Operations Research Letters ••• (••••) •••–•••

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Make waiting not to seem like waiting: Capacity management of waiting-area entertainment

Ke Sun [a], Yunan Liu [b,*], Xiang Li [a]

[a] School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China
[b] Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695, USA

ARTICLE INFO

Article history:
Received 12 April 2022
Received in revised form 27 October 2022
Accepted 31 October 2022
Available online xxxx

Keywords:
Queueing economics model
Psychology of waiting lines
Waiting-area entertainment
Strategic customers

ABSTRACT

To operationalize the psychological principle that "occupied time feels shorter than unoccupied time", it is common for service providers to offer entertainment options in the waiting areas. Typical examples that put in practice this mechanism include amusement parks, car dealers, airports, hospitals, restaurants, etc. In this paper, we study a queueing system where the server provides entertainment services to waiting customers. Assuming customers are strategic and delay-sensitive, we formulate a game-theoretical model and study customers' equilibrium behavior in response to this mechanism. Because offering waiting-area entertainment incurs extra operational costs, we discuss whether and when this option will benefit the service provider and obtain the optimal entertainment capacity that maximizes the system's profit. Our analysis reveals that this option is appealing if and only if the market size is intermediate and that the optimal capacity of the entertainment is a unimodal function in the market size. Our insights continue to remain valid when the service fee becomes endogenous.

1. Introduction

Waiting is frustrating and annoying. In service systems, customers' experience at the waiting lines can substantially influence their perception of the overall quality of the service. To reduce customers' waiting times, the majority of the extant queueing theory literature focuses on improving the management of service capacity and queueing discipline. Yet an equally important (albeit less explored) dimension in service operations puts in practice the psychological principles of the waiting line, which aims to reduce customers' perceived (not actual) waiting times. One of such principles is that *occupied time feels shorter than unoccupied time* [7].

According to [7], service providers can "fill up waiting times" by offering some activities that should either (a) be related to the subsequent service encounters, or (b) offer benefit in itself. Both cases have been predominantly operationalized in practice. As examples of Case (a), many restaurants hand out menus for waiting customers to peruse; Disney's amusement parks provide pre-ride video tutorials and instructions to waiting customers for the upcoming rides. As an example of Case (b), Haidilao, a popular hot pot restaurant chain, provides waiting customers with manicure and hand massage services, along with a variety of board games, snacks and drinks for free [2]. Similar waiting-line services have been adopted by many other restaurants (see Fig. 1 for an example of Chuanxi Bazi, another popular restaurant chain in China). For other examples, see [1,3].

Fig. 1. Waiting-area entertainment offered by Chuanxi Bazi, a popular restaurant chain in Sichuan, China.

* Corresponding author.
E-mail address: yliu48@ncsu.edu (Y. Liu).

https://doi.org/10.1016/j.orl.2022.10.015
0167-6377/

In this paper, we study the impact of *waiting-area entertainment* (WAE) on consumer behavior and system revenue by analyzing a queueing economics model. Of course, the right types of WAE depend on the context of the service; indeed, in different service organizations, WAE options vary from low-end services such as snacks and magazines, to high-end technologies such as computers and video game consoles, and to human services such as massaging, shoe shining and manicuring. Motivated by the above cited examples, we hereby consider two different mechanisms: (i) Type-1 WAE that requires human servers (e.g., manicurists and massagers), and (ii) Type-2 WAE that does not directly involve human resources (e.g., snacks, coffees and computers). Although the operational cost to maintain WAE activities apparently depends on the number of WAE "seats" (i.e., WAE's capacity), a major distinction of the two mechanisms lies in whether this cost also depends on the system state: in the former case the WAE cost incurs constantly regardless of how many customers are waiting in line (e.g., if 3 manicurists are hired then 3 salaries ought to be paid even when they may be idling), whereas in the latter case the cost is modulated by the system state (e.g., only 2 of 3 coffee stands will be used if there are only 2 waiting customers). Modeling assumptions and results of these two WAE mechanisms are detailed in later sections.

We characterize the customers' equilibrium behavior in response to both WAE mechanisms and establish the system-level performance functions such as throughput and profit. Our analysis reveals that WAE options can be used to improve the system's profit when the market size is intermediate and provide no benefits otherwise. In addition, the optimal number of WAE seats that maximizes the service provider's profit exhibits a unimodal shape in the market size. The above-mentioned phenomena are substantiated by both theoretical results and engineering confirmations via numerical examples. We also provide in-depth discussions to give additional insights. We also consider several extensions of the base model including the case of a random delay cost (to capture consumers' heterogeneous responses to WAE) and the case of endogenous pricing (to allow the service provider to jointly set the service fee and the WAE capacity).

Related literature. The queueing economics literature was pioneered by [8] where arriving customers decide on whether to join an $M/M/1$ queue based on the available queue length. Following [8], strategic customer behavior in queueing systems has been widely studied in the literature, see [5,4] for comprehensive reviews. Our work is also related to the small body of literature on WAE. A recent study by [9] investigated the co-opetition of multiple service providers in a service cluster where WAE is offered in a common space as a shared resource; they discovered that higher profitability may be achieved for all service providers by the choosing the right cost-allocation scheme that properly addresses an efficiency-fairness tradeoff. This stream of work has been extended by [6] with the additional consideration of how to optimally set the service capacity under WAE. The present work draws distinctions from the above-mentioned literature by focusing on the capacity sizing aspect of the management of the waiting-line entertainment. We aim to inform the service providers of whether and when this option can be an effective measure in terms of improving the system performance, and if yes, how many seats are to be invested and maintained.

2. Model descriptions

We model a service system as a single-server queue having customer arrives according to a Poisson process with rate Λ (also referred to the potential market size) and *independent and identically distributed* (I.I.D.) service times following an exponential distribution with rate μ (also referred to as the service capacity). In addition, the arrival and service processes are assumed to be mutually independent. Let $\rho \equiv \Lambda/\mu$ be the system's workload.

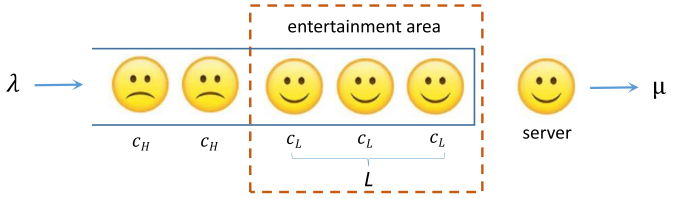

Fig. 2. A queueing model with waiting-area entertainment.

Strategic customers. Customers are delay-sensitive and demand service as soon as possible. Each customer receives a reward V after completing service and pays a fee P upon arriving at the service system. Customers are strategic and make their *joining* and *balking* decisions based on their ex ante utilities. All joining customers finding a busy server must wait in the queueing area and will be served according to the *first-come first-served* (FCFS) discipline. They incur a waiting cost c_H per time unit in the waiting area. Unlike the conventional queueing economics model in [8] where the waiting cost is generated throughout a customer's entire sojourn time, we hereby assume there is no waiting cost when a customer is already in service.

Waiting-area entertainment. To mitigate customers' waiting costs, the service provider offers WAE with a total capacity L. For example, L may mean the number of massaging chairs or the number of seats for manicure services. If there are less than L customers in the waiting area, all customers benefit from WAE; otherwise, customers receive the WAE services according to FCFS (that is, WAE is provided to the first L customers in the waiting line). See Fig. 2 for an illustration. We assume that waiting customers currently receiving the WAE service incur a waiting cost c_L per time unit, with $0 \le c_L < c_H$.

We assume that customers have the queue length information upon their arrivals, so that they are able to make their joining/balking decisions based on the knowledge of the total number of customers in the system including those in service, in the entertainment area, and outside of the entertainment area (if any).

Motivated from different types of WAE, we consider two mechanisms.

- **Inflexible (Type-1) WAE:** If the WAE requires extra human servers such as manicurists and massagers, the cost of maintaining the service is due to their hiring salary. Hence, the service provider incurs a cost at a fixed rate LK, with $K > 0$ representing the salary per time unit the service provider commits to each WAE server. This WAE cost is *inflexible* and independent of the number of waiting customers in the queue.
- **Flexible (Type-2) WAE:** The second kind of WAE does not directly involve any human resource. For example, if the service provider maintains a service area with L seats offering snacks and coffees, K means the cost per time unit these products are consumed. If the entertainment means L massaging chairs, K refers to the maintenance cost per chair per time unit. Whenever there are n customers waiting in line, the service provider incurs a cost $K \cdot \min(n, L)$ per time unit. This WAE cost is thus *flexible* and is adaptively coping with the system's state.

We treat the above-introduced inflexible (Type-1) WAE case as our base WAE model. In Section 3 we will carefully analyze the optimal capacity for the base WAE model that maximizes the system's profit. In Section 4 we will extend the results of the base

Table 1
Glossary of main notation.

Symbol	Definition
μ	Service capacity/rate
Λ	Market size
$\rho \equiv \Lambda/\mu$	System's workload
V	Service reward
c_L, c_H	Delay costs with and without WAE
p_L	Probability a customers enjoys WAE in an extension model
P	Service fee
L	WAE capacity
K	WAE cost of maintaining one seat per unit time
$\pi_n^0, \pi_n(L)$	Steady-state probabilities for models without and with WAE
$n_e^0, n_e(L)$	Joining thresholds without and with WAE
TH^0, TH	System throughput in models without and with WAE
Π^0, Π^1, Π^2	Net profit without WAE, with type-1 WAE, and type-2 WAE
$\mathbb{E}[X]$, $\mathbb{P}(A)$	Expectation of random variable X and probability of event A

model in three directions. First, in Section 4.1 we study the performance of the flexible (Type-2) WAE model which exhibits similar structure to the base WAE model. Second, in Section 4.2 we allow customers' delay cost to be heterogeneous. Finally, in Section 4.3 we consider a joint pricing and capacity sizing problem for the base WAE model. All notations are summarized in Table 1.

3. The base WAE model

We first review the case without WAE (with $L = 0$) which is referred to as the *no-entertainment* model. We next study the customer strategy and system performance in the base (Type-1) WAE Model.

3.1. The no-entertainment model

The no-entertainment model is a special case of the entertainment model (of either type) with $L = 0$. Apparently, it reduces to a standard $M/M/1$ queue with potential arrival rate Λ, service rate μ, and waiting cost c_H. (A subtle distinction from the classical Naor model [8] is that customers do not incur any waiting cost after entering service.) It is well known that when the queue length is observable, customers' joining decision follows a threshold-type strategy. That is, potential customers join the system only when the number of existing customers is below a certain threshold n_e^0; otherwise they balk.

We next provide the key performance functions of the no-entertainment case which will later be used as useful benchmarks for the entertainment models. We hereby append a superscript "0" to all notation. Specifically, let π_i^0, TH^0 and Π^0 denote the system's steady-state probability, throughput and profit, respectively.

Lemma 3.1 *(The no-entertainment case). In an $M/M/1$ queue without WAE, key performance measures are given as follows:*

$$n_e^0 = \left\lfloor \frac{\mu(V-P)}{c_H} \right\rfloor + 1, \qquad \pi_i^0 = \frac{\rho^i(1-\rho)}{1-\rho^{n_e^0+1}}, \quad i = 0, \ldots, n_e^0, \tag{1}$$

$$TH^0 = \sum_{n=0}^{n_e^0-1} \pi_i^0 \equiv \Phi_{n_e^0}(\rho), \qquad \Pi^0 = \Lambda \sum_{n=0}^{n_e^0-1} \pi_i^0 P \equiv \Phi_{n_e^0}(\rho) P,$$
$$\text{where} \quad \Phi_n(\rho) \equiv \mu \frac{\rho(1-\rho^n)}{1-\rho^{n+1}}. \tag{2}$$

3.2. The Type-1 (inflexible) WAE model

In this subsection, we study the model under type-1 WAE with capacity L, and we characterize the system performance in equilibrium.

We first describe the expected utility function of a "tagged" customer. For a fixed L, let $U(n)$ denote her expected utility for joining the system given that she observes n existing customers already in the system (excluding herself). Thus,

$$U(n) = \bar{U}(n)\mathbf{1}_{\{n \le L\}} + \widetilde{U}(n)\mathbf{1}_{\{n > L\}}, \tag{3}$$

where the indicator function $\mathbf{1}_A$ is 1 if condition A holds and 0 otherwise, and the two functions $\bar{U}(n)$ and $\widetilde{U}(n)$ are given by

$$\bar{U}(n) \equiv V - P - \frac{nc_L}{\mu}, \tag{4}$$

$$\widetilde{U}(n) \equiv V - P - \frac{(n-L)c_H}{\mu} - \frac{Lc_L}{\mu}. \tag{5}$$

Next, we compute the main performance features. With a given WAE capacity L, let $n_e(L)$, $\pi_n(L)$, $TH(L)$ and $\Pi^1(L)$ be the customers' joining threshold, system's steady-state probability, throughput and net profit (i.e., gross profit minus WAE cost). For this type-1 WAE, we append a superscript "1" only to the profit function $\Pi^1(L)$ but not to the other notation, because as we will see soon, type-1 and type-2 models coincide in all performance functions except for the system's net profit.

In parallel to (1), we define

$$\bar{n}_e \equiv \left\lfloor \frac{\mu(V-P)}{c_L} \right\rfloor + 1, \tag{6}$$

which is the joining threshold of an $M/M/1$ queue having a delay cost c_L. Throughout the paper, we assume that $n_e^0 < \bar{n}_e$ (i.e., c_H and c_L are not too close) so that the adoption of WAE can be beneficial.

Lemma 3.2 *(Type-1 WAE: Performance under a fixed L). Consider an $M/M/1$ model under type-1 WAE with a fixed L. We have*

$$n_e(L) = \begin{cases} \left\lfloor \frac{\mu(V-P)+L(c_H-c_L)}{c_H} \right\rfloor + 1, & \text{if} \quad L \le \bar{n}_e, \\ \bar{n}_e, & \text{if} \quad L > \bar{n}_e, \end{cases} \tag{7}$$

$$\pi_n(L) = \frac{\rho^n(1-\rho)}{1-\rho^{n_e(L)+1}}, \tag{8}$$

$$TH(L) = \Lambda \sum_{n=0}^{n_e(L)-1} \pi_n(L) = \Phi_{n_e(L)}(\rho), \tag{9}$$

$$\Pi^1(L) = \Phi_{n_e(L)}(\rho) P - KL, \tag{10}$$

where the function $\Phi_n(\cdot)$ is given in (2).

Using results in Lemma 3.2 with a fixed L, we next discuss how to select the optimal WAE capacity L_1^* that maximizes the system's net profit in (10). Although L is allowed to be any positive integer, to facilitate the theoretical analysis, we first truncate the support of L to a finite set in form of

$$\mathcal{D}_L \equiv \{\underline{L}, \underline{L}+1, \ldots, \bar{L}\} \qquad \text{for some} \quad 0 < \underline{L} \le \bar{L} < \infty, \tag{11}$$

without affecting the optimal profit. To see this, note that when L is already sufficiently large, further increasing L will have no impact on $n_e(L)$ or system-level performance but will incur a bigger WAE capacity cost (see Lemma 3.2). We call $\mathcal{D}_L$ the effective region of L.

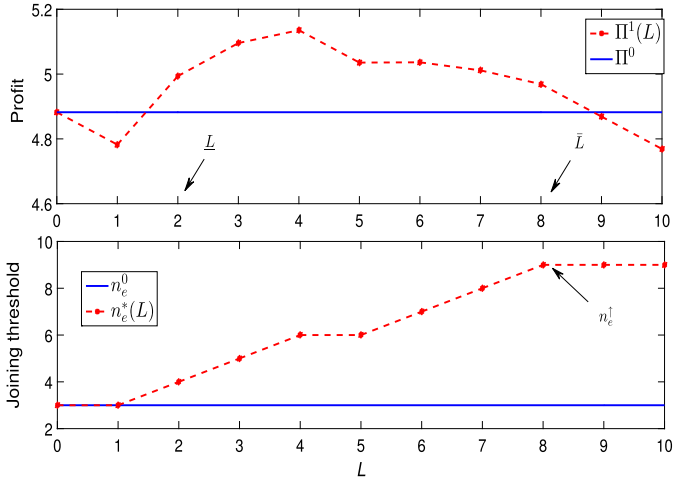

Fig. 3. Profit and customers' joining threshold under type-1 WAE, with $\Lambda = 1.2$, $\mu = 1$, $V = 10$, $c_L = 0.5$, $c_H = 2$, $P = 6$, $K = 0.1$, $\underline{L} = 2$, $\bar{L} = 8$, $n_e^0 = 3$ and $\bar{n}_e = 9$.

Lemma 3.3 *(Effective region of L). The optimal WAE capacity L_1^* is contained in the effective region $\mathcal{D}_L$ in (11) specified by $\underline{L}$ and $\bar{L}$, where*

$$\underline{L} \equiv \min\{L > 0 : n_e(L) = n_e^0 + 1\} \text{ and } \bar{L} \equiv \min\{L > 0 : n_e(L) = \bar{n}_e\}.$$

We supplement Lemma 3.3 using an example; in Fig. 3 we graph both the profit $\Pi^1(L)$ and the threshold $n_e(L)$ as L increases. We make the following observations: First, consistent with Lemma 3.3, customers' joining threshold $n_e(L)$ is nondecreasing in L and eventually capped at $\bar{n}_e$. Second, there exists an effective region (with $\underline{L} = 2 < 8 = \bar{L}$) in which the system is able to achieve improved profit relative to the no-entertainment model. (When $L = 1 < \underline{L}$, WAE is ineffective in improving the system's throughput because $n_e(L) = n_e^0$; when $L = 8 > \bar{L}$, the system reaches a saturation point with $n_e(L) = \bar{n}_e$.) Last, $\Pi^1(L)$ is unimodal in L within the effective region (for which we will provide theoretical justifications in what follows).

We are ready to characterize the optimal WAE capacity L_1^* that maximizes the system's net profit. In what follows, we consider two cases specified by the value of a constant

$$\bar{K} \equiv \sup_{0<\Lambda<\infty} P[\Phi_{n_e(\underline{L}+1)}(\Lambda/\mu) - \Phi_{n_e(\underline{L})}(\Lambda/\mu)], \tag{12}$$

where $\Phi_n(\cdot)$ is defined in (2). The next result describes the effectiveness of WAE and contrasts its performance to the case of no entertainment.

Theorem 3.1 *(Type-1 WAE: Profit-optimal capacity). There are two cases specified by $\bar{K}$ in (12) and the WAE cost K:*

(i) When $K \geq \bar{K}$, the optimal WAE capacity $L_1^*(\Lambda) = 0$ so that for all market size Λ, $\Pi^1(L_1^*(\Lambda)) = \Pi^0$.

(ii) When $K < \bar{K}$, there exists two thresholds of market size $\underline{\Lambda}_1$ and $\bar{\Lambda}_1$ with $\underline{\Lambda}_1 < \bar{\Lambda}_1$ being the two roots of the equation

$$\Delta_L(\Lambda) \equiv P[\Phi_{n_e(\underline{L})}(\Lambda/\mu) - \Phi_{n_e^0}(\Lambda/\mu)] - K\underline{L} = 0. \tag{13}$$

We further consider two subcases of the market size:

a. ***Small or large**: when $\Lambda \in [0, \underline{\Lambda}_1] \cup [\bar{\Lambda}_1, \infty)$, the optimal entertainment level $L_1^*(\Lambda) = 0$.*

b. ***Intermediate**: when $\Lambda \in (\underline{\Lambda}_1, \bar{\Lambda}_1)$, there exists a unique optimal entertainment capacity $0 < L_1^*(\Lambda) < \infty$ that guarantees that $\Pi^1(L_1^*(\Lambda)) > \Pi^0$, with*

$$L_1^*(\Lambda) = \min\left\{L > 0 : \Phi_{n_e(L+1)}\left(\frac{\Lambda}{\mu}\right) - \Phi_{n_e(L)}\left(\frac{\Lambda}{\mu}\right) < \frac{K}{P}\right\}.$$

In addition, $L_1^*(\Lambda)$ is first nondecreasing in Λ and then nonincreasing in Λ.

Remark 3.1. First, it is straightforward to see that WAE is ineffective if its capacity cost K is too large. When K is not too large, WAE can always help reduce the customers' waiting cost, which in turn boosts the system throughput. However, whether the profit gain from the incremented throughput can outweight the WAE cost (so that the overall profit can improve) will largely depend on the market size Λ. When Λ is sufficiently small, almost all customers will join for service because they rarely see a long queue; in this case WAE becomes less cost-effective for the system. Hence, the service provider should consider setting a small WAE capacity (if at all). As Λ increases, the waiting queue becomes longer which drives the service provider to expand the WAE capacity. However, when Λ becomes sufficiently large, the system becomes saturated with the throughput approaching the service capacity μ. In this case, it again becomes unworthy for the service provider to maintain any WAE activities (there is no room to further improve the system throughput).

To visualize our theoretical results, we next consider an example for which we numerically compute the profit, throughput, optimal WAE capacity and customers' joining threshold in the two models: no entertainment and base WAE; see Fig. 4. Consistent with Theorem 3.1, the optimal WAE capacity $L_1^*(\Lambda)$ is strictly positive only when Λ is not too large or too small, and it first increases and then decreases in Λ. So offering WAE is effective in boosting the system's profit only when the market size is intermediate. Moreover, in contrast to the case of no entertainment, the throughput under WAE is not always increasing in Λ, because the WAE capacity decreases when Λ is large due to its reduced cost effectiveness.

4. Extensions

4.1. The Type-2 (flexible) WAE model

We first extend our analysis to the flexible (type-2) WAE case. Unlike the base WAE model that continuously cost the service provider KL per time unit regardless of the number of the waiting customers, the capacity cost of type-2 WAE is dependent with the system state. In particular, let N_L be the steady-state number of customers waiting in line with a given WAE capacity L, the steady-state WAE cost is $K\mathbb{E}[L \wedge N_L]$ per time unit, where $x \wedge y \equiv \min(x, y)$. Hence, we modify (10) as below and append a superscript "2":

$$\Pi^2(L) = \Lambda \sum_{n=0}^{n_e(L)-1} \pi_n(L) P - K\mathbb{E}[L \wedge N_L], \tag{14}$$

Nevertheless, our new setting here has no impact on $n_e(L)$, $\pi_n(L)$ and $TH(L)$ (so their formulas remain the same as in (7)–(9)), but only on the system's net profit. In parallel to Theorem 3.1, we next study the performance of a type-2 WAE model under the profit-optimal L_2^* (the L that maximizes (14)). Our result exhibits similar structure and insights to that of the base WAE model.

Theorem 4.1 *(Type-2 WAE: Profit-optimal capacity). There exists a constant $\hat{K}$, such that:*

(i) When $K \geq \hat{K}$, the optimal WAE capacity $L_2^*(\Lambda) = 0$ so that for all $\Lambda > 0$, $\Pi^2(L_2^*(\Lambda)) = \Pi^0$;

(ii) When $K < \hat{K}$, there exists $\underline{\Lambda}_2$ and $\bar{\Lambda}_2$ with $\underline{\Lambda}_2 < \bar{\Lambda}_2$ such that,

- *If $\Lambda \in [0, \underline{\Lambda}_2] \cup [\bar{\Lambda}_2, \infty)$, the optimal WAE capacity $L_2^* = 0$, so that $\Pi^2(L_2^*(\Lambda)) = \Pi^0$;*

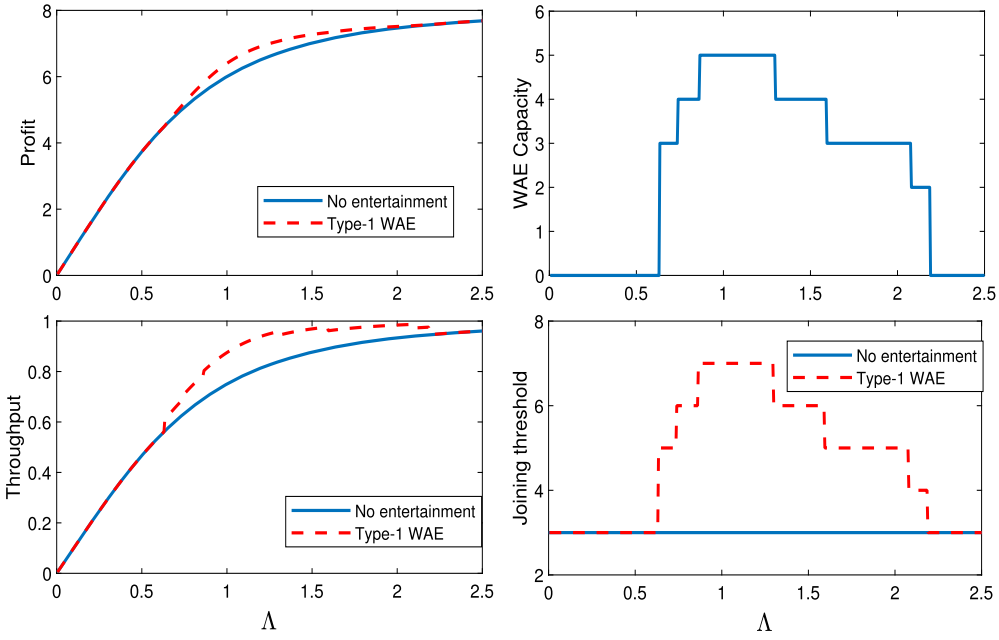

Fig. 4. Performance comparison of the two WAE models, with $\mu = 1$, $V = 12$, $c_L = 0.2$, $c_H = 2$, $P = 8$, $K = 0.12$, $\underline{\Lambda}_1 = 0.63$, $\bar{\Lambda}_1 = 2.19$.

– *If* $\Lambda \in (\underline{\Lambda}_2, \bar{\Lambda}_2)$, *the optimal WAE capacity* $L_2^*(\Lambda) > 0$ *and* $\Pi^2(L_2^*(\Lambda)) > \Pi^0$.

4.2. Heterogeneous delay cost

Although WAE can in general help reduce the perceived waiting times, not all customers will fully appreciate the WAE activities. To capture customers' heterogeneous responses to WAE, we now allow the waiting cost under WAE to be a random variable, denoted by the capital letter C_L (with the pre-WAE cost c_H remaining a deterministic number). For simplicity, we assume that C_L follows a two-point distribution with $\mathbb{P}(C_L = c_L) = 1 - \mathbb{P}(C_L = c_H) = p_L \in [0, 1]$, and $c_L < c_H$. Here event $\{C_L = c_L\}$ (event $\{C_L = c_H\}$) means that a customer enjoins (does not enjoin) the WAE activity. This setting divides all customers into two categories with the parameter p_L denoting the fraction of customers who benefit from WAE. (The delay cost reduces to that in the base WAE model when $p_L = 1$ and to that in the no-entertainment model when $p_L = 0$.)

For a fixed L, a customer having delay cost $C_L = c$ and observing a queue length n has utility

$$U(n|L, c) = V - P - \frac{(n - L)^+ c_H}{\mu} - \frac{(n \wedge L)c}{\mu}, \; c = c_L, c_H, \qquad (15)$$

and thus adopts a threshold joining strategy with a c-dependent threshold

$$n_e(L, c) = \max\{n : U(n|L, c) > 0\}. \qquad (16)$$

This threshold reduces to n_e^0 in (1) for customers having $C_L = c_H$ and to $n_e(L)$ in (7) for those having $C_L = c_L$. Hence, the system dynamics follows a birth-and-death process with birth rates $\lambda_i = \Lambda$ if $i < n_e^0$ and $\lambda_i = \Lambda p_L$ when $n_e^0 \le i < n_e(L)$, and death rates $\mu_i = \mu$.

The steady-state probabilities, system throughput, and partial structure of the WAE capacity are summarized in the proposition below.

Proposition 4.2 *(Heterogeneous delay cost under WAE). Consider the case of the two-point distributed delay cost under WAE:*

i. *For a given L, the system's steady-state probabilities are*

$$\widehat{\pi}_i(L) = \begin{cases} \widehat{\pi}_0(L)\rho^i, & \text{if} \quad 0 \le i < n_e^0 \\ \widehat{\pi}_0(L)\rho^{n_e^0}(p_L\rho)^{i-n_e^0}, & \text{if} \quad n_e^0 \le i \le n_e(L), \end{cases}$$

$$\widehat{\pi}_0(L) = \left(\frac{1 - \rho^{n_e^0}}{1 - \rho} + \rho^{n_e^0} \frac{1 - (p_L\rho)^{n_e(L) - n_e^0 + 1}}{1 - p_L\rho} \right)^{-1}.$$

The system's throughput is

$$\widehat{TH}(L) = \mu(1 - \widehat{\pi}_0(L)).$$

ii. *WAE is ineffective in improving the service provider's profit when the market size* Λ *is sufficiently small or large.*

In Fig. 5 we numerically compare the profits of a type-1 WAE model with homogeneous delay cost and heterogeneous delay cost. Unsurprisingly, the profit of the model under a heterogeneous delay cost is lower than that of the base WAE model, because WAE now becomes less effective in reducing customers' perceived waiting costs and improving the system throughput. Nevertheless, similar to the case of homogeneous costs, we see that WAE has benefits only when the market size is intermediate.

4.3. Joint pricing and capacity sizing

Finally, in our base WAE model, we allow the service provider to maximize the system's profit defined in (10) by jointly setting (i) the WAE capacity L and (ii) the service fee P. To understand the impact of WAE, we benchmark with the no-entertainment model where the service provider uses the service fee as the only profit-maximizing lever. Because the analysis of queueing economics models under optimal pricing is quite involved and in general admits no analytic solutions, we will gain qualitative insights via numerical experiments.

In Fig. 6 we compare the profits in the two models (top left panel) under their optimal prices (top right panel). Consistent with the case of exogenous price (Theorem 3.1), our results show that WAE can help improve the system's profit only when Λ is intermediate. Specifically, in this example, when Λ is sufficiently small

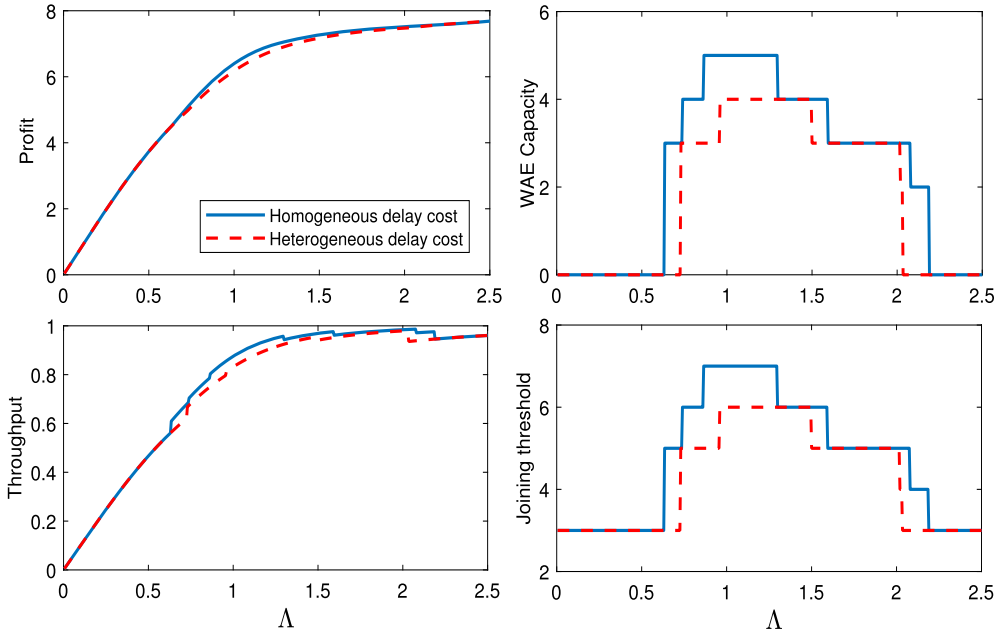

Fig. 5. Profit comparison of type-1 WAE with homogeneous delay cost and heterogeneous delay cost, with $\mu = 1$, $V = 12$, $c_L = 0.2$, $c_H = 2$, $K = 0.12$, $\underline{L} = 2$, $p_L = 0.8$, $\bar{L} = 8$, $n_e^0 = 3$ and $\tilde{n}_e = 9$.

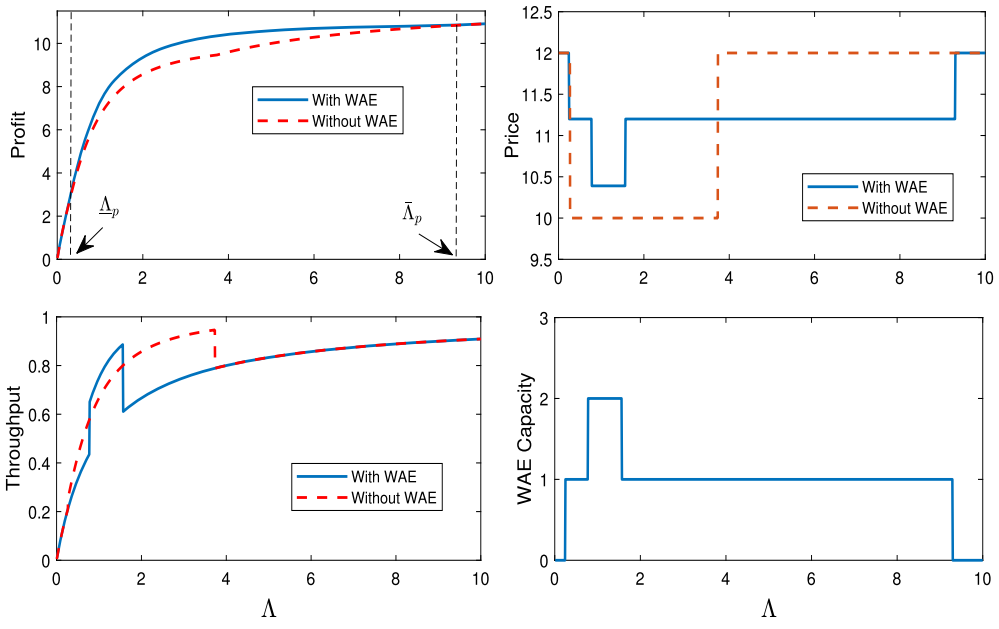

Fig. 6. Profit comparison under optimal service price for (a) Type-1 WAE model and (b) no-entertainment model, with $\mu = 1$, $V = 12$, $c_L = 0.8$, $c_H = 2$, $K = 0.25$.

(i.e., $\Lambda < \underline{\Lambda}_p = 0.24$) or large (i.e., $\Lambda > \bar{\Lambda}_p = 9.3$), the optimal WAE capacity $L_1^* = 0$ (and of course both the optimal prices and profits of the two models are identical). Also, when P is endogenous, the throughput is no longer monotone in Λ (even in the homogeneous-cost case, see the bottom left panel). In addition, we observe that, as Λ varies, the service fee and WAE capacity always work together by *moving to the same direction* (bottom panel): P decreases (increases) whenever L increases (decreases), aiming to entice (reject) more customer arrivals.

5. Conclusions

Motivated by the novel practice of WAE in service systems, we develop a queueing economics model to explore the impact of this mechanism on customer behavior and system revenue. We determine the optimal service capacity for the WAE by considering two mechanisms: inflexible WAE that involves human servers and flexible WAE that involves no human servers, both of which are relevant to various applications in practice. Our results reveal that, when WAE capacity cost is not too high, offering such a service to customers in the waiting area can help improve the system' net profit only when the system's load is intermediate, and in addition, the profit-optimal WAE capacity is unimodal in the system's load. We also consider two extensions including (i) random delay costs that account for customers' heterogeneous responses to WAE activities, and (ii) endogenous pricing that allows the service provider to jointly set the service fee and WAE capacity.

There are several venues for future research. One interesting direction is examine the mechanism where WAE is offered to

customers with a fee, so customers may choose to pay for the activities if they deem them to be worthy. A notable motivating example of this model is the airport VIP club. Another dimension is to investigate the *level* of WAE; for example, providing high-end WAE activity (e.g., game consoles) can apparently achieve a bigger reduction of the perceived waiting time as opposed to low-end options (e.g., magazines), but this will incur a higher cost for the service provider.

Acknowledgements

The authors would like to thank editors and referees for their constructive comments. This work was supported by NSFC (No. 71931001), and the Funds for First-class Discipline Construction (XK1802-5).

Appendix A

Proof of Lemmas 3.1 and 3.2. All results of Lemma 3.1 directly follow from [8] with a slight modification that the a customer in service incurs no waiting cost. To prove Lemma 3.2, first note that when L is large enough, all waiting customers incur the lower cost c_L, and each, when seeing n customers, has an expected utility $\bar{U}(n)$ as specified in equation (4).

The two thresholds $\bar{n}_e$ and $n_e(L)$ are the largest integers such that $\bar{U}(n) = 0$ and $U(n) = 0$, where $U(n)$ is given in (3). However, an $M/M/1$ model with customers following the joining threshold $n_e(L)$ is similar to the Naor model, with the first $n_e(L) \wedge L$ customers incurring a cost c_L and the next $(n_e(L) - L)^+$ incurring c_H. The formulas for the steady state probabilities, throughput and profit function in (8)-(10) naturally follow. □

Proof of Lemma 3.3. Due to the discrete nature of $n_e(L)$, a very small L will only increase the WAE cost (at a rate KL) without expending customers' joining threshold (with $n_e(L) = n_e^0$). Therefore, any $L < \underline{L}$ cannot be optimal. On the other hand, when L is already sufficiently large with $n_e(L)$ reaching its maximum value $\bar{n}_e$, investing any additional entertainment capacity will only increase the WAE cost which is detrimental to the profit. Hence, any $L > \bar{n}_e$ cannot be optimal. □

Proof of Theorem 3.1. We first investigate the monotonicity of the profit function $\Pi^1(L)$ (defined in (10)) with respect to L. For now we stipulate that the workload ρ and Λ remain fixed. Because μ is fixed, in what follows we work with ρ instead of Λ for the ease of notation. Note that

$$\Pi^1(L) - \Pi^1(L-1) \geq 0 \quad \Leftrightarrow \quad \Phi_{n_e(L)}(\rho) - \Phi_{n_e(L-1)}(\rho) \geq \frac{K}{P}, \tag{1}$$

where $\Phi_n(\cdot)$ is defined in (2). Taking the second-order derivative of $\Phi_{n_e(L)}(\rho)$ with respect to L (by treating L as a continuous variable) yields

$$\begin{aligned}\frac{\partial^2 \Phi_{n_e(L)}(\rho)}{\partial L^2} &= \frac{\partial^2 \Phi_{n_e(L)}(\rho)}{\partial n_e^2(L)} \cdot \frac{\partial n_e(L)}{\partial L} \\ &= \frac{(\rho-1)\rho^{n_e(L)}(\log\rho)^2(1+\rho^{n_e(L)+1})}{(1-\rho^{n_e(L)+1})^3} \cdot \frac{\partial n_e(L)}{\partial L} \leq 0,\end{aligned}$$

where the inequality holds because $n_e(L)$ is nondecreasing in L. Hence, we know that $\Phi_{n_e(L)}(\rho) - \Phi_{n_e(L-1)}(\rho)$ is nonincreasing in L. Next, we consider the following two cases:

(i) $K > P[\Phi_{n_e(\underline{L}+1)}(\rho) - \Phi_{n_e(\underline{L})}(\rho)]$ and

(ii) $K \leq P[\Phi_{n_e(\underline{L}+1)}(\rho) - \Phi_{n_e(\underline{L})}(\rho)]$.

In case (i) we must have that $\Phi_{n_e(L)}(\rho) - \Phi_{n_e(L-1)}(\rho) < K/P$ for all feasible $L \in \{\underline{L}, \cdots, \bar{L}\}$, which is the opposite of (1). Therefore, $\Pi^1(L)$ is decreasing in L so that it is optimal for the service provider set $L_1^* = 0$.

In case (ii), there exists a unique $L_1^* \geq 0$ such that

$$\Phi_{n_e(L_1^*+1)}(\rho) - \Phi_{n_e(L_1^*)}(\rho) < K/P \leq \Phi_{n_e(L_1^*)}(\rho) - \Phi_{n_e(L_1^*-1)}(\rho). \tag{2}$$

Thus the profit function $\Pi^1(L)$ first increases in $L \in \{\underline{L}, \cdots L_1^*\}$ and then decreases in $L \in \{L_1^*+1, \cdots, \bar{L}\}$. Therefore, for a given ρ, the maximum profit is achieved at the optimal entertainment capacity

$$L_1^* = \min\{L > 0 : \Phi_{n_e(L+1)}(\rho) - \Phi_{n_e(L)}(\rho) < K/P\}.$$

Next, we further establish the impact of system offered load ρ in case (ii). For $L > 0$, define

$$\Delta_L(\rho) \equiv \Pi_\rho^1(L) - \Pi_\rho^0 = P(\Phi_{n_e(L)}(\rho) - \Phi_{n_e^0}(\rho)) - KL,$$

where we have appended a subscript ρ to both $\Pi^1(L)$ and Π^0 in order to highlight the dependence on ρ. Taking the first-order and second-order derivatives with respect with ρ, we have that, for all $\rho > 0$,

$$\begin{aligned}&\frac{\partial \Phi_n(\rho)}{\partial \rho} = \mu \frac{(1-\rho)(1+\rho+\cdots+\rho^{n-1}-n\rho^n)}{(1-\rho^{n+1})^2} > 0, \\ &\frac{\partial \Phi_n^2(\rho)}{\partial \rho^2} = (1-\rho)\rho^{n-1}(n+1)\times \\ &\frac{[(\rho-\rho^{n+1})+\cdots+(\rho^n-\rho^{n+1})]+[(\rho-1)+\cdots+(\rho^n-1)]}{(1-\rho^{n+1})^3} < 0.\end{aligned}$$

In addition, we know that $\frac{\partial \Phi_n^2(\rho)}{\partial \rho^2}$ is decreasing in n because the numerator $[(\rho-\rho^{n+1})+\cdots+(\rho^n-\rho^{n+1})]+[(\rho-1)+\cdots+(\rho^n-1)]\rho^{n-1}(n+1) < 0$ and decreasing in n, and $(1-\rho)/(1-\rho^{n+1})^3 > 0$ increases in n. Therefore, $\partial \Phi^2_{n_e(L)}(\rho)/\partial \rho^2 \leq \partial \Phi^2_{n_e^0}(\rho)/\partial \rho^2 < 0$ due to $n_e(L) \geq n_e^0$ for any $L \geq 0$, which further implies that $\Delta_L(\rho)$ is concave in ρ. In addition, we focus on the two cases with $\rho \to 0$ and $\rho \to \infty$.

(ii.a) When the system load is sufficiently small, the optimal entertainment capacity $L_1^* = 0$, because $\lim_{\rho\to 0} \Pi_\rho^1(L) = -KL < 0 = \lim_{\rho\to 0} \Pi_\rho^0$.

(ii.b) When the system load is sufficiently large, the optimal entertainment capacity $L_1^* = 0$, because $\lim_{\rho\to\infty} \Pi_\rho^1(L) = \mu P - KL < \mu P = \lim_{\rho\to\infty} \Pi_\rho^0$.

Due to the fact that $\Delta_L(\rho)$ is concave in ρ, there exists two thresholds on the system load (market size), a lower bound $\underline{\rho}_1$ ($\underline{\Lambda}_1$) and an upper bound $\bar{\rho}_1$ ($\bar{\Lambda}_1$) that are the two roots of equation (13), such that $\Delta_L(\rho) > 0$ when $\rho \in (\underline{\rho}_1, \bar{\rho}_1)$. In this case there exists a unique $L_1^* > 0$ that guarantees that $\Pi^1(L_1^*) > \Pi^0$.

Finally, we establish the monotonicity of the optimal entertainment capacity $L_1^*(\Lambda)$ in case (ii). A continuous version of $L_1^*(\Lambda)$ can be determined by the first-order condition $\frac{\partial \Pi_\rho^1(L)}{\partial L} = 0$ with $\rho \in (\underline{\rho}_1, \bar{\rho}_1)$, which is equivalent to

$$\frac{\partial \Phi_{n_e(L)}(\rho)}{\partial n_e(L)} \cdot \frac{\partial n_e(L)}{\partial L} = \frac{K}{P} \quad \Leftrightarrow \quad \frac{\partial \Phi_{n_e(L)}(\rho)}{\partial n_e(L)} = \frac{K}{P} \cdot \frac{\partial L}{\partial n_e(L)}. \tag{3}$$

In addition, we have proved that $\frac{\partial \Phi_n^2(\rho)}{\partial \rho^2}$ is decreasing in n, which indicates that $\frac{\partial \Phi_{n_e(L)}(\rho)}{\partial n_e(L)}$ is concave in ρ. With the system load ρ

increases, for a given $n_e(L)$, $\frac{\partial \Phi_{n_e(L)}(\rho)}{\partial n_e(L)}$ is first increasing and then decreasing. Furthermore,

$$\frac{\partial^2 \Phi_n(\rho)}{\partial n^2} = \frac{\mu(\rho-1)(\log\rho)^2\rho^{n+1}(1+\rho^{n+1})}{(1-\rho^{n+1})^3} < 0,$$
$$\text{for all} \quad \rho > 0,$$

which means that $\frac{\partial \Phi_{n_e(L)}(\rho)}{\partial n_e(L)}$ is decreasing in $n_e(L)$. Note that the right-hand side of (3) is a constant, so customers' joining threshold must be first nondecreasing and then nonincreasing in Λ. The same property holds for the optimal entertainment capacity $L_1^*(\Lambda)$ (because these two quantities have a nondecreasing relationship). □

Proof of Theorem 4.1. Similar to the proof of Theorem 3.1, we define

$$\begin{aligned}\Delta^2(L) &\equiv \Pi^2(L) - \Pi^2(L-1)\\ &= P(\Phi_{n_e(L)} - \Phi_{n_e(L-1)})\\ &\quad - K(\mathbb{E}[L \wedge N(L)] - \mathbb{E}[L-1 \wedge N(L-1)]),\end{aligned} \tag{4}$$

which induces that

$$\Pi^2(L) - \Pi^2(L-1) \le 0$$
$$\Leftrightarrow \quad K \ge \frac{P(\Phi_{n_e(L)}(\rho) - \Phi_{n_e(L-1)}(\rho))}{\mathbb{E}[L \wedge N(L)] - \mathbb{E}[L-1 \wedge N(L-1)]} \ge \bar{K}.$$

When $K \ge \hat{K} = \max_{\rho,L}\left\{\frac{P(\Phi_{n_e(L)}(\rho)-\Phi_{n_e(L-1)}(\rho))}{\mathbb{E}[L\wedge N(L)]-\mathbb{E}[L-1\wedge N(L-1)]}\right\}$, $\Pi^2(L)$ decreases in L so that $L_2^*(\Lambda) = 0$. When $K < \hat{K}$, we consider the following two cases:

To show that the optimal entertainment level $L_2^* = 0$ when the workload is large, note that for any give $L \ge 0$, we have

$$\begin{aligned}\lim_{\rho\to\infty} \Pi^2(L) &= \mu P - K \lim_{\rho\to\infty} \mathbb{E}[L \wedge N(L)]\\ &= \mu P - K[L \wedge n_e(L)] \le \mu P = \lim_{\rho\to\infty} \Pi^0,\end{aligned}$$

where the second equality holds by (i) the dominated convergence theorem (which justifies the interchange of the limit and the expectation) and (ii) the fact that the queue length distribution reduces to a point mass at $n_e(L)$ as $\rho \to \infty$. Hence, when the system load is sufficiently large, the optimal entertainment level L must be 0.

To show that the optimal entertainment level $L_2^* = 0$ when the workload ρ is sufficiently small, we show that $\Pi^2(L)$ is nonincreasing in L when ρ is sufficiently small, that is, $\Delta^2(L) \le 0$. First, we show that $N(L)$ is *stochastically larger than* $N(L-1)$, i.e., $N(L) \ge_{\text{s.t.}} N(L-1)$. This can be directly verified because (i) $n_e(L) \ge n_e(L-1)$, and (ii) $\pi_n(L) = \frac{\rho^n(1-\rho)}{1-\rho^{n_e(L)+1}} \le \frac{\rho^n(1-\rho)}{1-\rho^{n_e(L-1)+1}} = \pi_n(L-1)$ for all $n = 0, 1, \ldots, n_e(L-1)$, so that $\mathbb{P}(N(L) > n) \ge \mathbb{P}(N(L-1) > n)$. Next, $N(L) \ge_{\text{st}} N(L-1)$ implies $L \wedge N(L) \ge_{\text{s.t.}} (L-1) \wedge N(L-1)$, which further induces that $\mathbb{E}[L \wedge N(L)] \ge \mathbb{E}[(L-1) \wedge N(L-1)]$. Finally, the desired result follows because the first term in (4) approaches 0 when ρ is sufficiently small since

$$\frac{\partial \Phi_{n_e(L)}}{\partial \rho} = \mu \frac{1 - (n_e(L)+1)\rho^{n_e(L)} + n_e(L)\rho^{n_e(L)+1}}{(1-\rho^{n_e(L)})^2},$$

which induces that $\frac{\partial \Phi_{n_e(L)}}{\partial \rho}|_{\rho\to 0} = \frac{\partial \Phi_{n_e(L-1)}}{\partial \rho}|_{\rho\to 0} = \mu$.

Since we have proved that $\Pi^1(L)$ is concave in ρ and $\Pi^2(L) \ge \Pi^1(L)$, thus there exist two thresholds $\underline{\rho}_2$ and $\underline{\rho}_2$ satisfying $\underline{\rho}_2 \le \bar{\rho}_2$, so that $L_2^*(\Lambda) = 0$ if $\rho \in [0, \underline{\rho}_2) \cup (\bar{\rho}_2, \infty)$ and $L_2^*(\Lambda) > 0$ otherwise. □

Proof of Proposition 4.2. The throughput formula easily follows from the steady-state distributions. Next, for a given L, it is evident that $\widehat{\pi}_0(L) > \pi_0(L) = \frac{(1-\rho)}{1-\rho^{n_e(L)+1}}$ for any $p_L < 1$, which indicates that $\widehat{\Pi}(L) = \mu(1-\widehat{\pi}_0(L))P - KL \le \mu(1-\pi_0(L))P - KL = \Pi^1(L)$. This means that the optimal profit under heterogeneous delay cost is less than that under homogeneous cost. In addition, for a given L, we must have $\Delta\widehat{\Pi}(L) \equiv \widehat{\Pi}(L) - \Pi^0 < \Pi^1(L) - \Pi^0 \equiv \Delta\Pi^1(L)$. Recall from Theorem 3.1, that when the market size Λ is sufficiently small or large, Type-1 WAE is ineffective with $L_1^* = 0$, indicating that $\Delta\widehat{\Pi}(L) < \Delta\Pi^1(L) \le 0$, which in turn implies that $\widehat{L}^* = 0$ as well for the case of heterogeneous cost when Λ is sufficiently small or large. □

References

[1] A. Borges, M.M. Herter, J.C. Chebat, It was not that long!: the effects of the in-store tv screen content and consumers emotions on consumer waiting perception, J. Retail. Consum. Serv. 22 (2015) 96–106.
[2] C.Y. Cai, Hai di Lao: Service Beyond Imaginations. Technology and Operations Management Assignments, 2015.
[3] M. Garaus, U. Wagner, Let me entertain you-increasing overall store satisfaction through digital signage in retail waiting areas, J. Retail. Consum. Serv. 47 (2019) 331–338.
[4] R. Hassin, Rational Queueing, CRC Press, Taylor and Francis Group, Boca Raton, 2016.
[5] R. Hassin, M. Haviv, To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, vol. 59, Springer Science & Business Media, 2003.
[6] X. Li, Z. Lian, Y. Shi, Capacity co-opetition in service clusters with waiting-area entertainment, Working Paper, 2021.
[7] D.H. Maister, The psychology of waiting lines, in: J.A. Czepiel, M.R. Solomon, C.F. Surprenant (Eds.), The Service Encounter, Lexington Books, Lexington, MA, 1985.
[8] P. Naor, The regulation of queue size by levying tolls, Econometrica 37 (1) (1969) 15–24.
[9] X. Yuan, T. Dai, L.G. Chen, S. Gavirneni, Co-opetition in service clusters with waiting-area entertainment, Manuf. Serv. Oper. Manag. 23 (1) (2021) 106–122.

Sponsor names

Do not correct this page. Please mark corrections to sponsor names and grant numbers in the main text.

NSFC, *country*=China, *grants*=71931001