

Impact of customer impatience on a production service system

Kangzhou Wang^a, Shulin Lan^b and Zhibin Jiang^{c*}

^a*School of Management, Lanzhou University, Lanzhou, China;* ^b*Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China;* ^c*Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai, China*

(Received 8 November 2014; accepted 22 November 2015)

In this paper, we explore the impact of customer impatience on the performance of a production service system that consists of one production inventory subsystem and one service subsystem. It is inevitable that the un-storability of service makes customers wait in the production service system. Waiting for service easily leads to impatient behaviors. The impact of two kinds of impatient behaviors balking and reneging on the production service system performance is examined. The problem is formulated as a Markov system, then the stationary probability of the system is derived to conduct the performance evaluation. The results show that customer impatience has the different impact patterns on the two subsystems (i.e. production inventory subsystem and service subsystem) of the production service system. Specifically, customer impatience leads to the performance improvement of one subsystem along with the performance deterioration of the other subsystem. Meanwhile, service capacity plays an important role in the influence extent of customer impatience on the system performance. In addition, lower customer impatience does not necessarily improve system profit. Therefore, the management of the production service system with impatient customers will be more intractable. Some managerial insights and suggestions are proposed.

Keywords: production service system; impatient behavior; balking; reneging; inventory service system; queueing system

1. Introduction

Along with the increased requirement of services, service-oriented manufacturing (SOM) arises and is adopted by many traditional manufacturing firms in decades. E.g. to offer a total solution to customers, logistics, installation, commissioning and technical training have been provided by almost of equipment manufacturing firms in China, such as Shanghai Electric Group, Shaanxi Blower Group, etc. The system operated by the SOM firms is named as production service system. A production service system is a system to provide customers with both product and product-related service (Wang et al. 2013). For example, Shanghai Biaowu High-Tensile Fasteners Co., Ltd. provides not only fasteners, but also logistics and distribution services. Therefore, the system operated by Shanghai Biaowu is a production service system consisting of a production inventory subsystem and a logistics and distribution services subsystem. In fact, it exists widely for manufacturers to provide logistics services in practice. There are a large number of suppliers providing distribution services for downstream customers. Logistics enhances the integration of manufacturing and product-based services, and promotes the development of both, to achieve a win-win goal. In addition, installation and commissioning services are very common in the elevators and equipment manufacturing industries. Meanwhile, in many service industries, such as after-sale service, health care service and catering service (Berman, Kaplan, and Shevishak 1993), the service providers provide material goods and intangible services simultaneously. Thus, the production service system is also common in the service industries. Meanwhile, it is clear that the regulation of the production service system needs not only an operational management of production inventory system, but also management of services system (Berman and Sapna 2002).

This paper considers a classical production service system consisting of one production inventory subsystem and one service subsystem in sequence. In the first stage of the system, there is a single make-to-stock manufacturing facility providing products as the ‘raw materials’ for the service orders processed in the second stage. Meanwhile, there is a service centre in the second stage fulfilling the external service orders based on the products in an order-driven strategy. It should be noted that demands for the products produced in the first stage is triggered by service events.

*Corresponding author. Email: zbjiang@sjtu.edu.cn

The dynamics of the production service system is different from that of the production inventory system purely providing material goods, and that of the service system purely providing services (Wang et al. 2013). With respect to the production inventory system, the product demand is determined jointly by service capacity and external demands in the production service system: the inventory decreases with service rate when there are customers waiting for service; the inventory decreases with demand rate when there are no customers waiting for service. And with respect to the service system, if there is no inventory in the production service system, the service rate is zero, even if the service centre is idle while customers are waiting for service.

The production service system fulfils external orders in the service centre. Because of the un-storability of service, it cannot buffer the mismatching between service capacity and demand through inventory like products. Therefore, customers waiting is inevitable. Waiting for service can easily lead to customers impatient behaviors such as balking and reneging (Hassin and Haviv 2003). Impatient behaviors make the actual demand change, and thereby have a significant impact on the system performance, thereby affecting the system management. Therefore, to deal with the effect of impatient customers on firms performance, the problem of how impact of customer impatience on the firms' operations management should be examined accurately. However, to our knowledge, there is few works devoting to this problem.

In fact, because inventory balances supply and demand have a buffering effect, there are a few of customers who show impatient behaviors in the production inventory systems (Benjaafar and Elhafsi 2012; Wang et al. 2014). However, the customers in the service systems are not the case. There is a lot of literature considering customer impatience in the service system (Hassin and Haviv 2003; Wang, Li, and Jiang 2010). The findings also show that customers impatient behaviors have a significant impact on the operation management of service systems (Rao 1965; van Tits and van der Veecken 1980). Thus, for the production service system with service subsystem and with different dynamic characteristics from the service system, the studies on the impact of impatient behaviors on the system performance have not only practical implications, but also important theoretical significance.

The rest of the paper is organised as follows. In Section 2, we review the researches related with this study. Section 3 presents the problem description and offers the quantified expressions of two impatient behaviors, i.e. balking and reneging. Section 4 establishes a Markov model to describe the production service system with balking and reneging and gives the solution procedure. As the main results of the paper, the impacts of customer impatience on the performance and optimal control of the production service system are explored in Sections 5 and 6, respectively. Section 7 summarises our findings.

2. Literature review

The research literature associated with this study can be divided into two categories: one is the research literature in production service system, the other is the research literature in impatient customers.

For the first category, it should be noted that both production service system and inventory service system provide customers with the package of product and service, rather than physical product or intangible service alone. In other words, two systems are characterised by that feature of simultaneously managing product and service order.

The study on inventory service system roots from the work by Berman, Kaplan, and Shevishak (1993). The authors study a system with deterministic demand rate, service and lead times and find the optimal order quantity. Subsequently, several extensions such as the systems with stochastic lead times and service times, retrial and jockeying are presented. For the studies on the optimal control, the optimal replenishment policies for Markovian models with instantaneous lead time process are analysed in Berman and Kim (1999) and He and Jewkes (2000). Berman and Sapna (2000) find an optimal stocking level for the finite queue model with Poisson arrivals, general service times and zero lead time. Berman and Kim (2001) obtain that the optimal replenishment policy is a monotonic threshold type when replenishment time is an Erlang distribution. He, Jewkes, and Buzacott (2002) focus on computing optimal replenishment policies. Berman and Sapna (2002) study the optimal service rate control of an inventory service system with perishables. Then several works devoted to the studies of steady-state probabilities, steady-state performance measures and optimisation of the control parameters. Most of them deal with the performance analysis when (r, Q) , (R, r) and (r, S) policies are used to manage inventory, such as Schwarz et al. (2006), Schwarz and Daduna (2006), Manuel, Sivakumar, and Arivarignan (2007), Deepak et al. (2008), Yadavalli, Sivakumar, and Arivarignan (2008) and Krishnamoorthy, Nair, and Narayanan (2012).

For the study on the production service system, (Krishnamoorthy and Narayanan 2011) consider the stability and performance evaluation of a production system in which production time follows a 'Markovian production scheme' and the on-off decision is determined by (s, S) policy. Wang et al. (2013) present the optimal production and admission control for a production service system that fulfils both demands for product and the package of product and service. The above existing studies on the production service systems do not consider impatient customers. In fact, two extensive

literature reviews Hassin and Haviv (2003) and Wang, Li, and Jiang (2010) point out, waiting for service can easily lead to customers impatient behaviors such as balking and reneging. Meanwhile, customers waiting for the product and product-based service arise inevitably in the production service systems. Therefore, this work studies production service system with impatient customers, and two impatient behaviors balking and reneging are considered.

As to the operations systems with impatient customers, two types of systems, inventory system and service system have received significant attention up to date in the literature. Several research studies focus on the inventory system with impatient customers. Example San José, Sicilia, and Garcia-Laguna (2005) study the lot size-reorder level inventory system with impatience. Toews, Pentico, and Drake (2011) propose an EOQ model with linear balking function. Economopoulos, Kouikoglou, and Grigoroudis (2011) investigate a manufacturing system with both balking and reneging behaviors. The literature (Benjaafar, Gayon, and Tepe 2010; Ioannidis et al. 2013; Benjaafar and Elhafsi 2012; Wang et al. 2014) study the optimal control of production inventory systems with impatient customers.

As we know, the lead time of delivery service in service systems is greater than that in production inventory systems. Thus, there are a large number of studies on the service system with impatient customers, and we refer the readers to literature reviews (Hassin and Haviv 2003) and (Wang, Li, and Jiang 2010).

In a production service system, waiting for service easily leads to impatient behaviors. However, the dynamics of the production service system differs largely from those of inventory system and service system. There is only one research, (Li and Jiang 2013), devoted to the production service system with impatient customers so far. The authors consider a system with additional service capacity, and explore the condition when the manager should lunch the additional service capacity. In other words, the authors focus on the capacity management, rather than the impact of customer impatience on the production service system. Generally, the additional service capacity is not available for a production service system during a middle-short horizon. Therefore, this paper focuses attention on the impact of different impatience degrees on the production service system with constant production and service capacity, and tries to obtain some managerial insights to deal with the system regulation problem under different impatience degrees.

3. Problem description

Consider a production service system consisting of one manufacturing facility and one service centre as shown in Figure 1, two kinds of impatient behaviors (i.e. balking and reneging) are considered in this paper.

In the first stage, products are produced in a make to stock (MTS) manner and are stored in the warehouse. Then the products can be as the 'raw materials' processed in the second stage to fulfil external orders. It is assumed that one unit of product is used for processing one external order. The first stage and the second stage coordinate their operations in the following way: if the warehouse is not empty and the service facility starts to process an order, a call for a unit of product is sent to the warehouse and a unit of product is sent immediately to the second stage and service begins.

Demands arriving is an independent Poisson process with rates λ . The production time of product and the processing time of external order are exponentially distributed with parameters μ_1 and μ_2 , respectively. Such assumptions are consistent with previous treatments of production–inventory and service systems, for example, (Stidham 1985; Veatch and Wein 1994; Ha 1997; Carr and Duenyas 2000; Iravani, Liu, and Simchi-Levi 2012). It is also precise in approximating the behavior of systems where variability is high. Actually, as the parameters of exponentially distributed production time and service time indicate the expected production time and service time are $\frac{1}{\mu_1}$ and $\frac{1}{\mu_2}$, respectively, parameters μ_1 and μ_2 represent the expected production throughput (i.e. production capacity) and service throughput (i.e. service capacity). We further assume that pre-emptions (i.e. interrupting production) are allowed during both production and service operations. If interruption occurs, it can be resumed the next time when production is initiated. Because of the

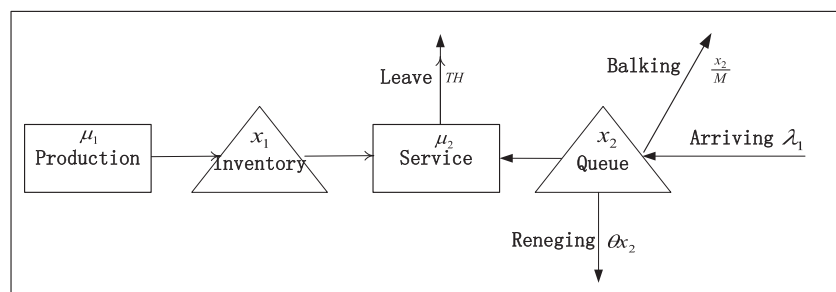


Figure 1. The production service system with balking and reneging.

memory-less property of exponential distribution, the probability distribution of the production time when the previous production is not interrupted is exactly the same as that of the production time when interrupting previous production occurs. Therefore, resuming production where it was interrupted is equivalent to initiating it from the start. That is, the memory-less property implies that the property of production time required to produce a product in future depend only on relevant information about the current time, not on information from further in the past. The cases of service and production are same.

The production service system can be decomposed into two subsystems: a MTS queue system and a classical M/M/1 queue system. The arrival of MTS queue system is the internal demands for products. Let x_1 denote the inventory level of products. x_2 is the number of waiting orders for service. $x_1, x_2 \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ is the set of all non-negative integer numbers. The system state can be described by a vector $x = (x_1, x_2) \in \Omega$, where $\Omega = \mathbb{Z}^+ \times \mathbb{Z}^+$. Noting that the pre-emptions are allowed during production and service periods, it is not necessary to include in the state description the numbers of products and waiting orders currently being produced and served, respectively.

Differing from production inventory systems, the production service system cannot buffer the mismatching between service capacity and demand through inventory due to the un-storability of service. Therefore, customers waiting is inevitable. Awaiting for service can easily lead to impatient behaviors. In this study, we consider two kinds of impatient behaviors, i.e. balking and reneging. When customers arrive, if the order backlog is beyond the expectation, customers may choose to quit the system without placing an order, namely balking. The probability function of balking is a linear function of the number of waiting orders. Linear balking function is the most concise positive correlation between balking probability and expected waiting time. This assumption is widely used in the studies on the balking behavior, as shown in Ancker and Gafarian (1963), Singer and Khmelnitsky (2010). Balking probability is:

$$G(x_2) = \frac{x_2}{M}. \quad (1)$$

Parameter M , a positive integer, measures the impatience degree of customers. The smaller M is, the greater the customer impatience in terms of refusing to join the waiting queue will be. Therefore, M is also known as balking parameter. The actual customers arrival rate can be written as:

$$\lambda(x_2) = \lambda * \left(1 - \frac{x_2}{M}\right). \quad (2)$$

Let S_2 denotes the maximum capacity of order waiting area, e.g. $x_2 \leq S_2$. To make sure that the value of balking probability is logical, we assume $S_2 < M$. Otherwise, balking probability may be greater than 1, or equal to 1. The former is illogical obviously. If $S_2 = M$, balking probability may equal 1, customer balking will make the waiting queue length always smaller than the maximum capacity of waiting area. Then the maximum capacity of waiting area will be meaningless. It should be noted that the balking probability in the case of $x_2 = S_2 - 1$ is still smaller than 1 and can be determined by Equation (1). However, in the case of $x_2 = S_2$, an arriving customer will never join the waiting queue because the waiting length reaches the maximum capacity of waiting area.

Balking cost c_b can be regarded as the sum of order loss cost and customers loss cost. Here, order loss cost refers to the opportunity cost gaining benefits by selling per unit of product and service portfolio; customers loss cost refers to the loss cost of losing the customers forever, which can be worked out through the sales statistics.

In addition, if the waiting time exceeds the upper limit T of customers expectation after entering the system, customers would choose to quit the waiting queue. T is an exponential random variable with mean of $1/\theta$. This assumption is widely used in the studies on the reneging behavior (Benjaafar, Gayon, and Tepe 2010), (Economopoulos, Kouikoglou, and Grigoroudis 2011), (Benjaafar and Elhafsi 2012), (Ioannidis et al. 2013). Then customers reneging rate can be expressed as:

$$r_\theta(x_2) = \theta x_2. \quad (3)$$

Reneging cost c_r covers balking cost and backlog cost. Backlog cost is the waiting cost of customers before leaving the system. Therefore, the inequality $c_b \leq c_r$ holds. For the detailed description of the influence of impatient customers on the operation cost, and the calculation of specific cost, we refer readers to Allon, Federguen, and Pierson (2011), Robinson and Chen (2011), and Blackburn (2012).

4. Markov modelling

In this paper, the base stock policy is used to manage the inventory. That is, when the inventory level falls below a fixed value S_1 , the system begins production and stops production otherwise. S_2 is the maximum capacity of the

system's order waiting area. Recall that $S_2 < M$. The system state space can be expressed as $\{(x_1, x_2) | x_i = 0, 1, 2, \dots, S_i, i = 1, 2\}$, and the size of state space is $L = (S_1 + 1)(S_2 + 1)$.

If the system's transition rate from the state x to x' is $q(x \rightarrow x')$, the unit vectors are $e_1 = (1, 0)$, $e_2 = (0, 1)$, so the equations set of the system state transition rate will be:

$q(x \rightarrow x + e_1) = \mu_1,$	$x_1 < S_1;$
$q(x \rightarrow x - e_1 - e_2) = \mu_2,$	$x_1 > 0, 0 < x_2 \leq S_2;$
$q(x \rightarrow x + e_2) = \lambda(1 - \frac{x_2}{M}),$	$x_2 < S_2;$
$q(x \rightarrow x - e_2) = \theta x_2,$	$0 < x_2 \leq S_2.$

Here, $q(x \rightarrow x + e_1) = \mu_1$ means that when the inventory level drops to a base stock level S_1 , the system begins to produce, then the inventory level will increase by one unit at the speed of μ_1 . $q(x \rightarrow x - e_1 - e_2) = \mu_2$ means that when there are orders waiting, and there is inventory in the system, the system begins to deal with the order, allowing the inventory to reduce by one unit at the speed of μ_2 , and the waiting orders to reduce by one unit at the speed of μ_2 . $q(x \rightarrow x + e_2) = \lambda(1 - \frac{x_2}{M})$ means that when the system state is (x_1, x_2) , and an arriving customer chooses to enter into the system with probability of $1 - \frac{x_2}{M}$, then the waiting orders in the system increase by one unit at the speed of $\lambda(1 - \frac{x_2}{M})$. $q(x \rightarrow x - e_2) = \theta x_2$ means that when the system state is (x_1, x_2) , the number of customers waiting for service decreases by one unit at the speed of θx_2 because of customer reneging.

Because the system state is a two-dimensional vector, the state transition rate matrix will be generated using the transfer methods from a multi-dimensional vector to a one-dimensional vector, which is a $(S_1 + 1)(S_2 + 1)$ -dimensional matrix. The specific method building the state transition rate function transition(a, b) between the two states is as follows. The position of state x is a , the position of state x' is b . The state transition rate matrix of a continuous-time Markov chain is H . The transition probability between system states can be expressed as follows.

- (i) If $b - a = 1, x_1 < S_1, H_{ab} = \mu_1$;
- (ii) If $a - b = (S_1 + 1) + 1, x_1 x_2 > 0, H_{ab} = \mu_2$;
- (iii) If $b - a = (S_1 + 1), x_2 < S_2, H_{ab} = \lambda(1 - \frac{x_2}{M})$;
- (iv) If $a - b = (S_1 + 1), 0 < x_2 \leq S_2, H_{ab} = \theta x_2$;
- (v) Otherwise, $H_{ab} = 0$.

Then the construction of the state transition velocity matrix H corresponding to a continuous-time Markov chain is completed.

Define $\pi(x_1, x_2)$ as the steady-state distribution probability of the state (x_1, x_2) . For the sake of convenience of description, to sort all the system states according to the dictionary sorting. The steady-state probability distribution is denoted as π_l , in which $l = 1, 2, \dots, L$, and $L = (S_1 + 1)(S_2 + 1)$ is the size of state space. Let:

$$\pi = (\pi_1, \pi_2, \dots, \pi_L),$$

then according to the steady propositions of continuous-time Markov chain, we have the following equations set.

$$\pi H = 0. \quad (4)$$

Add the normalised equation into the equations set:

$$\sum_{l=1}^L \pi_l = 1. \quad (5)$$

Then the matrix H is changed to the matrix Φ in the following manner:

$$\Phi(:, l) = \begin{cases} A(:, l), & l = 1, \dots, L - 1 \\ 1, & l = L \end{cases} \quad (6)$$

Then

$$\pi \Phi = \psi.$$

Here, L dimensional vector $\psi = (0, \dots, 0, 1)$. Therefore, π can be determined in the following manner:

$$\pi = \psi \Phi^{-1}. \quad (7)$$

Then, by the inverse transform between the system state (x_1, x_2) and the corresponding value l , the corresponding steady-state probability $\pi(x_1, x_2)$ of the system state (x_1, x_2) can be obtained. The continuous-time Markov chain in this study is irreducible (i.e. reaching another state from any state), and the state space size is finite, thus there is only one steady state solution (Hoel, Port, and Stone 1972). The specific solution procedure in Matlab is not discussed here. In Matlab, it is necessary to write the solution procedure as a function, so that we can continue to change system and economic parameters in the running programme to study the performance of the system under different scenarios.

5. Performance analysis

In this section, we analyse the impact of customer impatience on the system performance. To this end, we firstly give the expressions of key system performance indices, including throughput, expected inventory, expected queue length in the waiting area, expected balking rate, expected reneging rate, as well as system net profit. According to the system description, the expressions of performance indices are as follows.

The system throughput is expressed as:

$$TH = \mu_2 \sum_{x_2=1}^{S_2} \sum_{x_1=1}^{S_1} \pi(x_1, x_2). \quad (8)$$

The expected inventory level can be expressed as:

$$I = \sum_{x_1=1}^{S_1} x_1 \left(\sum_{x_2=0}^{S_2} \pi(x_1, x_2) \right). \quad (9)$$

The expected length of waiting queue is:

$$Q = \sum_{x_2=1}^{S_2} x_2 \left(\sum_{x_1=0}^{S_1} \pi(x_1, x_2) \right). \quad (10)$$

Here, we consider two kinds of impatient behaviors balking and reneging. When customers arrive and decide whether to enter the system according to the waiting queue length, the balking probability function is Equation (1). Therefore, the expected balking rate can be expressed as:

$$B_1 = \sum_{x_2=0}^{S_2} \left(\lambda \frac{x_2}{M} \left(\sum_{x_1=0}^{S_1} \pi(x_1, x_2) \right) \right). \quad (11)$$

When a customer joins the waiting queue and decides to renege after waiting for too long, according to the reneging rate as shown in Equation (3), customers expected reneging rate is:

$$B_2 = \sum_{x_2=0}^{S_2} \left(\theta x_2 \left(\sum_{x_1=0}^{S_1} \pi(x_1, x_2) \right) \right). \quad (12)$$

Serving customers leads to the system expected revenue $R * TH$. The system cost consists of inventory holding cost hI , customers waiting cost $c_w Q$, balking cost $c_b B_1$ and reneging cost $c_r B_2$. Thus, the expression of system net profit:

$$NP = R * TH - (hI + c_w Q + c_b B_1 + c_r B_2) \quad (13)$$

The numerical experiments are designed as follows. M and θ represent the degrees of two kinds of customer impatience (balking and reneging). The larger M is, the smaller the customer impatience will be; the larger θ is, the greater the customer impatience will be. To analyse the impact of customer impatience on the system, respectively, through the value selection changes of M and θ . When analysing one impatient behavior, it's better to neglect the impact of the other impatient behavior on the system performance by setting the parameter value of the other impatient behavior. Take reneging parameter $\theta = 10^{-6}$ to analyse the balking behavior, and take balking parameter $M = 10^6$ to analyse the reneging behavior. According to the non-heavy traffic range, the system parameters $\lambda = 0.8$, $\mu_1 = 1$ in the following analysis. Economic parameters are as follows: product and service order revenue $R = 200$, inventory holding cost per unit time for each item $h = 1.5$, waiting cost per unit time for each customer $c_w = 8$, balking cost for each customer $c_b = 15$, reneging cost for each customer $c_r = 17$. Maximum capacity of waiting area $S_2 = 10$. In order to simultaneously reflect the coordination between production capacity and service capacity in the production service system, when the

production capacity is given in all numerical experiments, by taking different sets of values for service capacity μ_2 , we can observe the impact of impatient behaviors under different service capacity on the system. In addition, by taking different sets of values for base stock level S_1 , we can observe the impact of impatient behaviors capacity on the system under different inventory level. The influence of production capacity on the system will be analysed in the optimal control section. For the sake of brevity, only four sets of results corresponding to the service capacity $\mu_2 = 1$, $S_1 = 10$ and $\mu_2 = 0.7$, $S_1 = 10$; $\mu_2 = 1$, $S_1 = 5$ and $\mu_2 = 0.7$, $S_1 = 5$ are retained in each numerical experiment, which can reflect the impact of the different service capacity on the system.

Firstly, we analyse the impact of balking behavior on the system. Change the value of the balking parameter M from $M = 10$ to $M = 50$, the value change of M reflects the decreasing of customer impatience, i.e. less and less impatient.

As it is shown in Figure 2, with the decreasing impatience in terms of customer balking, the inventory level in the system is gradually reduced. Due to the lower customer impatience, more customers are willing to wait for service, so their demand for products will increase, thereby reducing the inventory level.

It's also shown that the inventory levels are different under the different service capacity. When the service capacity is relatively low ($\mu_2 = 0.7$), the inventory level is relatively high. This is caused by the unique characteristics of the production service system, in which the demand for products is no longer fully determined by external demand, but jointly determined by service capacity and external demand. Intuitively, the product demand rate is the minimum between external demand rate and service capacity. Therefore, a lower service capacity reduces the product demand of the system, resulting in an increased inventory level.

It's also shown that the impact of impatient behaviors on the inventory changes under different service capacity. When $\mu_2 = 0.7$, the impact of value changes of the impatience parameter on the inventory changes is smaller than that when $\mu_2 = 1$. This phenomenon is also relevant to that the product demand is jointly determined by external demand and service capacity in the production service system, as we previously mentioned. The lower the service capacity is, the smaller the impact of external demand on product demand will be. Meanwhile, the balking behavior only affects the external demand. Therefore, the lower the service capacity is, the smaller the impact of balking behavior on inventory will be. This phenomenon indicates that the service in the production service system plays as a buffer on the impact of external product demand on the inventory. The lower the service capacity is, the smaller impact of external demand on product demand will be. In addition, the expected inventory level increases with the base stock level. The reason is clear.

As shown in Figure 3, with the customer impatience in terms of balking behavior is reduced, the number of customers waiting for service in the system will gradually be increased. According to the definition of balking probability, with the value of M increases, the probability of customer balking will be reduced. There will be more customers choosing to wait, thus increasing the queue length.

Figure 3 also indicates that the number of customers waiting for service is different under different service capacity. When the service capacity is relatively low, the number of customers waiting for service is relatively large. The reason

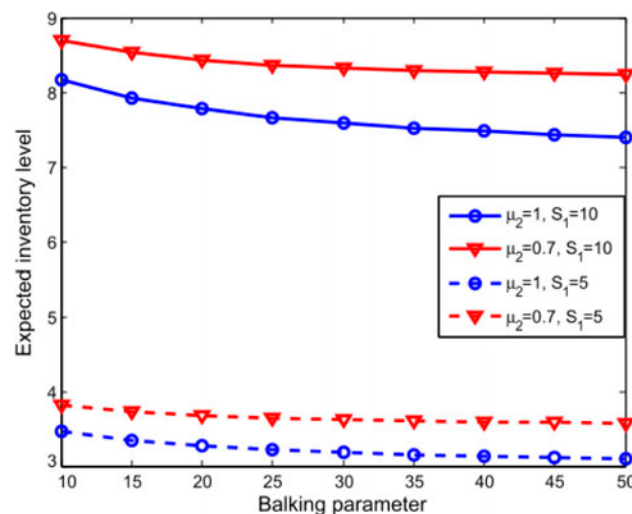


Figure 2. Impact of balking behavior on expected inventory level.

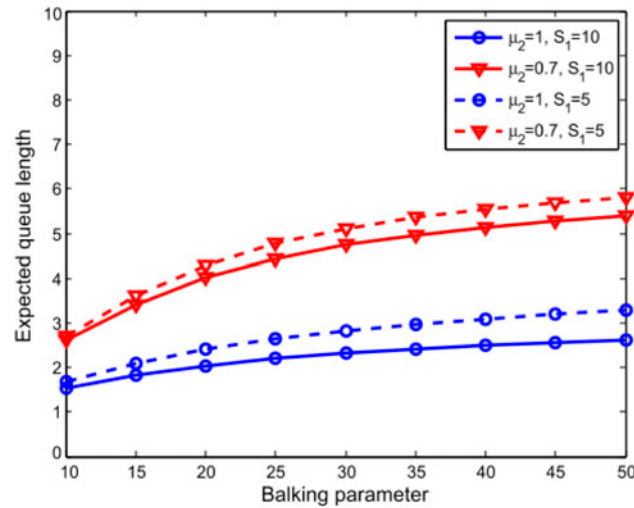


Figure 3. Impact of balking behavior on expected queue length.

is clear. Under the same inventory level and balking parameter, the lower service capacity reduces the system actual service capacity, thus increasing the number of customers waiting for service.

Moreover, the impact of balking on the queue length changes under different service capacity. When $\mu_2 = 0.7$, the impact of value changes of the balking parameter on the queue length changes is larger than when $\mu_2 = 1$. The reasons are as follows: the changes of balking parameter improve the actual arrival rate of customers in nature. When the service capacity is relatively high, the service centre can better respond to improve the actual arrival rate, with no large order backlog increased; when the service capacity is relatively low, the increased customers arrival rate will inevitably lead to a large backlog of service orders, and thus an increased waiting queue length. Moreover, the queue length decreases with the base stock level. This phenomenon reflects the influence of production inventory subsystem on the service subsystem. Higher base stock level leads to higher inventory level, and less waiting for product, then higher base stock results in shorter waiting queue. Keep in mind that the influence of service capacity on the inventory level observed in Figure 2, we conclude that two subsystems interact each other in the production service system. The interaction will be analysed in detail in the optimal control section.

Figure 3 also shows that more impatient the customers are, the smaller the influence of base stock level on the queue length will be. This is because more impatient customers lead to the actual demand for the product decreases. Therefore, lower base stock level has a little of influence on the queue length.

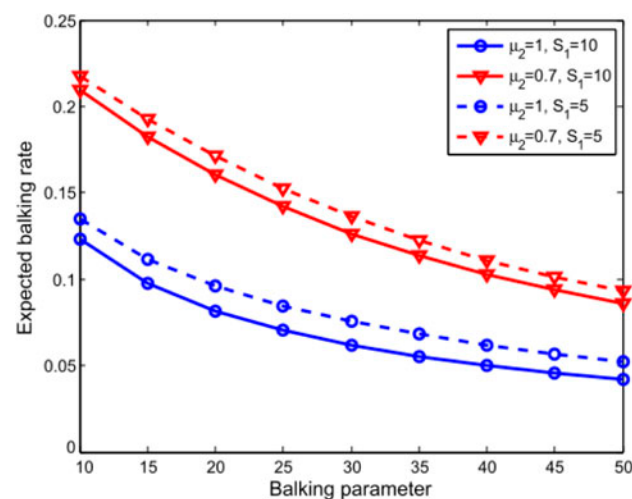


Figure 4. Impact of balking behavior on expected balking rate.

Figure 4 shows that as the degree of the customer impatience in terms of balking behavior is decreased, the expected balking rate will be gradually decreased. The reason is very obvious. It's noted that when the balking parameter changes, the expected balking rate is different under different service capacity. The balking rate is jointly determined by the queue length and balking parameter, so the balking rate is completely determined by queue length during the same change of balking parameter. As it is shown in Figure 3 when the service capacity is relatively low, the queue length is relatively high, resulting in the relatively high balking rate. In addition, higher base stock level, smaller the expected balking rate. This is because higher base stock level results in smaller queue length, and smaller balking rate.

As it is shown in Figure 5, as the degree of the customer impatience in terms of balking behavior is reduced, the system throughput will be gradually increased. This is because for a given production and service capacity, the reduced impatience degree improves the actual demand, resulting in an improved system throughput. In addition, higher service capacity and base stock level result in a higher throughput. The reason is clear.

From Figure 6, we can find the relationship between system net profit and balking behavior.

Figure 6 shows that as the customer impatience is reduced, the change pattern of system net profit under different service capacity is different. Figures 2–5 show that as the value of M is increased, the system throughput will be increased, the inventory level will be reduced, the customer balking rate will be reduced, whereas the customer queue length will be increased. Therefore, the system revenue is improved, the inventory holding cost and the customer balking cost are reduced, whereas the waiting cost is increased. Therefore, the profit change is uncertain, and is jointly determined by the size of performance indices of the system when the size of economic parameters are fixed. In addition, as shown from Figures 2 to 5, the size of performance indices is affected largely by the service capacity μ_2 . Therefore, the profit change pattern under the different service capacity is different (Figure 6). Meanwhile, the profit change pattern under different base stock level is also uncertain. The reason is similar to that related to service capacity since inventory is another form of production capacity.

Specifically, when $\mu_2 = 0.7$, it's shown in Figure 3 that the system queue length is increased significantly, making the waiting cost increase substantially, offsetting the increase range of system profit brought by the reduced balking rate and inventory level and the increased system throughput rate. Therefore, when $\mu_2 = 0.7$, as the value of M is increased, the system net profit will decline. However, when $\mu_2 = 1$, the system queue length is increased to a lesser extent (see Figure 3), making the waiting cost increase slightly, so along with the value of M is increased, the system net profit will rise.

The following is the analysis of the impact of customer reneging on the system. Change the value of θ from $\theta = 0.04$ to $\theta = 0.2$. The value range of θ is selected combined with the system state size, making the overall customer reneging rate always smaller than the actual customers arrival rate in the system, to ensure the effectiveness of the system state. The value change process of θ reflects the gradual increase in impatience degree in terms of customer reneging. Thus, customer impatience changes reflected by the change process of the value of reneging parameter θ and the change process of the value of balking parameter M is opposite. Figures 7–11 show the impact of reneging on the system performance.

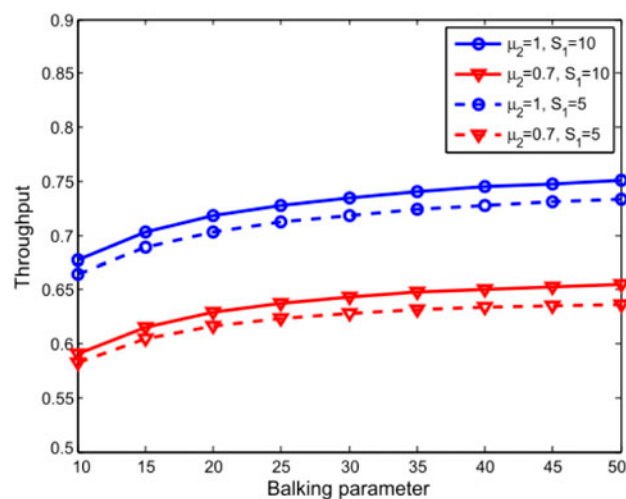


Figure 5. Impact of balking behavior on throughput.

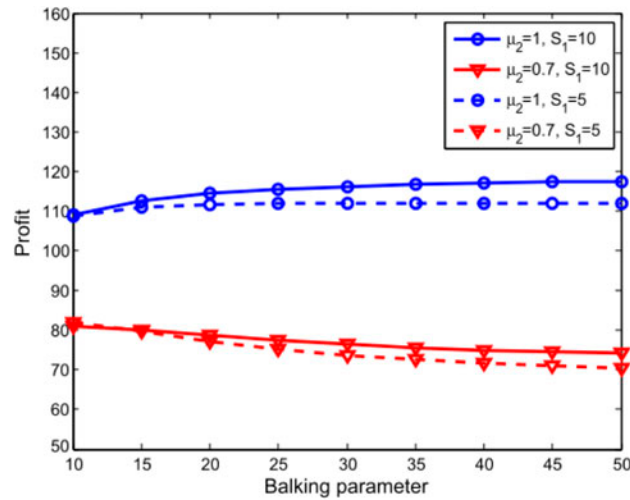


Figure 6. Impact of balking behavior on system net profit.

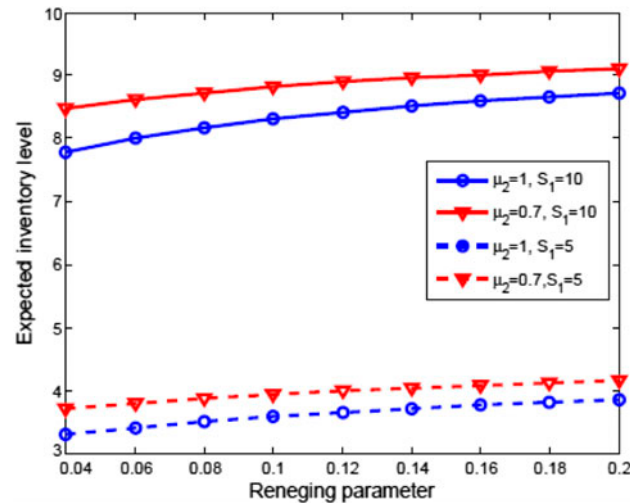


Figure 7. Impact of reneging behavior on expected inventory level.

Figures 7–10 show that the changes of system performance indices are similar to the changes of system performance indices when balking parameter is decreased. The details are as follows. (1) The inventory level increases with the degree of impatience, and a lower service capacity leads to a higher overall inventory level. (2) The number of customers waiting for service decreases with the degree of impatience; lower service capacity leads to greater impact of customer impatience on the queue length changes; higher degree of customers impatience results in smaller impact of base stock level on the queue length. (3) The expected reneging rate increases with the degree of impatience; lower service capacity leads to higher reneging rate; higher degree of customers impatience results in smaller impact of base stock level on the reneging rate because more impatient customers lead to a smaller impact of base stock level on the queue length. Meanwhile, the expected reneging rate is determined by the queue length completely. Therefore, higher degree of customers impatience results in smaller impact of base stock level on the reneging rate. (4) The system throughput decreases with the degree of impatience, higher service capacity leads to higher throughput. The above phenomenon is similar to what has been observed on the impact of balking behavior on the system performance shown in Figures 2–5. The explanation is also identical with what presented previously.

As shown in Figure 11, along with the increased customer impatience, the change patterns of system net profit under the different service capacity are different. Figures 7–10 show that as the value of reneging parameter, θ , is increased,

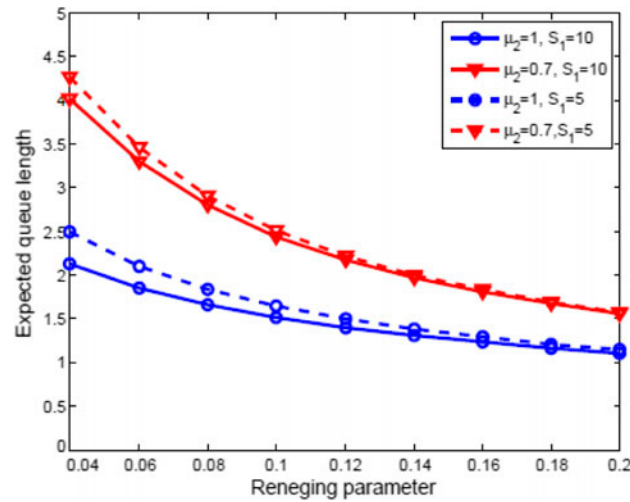


Figure 8. Impact of reneging behavior on expected queue length.

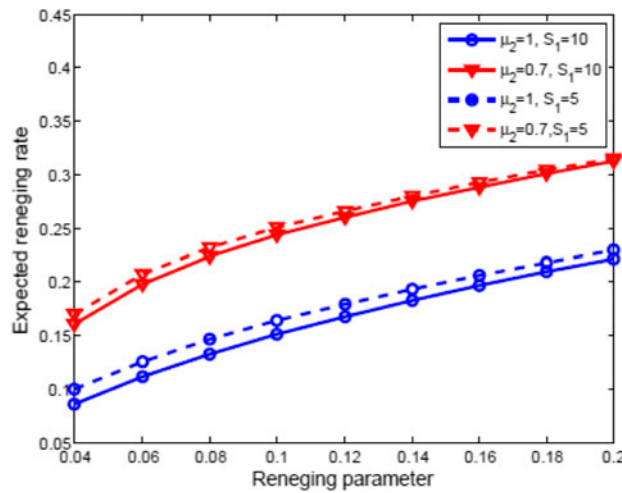


Figure 9. Impact of reneging behavior on expected reneging rate.

the inventory level will be increased, the queue length will be decreased, the reneging rate will be increased, and the system throughput will be reduced. Thus, the system profit is reduced. However, because of the lower waiting cost, the increased inventory holding cost and reneging cost, then the total cost change is uncertain. The corresponding suggestions for firms on the reneging impact are identical to that related to the balking behavior.

6. Optimal control

In this section, we study the impact of impatient behaviors on the system optimal control, by finding the optimal base stock level to allow the maximum net profit achieved.

The following theorem gives the upper bound of the optimal inventory level in order to determine the search scope.

Theorem 1. Optimal inventory level S^* satisfies the following conditions:

$$0 \leq S^* \leq \left\lfloor \frac{\mu_2 R - c_w}{h} \right\rfloor. \quad (14)$$

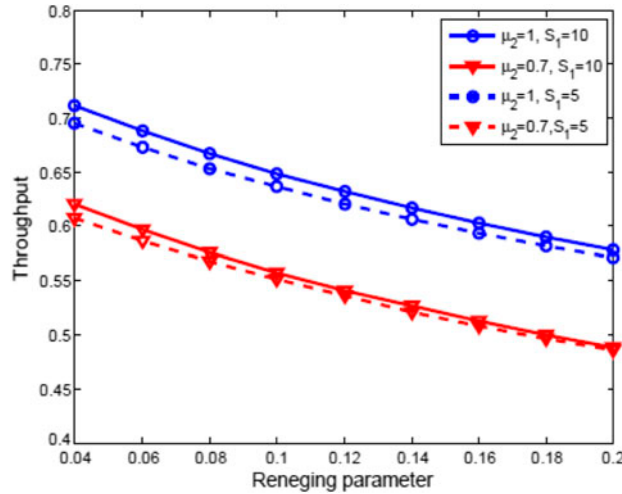


Figure 10. Impact of reneging behavior on throughput.

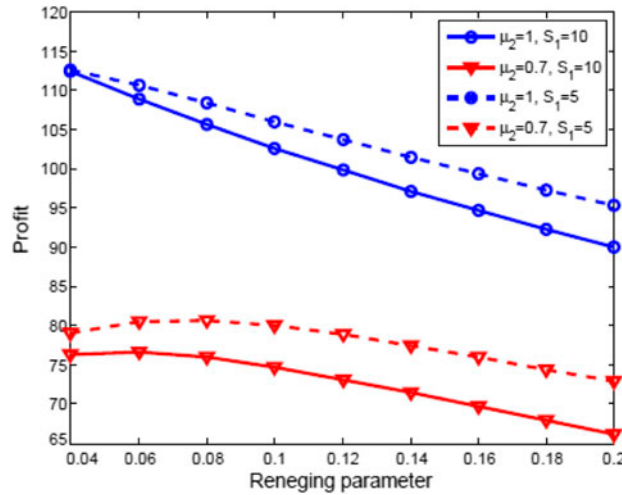


Figure 11. Impact of reneging behavior on system net profit.

where $\lfloor \cdot \rfloor$ is a floor function.

Proof: If serving orders is feasible, the system is always dealing with service orders. According to the characteristics of the production service system, the product demand rate is always not greater than the service capacity μ_2 . Therefore, when the inventory is $S^* - 1$, if one more product is produced, the minimum time of holding this product is no less than S^*/μ_2 . Then the holding cost of the product is greater than hS^*/μ_2 . If the expected service time to complete this order is $1/\mu_2$, the waiting cost for this order is c_w/μ_2 . If the revenue gained to complete this order is R , when the optimal inventory level is S^* , it shall meet the condition

$$hS^*/\mu_2 + c_w/\mu_2 < R. \quad (15)$$

i.e. $S^* < \frac{\mu_2 R - c_w}{h}$. Considering S^* is an integer, it is rewritten as $S^* \leq \lfloor \frac{\mu_2 R - c_w}{h} \rfloor$. Here $\lfloor \cdot \rfloor$ is rounded down to a floor integer value. Therefore, the inequality (14) holds.

For the sake of convenience of narration, write $\bar{S}^* = \lfloor \frac{\mu_2 R - c_w}{h} \rfloor$. Thus, simply by making a global search within the interval $[0, \bar{S}^*]$, we can get the optimal base stock level S^* and corresponding optimal profit NP^* . Here are the specific numerical experimental results and conclusions. Similar to the previous numerical experiments in the section of performance analysis, in order to eliminate the mutual influence between the two impatient behaviors, we take the reneging behavior parameter $\theta = 10^{-6}$ when analysing customer balking. In the analysis of the reneging behavior, the balking

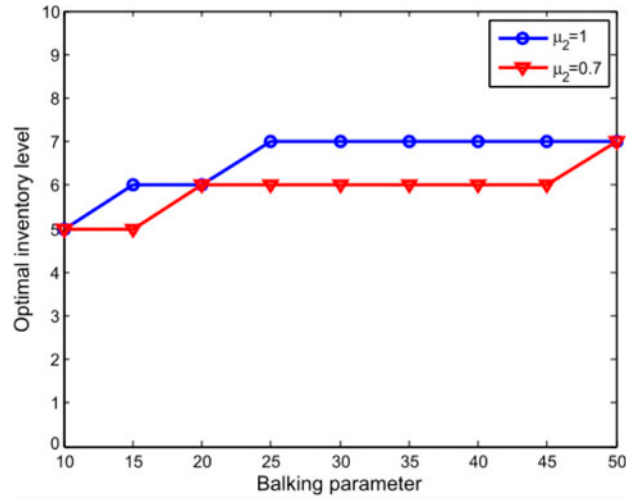


Figure 12. Impact of balking behavior on optimal inventory level.

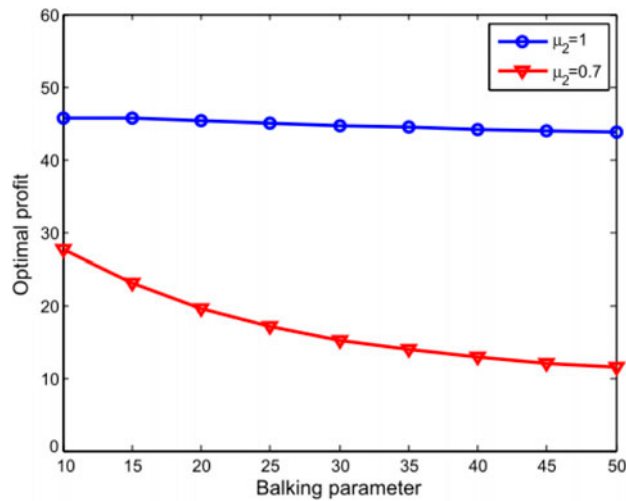


Figure 13. Impact of balking behavior on optimal profit.

parameter takes the value $M = 10^6$. In the following analysis, the system parameters are $\lambda = 0.8$, $\mu_1 = 1$ and the economic parameters are $R = 100$, $h = 1.5$, $c_w = 8$, $c_b = 15$, $c_r = 17$, maximum capacity of waiting area $S_2 = 10$. Through a global search, we can get the optimal base stock levels and the optimal profits of the system under two kinds of impatient behaviors, respectively, with the results shown in Figures 12–15.

Figures 12 and 13 show that as the impatience degree in terms of customer balking is decreased, the optimal base stock level of the system will be increased, and the optimal profit will be reduced. It indicates that under the optimal control, along with the increased actual demand of reduced customer impatience, firms need to improve the inventory level to respond. However, even along with the increased optimal inventory level, the system net profit will still decline as the reduced customer impatience. It is shown that although the increased system throughput rate will increase revenue, and the reduced impatience degree will reduce balking cost, the increased order waiting cost caused by the improved customer arrival is larger than the increased revenue and the reduced loss cost, for the increased inventory level will result in an increased holding cost.

It can be seen in Figure 12 that under two service capacity $\mu_2 = 1$ and $\mu_2 = 0.7$, the optimal inventory level is relatively high along with the higher service capacity. Higher service capacity will increase the demand for products, so the system needs to set a higher inventory level.

Table 1. Impact of balking behavior on optimal control under different economic and system parameters.

S^*											NP^*									
	$M=10$	15	20	25	30	35	40	45	50		$M=10$	15	20	25	30	35	40	45	50	
R	100	5	6	6	7	7	7	7	7		45.73	45.71	45.35	44.94	44.66	44.4	44.16	43.96	43.78	
	120	5	6	6	7	7	7	7	7		59.02	59.6	59.52	59.39	59.25	59.09	58.93	58.79	58.67	
	140	6	6	7	7	7	7	7	8		72.35	73.48	73.71	73.84	73.83	73.77	73.7	73.62	73.58	
	160	6	6	7	7	7	7	8	8		85.75	87.37	87.96	88.28	88.41	88.46	88.47	88.49	88.52	
h	1.1	6	6	7	7	8	8	8	8		47.29	47.4	47.2	46.91	46.6	46.41	46.23	46.07	45.93	
	1.3	5	6	7	7	7	7	8	8		46.42	46.56	46.21	45.93	45.63	45.36	45.12	44.95	44.81	
	1.5	5	6	6	7	7	7	7	7		45.73	45.71	45.35	44.94	44.66	44.4	44.16	43.96	43.78	
	1.7	5	6	6	6	7	7	7	7		45.03	44.87	44.53	44.08	43.69	43.44	43.21	43.01	42.84	
c_w	5	5	5	6	6	6	6	7	7		50.8	51.71	52.1	52.28	52.34	52.35	52.34	52.34	52.36	
	6	5	6	6	6	6	7	7	7		49.11	49.67	49.85	49.82	49.71	49.63	49.59	49.55	49.5	
	7	5	6	6	6	7	7	7	7		47.42	47.69	47.6	47.35	47.16	47.01	46.88	46.75	46.64	
	8	5	6	6	7	7	7	7	7		45.73	45.71	45.35	44.94	44.66	44.4	44.16	43.96	43.78	
c_b	9	5	6	6	7	7	7	7	7		46.54	46.35	45.89	45.4	45.06	44.76	44.49	44.26	44.06	
	11	5	6	6	7	7	7	7	7		46.27	46.14	45.71	45.25	44.93	44.64	44.38	44.16	43.97	
	13	5	6	6	7	7	7	7	7		46	45.93	45.53	45.09	44.79	44.52	44.27	44.06	43.87	
	15	5	6	6	7	7	7	7	7		45.73	45.71	45.35	44.94	44.66	44.4	44.16	43.96	43.78	
μ_1	0.8	7	9	10	10	11	11	12	12		42.58	41.63	40.52	39.58	38.82	38.2	37.68	37.27	36.93	
	0.9	6	7	7	8	8	8	9	9		44.56	44.24	43.61	43.11	42.64	42.23	41.94	41.69	41.48	
	1	5	6	6	7	7	7	7	7		45.73	45.71	45.35	44.94	44.66	44.4	44.16	43.96	43.78	
	1.1	5	5	6	6	6	6	6	6		46.5	46.67	46.36	46.15	45.92	45.7	45.5	45.33	45.18	
μ_2	0.8	5	6	6	6	7	7	7	7		35.44	33.21	31.25	29.62	28.42	27.53	26.81	26.24	25.76	
	0.9	5	6	6	7	7	7	7	7		41.31	40.47	39.5	38.61	37.99	37.47	37.03	36.66	36.36	
	1	5	6	6	7	7	7	7	7		45.73	45.71	45.35	44.94	44.66	44.4	44.16	43.96	43.78	
	1.1	5	6	6	7	7	7	7	7		49.12	49.58	49.56	49.43	49.34	49.24	49.14	49.04	48.95	

It can be observed in Figure 13 that along with the deceased impatience degree in terms of customer balking, the optimal profits under service capacity $\mu_2 = 0.7$, and the optimal profit declines sharply. It's verified by this phenomenon that, as explained previously, the reduced customer impatience will lead to a declined system optimal profit.

Table 1 shows the impact of balking behavior on the optimal control under different economic and system parameters. The set of test examples consists of six groups and in each group, and we change one parameter while fixing the rest. The standard system parameters are $\lambda = 0.8$, $\mu_1 = 1$, $\mu_2 = 1$ and the economic parameters are $R = 100$, $h = 1.5$, $c_w = 8$, $c_b = 15$, with maximum capacity of waiting area $S_2 = 10$. Table 1 exhibits that the optimal inventory level S^* increases with the balking parameter M regardless of economic parameters, system parameters. The phenomenon verifies the result shown in Figure 12 once again. The reason is similar to that following Figure 12. The observations of the optimal inventory management with respect to economic and system parameters are as follows. The optimal inventory level S^* increases with waiting cost c_w , decreases with holding cost h . This phenomenon is identical with our intuition. Higher order waiting cost c_w results in a higher inventory level, which avoids a long wait and thereby reduces the waiting cost. Higher holding cost makes the manager set a lower inventory level to reduce holding cost. Meanwhile, the optimal inventory level S^* increases with revenue R . This phenomenon is caused by customers impatience, which will result in a part of impatient customers balk and hence lead to a higher opportunity cost. Thus, impatience makes the manager take the probable order loss into account, whereas it is not the case when customers are perfectly patient. Although the optimal inventory level S^* under different balking costs are same in Table 1, S^* will be increased when c_b is large enough, e.g. $c_b = 40$. This phenomenon is also identical with our intuition. Higher balking cost makes the manager to set a higher inventory level, which reduces customers waiting time. Therefore, it reduces balking cost. For the influence of system capacity parameters on the optimal inventory management, the last two groups in Table 1 shows that the optimal inventory level S^* decreases with production capacity μ_1 sharply, whereas increases with service capacity μ_2 mildly. Higher production capacity leads to replenishing stock quickly is probable, therefore the manager can set a lower base stock level to avoid a higher inventory holding cost. Actually, inventory is as the other form of the

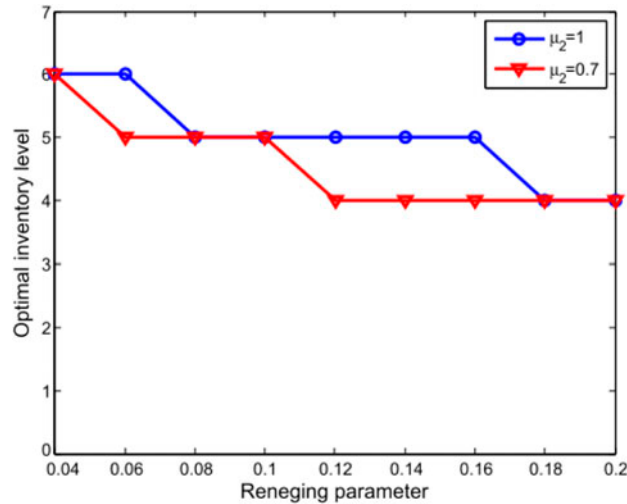


Figure 14. Impact of reneging behavior on optimal inventory level.

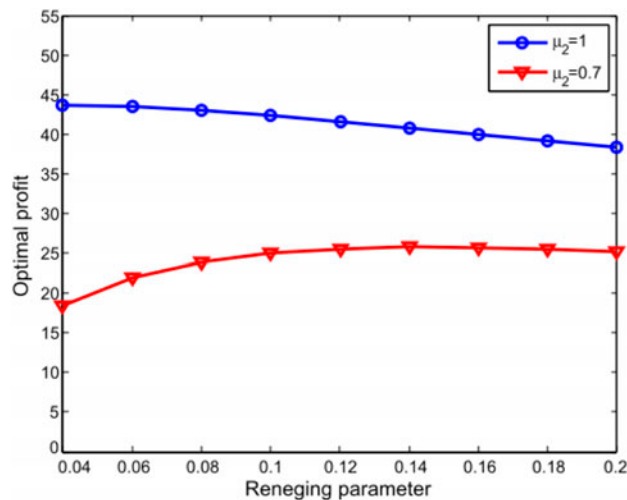


Figure 15. Impact of reneging behavior on optimal profit.

production capacity in the production inventory subsystem. So a higher one is with a lower other one, and *vice versa*. The relationship between the optimal base stock level S^* and service capacity μ_2 reflects the importance of coordinating production inventory subsystem and service subsystem management. For a higher service capacity, as we analysed previously, the internal demand for the product will be higher. Thus, the inventory level should be higher to meet the internal demand. Meanwhile, because the internal demand is determined by service capacity and external demand jointly, then a higher service capacity results in a mildly higher optimal inventory level.

The observations of the optimal profit with respect to economic and system parameters are as follows. Firstly, for all of constant customers impatience degrees, the optimal net profit NP^* increases with revenue R , production capacity μ_1 and service capacity μ_2 , and decreases with various costs h , c_w , c_b . The reason of the above phenomenon is clear. The profit increases with revenue and decreases with cost when all other system and economic parameters are fixed. The profit increases with production capacity since, for the higher production capacity, and the optimal inventory level will be lower as the Table 1 shown. Similarly, a higher service capacity will lead to a shorter waiting queue. Thus, the waiting cost will be less. It should be pointed that we do not take the fixed investment costs of production and service capacity into account here. The profit can be seen as the short-term goal of the firm. Secondly, differing from the

Table 2. Impact of renegeing behavior on optimal control under different economic and system parameters.

	S^*										NP^*									
	$\theta = 0.04$					$\theta = 0.04$					$\theta = 0.04$					$\theta = 0.04$				
	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2			0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2		
R	100	6	6	5	5	5	4	4			43.66	43.49	42.95	42.35	41.61	40.79	39.94	39.17	38.43	
	120	6	6	6	5	5	5	4			57.71	57.08	56.08	55.08	54.03	52.93	51.82	50.72	49.69	
	140	7	6	6	5	5	5	5			71.83	70.66	69.27	67.81	66.45	65.07	63.7	62.36	61.06	
	160	7	6	6	6	5	5	5			85.96	84.25	82.45	80.59	78.87	77.21	75.58	74.01	72.48	
h	1.1	7	6	6	5	5	5	5			45.57	45.2	44.65	43.88	43.07	42.28	41.45	40.61	39.77	
	1.3	7	6	6	5	5	5	4			44.57	44.35	43.77	43.07	42.34	41.53	40.69	39.84	39.02	
	1.5	6	6	5	5	5	4	4			43.66	43.49	42.95	42.35	41.61	40.79	39.94	39.17	38.43	
	1.7	6	6	5	5	5	4	4			42.84	42.64	42.25	41.64	40.88	40.05	39.33	38.59	37.84	
c_w	5	6	5	5	4	4	4	4			50.7	49.55	48.45	47.27	46.09	44.94	43.93	42.94	41.98	
	6	6	6	5	5	4	4	4			48.36	47.49	46.61	45.63	44.6	43.55	42.59	41.68	40.79	
	7	6	6	5	5	4	4	4			46.01	45.49	44.78	43.99	43.1	42.17	41.25	40.43	39.61	
	8	6	6	5	5	5	4	4			43.66	43.49	42.95	42.35	41.61	40.79	39.94	39.17	38.43	
c_r	11	6	6	5	5	4	4	4			44.23	44.21	43.83	43.34	42.69	41.95	41.19	40.53	39.85	
	13	6	6	5	5	5	4	4			44.04	43.97	43.54	43.01	42.33	41.56	40.77	40.08	39.37	
	15	6	6	5	5	5	4	4			43.85	43.73	43.24	42.68	41.97	41.18	40.35	39.63	38.9	
	17	6	6	5	5	5	4	4			43.66	43.49	42.95	42.35	41.61	40.79	39.94	39.17	38.43	
μ_1	0.8	10	8	8	7	6	6	5			38.98	39.74	39.83	39.63	39.13	38.63	38.02	37.33	36.65	
	0.9	8	7	6	6	5	5	5			41.95	42.09	41.8	41.28	40.58	39.93	39.22	38.48	37.72	
	1	6	6	5	5	5	4	4			43.66	43.49	42.95	42.35	41.61	40.79	39.94	39.17	38.43	
	1.1	6	5	5	5	4	4	4			44.72	44.43	43.82	43.03	42.2	41.43	40.62	39.8	38.98	
μ_2	0.8	6	5	5	5	4	4	4			29.52	31.13	32.03	32.24	32.06	31.82	31.44	30.96	30.41	
	0.9	6	6	5	5	5	4	4			37.7	38.2	38.21	37.91	37.38	36.71	36.11	35.47	34.79	
	1	6	6	5	5	5	4	4			43.66	43.49	42.95	42.35	41.61	40.79	39.94	39.17	38.43	
	1.1	6	6	6	5	5	5	4			48.04	47.49	46.66	45.86	45.01	44.11	43.2	42.29	41.46	

optimal inventory level's uniform change pattern with respect to impatience degree (recall that the optimal inventory level S^* increases with the balking parameter M in each test example), the optimal profit has no uniform change pattern. The reason is similar to that in Section 5, e.g. although the increased customer impatience will lead to the increased order loss, it will reduce the waiting cost and the inventory holding cost at the same time. Thus, it is possible to increase profits instead. So, it's not necessarily better with lower customer impatience in the production service system.

From Figures 14 and 15 we can see that with an increased customer impatience in the renegeing behavior, the optimal inventory level will decline while the optimal profit has no uniform change pattern. The reason is consistent with that of no uniform change pattern of the optimal profit under the balking behavior, i.e. affected by economic parameters.

Table 2 shows the impact of renegeing behavior on optimal control under different economic and system parameters. The set of test examples is the same as that used in Table 1, except that $c_r = 17$ replaces $c_b = 15$. The impact of renegeing behavior on the optimal control, and the change patterns of optimal inventory level S^* and the optimal net profit NP^* under different economic and system parameters are consistent with the findings from Table 1.

7. Conclusions

This paper employs a classical production service system modelling service-oriented manufacturing systems. The production service system delivers customers with both material goods and intangible services. The most significant feature of production service systems is a combination of production and services as considered in this study. The system is modelled as a two-stage hybrid serial system. The problem modelling provides a basic method in the quantitative analysis of the production service system. Two impatient behaviors (balking and renegeing) appeared in waiting for products and service are considered. A Markov model is established to describe the random dynamics of the system. The impacts of customer impatience on system performance and optimal management are analysed.

The results show that customer impatience has a great impact on the system performance. The impacts of customer impatience on the changes of the inventory level and the number of customers waiting for service are opposite. The impatient behaviors have the different impacts on the two subsystems (i.e. production inventory subsystem and service subsystem) of the production service system. Specifically, customer impatience makes the improvement of one subsystem's performance along with the performance deterioration of the other subsystem. Therefore, if we pay close attention to the system performance indices, the system management will be more intractable when customers are impatient. Meanwhile, the different configurations of service and production capacity affect the influence extent of customer impatience on the system. In addition, the service in the production service system plays as a buffer on the impact of external product demand on the inventory. The lower the service capacity is, the smaller impact of external demand (or customer impatience) on the inventory will be.

The system net profit consists of revenue and costs related to all system performance indices. Therefore, net profit comprehensively reflects the system performance. With the profit to be of main interest for practitioners, the profit change pattern under the different capacity settings, unfortunately, is also different. That is, there is no uniform change pattern of the system profit when customer impatience changes. This conclusion indicates that for the production service system with the main feature of coexisting for inventory and orders waiting, the reduced customer impatience will not necessarily improve the firm profits, e.g. although the increased customer impatience will lead to the increased order loss, it will reduce the waiting cost and the inventory holding cost at the same time. Thus, it is possible to increase profits instead. In other words, it's not necessarily better with lower customer impatience in the production service system. There are two ways to deal with customers impatience: inducing customers impatience within a reasonable range and/or setting production and service capacity economically, i.e. matching supply with demand.

For the first one, the manager should value the degree of customers impatience accurately, then induce customers to be in a 'right' impatience degree, e.g. information about the anticipated wait is important to a customer in queue. Different levels of waiting information have different effects on customers' decision whether to balk (or renege) or not, and on the system. Therefore, in many industries, the manager provides delay information with different level of precision is a means to induce customers impatience degree (Guo and Zipkin 2007; Guo and Zipkin 2009). For the second one, the manager needs to have a more accurate understanding of customer impatience, and set the reasonable production and service capacity, making the costs and benefits change caused by customer impatience at a more reasonable range according to different levels of costs (such as waiting cost, holding cost, balking or renegeing cost), thereby making customers impatience becomes a factor for increasing firm profits by managing firm production and service capacity reasonably depending on the size of various costs.

Funding

The work described in this paper was supported by the Fundamental Research Funds for the Central Universities [15LZUJBWZY121], Research Grant from National Natural Science Foundation of China [71471114, 70932004].

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Allon, G., A. Federgruen, and M. Pierson. 2011. "How Much is a Reduction of Your Customers' Wait worth? An Empirical Study of the Fast-food Drive-thru Industry Based on Structural Estimation Methods." *Manufacturing & Service Operations Management* 13 (4): 489–507.
- Ancker, C., and A. Gafarian. 1963. "Some Queuing Problems with Balking and Reneging. I." *Operations Research* 11 (1): 88–100.
- Benjaafar, S., and M. Elhafsi. 2012. "A Production–inventory System with Both Patient and Impatient Demand Classes." *IEEE Transactions on Automation Science and Engineering* 9 (1): 148–159.
- Benjaafar, S., J.-P. Gayon, and S. Tepe. 2010. "Optimal Control of a Production–inventory System with Customer Impatience." *Operations Research Letters* 38 (4): 267–272.
- Berman, O., and E. Kim. 1999. "Stochastic Models for Inventory Management at Service Facilities." *Communications in Statistics. Stochastic Models* 15 (4): 695–718.
- Berman, O., and E. Kim. 2001. "Dynamic Order Replenishment Policy in Internet-based Supply Chains." *Mathematical Methods of Operations Research* 53 (3): 371–390.
- Berman, O., and K. P. Sapna. 2000. "Inventory Management at Service Facilities for Systems with Arbitrarily Distributed Service times." *Communications in Statistics. Stochastic Models* 16 (3–4): 343–360.
- Berman, O., and K. P. Sapna. 2002. "Optimal Service Rates of a Service Facility with Perishable Inventory Items." *Naval Research Logistics (NRL)* 49 (5): 464–482.
- Berman, O., E. H. Kaplan, and D. G. Shevishak. 1993. "Deterministic Approximations for Inventory Management at Service Facilities." *IIE Transactions* 25 (5): 98–104.
- Blackburn, J. 2012. "Valuing Time in Supply Chains: Establishing Limits of Time-based Competition." *Journal of Operations Management* 30 (5): 396–405.
- Carr, S., and I. Duenyas. 2000. "Optimal Admission Control and Sequencing in a Make-to-Stock/Make-to-Order Production System." *Operations Research* 48 (5): 709–720.
- Deepak, T., A. Krishnamoorthy, V. Narayanan, and K. Vineetha. 2008. "Inventory with Service Time and Transfer Of Customers and Inventory." *Annals of Operations Research* 160 (1): 191–213.
- Economopoulos, A. A., V. S. Kouikoglou, and E. Grigoroudis. 2011. "The Base Stock/Base Backlog Control Policy for a Make-to-Stock System with Impatient Customers." *IEEE Transactions on Automation Science and Engineering* 8 (1): 243–249.
- Guo, P., and P. Zipkin. 2007. "Analysis and Comparison of Queues with Different Levels of Delay Information." *Management Science* 53 (6): 962–970.
- Guo, P., and P. Zipkin. 2009. "The Effects of the Availability of Waiting-time Information on a Balking Queue." *European Journal of Operational Research* 198 (1): 199–209.
- Ha, A. Y. 1997. "Optimal Dynamic Scheduling Policy for a Make-to-Stock Production System." *Operations Research* 45 (1): 42–53.
- Hassin, R., and M. Haviv. 2003. *To Queue or Not to Queue: An Equilibrium Behavior in Queueing Systems*. Boston, MA: Kluwer Academic.
- He, Q.-M., and E. M. Jewkes. 2000. "Performance Measures of a Make-to-Order Inventory-production System." *IIE Transactions* 32 (5): 409–419.
- He, Q.-M., E. M. Jewkes, and J. Buzacott. 2002. "Optimal and Near-optimal Inventory Control Policies for a Make-to-Order Inventory-production System." *European Journal of Operational Research* 141 (1): 113–132.
- Hoel, P. G., S. C. Port, and C. J. Stone. 1972. *Introduction to Stochastic Processes*. Boston, MA: Houghton Mifflin.
- Ioannidis, S., O. Jouini, A. Economopoulos, and V. Kouikoglou. 2013. "Control Policies for Single-stage Production Systems with Perishable Inventory and Customer Impatience." *Annals of Operations Research* 209 (1): 115–138.
- Iravani, S. M. R., T. Liu, and D. Simchi-Levi. 2012. "Optimal Production and Admission Policies in Make-to-Stock/Make-to-Order Manufacturing Systems." *Production and Operations Management* 21 (2): 224–235.
- Krishnamoorthy, A., and V. C. Narayanan. 2011. "Production Inventory with Service Time and Vacation to the Server." *IMA Journal of Management Mathematics* 22 (1): 33–45.
- Krishnamoorthy, A., S. Nair, and V. Narayanan. 2012. "An Inventory Model with Server Interruptions and Retrials." *Operational Research* 12 (2): 151–171.
- Li, N., and Z. Jiang. 2013. "Modeling and Optimization of a Product-service System with Additional Service Capacity and Impatient Customers." *Computers & Operations Research* 40 (8): 1923–1937.

- Manuel, P., B. Sivakumar, and G. Arivarignan. 2007. "A Perishable Inventory System with Service Facilities, MAP Arrivals and PH – Service times." *Journal of Systems Science and Systems Engineering* 16 (1): 62–73.
- Rao, S. S. 1965. "Queueing Models with Balking, Reneging, and Interruptions." *Operations Research* 13 (4): 596–608.
- Robinson, L. W., and R. R. Chen. 2011. "Estimating the Implied Value of the Customer's Waiting Time." *Manufacturing & Service Operations Management* 13 (1): 53–57.
- San José, L. A., J. Sicilia, and J. Garcia-Laguna. 2005. "The Lot Size-reorder Level Inventory System with Customers Impatience Functions." *Computers & Industrial Engineering* 49 (3): 349–362.
- Schwarz, M., and H. Daduna. 2006. "Queueing Systems with Inventory Management with Random Lead times and with Backordering." *Mathematical Methods of Operations Research* 64 (3): 383–414.
- Schwarz, M., C. Sauer, H. Daduna, R. Kulik, and R. Szekli. 2006. "M/M/1 Queueing Systems with Inventory." *Queueing Systems* 54 (1): 55–78.
- Singer, G., and E. Khmelnitsky. 2010. "A Finite-horizon, Stochastic Optimal Control Policy for a Production–inventory System with Backlog-dependent Lost Sales." *IIE Transactions* 42 (12): 855–864.
- Stidham, S. 1985. "Optimal Control of Admission to a Queueing System." *IEEE Transactions on Automatic Control* 30 (8): 705–713.
- Toews, C., D. W. Pentico, and M. J. Drake. 2011. "The Deterministic EOQ and EPQ with Partial Backordering at a Rate that is Linearly Dependent on the Time to Delivery." *International Journal of Production Economics* 131 (2): 643–649.
- van Tits, M., and H. van der Veecken. 1980. "Simulation of a Queueing Problem with Balking." *Simulation* 35 (3): 88–93.
- Veatch, M. H., and L. M. Wein. 1994. "Optimal Control of a Two-station Tandem Production/Inventory System." *Operations Research* 42 (2): 337–350.
- Wang, K., N. Li, and Z. Jiang. 2010. "Queueing System with Impatient Customers: A Review." *Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics* 82–87.
- Wang, K., Z. Jiang, N. Li, and N. Geng. 2013. "Optimal Production and Admission Control for a Stochastic SOM System with Demands for Product and PSS." *International Journal of Production Research* 51 (23–24): 7270–7288.
- Wang, K., Z. Jiang, N. Li, and N. Geng. 2014. "Optimal Production Control of a Service-oriented Manufacturing System with Customer Balking Behavior." *Flexible Services and Manufacturing Journal* 26 (3): 387–407.
- Yadavalli, V. S. S., B. Sivakumar, and G. Arivarignan. 2008. "Inventory System with Renewal Demands at Service Facilities." *International Journal of Production Economics* 114 (1): 252–264.

Copyright of International Journal of Production Research is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.