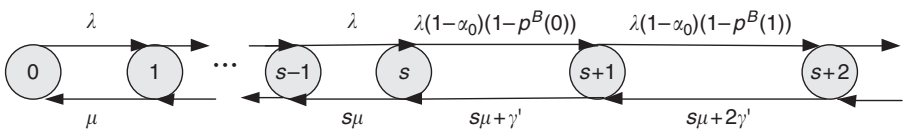


Figure 2.
Birth-death process in the model

Moreover since the arrival process of a customer follows a Poison process, we use the PASTA property to state that the stationary probabilities seen by a new arrival coincide with those seen at an arbitrary instant. Hence, it is straightforward to drive the probability of immediate service P^I that a new arrival get the service without waiting is $P^I = \sum_{i=0}^{s-1} p(i)$. Thus, the mean number of customer in queue L_q is $L_q = \sum_{i=1}^{\infty} ip(s+i)$.

Jouini *et al.* (2011) derive the performance of the conditional probability that a customer will renege $r_n(\theta)$:

$$r_n(\theta) = 1-\beta-\sum_{i=0}^{n}\left(\prod_{j=0, j\neq i}^{n}\frac{s\mu+j\gamma'}{(j-i)\gamma'}\right)\frac{s\mu+i\gamma'}{s\mu+\frac{\gamma}{\theta}+i\gamma'}e^{-(s\mu+i\gamma')d_n}, \tag{9}$$

In this way above, the new patience parameter θ is taken as a single value since customers are assumed to react in the same way to delay information. Here, we make the extent to another case of heterogeneous customer reactions by substituting that new patience parameter θ by $\overline{\theta}$, which is the random variable considering various scenarios of updated patience rather than a constant. Therefore, we calculate the conditional probability $r_n(\overline{\theta})$ as the expected value, where $r_n(\overline{\theta})$ denotes the impact of having different customer reactions. For simplicity and ease of computation, we consider the example that the new patience parameter θ is uniformly distributed and rescaled within the range $[a, b]$, hence, coming back to Equation (9), we have:

$$\begin{aligned} r_n(\overline{\theta}) &= \frac{1}{b-a}\int_a^b\left(1-\beta-\sum_{i=0}^{n}\left(\prod_{j=0, j\neq i}^{n}\frac{s\mu+j\gamma'}{(j-i)\gamma'}\right)\frac{s\mu+i\gamma'}{s\mu+\frac{\gamma}{\theta}+i\gamma'}e^{-(s\mu+i\gamma')d_n}\right)d\theta \\ &= 1-\beta-\sum_{i=0}^{n}\left(\prod_{j=0, j\neq i}^{n}\frac{s\mu+j\gamma'}{(j-i)\gamma'}\right)\frac{s\mu+i\gamma'}{s\mu+\frac{\gamma}{\theta}+i\gamma'}e^{-(s\mu+i\gamma')d_n} \\ &\quad\times\left(1-\frac{\gamma}{(b-a)\times(s\mu+i\gamma')}\times\ln\left(\frac{((b-a)\times(s\mu+i\gamma')+\gamma)}{\gamma}\right)\right). \end{aligned} \tag{10}$$

Specially, when $a=b=0$, the special scenario is the complete update case, which represents the special customer behavior that they update their initial patience to the announce delay. The quantity $r_n^0(\overline{\theta})$ is:

$$r_n^0(\overline{\theta}) = P(d_n < D_n | T > d_n). \tag{11}$$

Because the initial patience T and D_n are independent, we can finally state:

$$r_n^0(\overline{\theta}) = P(d_n < D_n) = 1-\beta. \tag{12}$$

With the intuitive understanding of the conditional probability $r_n^0(\overline{\theta})$: a customer once in queue, she or he abandons if and only if the delay information finally turns out to shorter than the actual delay.

In what follows, the new reneging rate γ' can be calculated by applying the fixed point algorithm (Karamardian and Garcia, 1977), where r_n is the general denotation of the conditional probability including the case of homogenous or heterogeneous customer reactions. Assume the system reaches the stationary regime, we denote the

mean rate of abandoning rate by λ_R. Apply PASTA, this quantity equals that seen by a new arrival, we obtain:

$$\lambda_R = \sum_{n=0}^{\infty} \lambda(1-\alpha)\left(1-p^B(n)\right)p(s+n)r_n. \tag{13}$$

From the exponential distribution of customer updated patience, we can write:

$$\lambda_R = \gamma' L_q. \tag{14}$$

Hence, we can numerically computer the rate γ':

$$\gamma' = \frac{\lambda}{L_q}\sum_{n=0}^{\infty}(1-\alpha)\left(1-p^B(n)\right)p(s+n)r_n. \tag{15}$$

The both cases of $r_n(\theta)$ and $r_n\left(\overline{\theta}\right)$ will be, respectively used for the numerical illustrations.

Finally, we can get other performance, the probability of a new arrival to balk is denoted by P^B, the probability of reneging by P^R, the probability of entering service by P^S:

$$P^B = \sum_{n=0}^{\infty}\left(\alpha_0+(1-\alpha_0)p^B(n)\right)p(s+n),\ P^R = \frac{\gamma' L_q}{\lambda_e}$$

$$P^S = 1-P^B-P^R. \tag{16}$$

3.3 Characterization of satisfaction with delay information

We define the satisfaction index P_c as the service level that customers react to announced delay upon entering the service. It is true in practice that these customers still feel dissatisfied even if they enter the service, because they have experienced a delay that perceived to be longer than initial announced delay. These customers with delay information satisfaction has no influence on this experience of service, however, P_c influences whether the customer will return later or choose to leave forever. We can write this delay information satisfaction as:

$$P_c = P\left(\frac{\text{virtual delay}}{\text{announced delay}} \leqslant 1 | \text{entering the service}\right). \tag{17}$$

Next, we derive the expression of the delay information satisfaction P_c. The conditional probability that a customer is satisfied with the announced delay upon experiencing the service in the system, given that the initial patience exceeds the delay information. Particularly, we take customers of immediate service as satisfying with delay information, so it is given by with two parts, P_{c1} and P^I:

$$P_c = P^I + P_{c1}. \tag{18}$$

Because the customers of entering the service do not balk with $d_n < T$, calculate further, we obtain:

$$P\left(D_n \leqslant d_n \middle| \theta t_k + (1-\theta)d_n \geqslant D_n\right) = P(D_n \leqslant d_n) = \beta \tag{19}$$

Next, we finally state that:

$$P_c = P^I + \beta \cdot \sum_{n=0}^{\infty} (1-\alpha_0)\left(1-p^B(n)\right)p(s+n). \tag{20}$$

3.4 Characterization of the new service levels in waiting experiences

For the model with delay information, there are a lot particular service levels that reflect customer satisfaction with waiting experiences other than satisfaction with delay information. And there is no single perfect or complete list of performance metric for all call centers. Disregarding some process-related measurements may have an adverse effect on customer psychology and experience. Hence, we focus on these new metrics related to queueing delays with the feature of delay information, and we will introduce two definitions of process-related metrics including quick answers and short abandonment.

First, quick answers is that customers immediately get service upon arrival or get service in the acceptable waiting time very quickly. These customers are really considered as being very satisfied so that they would prefer to service of the call center. We define τ denotes the acceptable delay with high satisfaction level of quick answers. Customers who can enter service before τ are really considered as being satisfied. A reasonable value of τ is approximately 20 seconds according to Jouini *et al.* (2013). Second, short abandonment represents customers abandon before a specified short time. We define π denotes another threshold of short abandonment, customers who abandon before π are considered as short abandonment. This is not considered a sign of bad service. Call centers usually count short abandonments differently and limit other regular abandonment. Thus, this is a type of metric that reflects abandonment psychology, and a reasonable value of π is approximately five seconds according to Jouini *et al.* (2013).

Next, we define three service levels that are useful in practice as mentioned above, and denote them by SL_i, for $i = 1,2,3$. We present them in terms of the number of calls that arrive in a certain time period, including:

$$SL_1 = \frac{\text{Number answered} \leqslant \tau}{\text{Number offered}}. \tag{21}$$

$$SL_2 = \frac{\text{Number answered} \leqslant \tau}{\text{Number answered}}. \tag{22}$$

$$SL_3 = \frac{\text{Number abandonments} \geqslant \pi}{\text{Number offered}}. \tag{23}$$

What should be the right metric? SL_1 and SL_2 both do not give information about abandonments, which entice managers to give priority to callers who have enter service before the acceptable time. SL_3 does not give information about waiting of entering service, but SL_3 gives full information on how long callers that have exceeded π have to renege. Taking fully abandonment psychology into account, SL_3 is a more detailed metric than reneging probability P^R. Even if these three levels have perverse effects,

they are regularly used in practice. These service levels including satisfaction with delay information help managers avoid some unwanted behavior by adding them as the objective or constraint condition.

In what follows, we first give expressions for these three service levels as a function of the random variables τ, π, D_n and T'. SL_1 can be taken as a function of the random variables τ, D_n and T'. We can obtain:

$$SL_1 = P(D_n \leqslant \tau, D_n < T'). \tag{24}$$

The second service level is:

$$SL_2 = \frac{P(D_n \leqslant \tau, D_n < T')}{P(D_n < T')}. \tag{25}$$

Similarly, SL_3 is given by:

$$SL_3 = P(D_n \geqslant \pi, D_n > T'). \tag{26}$$

Next, we will explicitly drive the three expressions for these service levels. The first service level can be obtained by the conditional probability σ_n, that customers does not renege while waiting in queue, given that she finds all servers busy, n waiting customers ahead of her and does not balk. Thus, the conditional probability σ_n is expressed as:

$$\sigma_n = P(T' > D_n | T \geqslant d_n). \tag{27}$$

Calculating further, we can obtain:

$$\begin{aligned}\sigma_n &= \frac{P(\theta T + (1-\theta)d_n > D_n | T \geqslant d_n)}{P(T \geqslant d_n)} = \frac{P(\theta T + (1-\theta)d_n > D_n, T \geqslant d_n)}{P(T \geqslant d_n)} \\ &= \frac{P\left(T > (D_n - (1-\theta)d_n)/\theta, T \geqslant d_n\right)}{P(T \geqslant d_n)}.\end{aligned} \tag{28}$$

As for the numerator, it is obtained by two parts:

For the first case $(D_n - (1-\theta)d_n)/\theta \geqslant d_n$, while be equivalent to the probability of $D_n \geqslant d_n$, we get:

$$\begin{aligned}&P\left(T > (D_n - (1-\theta)d_n)/\theta, T \geqslant d_n, D_n \geqslant d_n\right) \\ &\quad = P\left(T > (D_n - (1-\theta)d_n)/\theta, D_n \geqslant d_n\right) \\ &\quad = P\left(T > (D_n - (1-\theta)d_n)/\theta | D_n \geqslant d_n\right) P(D_n \geqslant d_n) \\ &\quad = \int_{d_n}^{\infty} e^{-\gamma\left((t-(1-\theta)d_n)/\theta\right)} g_n(t)dt.\end{aligned} \tag{29}$$

For the second case $(D_n - (1-\theta)d_n)/\theta \leqslant d_n$, while be equivalent to the probability of $D_n \leqslant d_n$, we get:

$$\begin{aligned}&P\left(T > (D_n - (1-\theta)d_n)/\theta, T \geqslant d_n, D_n \leqslant d_n\right) \\ &= P(T \geqslant d_n, D_n \leqslant d_n) = \int_0^{d_n} g_n(t)dt P(T \geqslant d_n).\end{aligned} \tag{30}$$

Combining with the denominator, simply $e^{-\gamma d_n}$, we can obtain:

$$\sigma_n = e^{\gamma(d_n/\theta)} \int_{d_n}^{\infty} e^{-\gamma t/\theta} g_n(t)dt + \int_0^{d_n} g_n(t)dt$$
$$= \sum_{i=0}^{n} \left(\prod_{j=0, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \frac{s\mu + i\gamma'}{s\mu + \frac{\gamma}{\theta} + i\gamma'} e^{-(s\mu + i\gamma')d_n} + \beta. \quad (31)$$

Incorporating to the case of the random variables τ, $\sigma_n(\tau)$, and our customers can get service while n waiting customers ahead of her and before τ, can be calculated as in the following equation:

$$\sigma_n(\tau) = e^{\gamma(d_n/\theta)} \int_{d_n}^{\tau} e^{-\gamma t/\theta} g_n(t)dt + \beta \left(1 - \sum_{i=0}^{n} \left(\prod_{j=0, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) e^{-(s\mu + i\gamma')\tau} \right)$$
$$= \sum_{i=0}^{n} \left(\prod_{j=0, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \frac{s\mu + i\gamma'}{s\mu + \frac{\gamma}{\theta} + i\gamma'} \left(e^{-(s\mu + i\gamma')d_n} - e^{-\left(s\mu + \gamma/\theta + i\gamma'\right)\tau + \gamma\left(d_n/\theta\right)} \right)$$
$$+ \beta \left(1 - \sum_{i=0}^{n} \left(\prod_{j=0, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) e^{-(s\mu + i\gamma')\tau} \right). \quad (32)$$

And coming back to SL_1, including the part customers of immediate service, we finally state that:

$$SL_1 = P^I + \sum_{n=0}^{\infty} (1-\alpha_0)\left(1-p^B(n)\right)\sigma_n(\tau)p(s+n). \quad (33)$$

Observing the relationship between σ_n and $\sigma_n(\tau)$, the second service level SL_2 can be obtained:

$$SL_2 = \frac{P^I + \sum_{n=0}^{\infty} (1-\alpha_0)\left(1-p^B(n)\right)\sigma_n(\tau)p(s+n)}{P^I + \sum_{n=0}^{\infty} (1-\alpha_0)\left(1-p^B(n)\right)\sigma_n p(s+n)}. \quad (34)$$

The third service level can be obtained by the conditional probability $r_n(\pi)$, which denotes the part of customers with short abandonment:

$$r_n(\pi) = 1 - \beta - \sum_{i=0}^{n} \left(\prod_{j=0, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) e^{-(s\mu + i\gamma')\pi}$$
$$- \sum_{i=0}^{n} \left(\prod_{j=0, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \frac{s\mu + i\gamma'}{s\mu + \frac{\gamma}{\theta} + i\gamma'}$$
$$\times \left(e^{-(s\mu + i\gamma')d_n} - e^{-\left(s\mu + \gamma/\theta + i\gamma'\right)\pi + \gamma\left(d_n/\theta\right)} \right). \quad (35)$$

Hence, considering the limitation on regular abandonment but not short abandonment, the new service lever SL_3 can be obtained in the following equation:

$$SL_3 = \sum_{n=0}^{\infty} (1-\alpha_0)(1-p^B(n))(r_n(\theta)-r_n(\pi))p(s+n). \tag{36}$$

Based on the complexity of calculation of these service levels, we only take the new patience parameter θ as a single value for these metrics. We now have expressions for these new service levels of system with delay information. These expressions will be used for the numerical illustrations.

3.5 The optimal design of announcement coverage

From Jouini *et al.* (2011), it is necessary to understand the relationship between announcement coverage and customer reaction performance:

P1. Consider the model with delay information under any case. For two systems x and y having identical parameters but with different delay information coverage $\beta_x > \beta_y$, we have $P_x^B > P_y^B$ and $P_x^R < P_y^R$.

Using *P1*, we can determine the best announcement coverage. However, stated this way, one should avoid the extreme case of $\beta = 0$ percent or $\beta = 100$ percent because balking or reneging is too high. The extreme case is not desirable from customer service or revenue standpoint.

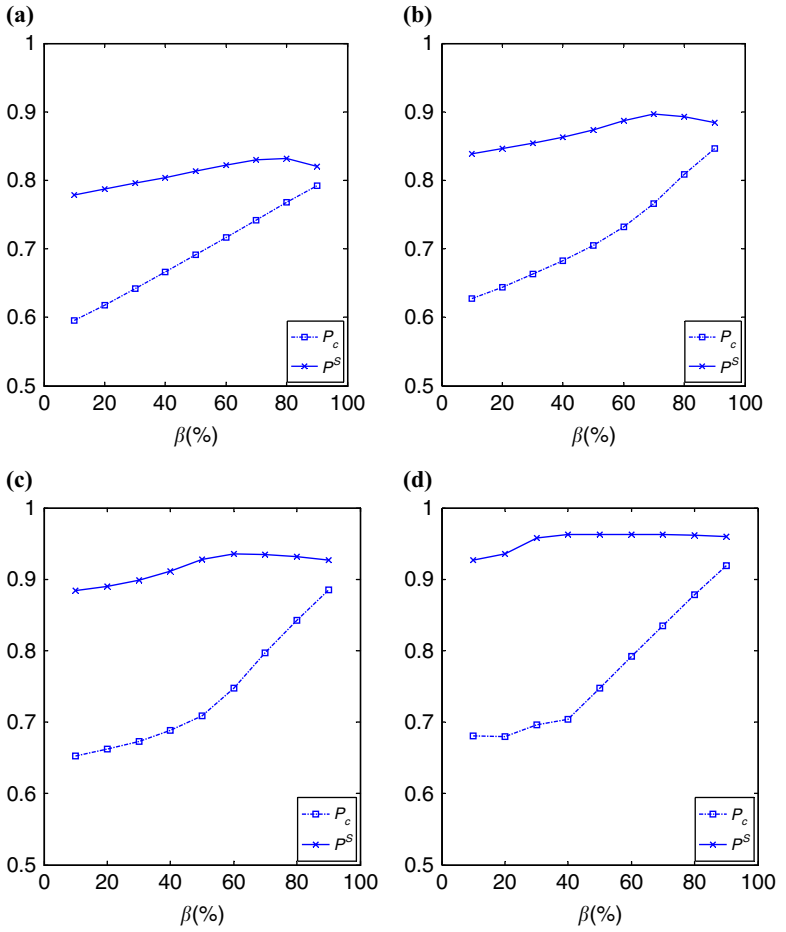
4. Numerical analysis of satisfaction with delay information and call center metrics

In practice, the most effective way of controlling these service-level metrics is staffing. In this section, we mainly emphasize the impact of delay information on these metrics ignoring staffing level, while the system is in a quality- and efficiency-driven regime. We first explore the effect of announcement coverage β on satisfaction with delay information. Then we explore the effect of announcement coverage β on metrics structured in this paper, and further confirm the interest of all metrics. The common parameters are $\alpha_0 = 0.05$, $\gamma = 0.5$, $\mu = 1$. For every set of parameters, the announcement coverage β values are given in percentages and rounded up to integer values.

4.1 Effect of satisfaction with delay information on call center

To illustrate the performance of customer satisfaction with delay information, we first consider the two performance parameters P_c and P^S as functions of the system pooling $\lambda = s$, which means the system is in a quality- and efficiency-driven regime. These performances are for each $\lambda \in \{5,10,20,50\}$ and the patience update interval value (a,b) is (0,1/3).

In Figure 3(a)-(d), we show a comprehensive analysis of delay information satisfaction as functions of β. The changes of P_c are particularly apparent, but the pooling effect is absent. Delay information satisfaction P_c does vary with β with apparently increasing regularity. We also show the probability of entering service p^S, which is increased with β until it reaches its maximum, then is gradually decreased. The system size determines the maximum value of the probability of service p^S. This rule shows that a manager attempting to maximize revenue should choose a value for β because of its role in customer behavior, and the relationship between satisfaction with delay information and staffing will be further discussed in detail in next section.



Notes: (a) $s=\lambda=5$; (b) $s=\lambda=10$; (c) $s=\lambda=20$; (d) $s=\lambda=50$

Figure 3.
Impact of β on delay information satisfaction

4.2 Effect of the metrics on call center

To illustrate the effect of the metrics on call center, we consider these new service levels SL_1, SL_2, and SL_3 as functions of the system pooling $\lambda = s$, which means the system is in a quality- and efficiency-driven regime. These metrics are for each $s \in \{5,10,20\}$ and the patience update value $\theta = 1/3$.

First, we show a comprehensive analysis of SL_1 and SL_2 as functions of β from Figure 4. Note that there are two roles that β plays in the both metrics, which are different from monotonic effect of β on balking and reneging. Though increasing β reduces the reneging of customers at the expense of additional balking of customers, SL_1 and SL_2 both have large values in lower and higher announcement coverage. The numerical analysis shows that in such a quality- and efficiency-driven regime, the delays are relatively short in spite of less balking and most customers tagged in the queue in lower announcement, however, under the tightest control of balking most waiting customers can get service before short time τ, that is to say that the optimal SL_1 and SL_2 are in lower announcement coverage.

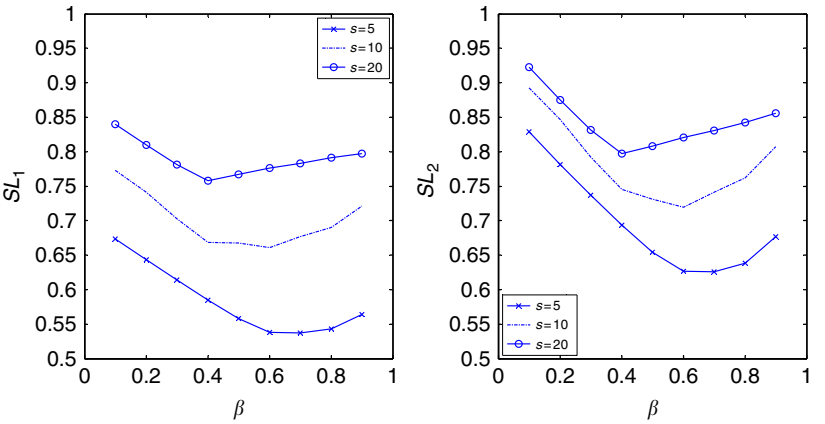


Figure 4.
Impact of β on SL_1 and SL_2

Second, Figure 5 illustrates the impact of announcement coverage in SL_3. Note that the higher β is, the less SL_3 is. The reason is that reneging of customers plays a decisive role in SL_3, which illustrates that most reneging customers wait for more than the value of π. Hence, SL_3 keeps the similar tendency of reneging probability P^R as functions of announcement coverage β. In all, managers can avoid reneging behavior with long waiting time by adding this objective with more accurate customer psychology, according to the practical situation of call center.

5. The optimal design of call center with delay information satisfaction

In this section, we first propose the staffing problem, which is used to make the link between staffing and revenue according to the performance metric of customer loss. Call center revenue is generated by serving a customer. The call center incurs a revenue loss once a customer is lost each time. Therefore, it is reasonable that we characterize the customer loss as a function of the number of servers in order to make the link between staffing and revenues. At the same time, it is true in practice that customers who have reneged have experienced longer delay so that they would leave with relatively lower satisfaction than balking customers. Such a reneged customer possibly

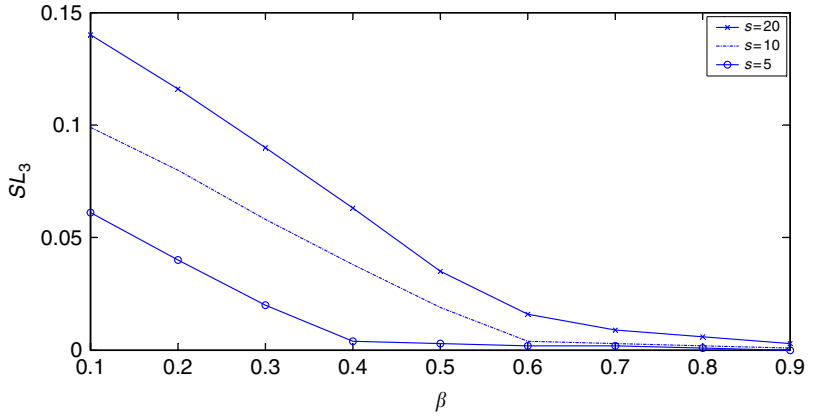


Figure 5.
Impact of β on SL_3

loses trust in the call center upon experience frustration due to time lost, thus, the first set of constraints in the model ensures the dissatisfaction of reneging is not high. In addition, a customer reacts by satisfaction upon entering the service with full perception of delay information. Delay information may further modulate these customer reactions in next choice. Thus, this paper takes the delay information satisfaction as another constraint. More specifically, the model can be stated as follows:

$$\begin{cases} \max TP = c_1\lambda \cdot P_S - c_2 s \\ \text{subject to } P^R \leqslant \delta_1 \\ \qquad\qquad P_c \geqslant \delta_2 \end{cases} \tag{37}$$

The first term is the customer expected service revenue, which can be expressed as $c_1 \cdot \lambda \cdot P_S$, where c_1 is the customer service revenue parameter; Given an average revenue per customer served, such that system revenues net of abandonments loss are $\lambda \cdot P_S$. The second term is the human resource cost $c_2 \cdot s$, which is in the form of salaries. For the current analysis, it is necessary to determine the optimal design with the link between staff and announcement control. This paper uses the enumeration method to obtain the desired economically optimal staffing solution. In addition, another decision variable embedded in the above formulated sizing model is β, which ensures that no more than δ_1 of the customers that enter the queue should renege and no less than δ_2 of the customers that are considered be satisfied with delay information.

Next, we explore how to design the model with delay information satisfaction and impatient customers when various system parameters of pooling and customer patience are varied. Then, we explore the optimal design of call center with delay information for optimal system revenues. For all cases, $c_1 = 5$, $c_2 = 2$, and all parameters not mentioned take the same values as earlier in this section.

For the system with delay information, we consider five systems with increasing levels of pooling. The common parameters are $(s,\lambda) = (3,3)$, (5,5), (10,10), (20,20), and (50,50). At the same time, we also consider five customer reactions to announcement by choosing patience update interval value $(a,b) = (0,0)$, (0,1/3), (0,2/3), (0,1) and (5/3). In addition, the first five models are working under the optimal announcement coverage β^*. These results of the comparison are shown in Table I for cases when $\delta_2 = 0.6$, and Figure 6 for the case when imposing stricter constraint of satisfaction with delay information. All reneging level constraints are $\delta_1 = 0.05$.

When we analyze the results in Table I, staffing level s is increased currently with λ for ensuring a quality- and efficiency-driven regime. We can clearly observe the optimal design of announcement coverage β^*. First, the range of patience update values (a,b) has an important effect on the results. When $b = 0$, β^* is the highest, that is customers are updating their patience to announced delay. Larger b means customers are more and more patience, which leads to smaller optimal coverage announcement β^*. Furthermore, the system achieves higher revenues for a wider range of patience update values (a,b). We note when $b = 0$, the probability of satisfaction with delay information is equal to the probability of entering service. In fully updating case, all customers of entering service are satisfied with announced delay because their patience is updated to delay information. The abandonment behavior rule under the given coverage probability is consistent with the research of Jouini *et al.* (2011) even if take account of the uniform distribution of the updated patience threshold parameter θ.

Table I.
Optimal announcement coverage β^*

$s=\lambda$	3	5	10	20	50	3	5	10	20	50
			$(a, b)=(0, 1/3)$					$(a, b)=(0, 2/3)$		
$\beta^*(\%)$	74	73	64	51	45	67	60	44	33	18
P^R	0.049	0.032	0.022	0.016	0.007	0.043	0.031	0.022	0.014	0.007
P^B	0.18	0.134	0.08	0.047	0.03	0.172	0.12	0.068	0.043	0.027
P_{ζ}	0.719	0.746	0.738	0.704	0.726	0.689	0.683	0.627	0.603	0.601
P^S	0.771	0.834	0.898	0.936	0.963	0.785	0.849	0.91	0.943	0.966
TP	5.57	10.85	24.91	53.64	140.7	5.77	11.24	25.51	54.35	141.44
			$(a, b)=(0, 1)$					$(a, b)=(0, 5/3)$		
$\beta^*(\%)$	60	50	41	34	19	50	47	41	34	20
P^R	0.038	0.03	0.018	0.01	0.005	0.033	0.021	0.012	0.007	0.003
P^B	0.164	0.11	0.067	0.044	0.028	0.152	0.109	0.068	0.044	0.028
P_{ζ}	0.655	0.628	0.605	0.604	0.602	0.601	0.604	0.6	0.601	0.604
P^S	0.798	0.861	0.916	0.947	0.967	0.815	0.87	0.921	0.949	0.969
TP	5.97	11.52	25.79	54.65	141.85	6.23	11.75	26.03	54.94	142.2
$s=\lambda$	3	5	10	20	50					
			$(a, b)=(0, 0)$							
$\beta^*(\%)$	85	90	95	97	99					
P^R	0.049	0.036	0.019	0.012	0.004					
P^B	0.199	0.158	0.115	0.077	0.046					
P_{ζ}	0.752	0.807	0.866	0.911	0.95					
P^S	0.752	0.807	0.866	0.911	0.95					
TP	5.27	10.16	23.31	51.08	137.44					

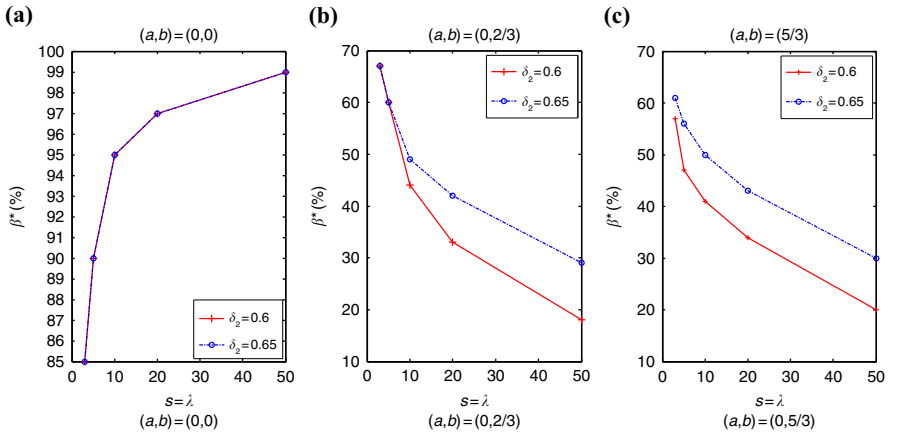


Figure 6. Optimal announcement coverage β^* under stricter constraint

The pooling of system has a distinctive effect on these performances, especially the probability of a new arrival to enter service P^S. When $b > 0$, as the pooling of system increases, β^* decreases. The pooling increases the probability P^S, allows for a lower announcement coverage without violating both constraints of the reneging probability and the probability of satisfaction with delay information. This rule shows that a manager attempting to maximize revenue should choose a value for β^* in different pooling. For observing the role of satisfaction with delay information, the constraint on P_c is tight in contrast to not tight P^R. The differences in optimal β with the research of Jouini *et al.* (2011) are observed to be strong when the system is larger pooling, correspondingly, the optimal announce coverage in our results keeps a relatively high point in contrast to very low β^* of the research of Jouini *et al.* (2011). That is because of the case that TP even has reached its maximum in lower β^* values where the constraint on P_c is not satisfied (e.g. $b = 5/3$, $s = \lambda = 50$), then this system must allow for higher coverage announcement. It is very important for optimal design with satisfaction with delay information.

In sum, all abandonment behavior decreases for larger system operating under a quality- and efficiency-driven regime, which diminishes the importance of announcing delay because of staffing role. Hence, in such scenario, managers have to control the announcement coverage with the key consideration of satisfaction with delay information.

For highlighting the role of satisfaction with delay information in optimal announcement, we impose a stricter delay information constraint ($\delta_2 = 0.65$) in $(a,b) = (0,0)$, $(0,2/3)$, $(0,5/3)$. From Figure 6, we observe different roles that stricter delay information constraint plays under different customer patience reaction. First, the curve of $\delta_2 = 0.6$ coincides exactly with the curve of stricter delay information when $(a,b) = (0,0)$, that is because β^* has been highest when customers are updating their patience to the delay information. Any customer who enters the service is satisfied with the delay information, resulting in that the constraint loses the role. Second, the stricter constraint results in higher optimal announcement values throughout when $b > 0$, especially for large pooling. That is because customers are more sensitive to delay information when the patience reaction is relative increasing. This case shows mangers needs to attach importance to the role of satisfaction with delay information in this system.

In the previous analysis, the system is always set in a quality- and efficiency-driven regime. We further explore optimal staffing level s^* with different β values while

keeping arrival rate λ fixed, for $\lambda = 10$, $\delta_1 = 0.05$, and $\delta_2 = 0.6$. Then the system may allow for the efficiency driven regime.

From Figure 7 we can observe that optimal staffing level s^* need be made highest when $(a,b) = (0,0)$, that is when customers are updating their patience to announced delay. In such case of $(a,b) = (0,0)$, as the announcement coverage β increases, s^* could be decreased. This is due to the constraint of delay information satisfaction and reneging, for higher β value the constraint is easily satisfied so that less staffing could be taken and corresponding TP could also be increased.

As b increases, that is when customers are updating their patience to larger value. Hence, higher b value will lead to lower staffing level. Note that staffing level plays both roles with different β. First, as β increases and begins to reach about 80 percent, s^* decreases as a result of not tight need to satisfy reneging and delay information satisfaction constraints. However, as β is further increased, too many customers balk and only a small number of customers would renege, which reduce the entering service probability P^S compared to the system with small β. Therefore, the inflection point emerges. At the same time, corresponding TP also changes similarly because of the entering service probability.

In this paper, the general objective of the system is to optimize the revenue function with respect to the two decision variables β^* and s^*. Here, we give optimal design of call center using enumeration method, for different patience update interval values, that is fixing $a = 0$ while varying the value b, still given in $\delta_1 = 0.05$ and $\delta_2 = 0.6$ for the arrival rate $\lambda = 10$.

The results for optimal design of staffing level and announcement coverage are presented in Figure 8. Especially, Figure 8 apparently shows optimal staffing level varies along a U-shaped curve with the enlargement of customer patience reaction, and the curve of optimal announcement coverage is shifted downward because the increased patience reactions relax the constraint of satisfaction with delay information. First, more coverage in combination with higher staffing level is necessarily better for the mangers near $(a,b) = (0,0)$. As b starts to increase, staffing level increases as a result of a need to satisfy the relatively short patience. Second, the optimal choice of staffing and coverage are almost the same in the presence of smaller customer patience reactions. Third, when for larger customer patience reactions, less coverage in combination with higher staffing level is properly used. In this situation, though

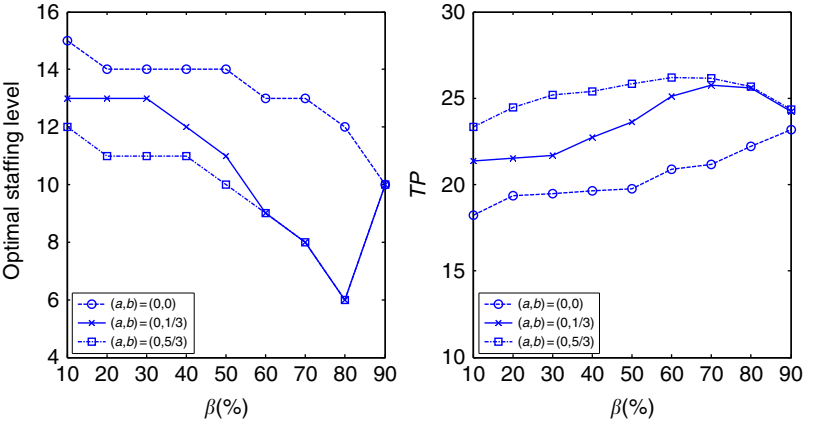


Figure 7.
Optimal staffing level s^* with different announcement coverages

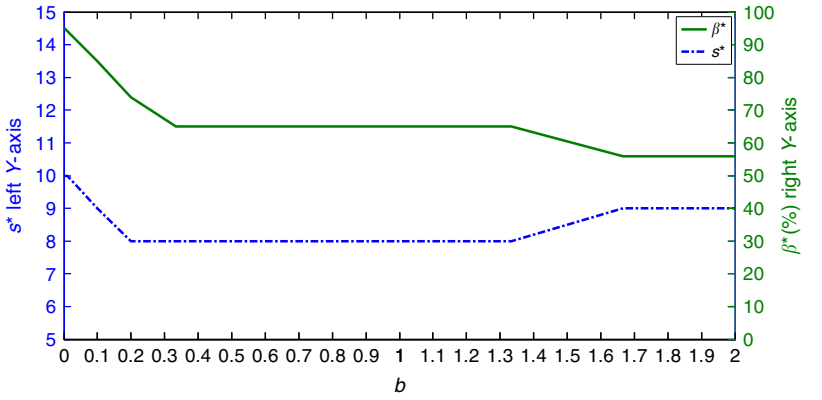


Figure 8.
Optimal design of staffing level and announcement coverage

customers have relative long patience reaction, staffing level still has to been made higher because of the constraints of reneging and satisfaction with delay information. In sum, all the decisions have to be carefully made under both the roles of announcement coverage and staffing level, particularly if customer patience reaction to delay information is strong.

6. Extension to the optimal design of the model with new service levels

In this section, we extend the optimal design by allowing three service level constraints. In practice, managers want to avoid some unwanted behavior by adding some special objective or constraint condition. Hence, the model can be stated in contrast to the formulation (37):

$$\begin{cases} \max\limits_{s} TP = c_1\lambda \cdot P_S - c_2 s \\ \text{subject to } SL_i \geqslant \delta_{3i} \\ \qquad\qquad \text{or} \quad SL_3 \leqslant \delta_{33}. \end{cases} \tag{38}$$

Then, we explore the optimal design of call center with service level constraints for optimal system revenues, similarly solving the formulation (38). For all cases, $c_1 = 5$, $c_2 = 2$, and the patience update value $\theta = 1/3$.

First, we show a comprehensive analysis of SL_1 and SL_2 as functions of system pooling for ensuring a quality- and efficiency-driven regime. For highlighting the role of both service levels in optimal announcement, we impose two constraints $\delta_{31} = 0.6$, $\delta_{31} = 0.65$ on SL_1, and we impose two constraints $\delta_{32} = 0.7$, $\delta_{32} = 0.75$ on SL_2. From Figure 9, we can clearly observe the optimal design of announcement coverage β^* in the influence of the two service levels. First, this figure shows the delay announcement is more important for satisfying the constraints of two new service levels when the system is small. On the other side, it conforms to the law of diminishing returns (see e.g Koole and Pot, 2011), which states that the marginal increase in service level declines in the staffing level. Second, when the delay announcement is set to maximize self-profit of call center, delay announcement played a more significant role in satisfying the service level SL_2, since SL_2 is a better metric that removes the abandonment thereby more accurately penalizing customers who are very impatient.

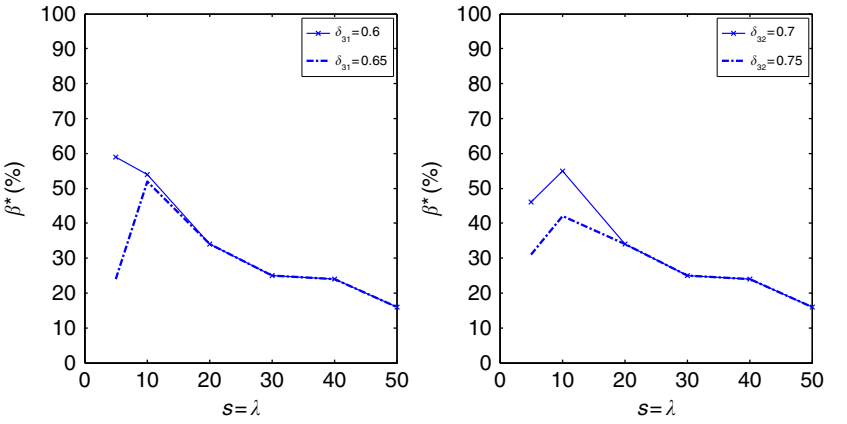


Figure 9.
Optimal announcement coverage β^* under stricter SL_1 and SL_2

Second, for highlighting the role of short abandonment in optimal announcement, we impose two constraints $\delta_{33} = 0.01$ and $\delta_{33} = 0.03$ on SL_3. The results are given in Figure 10. The higher β^* values in this constraint show the role that delay announcement coverage plays in satisfying much stricter SL_3 of $\delta_{33} = 0.01$. Similarly, the marginal increase of β^* in service level declines in the system pooling.

In what follows, we want to numerically study the impact of three new service levels on the optimal system revenues. In previous three settings, the optimal system revenues are all influenced by the service levels constrains. Then, we will compare these three system revenue with service levels constrains with the optimal system revenue TP^* without any constraint. Using the results in Figures 9 and 10 we can calculate revenue margins from service level constraint by a relative percentage $\Delta TP(sl)$, which is equal to $\left(\left|TP^* - TP^*_{sl}\right| / TP^*\right) \times 100\%$, where TP^*_{sl} is the optimal revenue under some service level constraint. The results of profit margins are given in Figure 11. Note that $\Delta TP(sl)$ is zero means that the optimal system revenue gets to TP^* without any constraint. We further show that these service levels become ineffective when the system pooling is more than 20. Overall, SL_1 and SL_2 have more influence on system revenue than SL_3 when the system pooling is small.

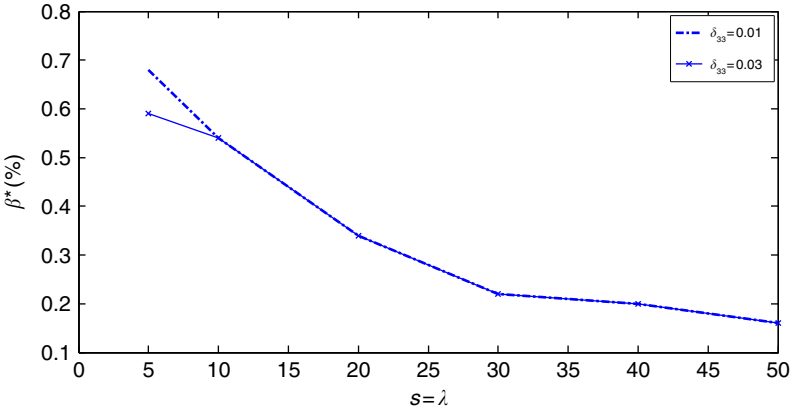


Figure 10.
Optimal announcement coverage β^* under stricter SL_3

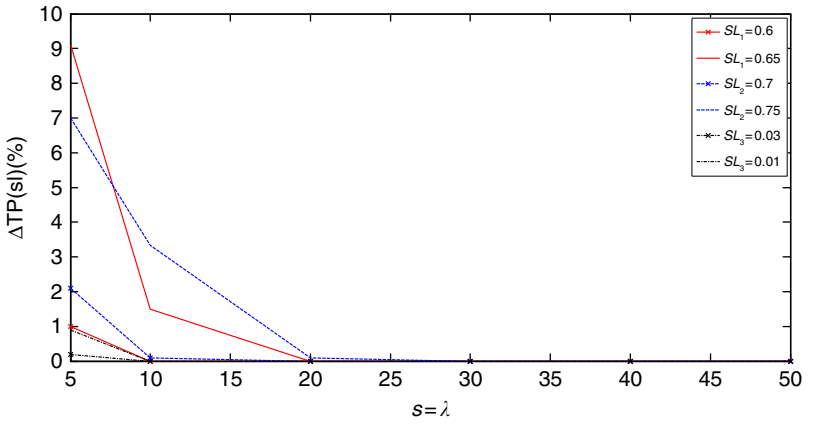


Figure 11.
Results for optimal system revenues with new service levels

7. Conclusion

In this paper, we have formulated and analyzed a call center queue with delay information and impatient customers. The main decisions in the design of such systems are staffing levels with combinations of the appropriate control to announce delay anticipated. The satisfaction with delay information and a number of different service levels are the key distinguishing features of this model.

The numerical analysis illustrates that choosing an optimal announcement coverage and staffing level together enables the highest revenue for the system. Announcements with higher coverage are more important when customer reaction range of (a,b) is narrow, and when the pooling of systems are small. In particular, relatively high coverage is still required in large pooling of systems for the avoidance of dissatisfaction with delay information, which is distinguishing difference with Jouini *et al.* (2011). Because of satisfaction with delay information, announcement coverage should be carefully made in the presence of large pooling and wide customer reaction range.

In future work, first, it will be useful to investigate customer behavior with a field data set of call centers. Such work allows a direct comparison between practice and theory approximation. Second, in terms of the systems for different markets, our implementation may encounter problems with regard to customer satisfaction, and further refinements of the method would need to be investigated.

References

Aksin, O.Z., Ata, B., Emadi, S. and Su, C.-L. (2013), "Structural estimation of callers' delay sensitivity in call centers", *Management Science*, Vol. 59 No. 12, pp. 2727-2746.

Aksin, O.Z. and Harker, P.T. (2003), "Capacity sizing in the presence of a common shared resource: dimensioning an inbound call center", *European Journal of Operational Research*, Vol. 147 No. 3, pp. 464-483.

Allon, G., Bassambo, A. and Gurvich, I. (2012), "We will be right with you: managing customer with vague promises", *Operations Research*, Vol. 59 No. 6, pp. 1382-1394.

Anderson, E.W. and Sullivan, M.W. (1993), "The antecedents and consequences of customer satisfaction for firms", *Marketing Science*, Vol. 12 No. 2, pp. 125-143.

Armony, M., Shimkin, N. and Whitt, W. (2009), "The impact of delay announcements in many-server queues with abandonment", *Operations Research*, Vol. 57 No. 1, pp. 66-81.

Au, N., Ngai, E.W.T. and Cheng, T.C.E. (2002), "A critical review of end-user information system satisfaction research and a new research framework", *Omega*, Vol. 30 No. 6, pp. 451-478.

Baron, O. and Milner, J. (2009), "Staffing to maximize profit for call centers with alternate service-level agreements", *Operations Research*, Vol. 57 No. 3, pp. 685-700.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005), "Statistical analysis of a telephone call center: a queueing-science perspective", *Journal of the American Statistical Association*, Vol. 100 No. 469, pp. 36-50.

Dean, A.M. (2002), "Service quality in call centers: implications for customer loyalty", *Managing Service Quality*, Vol. 12 No. 6, pp. 414-423.

Feigin, P. (2005), "Analysis of customer patience in a bank call center", working paper, Technion, Haifa.

Gans, N., Koole, G. and Mandelbaum, A. (2003), "Telephone call centers: tutorial, review, and research prospects", *Manufacturing & Service Operations Management*, Vol. 5 No. 2, pp. 79-141.

Garnett, O., Mandelbaum, A. and Reiman, M. (2002), "Designing a call center with impatient customers", *Manufacturing & Service Operations Management*, Vol. 4 No. 3, pp. 208-227.

Guo, P.F. and Zipkin, P. (2007), "Analysis and comparison of queues with different levels of delay information", *Management Science*, Vol. 53 No. 6, pp. 962-970.

Halfin, S. and Whitt, W. (1981), "Heavy-traffic limits for queues with many exponential servers", *Operations Research*, Vol. 29 No. 3, pp. 567-588.

Hassin, R. (1986), "Consumer information in markets with random product quality: The case of queues and balking", *Econometrica*, Vol. 54 No. 5, pp. 1185-1195.

Hui, M.K. and Tse, D.K. (1996), "What to tell customer in waits of different lengths: an integrative model of service evaluation", *Journal of Marketing*, Vol. 60 No. 2, pp. 81-90.

Ibrahim, R. and Whitt, W. (2009a), "Real-time delay estimation based on delay history", *Manufacturing & Service Operations Management*, Vol. 11 No. 3, pp. 397-415.

Ibrahim, R. and Whitt, W. (2009b), "Real-time delay estimation in overloaded multiserver queues with abandonments", *Management Science*, Vol. 55 No. 10, pp. 1729-1742.

Ibrahim, R. and Whitt, W. (2011), "Real-time delay estimation based on delay history in many-server queue with time-varying arrivals", *Production Operations Management Society*, Vol. 20 No. 5, pp. 654-667.

Jouini, O., Aksin, Z., Aguir, M.S., Karaesmen, F. and Dallery, Y. (2014), "Call center delay announcement using a newsvendor-like performance criterion", *Production Operations Management Society*, Vol. 24 No. 4, pp. 587-604.

Jouini, O., Aksin, Z. and Dallery, Y. (2011), "Call centers with delay information: models and Insights", *Manufacturing & Service Operations Management*, Vol. 13 No. 4, pp. 534-548.

Jouini, O., Koole, G. and Roubos, A. (2013), "Performance indicators for call centers with impatient customers", *IIE Transactions*, Vol. 45 No. 3, pp. 341-354.

Karamardian, S. and Garcia, C.B. (1977), *Fixed Points: Algorithms and Applications*, Academic Press, New York, NY.

Kim, W.J. and Ha, S.H. (2010), "Consecutive staffing solution using simulation in the contact center", *Industrial Management & Data Systems*, Vol. 110 No. 5, pp. 718-730.

Koole, G. and Pot, A. (2011), "A note on profit maximization and monotonicity for inbound call centers", *Operations Research*, Vol. 59 No. 5, pp. 1304-1308.

Luo, J. and Zhang, J. (2013), "Staffing and control of instant messaging contact centers", *Operations Research*, Vol. 61 No. 2, pp. 328-343.

Mandelbaum, A., Massey, W., Reiman, M., Rider, B. and Stolyar, A. (2002), "Queue lengths and waiting times for multiserver queues with abandonment and retrials", *Telecommunication Systems*, Vol. 21 Nos 2-4, pp. 149-171.

Mandelbaum, A. and Zeltyn, S. (2009), "Staffing many-server queues with impatient customers: constraint satisfaction in call centers", *Operations Research*, Vol. 57 No. 5, pp. 1189-1205.

Munichor, N. and Rafaeli, A. (2007), "Numbers or apologies? Customer reactions to telephone waiting time fillers", *Journal of Applied Psychology*, Vol. 92 No. 2, pp. 511-518.

Nakibly, E. (2002), "Predicting waiting times in telephone service systems", MS thesis, The Technion, Haifa.

Nie, W. (2000), "Waiting: integrating social and psychological perspectives in operations management", *Omega*, Vol. 28 No. 6, pp. 611-629.

Taylor, S. (1994), "Waiting for service: the relationship between delays and evaluations of service", *Journal of Marketing*, Vol. 58 No. 2, pp. 56-69.

Tezcan, T. and Zhang, J. (2014), "Routing and staffing in customer service chat systems with impatient customers", *Operations Research*, Vol. 62 No. 4, pp. 943-956.

Whitt, W. (1999), "Improving service by informing customers about anticipated delays", *Management Science*, Vol. 45 No. 2, pp. 192-207.

Wolff, R. (1982), "Poisson arrivals see time averages", *Operations Research*, Vol. 30 No. 2, pp. 223-231.

Yu, Q., Allon, G. and Bassamboo, A. (2015), "How do delay announcements shape customer behavior? An empirical study", *Management Science*, forthcoming.

Zhang, J., Dai, J.G. and Zwart, B. (2011), "Diffusion approximations of limited processor sharing queues in heavy traffic", *Annals of Applied Probability*, Vol. 21 No. 2, pp. 745-799.

Further reading

Hui, M. and Zhou, L. (1996), "How does waiting duration information influence customers' reactions to waiting for services?", *Journal of Applied Social Psychology*, Vol. 26 No. 19, pp. 1702-1717.

Corresponding author

Jun Gong can be contacted at: gongjun@ise.neu.edu.cn