Mitigating Service Incompletion via Sunk Costs: A Queueing Approach

Jimmy Qin
Naveen Jindal School of Management, UT Dallas, jimmy.qin@utdallas.edu

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School, cwchan@gsb.columbia.edu

Jing Dong
Decision, Risk, and Operations, Columbia Business School, jing.dong@gsb.columbia.edu

Sunk-cost bias occurs when decisions are influenced by the time, energy, and money already invested, rather than considering the future costs necessary to achieve success. This phenomenon of 'irrational behavior' is well-documented in decision-making studies and is generally recognized as a factor that can lead to suboptimal decisions. In this work, we investigate how sunk cost (and the behavioral bias associated with it) can be used as an operational lever to increase service completion rates in a congested service system. We run a controlled online experiment and find that the abandonment rate is significantly reduced for the group of participants who incur a larger sunk cost. To better capture the dynamics of service systems and their impact on customers' behavior, we study a queueing model with sunk cost and strategic customers, where customers experience a disutility of balking that is proportional to the sunk cost they incur. We characterize the equilibrium behavior of the customers, from which we further derive the optimal strategy for the service provider in terms of whether to provide real-time queue length information to customers as well as the optimal level of sunk cost to impose. Our results show that the sunk cost strategy is effective only when waiting information is provided and that using a non-zero sunk cost is optimal when the queueing system is moderately congested. Through a comprehensive numerical study, we demonstrate that implementing a non-zero sunk cost can substantially improve the throughput of the system. In addition, we reveal an interesting asymmetric pattern in the robustness of the service provider's optimal policy when the customers' sensitivity to sunk cost cannot be accurately estimated, which suggests that if the service provider cannot accurately estimate the customer's sensitivity to sunk cost, using an underestimated value will give more robust performance improvements.

Key words: service operations, queueing system, sunk cost effect, delay announcement, balking and abandonment

1. Introduction

Abandonment and/or balking, which refers to the phenomenon that a customer arrives at a system but decides to leave without receiving service, is widely observed in congested service systems (Mandelbaum and Shimkin 2000). Abandonment and balking result in incomplete service, which not only harms the revenue of service providers but can also lead to other undesirable consequences.

For example, patients who present to a hospital's emergency department (ED) but leave without being seen may be at increased risk of adverse medical events (Li et al. 2019). In outpatient clinics, patient no-shows and abandonment can interrupt patient adherence to their care plan (Schectman et al. 2008) and pose challenges for appointment scheduling (Zacharias and Pinedo 2014). Service incompletion can also affect customer satisfaction and customer loyalty (Dube and Renaghan 1994).

Many operational strategies have been proposed to mitigate abandonment. These include policies to directly reduce the congestion of the system by increasing the service capacity (Lee and Ward 2019); smoothing demand over time (Green et al. 2007); improving customer experience through operational transparency (Buell 2021); creating a more fair waiting experience (Larson 1987); reducing the perceived duration of the wait (Hui and Tse 1996); or influencing customers' expectations through delay announcements (Armony et al. 2009, Jouini et al. 2011, Batt and Terwiesch 2015, Akşin et al. 2017, Ibrahim et al. 2017, Yu et al. 2017, 2021). In this study, we investigate how to use *sunk* cost in combination with delay announcements to increase service completion. To the best of our knowledge, this is the first study that looks into how sunk cost can be used to mitigate service incompletion.

The sunk cost effect/bias is described as a greater tendency to continue an endeavor once an investment in money, effort, or time has been made (Arkes and Blumer 1985). This effect has been widely recorded experimentally and empirically (Staw 1981, Thaler 1980, Nozick 1994, Dick and Lord 1998). While most studies show that sunk cost leads to worse outcomes, e.g., overinvestment or overconsumption (Ho et al. 2018), there are also recent studies showing that the sunk cost effect could play a beneficial role (Nozick 1994, Jain and Chen 2023, Zhang et al. 2023). For example, Jain and Chen (2023) and Zhang et al. (2023) show that the sunk cost effect could benefit customers who suffer from self-control problems. In this work, we study how the sunk cost could help increase service completion and, subsequently, system throughput, which is beneficial from the perspective of service providers and resource utilization, and could also lead to better outcomes for customers, especially in healthcare contexts.

We develop an online experiment where participants are randomly assigned to a high sunk-cost group by doing a long screening task and a low sunk-cost group by doing a short screening task. We estimate the effect of the sunk cost on participants' willingness to wait for the main task, which is associated with a substantial reward. We provide experimental evidence that injecting a higher sunk cost significantly decreases the likelihood of abandonment. The experiment results offer a promising operational tool for a service provider to control and optimize the queueing system. In practice, sunk costs could take the form of a non-refundable advance payment, admission fee, or non-monetary costs such as effort or time (e.g., completing a long survey about medical history before seeing a doctor online).

While injecting sunk cost can potentially increase customers' willingness to wait for service, service providers should be cautious with this strategy, as customers may reject incurring such sunk cost in the first place. For example, if a customer has to pay a non-refundable fee to enter the waitlist, they may choose an alternative provider or channel to get the service. In this sense, introducing sunk cost poses a tradeoff: On one hand, a higher sunk cost can deter customers from joining the system in the first place. On the other hand, once customers incur the sunk cost and join the system, they are more likely to stay. To study the effect of sunk cost in a service system setting with unobservable queues and repeated interactions, we consider a game-theoretic queueing model. At the beginning of the game, the service provider determines the sunk cost and whether to provide real-time queue length information. The service provider aims to maximize the system's throughput. Based on sunk cost and expected waiting time in the system, the customer decides whether to incur the sunk cost in the first stage and whether to wait for service in the second stage. In the customer's second-stage decision, sunk cost imposes a disutility of balking that is proportional to the sunk cost incurred in the first stage. We study the equilibrium that arises in the queueing game and the service provider's optimal policy.

Our analytical results show that sunk cost always decreases throughput when no real-time waiting information is provided. On the other hand, when real-time waiting information is provided, a properly tuned non-zero sunk cost can lead to higher throughput than no sunk cost. The service provider's optimal policy depends on the congestion level of the system. If the system is very underloaded, it is usually optimal not to provide waiting information and not to use any sunk cost. If the system is heavily overloaded, it is usually optimal to provide waiting information but not use any sunk cost. If the system is moderately loaded, it is usually optimal to provide waiting information and use a non-zero sunk cost. Extensive numerical studies further demonstrate that using a non-zero sunk cost could lead to substantial improvement in throughput, especially when the system is moderately loaded and customers' sensitivity to sunk cost is moderate to high. We also show that when it is hard to accurately estimate customers' sensitivity to sunk cost, using an underestimated sensitivity parameter tends to provide more robust performance improvements.

The remainder of the paper is organized as follows. In Section 2, we review the relevant literature. In Section 3, we present the details of the controlled experiment. In Section 4, we introduce the game-theoretic queueing model, which incorporates the sunk cost effect as a disutility of balking. In Section 5, we derive customers' equilibrium behavior given the sunk cost and available waiting information. In Sections 6 and 7, we derive the service provider's optimal policy that jointly optimizes over the sunk cost and delay announcement decisions. In Section 8, we run an extensive numerical study to complement our theoretical results. In particular, we demonstrate the magnitude of performance improvement brought by sunk cost and study how sensitive the system performance is to mis-specified sunk-cost effects. We conclude in Section 9.

2. Literature Review

Our work is related to the literature on behavioral queueing, particularly those that combine laboratory experiments with modeling approaches. Kremer and Debo (2016) experimentally test the impact of wait time on customers' purchasing behavior where wait times act as a signal of quality. Shunko et al. (2018) study the impact of queue structure (i.e., parallel versus single queue) and queue-length visibility on worker productivity. Buell (2021) documents a last-place aversion effect in queues and explores its implications for customer experiences. Kim et al. (2020) identify cognitive and environmental factors that drive systematic bias in admission decisions. Ülkü et al. (2020) challenge the assumption that purchase decisions are independent of the waiting time. Luo et al. (2022) study how customers form their completion costs based on the length, speed, and wait time of a queue. Althenayyan et al. (2022) investigate how line-sitting and express lines affect customers' satisfaction and fairness perception about queues. Kremer and de Véricourt (2023) study how congestion influences the decision to gather more information. Hathaway et al. (2023) study how agents react to incentives and congestion in deciding when to escalate a call. Ibrahim et al. (2024) propose a queueing-game-theoretic model to study the service provider's optimal scheduling policy when strategic customers may be dishonest about their claim type, and use controlled experiments to validate the model assumptions. See Allon and Kremer (2018) for a review of this topic. Our work contributes to this literature by studying a new and practically relevant setting, i.e., how to use sunk cost to reduce service incompletion.

There is also extensive literature that studies how to manage customers' abandonment/balking behavior in queues through the control of service capacity, pricing, priority rules, etc. One of the most common ways for a service provider to nudge customers' behavior is through sharing information about waiting with customers. Sharing waiting information has been widely seen in applications such as amusement parks, call centers, hospitals, etc. The core questions in this stream of research focus on whether to communicate waiting information to customers, and what types of information or what granularity of information to share (Ibrahim 2018b). Interestingly, it is not always optimal to provide waiting information or to provide the most granular information (Naor 1969, Hassin 1986, Guo and Zipkin 2007, He and Down 2009, Allon et al. 2011, Hu et al. 2018, Armony et al. 2009, Lingenbrink and Iyer 2019, Hassin and Roet-Green 2020). Sharing waiting information can be used together with other control levers to manage abandonment. For example, Armony and Maglaras (2004) study joint routing and delay announcement decisions in the context of a call center that offers a call-back option to delayed customers. Yu et al. (2018) consider a setting where a profit-maximizing firm uses delay announcements alongside an optimized routing rule. Ibrahim (2018a) examines how to control delay announcements, compensation offered to agents, and staffing levels to minimize costs incurred by the service provider. Most of these joint

optimization studies focus on staffing or scheduling decisions, but none of them have considered sunk cost as a lever of control in queuing systems. To the best of our knowledge, we are the first to propose a sunk-cost strategy to mitigate customer balking behavior and explore the joint optimization of sharing waiting information with sunk cost.

There is some queueing literature that conceptually introduces the idea of sunk cost. For example, customers need to make an effort or pay a fee to receive the delay information or find out about the service quality (Cui et al. 2019, Yang et al. 2019, Hassin and Roet-Green 2020, Wang and Hu 2020). In these models, customers are assumed to be fully rational, i.e., the cost that has already been incurred and cannot be recovered does not affect their future decision-making. Liu et al. (2023) study the joint optimization of service capacity allocation and appointment delay information revelation in an outpatient service with strategic walk-in patients. In their model, the patient incurs a disengagement cost if they engage in appointment booking at the first stage but switch to walk-in or balking after acquiring the exact appointment delay in the second stage. Despite its similarity to the sunk cost in our setting, there are notable differences that make our setting unique. First, the service provider in Liu et al. (2023) can only determine whether the customer incurs the disengagement cost by deciding which appointment scheduling system to use, i.e., they do not get to control/vary the size of the disengagement cost. Second, Liu et al. (2023) focuses on the service provider's capacity allocation decision, where we assume capacity is fixed and focus on optimizing the sunk cost.

While it is typically assumed that customers are fully rational, modeling customers' bounded rationality has been gaining increasing interest. Huang et al. (2013) capture bounded rationality using a model in which customers are incapable of accurately estimating their expected waiting time. Plambeck and Wang (2013) propose a tractable quasi-hyperbolic discounting model and show how customers' lack of self-control and naivete affect the optimal pricing and scheduling strategy. See Hassin (2016) Chapter 11 and Ren and Huang (2018) for a review of the bounded rationality literature in queueing and service operations management. Overall, there is very limited study on the sunk cost effect in queueing games, and our research fills this gap.

There is a broad range of psychology, consumer behavior, and behavioral economics literature that documents the sunk cost effect and empirically investigates its impact on decision-making. Most of the work shows that the sunk cost effect leads to suboptimal decisions such as overconsumption or escalation of commitment (Thaler 1980, Staw 1981, Arkes and Blumer 1985, Dick and Lord 1998). Our experiment is related to the sunk cost effect in the temporal context or time domain. Ülkü et al. (2020) use laboratory and field experiments to show that consumption increases with the time spent in line. Instead of studying consumption behavior, we focus on the willingness to wait for a reward after incurring the sunk cost. Most of the experimental literature describes

hypothetical settings (i.e., scenario-based studies), while very few experiments are real-effort tasks. As far as we know, only one study (Navarro and Fantino 2009) explores the sunk-cost effect in the temporal domain and is implemented using real-effort tasks. Navarro and Fantino (2009) require subjects to work on a 500-piece jigsaw puzzle, and ask them to choose whether to continue or quit after performing the puzzle for some time. Subjects in the "short" condition do the main puzzle for 10 minutes. Subjects in the "long" condition do the main puzzle for 50 minutes. The authors find out that subjects in the "long" condition are more likely to "persist" than "quit", showing a sunk-time effect. This escalation of commitment or continuum of effort is different from what we test – waiting for a reward. In Navarro and Fantino (2009), regardless of their decision to continue or not with the puzzle, the participants will receive the same reward.

Lastly, our work is related to the literature that explores incorporating sunk-cost effects into operational decisions. Jain and Chen (2023) and Zhang et al. (2023) study customers with sunk cost bias and time-inconsistent preferences. They find that the sunk-cost effect can make customers better off by inducing higher efforts. Jain and Chen (2023) study a firm's pricing decisions and show that the sunk-cost effect can sometimes improve a firm's profit, while Zhang et al. (2023) study optimal contract design. Both Jain and Chen (2023) and Zhang et al. (2023) focus on the setting of investment of goods such as gym membership or online education where there are no interactions between customers. The distinct feature of our work is that we study queue, which is a limited resource environment and customers' decisions interact with each other, e.g., one customer deciding to join the queue will affect another customer's waiting time.

3. Online Experiment

We first describe an experiment that tests the hypothesis that increasing sunk cost increases customers' likelihood of waiting for the reward. The online experiment consists of four stages: a qualification task, a manipulated waiting period, an attention checking buffer, and a main task (see Figure 1). Participants receive ten cents after finishing the qualification task and an additional one dollar after finishing the main task. We vary the length of the qualification task and measure its impact on the proportion of participants who abandon while waiting to start the main task.

3.1. Experimental Design

We begin by describing our design and detailed procedures for the online experiment.

Participants, payments, and exclusions: We recruit participants on Amazon Mechanical Turk (MTurk) and advertise the experiment as an image classification task. Participants with at least 0.95 HIT approval ratio (proportion of completed tasks) and located in the United States were recruited to take part in the online experiment. Participants are told that they will first complete some initial task, i.e., the qualification task, which will be processed to determine whether they

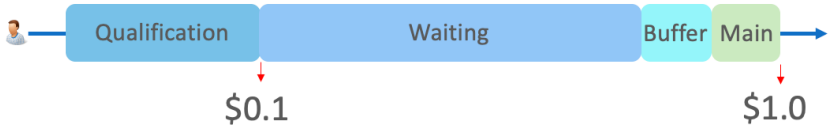


Figure 1 Four stages of the online experiment and payment scheme. If the participants quit at any time before finishing the qualification task, they receive zero payment. If the participants finish the qualification task, they receive ten cents. If the participants finish the main, they receive a dollar.

will proceed to the second task, i.e., the main task. The second task consists of five questions that are similar to what they have done in the initial task but slightly more challenging. Participants understand that they will be paid ten cents if they finish the initial task and an additional one dollar if they finish the second task. We inform the participants that the verification of their qualification for the second task may take some time, and they can choose to leave after the initial task. They have to finish the whole experiment in one sitting without interruption. For details of the interface, see Figure 13 in Appendix A.

Experimental setup: The first stage is the qualification task where participants are asked to categorize images into distinct classes. In each question, we show one picture and ask the participants to choose an appropriate category for this picture from the ten categories we provided. Figure 14 in Appendix A shows an example of the questions in the first stage. Participants are randomly assigned to either a four-question group or a thirty-question group. The four-question group is referred to as the low sunk cost group and the thirty-question group is referred to as the high sunk cost group. The second stage is the manipulated waiting period where participants are told to wait for the verification of their eligibility for the second task (see Figure 15 in Appendix A for the interface). This waiting period lasts for exactly 6 minutes for all participants. Participants are unaware of this "manipulated" waiting but are told that the waiting is the time it takes to verify their eligibility to proceed to the main task. In addition, participants only know that "the verification process can take some time" but do not know exactly how long it will take. We allow the participants to "quit the study" (by clicking a button) at any time during the waiting, receiving only ten cents as a baseline payment and forfeiting the bonus payment of one dollar, which the participants are fully aware of. To make sure that the participants finish the online experiment in one sitting and prevent them from leaving the page and checking back much later, there is a statement indicating that they need to proceed to the main task within one minute after the verification is completed. The third stage is the attention checking buffer where participants see a one-minute countdown clock (see Figure 16 in Appendix A for the interface), and they have to proceed within that one minute. If the participants fail to proceed within one minute, they

will be automatically disqualified and only receive ten cents. The last stage is the "main task". Although we have already collected sufficient information to test the hypothesis, we still created this task (Chihuahua versus Muffin, which is a bit more challenging than the qualification task, see Appendix A Figure 17) to keep the online experiment engaging and to prevent confusing the participants. Our goal is to understand the impact of sunk cost on the likelihood of waiting for the main task which is associated with a one-dollar reward.

Discussions of the Experimental Design: MTurk is a crowdsourcing platform where remotely located crowd workers are hired to perform on-demand tasks. Common tasks range from transcribing text in images to labeling or categorizing images. Often the results are used to build training datasets for machine learning algorithms. Workers on the platform are familiar with the type of task we advertise for (i.e., image classification) and they do the task mainly for the one-dollar reward. The qualification section serves as the sunk cost where participants put in a lot of effort but only get a little reward (i.e., ten cents, in the form of monetary payment). In an ideal experiment, participants should get nothing after finishing the qualification section. But for ethical reasons, we cannot do that. Instead, we choose a reasonably small amount of ten cents. The total payment, 1.1 dollars, matches the regular hourly rate on the platform. Since we try to use the experiment to understand synchronous online interactions rather than asynchronous interactions, we create the attention checking buffer to ensure that the participants stay on the waiting page or at least check the page regularly. Those who do not pass the attention checking are considered to abandon the experiment even though they did not proactively "quit the study."

This experiment was approved by the Columbia University Institutional Review Board.

3.2. Empirical Findings

A total of 425 participants were recruited on the MTurk platform, 212 of which are in the low sunk cost group (four questions in the qualification section) and 213 of which are in the high sunk cost group (thirty questions in the qualification section). Table 1 shows the number of participants entering each stage of the experiment. In particular, 5 and 13 participants in the low and high sunk cost group, respectively, left the experiment during the qualification section and are thus excluded from the following analysis.[1] We are interested in the proportion of participants who waited for the whole manipulated waiting period, excluding those who did not proceed to the main task within one minute, i.e., within the attention checking buffer. To test the hypothesis that higher sunk cost increases customers' likelihood of waiting for a reward, we conduct two types of analysis.

[1] These participants did not legitimately quit the online experiment (i.e., through selecting the "quit the study" button) and are, thus, not paid the ten-cent base payment. It is likely that they directly closed the window at the beginning or in the middle of the qualification stage. We discuss possible selection bias in Appendix A.

Table 1 Number of participants entering each of the four stages of the online experiment for the low and high sunk cost groups: the number of participants who 1) were recruited, 2) finished the qualification section, 3) passed the manipulated waiting, and 4) proceeded to the main task within one minute.

Sunk Cost	# Recruited	# Finish Qualification	# Finish Waiting	# Proceed to Main
Low	212	207	104	88
High	213	200	126	116

First, we use a two-proportion z-test to compare the proportion of abandoned participants from the low and high sunk costs groups. 57.5% (119/207) of the participants who incurred a low sunk cost abandoned while waiting and 42.0% (84/200) of the participants who incurred a high sunk cost abandoned while waiting. In other words, injecting a higher sunk cost decreases the abandonment rate by 27.0% (CI: [10.3%,56.3%], p-value: 0.002).

We also estimate the patience time distribution. There are three types of observations. The first type is those who abandoned (i.e., click the "quit the study" button) during the manipulated waiting period (stage two in Figure 1). In this case, we observe their patience exactly. The second type is those who passed the manipulated waiting but did not proceed to the main task within one minute. In this case, we assume they "abandon" at some point during the manipulated waiting, but we do not know when exactly. As such, their patience time is interval-censored between zero and six minutes. The third type is those who passed the manipulated waiting and proceeded to the main task. We only know their patience time is beyond six minutes. In this case, we have right-censored data. We employ a proportional hazard model (Cox 1972) to study the impact of sunk cost. In particular, we consider the following hazard rate function

$$h(t; Sunk) = h_0(t) \exp(\beta \times Sunk) \tag{1}$$

where $h_0(t)$ is the baseline hazard rate function, and β measures how the hazard rate varies in response to $Sunk$, which is a binary variable indicating whether the participant is in the High sunk cost group. This is a semi-parametric model because we do not impose any parametric assumptions on the baseline hazard rate function h_0. We use maximum likelihood estimation that takes care of the interval-censored data and right-censored data.[2] The estimation result shows that a higher sunk cost decreases the hazard rate by 35.3%, i.e., $1 - \exp(\hat{\beta})$ ($\hat{\beta}$: -0.434, 95% CI: [-0.710, -0.158], p-value: 0.002).

Using the online experiment, we show that participants are more likely to wait for a reward if they have incurred a larger sunk cost. If we extrapolate the experimental findings to a service setting, we may infer that customers are more likely to wait for a service if they have incurred a larger sunk cost. However, these conclusions assume that customers have already incurred or are willing

[2] We use the R package "icenReg" (Anderson-Bergman 2017).

to incur the sunk cost, which might not be the case in a real-world service setting. For example, if the sunk cost is too high, customers may be reluctant to incur it in the first place, especially after multiple interactions with the system that allow them to learn about the equilibrium waiting time. In addition, the customers' decisions interact with each other: one customer's decision to wait for the service will increase the waiting time of some other customers. Simulating multiple interactions with the queueing system and observing participants' joining or balking behaviors in equilibrium is challenging in a lab experiment. Therefore, we turn to a game-theoretic queueing model, presented in the next section.

4. Model: Two-Stage Queueing System

To understand customers' equilibrium behavior, we introduce a two-stage queueing model with the sunk cost effect on the willingness to wait. Various mechanisms of the sunk-cost effect have been proposed in the literature, including the desire to avoid waste (Arkes and Blumer 1985), the need for self-justification (Staw 1976, Schulz and Cheng 2002), information asymmetries (Kanodia et al. 1989), and mental accounting effects (Thaler 1980, Soman and Cheema 2001). Some recent literature also attempts to provide rationales for the sunk cost effect (Baliga and Ely 2011, Hong et al. 2019), e.g., as a self-management device. Our modeling approach of the sunk cost effect is similar to those in Jain and Chen (2023) and Zhang et al. (2023) where a sunk cost imposes a disutility of changing a chosen plan of action, i.e., balking instead of waiting in our case. We assume that a larger sunk cost leads to a larger disutility. In particular, we assume the disutility takes a linear form: $-\theta s$, where θ is a sensitivity parameter and s is the amount of sunk cost incurred.

The core trade-off we wish to explore is that a higher sunk cost increases the likelihood that customers will wait for the service; however, it also decreases the likelihood that customers will join the system in the first place. To maximize system throughput, service providers must set a sunk cost that balances these competing effects. Based on that we can develop a better understanding of a service provider's optimal strategy for using the sunk cost, together with whether to provide real-time queue length information.

We first provide more details about the two-stage queueing model, which is based on an $M/M/1$ queue. Homogeneous customers arrive at the system according to a Poisson process with rate λ_0. Customers make decisions in two sequential stages (see, Figure 2).

First Stage: In the first stage, based on the revealed information about the system, a customer decides whether to join the system and incur a fixed cost of $s \geq 0$ or balk, which has zero cost. We denote the equilibrium, or effective, arrival rate after the first stage decision as $\lambda_{e1} = q_1 \lambda_0$. This is also the initial arrival rate to the second stage.

Second Stage: In the second stage, based on updated information about the system, the customer can choose between waiting to get service or balking. Customers who decide to wait are

served on a first-come-first-served (FCFS) basis and the service times are assumed to be independent and identically distributed exponential random variables with rate μ. Without loss of generality, we assume the service rate $\mu \equiv 1$. Moreover, we assume that customers do not renege if they have decided to wait in the queue. In the second stage, the fixed cost incurred in the first stage, s, is a sunk cost because it cannot be recovered regardless of whether the customers wait or balk. Examples of this type of sunk costs include non-refundable advance payment (i.e., a due sum that is paid in advance for services), non-refundable admission fees, and non-monetary costs such as time spent or efforts made. To capture the customer's behavioral bias towards the sunk cost, we assume that if a customer decides to balk, the customer incurs a cost of θs. If a customer decides to wait, a linear waiting cost with the marginal rate $c > 0$ is incurred, and upon service completion, the customer receives a reward of R. We denote the equilibrium, or effective, arrival rate to the queue after the second stage decision as $\lambda_{e2} = q_2 \lambda_{e1}$, which can also be interpreted as the throughput of the system.

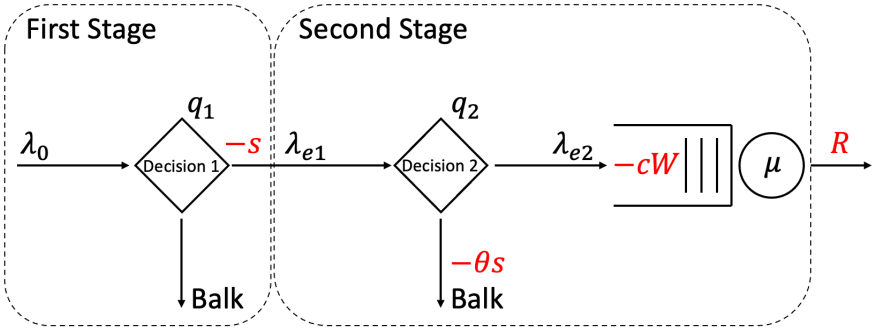


Figure 2 Two-stage queueing model: Customers arrive at the system according to a Poisson process with rate λ_0 and make a decision between joining the system and balking. If a customer joins, she will incur a fixed cost and go to the second stage. If a customer balks, she leaves the system and gets zero utility. The proportion of customers that decide to join in equilibrium is denoted by q_1 and the resulting equilibrium arrival rate is denoted by λ_{e1}. In the second stage, customers make a decision between waiting in the queue or balking. If a customer waits, she will join the queue and get service on a first-come-first-serve basis. If a customer balks, she leaves the system and incurs a disutility proportional to the cost she incurred in the first stage. The proportion of customers that decide to wait in equilibrium is denoted by q_2 and the resulting equilibrium arrival rate after the second stage decision is denoted by λ_{e2}. Customers receive a reward upon completion of the service and incur a linear cost while waiting in the second stage. The reward and associated costs are marked in red.

Both the service provider and the customers are strategic, and a queueing game is forged as follows. The service provider first observes the exogenous system parameters and has two control levers: i) how much sunk cost s (if any) to impose and ii) whether to provide real-time queue length information to customers in the second stage. All the system parameters, including the exogenous

system parameters and the service provider's decisions, are common knowledge to the customers. The service provider understands that the customers are strategic and that the customers will take the best responses to the service provider's decisions to maximize their utility, which we will discuss in more detail later in this section. The objective of the service provider is to maximize the system throughput, i.e., minimize the service incompletion rate. Note that the throughput in equilibrium is jointly determined by the service provider's decisions and the best responses of utility-maximizing customers.

Depending on whether or not the queue length information is provided, there are two types of systems: *informed system* (I-System) and *uninformed system* (U-System). Customers in the informed system are called *informed customers* and they have access to real-time queue length information, N, when arriving at the second stage. Note that informed customers still do not observe the queue length in the first stage. Customers in the uninformed system are called *uninformed customers* and they do not have access to real-time queue length information. Instead, they have to rely on the expected queue length to decide whether to wait or balk. Customers know the system type when they arrive. We use the terminology (un)informed system and (un)informed customer interchangeably. The system type $\gamma \in \{I, U\}$ is again predetermined by the service provider. The service provider either provides the queue length information to all customers or does not provide such information to any customer. To mark the dependence on the system type explicitly, for γ-System, we denote the initial arrival rate to the first stage by λ_0^γ, the equilibrium arrival rate after the first stage decision by λ_{e1}^γ, and the equilibrium arrival rate after the second stage decision by λ_{e2}^γ. We assume $\lambda_0^I = \lambda_0^U = \lambda_0$, i.e., the initial arrival rate is exogenous.

Note that in case of a unique equilibrium, which we will show in Sections 5.1 and 5.2, the equilibrium arrival rate after the second stage decision (or, the throughput of the system) can be viewed as a function of the system parameters (including the exogenous system parameters $\lambda_0, \mu, R, c, \theta$ and the service provider's decisions s, γ). For the analysis in Sections 5 and 6, we will assume that μ, R, c, θ are fixed. As such, when viewed as a function, we write the equilibrium arrival rate after the second stage decision as $\lambda_{e2}^\gamma(s, \lambda_0)$, i.e., a function of the sunk cost s and the initial arrival rate λ_0. With a little abuse of notation, when the context is clear, we write the equilibrium arrival rate after the second stage decision as $\lambda_{e2}^\gamma(s)$ when λ_0 is fixed. Similarly, when viewed as functions, we write the equilibrium joining and waiting probabilities as $q_1^\gamma(s, \lambda_0)$ (or, $q_1^\gamma(s)$) and $q_2^\gamma(s, \lambda_0)$ (or, $q_2^\gamma(s)$).

5. Customers' Optimal Policy

In this section, we discuss the customer's optimization problem and the corresponding equilibrium given a fixed system type γ and sunk cost s. The optimization problem of the service provider will be analyzed in Section 6.

5.1. Modeling Uninformed Customers' Decision

Customers are uninformed if they do not get the real-time queue length information in the second stage. We analyze the uninformed customers' utility and their equilibrium behavior starting with the second stage. Recall that all the system parameters (including the exogenous system parameters $\lambda_0, \mu, R, c, \theta$ and the service provider's decisions s, γ) are common knowledge to the customers.

In the second stage, the uninformed customers make a decision between waiting and balking. We follow the established convention by focusing exclusively on symmetric strategies (Hassin and Haviv 2003). A pure or mixed strategy for an uninformed customer in the second stage can be described by a fraction $q_2^U \in [0,1]$, which represents the probability that an uninformed customer will wait in the queue, or equivalently, the proportion of uninformed customers who will wait in the queue. The uninformed customer's utility of waiting is $R - c\bar{W}^U$ where $\bar{W}^U$ is the expected waiting time, endogenously determined by the customer's strategy. The uninformed customer's utility of balking is $-\theta s$. Recall that s is the sunk cost incurred in the first stage and θ measures customers' sensitivity to sunk costs. The uninformed customer's optimization problem in the second stage is

$$U_2^U = \max_{q\in[0,1]} q(R - c\bar{W}^U) + (1-q)(-\theta s) \text{ and } q_2^U = \arg\max_{q\in[0,1]} q(R - c\bar{W}^U) + (1-q)(-\theta s),$$

where U_2^U denotes the expected utility in the second stage. We first evaluate uninformed customers' behavior in equilibrium in the second stage given the equilibrium arrival rate after the first stage decision, λ_{e1}^U, which can also be viewed as the initial arrival rate to the second stage.

Lemma 1. *For fixed R, c, θ, and s, given the equilibrium arrival rate after the first stage decision λ_{e1}^U, the equilibrium arrival rate following the second stage decision is*

$$\lambda_{e2}^U = \min\left\{\lambda_{e1}^U, \frac{R+\theta s}{c+R+\theta s}\right\};$$

the proportion of uninformed customers that wait in the second stage is

$$q_2^U = \min\left\{1, \frac{R+\theta s}{\lambda_{e1}^U(c+R+\theta s)}\right\};$$

and the expected waiting time in the second stage is

$$\bar{W}^U = \min\left\{\frac{\lambda_{e1}^U}{1-\lambda_{e1}^U}, \frac{R+\theta s}{c}\right\}.$$

The proof of Lemma 1 is in Appendix B. Lemma 1 shows that if $\lambda_{e1}^U \le (R+\theta s)/(c+R+\theta s)$, then all customers will choose to wait; if $\lambda_{e1}^U > (R+\theta s)/(c+R+\theta s)$, then customers will use a mixed strategy so that only a proportion of customers will choose to wait in equilibrium and the equilibrium arrival rate following the second stage decision is $(R+\theta s)/(c+R+\theta s)$.

In the first stage, the uninformed customers make a decision between joining or balking. Recall that all the system parameters are common knowledge to the customers. However, instead of using

the true sensitivity parameter for sunk cost, θ, we assume that the customers have a *belief* that $\hat{\theta}=0$, i.e., the customers *believe* themselves to be fully rational in the second stage. This follows a notion of bounded rationality (i.e., rational in the first stage but irrational in the second stage).[3] Indeed, people often overestimate their ability or rationality (Kruger and Dunning 1999). They do not believe in the existence of cognitive and motivational biases in themselves even though they believe in the existence of such biases in other people (Pronin et al. 2002, Pronin 2007). We again restrict our attention to symmetric strategies. A pure or mixed strategy for an uninformed customer in the first stage can be described by a fraction $q_1^U \in [0,1]$, which represents the probability that an uninformed customer will join the queue, or equivalently, the proportion of uninformed customers who will join the queue. The uninformed customer's utility of joining is $-s+\hat{U}_2^U$ where $\hat{U}_2^U$ is the *belief* about the expected utility in the second stage. The uninformed customer's utility of balking is zero. Then, the uninformed customer's optimization problem in the first stage is

$$U_1^U = \max_{q\in[0,1]} q(-s+\hat{U}_2^U) \text{ and } q_1^U = \arg\max_{q\in[0,1]} q(-s+\hat{U}_2^U)$$

where

$$\hat{U}_2^U = \max_{q\in[0,1]} q(R-c\bar{W}^U).$$

In particular, the uninformed customer's *belief* about the utility of waiting in the second stage is $R-c\bar{W}^U$ where $\bar{W}^U$ is the *actual* expected waiting time in the second stage that uninformed customers learn from repeated interactions with the system, which is characterized in Lemma 1. The uninformed customer's *belief* about the utility of balking in the second stage is 0, rather than $-\theta s$, because they believe themselves to behave rationally in the second stage. We also denote $\hat{q}_2^U = \arg\max_{q\in[0,1]} q(R-c\bar{W}^U)$. $\hat{q}_2^U$ is the uninformed customers' *belief* about their decision in the second stage, which might be different from their *actual* decision q_2^U characterized in Lemma 1. Next, we evaluate uninformed customers' behavior in equilibrium in the first stage when the initial arrival rate to the first stage is λ_0.

Lemma 2. *For fixed R, c, θ, and s, given the initial arrival rate to the first stage λ_0, in the unique equilibrium of an uninformed system, the equilibrium arrival rate following the first stage decision is*

$$\lambda_{e1}^U = \min\left\{\lambda_0, \frac{R-s}{c+R-s}\right\};$$

the proportion of uninformed customers who choose to incur the sunk cost in the first stage is

$$q_1^U = \min\left\{1, \frac{R-s}{\lambda_0(c+R-s)}\right\}.$$

[3] Denote the belief of sensitivity to sunk cost in the first stage as $\hat{\theta}$ and the actual sensitivity to sunk cost in the second stage as θ. Customers are fully rational if $\hat{\theta}=\theta=0$. Customers are fully irrational if $\hat{\theta}=\theta>0$. In the fully irrational case, i.e., $\hat{\theta}=\theta>0$, we get analytical results similar to the main findings of this paper.

The proof of Lemma 2 is in Appendix B. Lemma 2 shows that if $\lambda_0 \leq (R-s)/(c+R-s)$, then all customers will choose to join; if $\lambda_0 > (R-s)/(c+R-s)$, then customers will use a mixed strategy so that only a proportion of customers will choose to join in equilibrium and the equilibrium arrival rate following the first stage decision is $(R-s)/(c+R-s)$.

Combining the two stages of decisions, we have the following theorem characterizing the equilibrium strategy of uninformed customers.

Theorem 1. *(Equilibrium Strategy of Uninformed Customers) For fixed R, c, θ, and s, given the initial arrival rate λ_0, in the unique global equilibrium of an uninformed system, the proportions of uninformed customers that join in the first stage and wait in the second stage are*

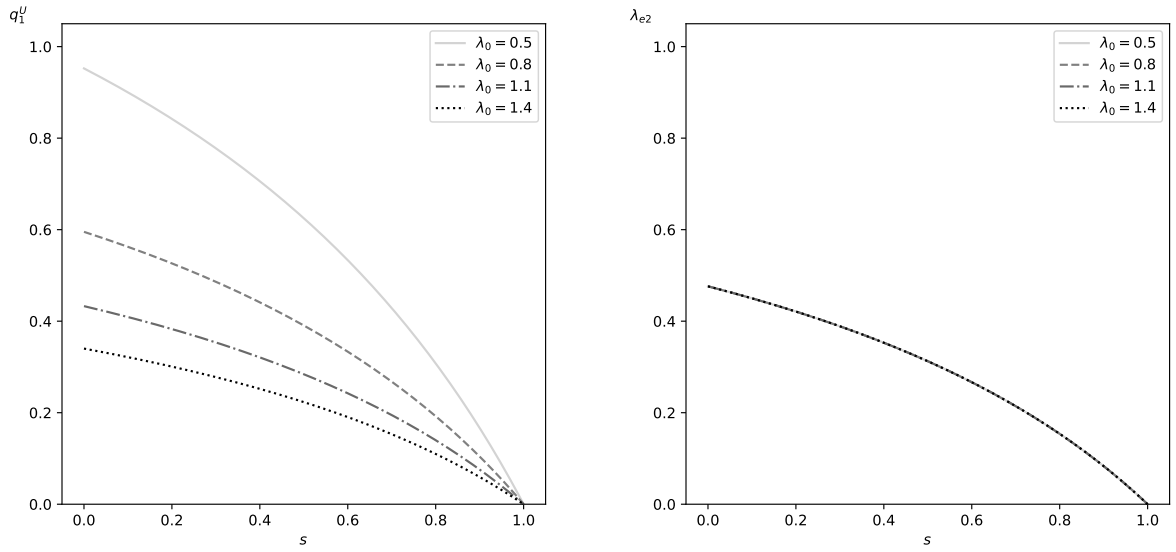
$$(q_1^U, q_2^U) = \begin{cases} (1,1) & \text{if } \lambda_0 \leq \frac{R-s}{c+R-s} \\ \left(\frac{R-s}{\lambda_0(c+R-s)}, 1\right) & \text{if } \lambda_0 > \frac{R-s}{c+R-s} \end{cases}$$

We note from Theorem 1 that the equilibrium joining probability in the first stage is identical to what we have in Lemma 2, while the equilibrium waiting probability in the second stage always equals one, i.e., all customers will continue to wait in the second stage. This is because there is no information update from the first stage to the second stage: uninformed customers use the expected queue length information to make the decision in both stages. If the uninformed customers are fully rational in the second stage, as they are in the first stage, their behavior should be consistent across the two stages, i.e., if they decided to join the system in the first stage, they should not balk in the second stage. The existence of the sunk cost effect will only make the uninformed customers more reluctant to balk in the second stage. Figure 3 (a) plots the first stage equilibrium strategy of uninformed customers q_1^U as a function of s, for various λ_0. For a fixed s, q_1^U is decreasing in λ_0. For a fixed λ_0, q_1^U is decreasing in s.

Combining the two stages, we can write the throughput (i.e., the equilibrium arrival rate λ_{e2}^U following the second stage decision) of the uninformed system as a function of s and λ_0:

$$\lambda_{e2}^U(s, \lambda_0) = \begin{cases} \lambda_0 & \text{if } \lambda_0 \leq \frac{R-s}{c+R-s} \\ \frac{R-s}{c+R-s} & \text{if } \lambda_0 > \frac{R-s}{c+R-s}. \end{cases} \tag{2}$$

Note that $\lambda_{e2}^U(s, \lambda_0)$ is decreasing in s. See Figure 3 (b) for a pictorial illustration. (Note that in this figure, $R/(c+R) = 0.48$, which is smaller than $\lambda_0 = 0.5$. Thus, we only see the case where $\lambda_{e2}^U(s, \lambda_0) = (R-s)/(c+R-s)$ in this figure.) In Proposition 1, we will formally show that the throughput is the largest when $s = 0$ for the uninformed system. That is to say, the service provider should never impose sunk costs on uninformed customers.



(a) Equilibrium first stage strategy (b) Equilibrium throughput

Figure 3 Equilibrium strategy of uninformed customers in the first stage q_1^U (left) and equilibrium throughput of uninformed customers (right), as a function of s, for various λ_0. $R = 1$, $c = 1.1$, $\theta = 1.1$.

5.2. Modeling Informed Customers' Decision

Customers are informed if they get the real-time queue length when they reach the second stage. We analyze the informed customers' utility and equilibrium behavior starting at the second stage.

In the second stage, all the system parameters are common knowledge to the informed customers, who make a decision between waiting and balking. After observing the current queue length N, the informed customer's utility of waiting is $R - cN/\mu = R - cN$. The informed customer's utility of balking is $-\theta s$. The informed customers' optimization problem in the second stage is

$$U_2^I = \max\{R - cN, -\theta s\},$$

where U_2^I denotes the utility in the second stage. We first evaluate informed customers' behavior in equilibrium in the second stage when the equilibrium arrival rate after the first stage decision is λ_{e1}^I. It is easy to see that the informed customer chooses to wait if $R - cN > -\theta s$; the informed customer chooses to balk if $R - cN < -\theta s$; and the informed customer is indifferent between waiting and balking if $R - cN = -\theta s$. Without loss of generality, we assume the informed customer will always wait when $R - cN = -\theta s$. This implies that the informed customers use a pure threshold strategy with the threshold

$$n_s \equiv \left\lfloor \frac{R + \theta s}{c} \right\rfloor + 1,$$

which is independent of λ_{e1}^I. Note that a customer will continue to wait if the observed queue length $N < n_s$ and will balk if $N = n_s$. The following lemma gives the steady-state distribution of the queue length in the second stage of an informed system.

Lemma 3. *For fixed R, c, θ, and s, given the equilibrium arrival rate from the first stage to the second stage λ_{e1}^I, when $\lambda_{e1}^I \neq 1$, the steady-state distribution of the queue length in the second stage is*

$$\pi_i = (\lambda_{e1}^I)^i \frac{1-\lambda_{e1}^I}{1-(\lambda_{e1}^I)^{n_s+1}} \quad \textit{for} \quad i = 0, 1, \cdots, n_s,$$

and the proportion of informed customers that wait in the second stage is

$$q_2^I = 1 - \pi_{n_s} = \frac{1-(\lambda_{e1}^I)^{n_s}}{1-(\lambda_{e1}^I)^{n_s+1}}.$$

When $\lambda_{e1}^I = 1$,

$$\pi_i = \frac{1}{n_s+1} \quad \textit{for} \quad i = 0, 1, \cdots, n_s,$$

and the proportion of informed customers that wait in the second stage is

$$q_2^I = 1 - \pi_{n_s} = \frac{n_s}{n_s+1}.$$

The proof of Lemma 3 is in Appendix B.

In the first stage, informed customers make a decision between joining or balking. The informed customer's utility of joining is $-s + \mathbb{E}[\hat{U}_2^I]$ where $\mathbb{E}[\hat{U}_2^I]$ is the *belief* about the expected utility in the second stage, and the informed customer's utility of balking is zero. Again, we assume that informed customers believe themselves to be fully rational in the second stage, i.e., $\hat{\theta} = 0$. In particular, given the equilibrium arrival rate after the first stage decision, i.e., λ_{e1}^I, the *belief* about the expected utility in the second stage is

$$\mathbb{E}[\hat{U}_2^I] = \sum_{i=0}^{n_0-1} (R - ci)\pi_i$$

where π_i, which depends on λ_{e1}^I, is characterized in Lemma 3 and $n_0 = \lfloor R/c \rfloor + 1$. Note that this is different from their actual expected utility in the second stage,

$$\mathbb{E}[U_2^I] = \sum_{i=0}^{n_s-1} (R - ci)\pi_i - \theta s \pi_{n_s},$$

which includes an "irrational" (dis)utility of balking from the sunk cost. Note that the same distribution π_i under the irrational second stage behavior is used in both cases because customers can learn such a distribution of queue length from repeated interactions with the system. The informed customer's utility in the first stage is then

$$U_1^I = \max_{q \in [0,1]} q(-s + \mathbb{E}[\hat{U}_2^I]).$$

We also write $q_1^I = \arg\max_{q \in [0,1]} q(-s + \mathbb{E}[\hat{U}_2^I])$.

Define

$$f(\lambda; s) = -s + \sum_{i=0}^{n_0-1} (R - ci)\pi_i, \tag{3}$$

where π_i depends on λ as characterized in Lemma 3 with $\lambda_{e1}^I = \lambda$. $f(\lambda; s)$ can be interpreted as the utility of joining in the first stage when $\lambda_{e1}^I = \lambda$. For a fixed s, in Lemma 6 in Appendix B, we show that when $s = 0$, $f(\lambda; s) \geq 0$ for all $\lambda > 0$. When $s > 0$, $f(\lambda; s) = 0$ has a unique solution. We define $\lambda_f(s)$ as the solution of $f(\lambda; s) = 0$ when $s > 0$ and $\lambda_f(s) = \lambda_0$ when $s = 0$. To simplify notation, we may omit the dependence of $\lambda_f(s)$ on s when it is clear from the context. In what follows, we first show that $\lambda_f(s)$ satisfies a monotonicity property (Lemma 4) and then characterize the equilibrium behavior in the first stage (Lemma 5). The proofs of the results are provided in Appendix B.

Lemma 4. *$\lambda_f(s)$ is decreasing in s.*

Lemma 5. *For fixed R, c, θ, and s, given the initial arrival rate to the first stage λ_0, the equilibrium arrival rate after the first stage decision is*

$$\lambda_{e1}^I = \min\{\lambda_0, \lambda_f\};$$

the proportion of informed customers who choose to incur the sunk cost in the first stage is

$$q_1^I = \min\left\{1, \frac{\lambda_f}{\lambda_0}\right\}.$$

Note that if $\lambda_0 \leq \lambda_f$, then $f(\lambda_0; s) > 0$. In this case, all customers will choose to join in the first stage. If $\lambda_0 > \lambda_f$, then $f(\lambda_0; s) < 0$. Customers will use a mixed strategy and join with probability λ_f/λ_0. Combining Lemmas 3 and 5, we summarize the equilibrium strategy of the informed customers in the following theorem.

Theorem 2. *(Equilibrium Strategy of Informed Customers) For fixed R, c, θ, and s, given the initial arrival rate λ_0, in the unique global equilibrium of an informed system, the proportions of informed customers that join in the first stage and wait in the second stage are*

$$(q_1^I, q_2^I) = \left(\min\left\{1, \frac{\lambda_f}{\lambda_0}\right\}, \frac{1 - \min\{\lambda_f, \lambda_0\}^{n_s}}{1 - \min\{\lambda_f, \lambda_0\}^{n_s+1}}\right).$$

Figure 4 plots the informed customers' first-stage equilibrium strategy q_1^I (left plot), second-stage equilibrium strategy q_2^I (middle plot), and the product of the two stages' equilibrium strategies (right plot), as functions of s, for different values of λ_0. For a fixed s, both q_1^I and q_2^I are non-increasing in λ_0. For a fixed λ_0, we observe that q_1^I is non-increasing in s (left plot) while q_2^I is non-decreasing in s (middle plot). This indicates that there is a trade-off between the first and the second stage equilibrium strategies when increasing sunk cost. In particular, increasing sunk cost discourages informed customers from joining the system in the first stage, but once they join, they are more likely to wait in the second stage. From the middle plot of Figure 4, we also observe that

q_2^I "jumps" at particular sunk cost levels. The jumps are due to the quantization from the floor function in the definition of n_s and happen when n_s reaches the next integer level (as s increases).

Combining the two stages, we can write the throughput (i.e., the equilibrium arrival rate λ_{e2}^I following the second stage decision) of the informed system as a function of s and λ_0:

$$\lambda_{e2}^I(s,\lambda_0) = \underbrace{\min\{\lambda_f(s),\lambda_0\}}_{\text{first-stage effect}} \cdot \underbrace{\frac{1-\min\{\lambda_f(s),\lambda_0\}^{n_s}}{1-\min\{\lambda_f(s),\lambda_0\}^{n_s+1}}}_{\text{second-stage effect}}. \tag{4}$$

With Lemma 4, we note that the first-stage effect of $\lambda_{e2}^I(s,\lambda_0)$ is non-increasing in s, while the second-stage effect of $\lambda_{e2}^I(s,\lambda_0)$ is non-decreasing in s. The product of the two effects can be non-monotone in s, see Figure 5 for an illustration. In particular, when s is small, the second-stage effect dominates and the throughput increases with s. When s is large, the first-stage effect dominates and the throughput decreases with s. As such, using a non-zero sunk cost increases throughput in this case. In Proposition 2, we formally show that under certain conditions, using a non-zero sunk cost increases the throughput of an informed system.

6. Service Provider's Optimal Policy

In this section, we discuss the optimal policy for the service provider. In particular, the service provider makes two decisions: how much sunk cost customers will have to incur in the first stage and whether or not to provide the queue length information in the second stage. The objective of the service provider is to maximize the throughput of the system

$$\max_{s\geq 0,\gamma\in\{I,U\}} \lambda_{e2}^{\gamma}(s),$$

where λ_{e2}^{γ}, characterized in (2) and (4), is the equilibrium arrival rate following the second stage decision of the γ-System.

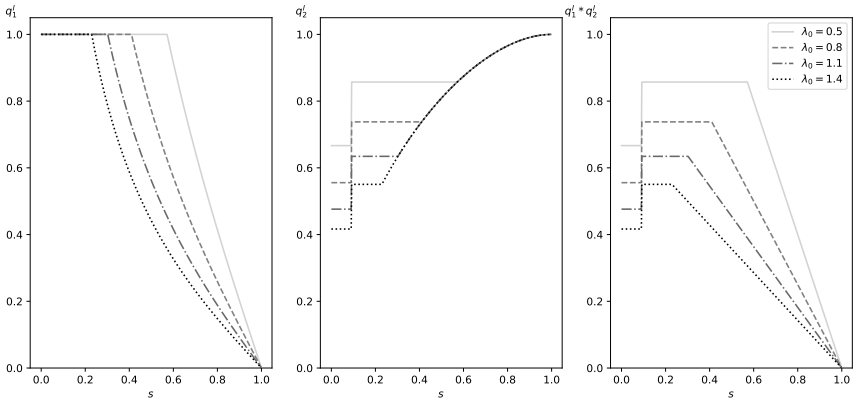


Figure 4 **Equilibrium strategy of informed customers in the first stage q_1^I (left), second stage q_2^I (middle), and the product of two stages $q_1^I * q_2^I$ (right), as functions of s, for various λ_0. $R = 1$, $c = 1.1$, $\theta = 1.1$.**

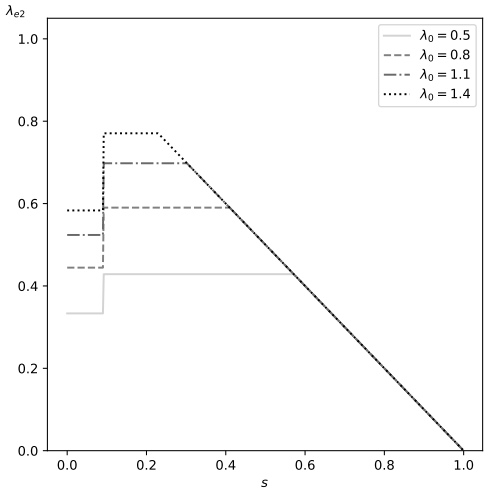


Figure 5 **Equilibrium throughput of informed customers as a function of s, for various λ_0. $R = 1$, $c = 1.1$, $\theta = 1.1$.**

6.1. Optimal Policy for an Uninformed System

We start by analyzing the service provider's optimal policy when the system is uninformed. In this case, the service provider only decides how much sunk cost the uninformed customers will have to incur in the first stage, i.e., $\max_{s\geq 0} \lambda_{e2}^U(s)$.

Proposition 1. *(Optimal Sunk Cost for Uninformed Customers) For fixed λ_0, R, c, and θ, the throughput is the largest when $s = 0$ for an uninformed system, i.e.,*

$$\max_{s\geq 0} \lambda_{e2}^U(s) = \lambda_{e2}^U(0).$$

Proposition 1 is a direct consequence of Theorem 1 and the proof is in Appendix C. In particular, we note from Theorem 1 that the sunk cost decreases the joining probability in the first stage, q_1^U, while having no effect on the second stage waiting probability, q_2^U. Thus, the service provider should never impose sunk costs on uninformed customers.

6.2. Optimal Policy for an Informed System

We next analyze the service provider's optimal policy when the system is informed. In this case, the service provider only decides how much sunk cost the informed customers will have to incur in the first stage, i.e., $\max_{s\geq 0} \lambda_{e2}^I(s)$.

Define

$$s_k \equiv \frac{c(k-1+n_0)-R}{\theta}, \quad k \in \mathbb{N}^+. \tag{5}$$

Note that s_k is the point at which n_s, as a function of s, jumps from $n_0 + k - 1$ to $n_0 + k$. We note from the characterization of $q_2^I(s)$ in Theorem 2 that $q_2^I(s)$ "jumps" upwards at s_k's. See, for example, the middle plot in Figure 4 where we see a jump at s_1. In this plot, $s_2 > 1$, so we only see one jump for $s \in [0, 1]$.

Proposition 2. *(Optimal Sunk Cost for Informed Customers) For fixed R, c, and θ, there exists a unique critical value*

$$T^I \equiv \sup\left\{\lambda : \max_{k\in\mathbb{N}^+} \lambda_{e2}^I(s_k,\lambda) > \lambda_{e2}^I(0,\lambda)\right\},$$

such that if $\lambda_0 \le T^I$, then there exists $s^\star > 0$ such that

$$\lambda_{e2}^I(s^\star) > \lambda_{e2}^I(0).$$

If $\lambda_0 > T^I$, then

$$\max_{s\ge 0} \lambda_{e2}^I(s) = \lambda_{e2}^I(0).$$

The proof of Proposition 2 is in Appendix C. Proposition 2 considers the combined effect of sunk cost on the first and second stages and shows that it depends on the initial arrival rate λ_0. In particular, there exists a unique critical value such that if the initial arrival rate is less than this value, it is optimal to have a positive sunk cost; if the initial arrival rate is greater than this value, it is optimal to have zero sunk cost. One critical difference here from the uninformed system is that customers have the queue length information updated from the first to the second stage. In particular, informed customers rely on the expected queue length to make a decision in the first stage, and they observe the real-time queue length in the second stage and make an updated decision accordingly. When the system is less congested (i.e., small λ_0), all customers shall choose to join the system in the first stage since the expected waiting time is short. In this case, having a properly designed positive sunk cost in the first stage keeps more customers staying in the system in the second stage while not making customers balk in the first stage. When the system is congested (i.e., large λ_0), there are already some customers balking in the first stage in equilibrium. In this case, while increasing sunk cost potentially keeps more customers waiting in the second stage, it also makes more customers balk in the first stage. The combined effect is negative.

6.3. Comparing Uninformed and Informed Systems

In this section, we compare the throughput of informed and uninformed systems. We first analyze the service provider's optimal policy when the sunk cost is zero. In this case, the service provider only decides whether or not to provide queue length information in the second stage, i.e., $\max_{\gamma\in\{I,U\}} \lambda_{e2}^\gamma(0)$.

Define the function

$$f_0^{UI}(\lambda) := \begin{cases} \frac{R}{c+R} - \frac{\lambda-\lambda^{n_0+1}}{1-\lambda^{n_0+1}} & \text{if } \lambda \ne 1 \\ \frac{R}{c+R} - \frac{n_0}{n_0+1} & \text{if } \lambda = 1. \end{cases}$$

Note that $R/(c+R)$ is the cap of the throughput of an uninformed system without sunk cost; $(\lambda-\lambda^{n_0+1})/(1-\lambda^{n_0+1})$ is the throughput of an informed system without sunk cost when $\lambda_0 = \lambda \ne 1$ and $n_0/(n_0+1)$ is the throughput when $\lambda_0 = \lambda = 1$. Thus, $f_0^{UI}(\lambda)$ is the difference between the uninformed and informed throughput when there is no sunk cost. In Lemma 9 in Appendix C, we show that $f_0^{UI}(\lambda) = 0$ has a unique solution and we denote this solution as $\bar{T}$.

Proposition 3. *(Comparing Uninformed and Informed Customers Without Sunk Cost) For fixed R, c, θ, and λ_0, if $\lambda_0 > \bar{T}$, then $\lambda^I_{e2}(0) > \lambda^U_{e2}(0)$. If $\lambda_0 < \bar{T}$, then $\lambda^I_{e2}(0) < \lambda^U_{e2}(0)$. If $\lambda_0 = \bar{T}$, then $\lambda^I_{e2}(0) = \lambda^U_{e2}(0)$.*

The proof of Proposition 3 is in Appendix C. Proposition 3 shows that the impact of revealing queue length information can vary based on the overall congestion of the system. In lightly congested systems (i.e., $\lambda_0 < \bar{T}$), disclosing the queue length information might deter some customers who might otherwise join under a short expected waiting time but now observe a long queue due to the stochastic fluctuation of the system. Conversely, in heavily congested systems (i.e., $\lambda_0 > \bar{T}$), revealing the queue length encourages more customers to join when the actual queue length is small. Some of these customers would otherwise not join due to a long expected waiting time. As such, in cases where the arrival rate is low, real-time queue length information might reduce the system throughput compared to an uninformed system where decisions are based on low average waiting times. However, if the arrival rate is high, revealing the queue length can lead to a higher system throughput than an uninformed system. Similar arguments and intuitions have been shown in the literature under different model settings (Hassin 1986, Chen and Frank 2004).

Next, we characterize the optimal policy for the service provider for different initial arrival rates.

Theorem 3. *For fixed R, c, and θ, there exist two unique critical values $T^{UI} \in \left[\frac{R}{c+R}, \bar{T}\right]$ and T^I such that the following hold:*

1. *When $T^I > T^{UI}$:*
   (a) *If $\lambda_0 \le T^{UI}$, $\max_{s\ge 0, \gamma\in\{I,U\}} \lambda^\gamma_{e2}(s) = \lambda^U_{e2}(0)$.*
   (b) *If $\lambda_0 \ge T^I$, $\max_{s\ge 0, \gamma\in\{I,U\}} \lambda^\gamma_{e2}(s) = \lambda^I_{e2}(0)$.*
   (c) *If $T^{UI} < \lambda_0 < T^I$, there exists $s^\star > 0$ such that $\max_{s\ge 0, \gamma\in\{I,U\}} \lambda^\gamma_{e2}(s) = \lambda^I_{e2}(s^\star)$.*
2. *When $T^I \le T^{UI}$, $T^{UI} = \bar{T}$:*
   (a) *If $\lambda_0 \le T^{UI}$, $\max_{s\ge 0, \gamma\in\{I,U\}} \lambda^\gamma_{e2}(s) = \lambda^U_{e2}(0)$.*
   (b) *If $\lambda_0 > T^{UI}$, $\max_{s\ge 0, \gamma\in\{I,U\}} \lambda^\gamma_{e2}(s) = \lambda^I_{e2}(0)$.*

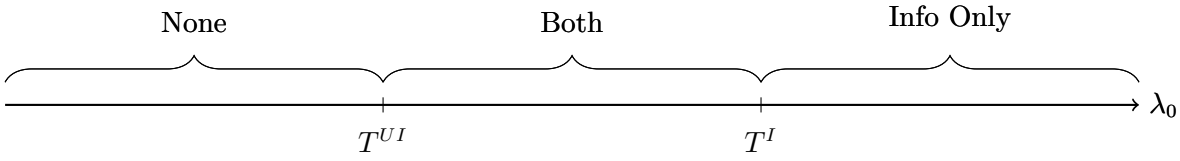


Figure 6 Service provider's optimal policy of using sunk cost together with providing real-time queue length information, for different initial arrival rates.

The proof of Theorem 3 is in Appendix C. Theorem 3 shows that if the initial arrival rate to the system is low (i.e., below T^{UI}), then the optimal strategy for the service provider is to neither provide queue length information nor inject sunk cost, which we refer to as the **None** policy. If the

initial arrival rate to the system is high (i.e., above T^I), then the optimal strategy for the service provider is to provide real-time queue length information only and use a zero sunk cost, which we refer to as the **Info Only** policy. If the initial arrival rate to the system is moderate (i.e., in between T^{UI} and T^I), then the optimal strategy for the service provider is to provide both real-time queue length information and use a non-zero sunk cost, which we refer to as the **Both** policy. See Figure 6 for an illustration. Note that T^I can be smaller than T^{UI}, in which case, $T^{UI} = \bar{T}$ and it is optimal not to use any sunk cost. As we will demonstrate in the numerical experiments below, this situation typically arises when θ is very small.

Next, we numerically demonstrate the structures of optimal policies for the service provider. In Figure 7, we show the optimal policy for the service provider when $R = 1.0$ and $c = 1.1$, while θ ranges from 0.01 to 1.5 and λ_0 ranges from 0.5 to 2. This figure represents a wide range of possible yet reasonable choices of θ and λ_0. From left to right, customers become more sensitive to the sunk cost. From bottom to top, the service system becomes more congested. The lighter blue without any numbers represents the region where the optimal policy is not to provide queue length information and not to use any sunk cost (i.e., **None**). The darker blue without any numbers represents the region where the optimal policy is to provide queue length information but not to use any sunk cost (i.e., **Info Only**). The remaining parts represent the region where the optimal policy is to provide queue length information and to use a non-zero sunk cost (i.e., **Both**), with the numbers in the cell showing the optimal level of sunk cost (relative to a unit reward). We observe that the range of θ and λ_0 values for which using a non-zero sunk cost is optimal is quite large. This indicates that in most practically relevant operating regions, it is beneficial to inject some sunk cost to mitigate service incompletion. We note that for the systems studied in Figure 7, when $\theta \leq 0.14$, $T^I < T^{UI}$. Thus, the optimal policy switches from **None** to **Info Only** when $\theta \leq 0.14$. When $\theta \geq 0.32$, $T^I > 2$. Thus, we do not get to see the **Info Only** region for $\theta \geq 0.32$ in the figure.

We also make several observations about the optimal sunk cost to use, which we denote as $s^\star(\lambda_0, \theta)$ to mark its dependence on λ_0 and θ explicitly. In Figure 7, for a fixed value of θ, when it is optimal to use a non-zero sunk cost, the optimal sunk cost is

$$s^\star(\lambda_0, \theta) = s_1 = \frac{cn_0 - R}{\theta}. \tag{6}$$

Note that $(cn_0 - R)/\theta$ does not vary with λ_0 and it decreases as θ increases. In Appendix E, we test a larger range of λ_0 and θ values, probably beyond what one would expect to see in practice (see Figure 19). We also test other values for the waiting cost c (see Figure 20). We observe that in all cases, as θ increases, T^{UI} (i.e., the boundary between **None** and **Both**) is non-increasing and T^I (i.e., the boundary between **Both** and **Info Only**) is non-decreasing. In other words, the region of

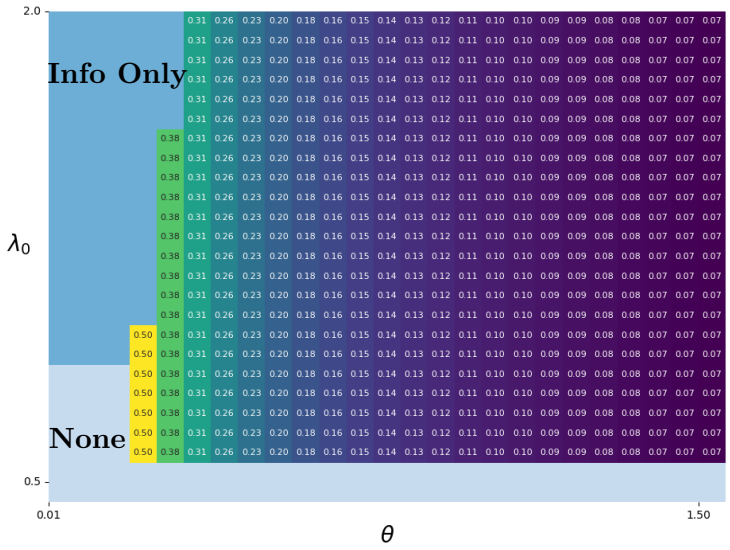


Figure 7 **Optimal policy for the service provider when $R = 1$ and $c = 1.1$. The sensitivity to sunk cost parameter θ ranges from 0.01 to 1.50. The initial arrival rate λ_0 ranges from 0.5 to 2.0. Lighter blue without numbers indicates the optimal policy is None. Darker blue without numbers indicates the optimal policy is Info Only. The remaining region indicates the optimal policy is Both and the numbers indicate the optimal sunk cost.**

λ_0 for which the service provider should use a non-zero sunk cost to maximize the throughput is expanding as customers are more sensitive to sunk cost. For $\lambda_0 \in (T^{UI}, T^I)$, the optimal sunk cost may first decrease with λ_0 and then stay at s_1. In particular, we observe that for $\lambda_f(s_1) \le \lambda_0 < T^I$, $s^*(\lambda_0, \theta) = s_1$. For $T^{UI} < \lambda_0 < \lambda_f(s_1)$, $s^*(\lambda_0, \theta)$ is non-increasing in λ_0. (For the systems studied in Figure 7, when $\theta \ge 0.14$, $\lambda_f(s_1) < T^{UI}$. Thus, we see that when it is optimal to use a non-zero sunk cost, $s^*(\lambda_0, \theta) = s_1$ in the figure.)

7. Heterogeneous Customers

Our main model assumes that customers are homogeneous. In this section, we extend the model to include two classes of customers: those who receive a high reward R^H and those who receive a low reward R^L after receiving the service, where $R^H > R^L$. We further assume that all customers share the same waiting cost c and sensitivity to sunk cost θ. We use λ_{0H} (λ_{0L}) to denote the initial arrival rate of high (low) reward customers, λ_{e1H} (λ_{e1L}) to denote the equilibrium arrival rate of high (low) reward customers after the first stage decision, and λ_{e2H} (λ_{e2L}) to denote the equilibrium arrival rate of high (low) reward customers after the second stage decision. In addition, we denote $\lambda_0 = \lambda_{0H} + \lambda_{0L}$, $\lambda_{e1} = \lambda_{e1H} + \lambda_{e1L}$, and $\lambda_{e2} = \lambda_{e2H} + \lambda_{e2L}$.

We assume that the policy implemented is independent of customer type, i.e., we have to apply the same policy to both classes of customers. We characterize the optimal policy for the service provider for different initial arrival rates in the following theorem.

Theorem 4. *For fixed R^H, R^L, c, and θ, the following hold:*

1. *If $\lambda_0 \leq \frac{R^L}{c+R^L}$, $\max_{s\geq 0,\gamma\in\{I,U\}} \lambda_{e2}^{\gamma}(s) = \lambda_{e2}^{U}(0)$.*
2. *If $\lambda_{0H} \geq 1$, $\max_{s\geq 0,\gamma\in\{I,U\}} \lambda_{e2}^{\gamma}(s) = \max_{s\geq 0} \lambda_{e2}^{I}(s)$.*

The proof of Theorem 4 is in Appendix D. Theorem 4 shows that if the initial arrival rate to the system is low (i.e., below $R^L/(c+R^L)$), then the optimal strategy for the service provider is to neither provide queue length information nor inject sunk cost, which we refer to as the **None** policy. If the initial arrival rate to the system of the high reward customers is high (i.e., above 1), then the optimal strategy for the service provider is to provide real-time queue length information. In this case, the optimal sunk cost may or may not equal zero. Figure 8 provides a numerical demonstration of the structure of the optimal policy for different system loads. In this figure, we set $R^H = 1.5$, $R^L = 1.0$, $c = 1.1$, and $\theta = 0.4$, while have both λ_{0H} and λ_{0L} ranging from 0.2 to 2. We observe that if the initial arrival rate to the system is very low, then the optimal strategy for the service provider is **None**. If the initial arrival rate to the system is very high, then the optimal strategy is to provide real-time queue length information only and do not impose any sunk cost, which we refer to as the **Info Only** policy. If the initial arrival rate to the system is moderate, then the optimal strategy for the service provider is to both provide real-time queue length information and use a non-zero sunk cost, which we refer to as the **Both** policy. The optimal sunk cost is identical (and equal to 0.25) in all the cases shown in the figure where **Both** is optimal. We also note that the region where **Both** is optimal is quite large, suggesting that in many practically

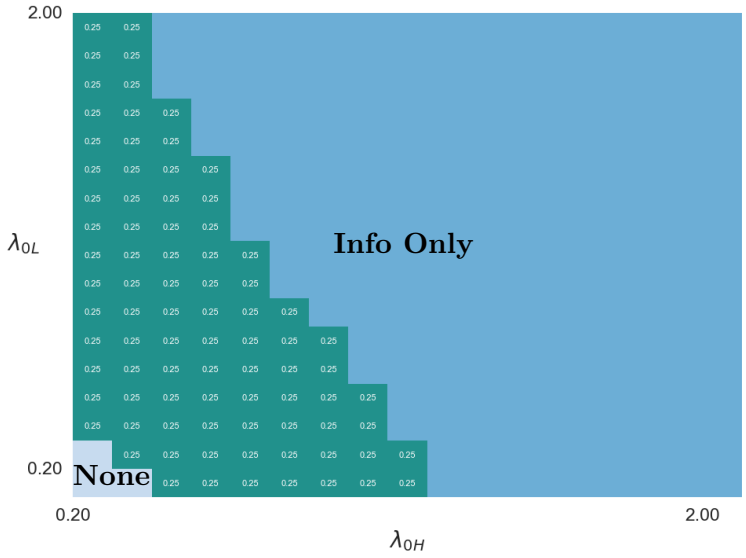


Figure 8 Optimal policy for the service provider when $R^H = 1.5$, $R^L = 1$, $c = 1.1$, and $\theta = 0.4$. The initial arrival rates of both classes of customers λ_{0H} and λ_{0L} ranges from 0.2 to 2.0. Lighter blue without numbers indicates the optimal policy is None. Darker blue without numbers indicates the optimal policy is Info Only. The region with numbers indicates the optimal policy is Both and the numbers indicate the optimal sunk cost.

relevant settings, it is beneficial to inject some sunk cost to mitigate service incompletion. These insights are similar to what we have derived for homogeneous customers in Section 6. We also try other parameters of R^H, R^L, c, θ, and observe similar trends in the results. Notably, when θ is smaller, the range of $(\lambda_{0L}, \lambda_{0H})$ values for which using a non-zero sunk cost is optimal narrows; conversely, as θ increases, the range of $(\lambda_{0L}, \lambda_{0H})$ values for which using a non-zero sunk cost is optimal expands.

8. Numerical Experiments

In this section, we demonstrate the performance gain from using a non-zero sunk cost. We also conduct sensitivity analysis on how estimation error in θ affects the performance of the sunk cost strategy and provide insights for managers on how to use the sunk cost in practice when one cannot estimate θ accurately.

8.1. Benefits of Sunk Cost

Figure 9 shows the percentage of improvement in the throughput when using a non-zero sunk cost. For the region where the optimal policy is **None** or **Info Only**, the improvement is defined to be zero. For the region where the optimal policy is **Both**, the improvement is defined as

$$\frac{\max_{s\geq 0} \lambda_{e2}^{I}(s) - \max_{\gamma\in\{I,U\}} \lambda_{e2}^{\gamma}(0)}{\max_{\gamma\in\{I,U\}} \lambda_{e2}^{\gamma}(0)}.$$

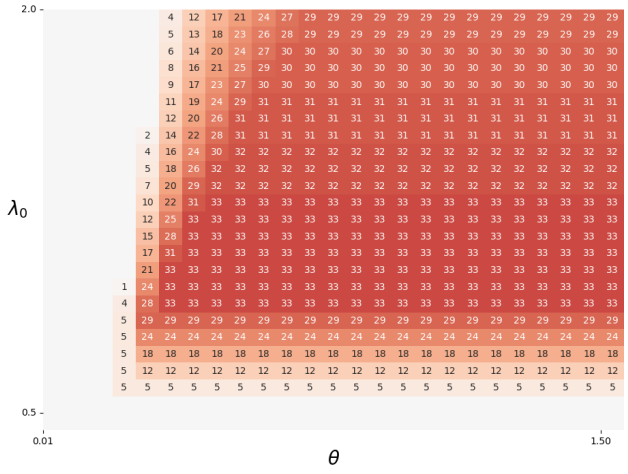


Figure 9 Improvement to the system throughput of the optimal policy against the suboptimal policy using zero sunk cost when $R=1$ and $c=1.1$. Numbers are in percentage scale.

For a large range of λ_0 and θ, the benefit of using a non-zero sunk cost is substantial: throughput can increase by as much as 33%. The sunk cost strategy is especially useful when the system is moderately congested and customers are sensitive to sunk cost (i.e., when $\theta \geq 0.2$). Figure 21 in Appendix E expands the range of λ_0 and the benefit of sunk cost is substantial even for larger λ_0. Figure 22 in Appendix E shows the benefit when $c=0.7$ and $c=1.5$ and the conclusions are similar. Using sunk costs as an operational lever can have meaningful benefits in many cases.

8.2. Sensitivity Analysis

The sunk cost fallacy is a psychological phenomenon and there is no consensus on how to accurately measure customers' sensitivity to sunk cost (i.e., parameter θ in our study). Experiments have been used to measure customers' relative susceptibility to the sunk cost fallacy (Soman 2001, Navarro and Fantino 2009, Ülkü et al. 2020). None of these approaches are directly applicable to our setting because they do not have a utility-based model to quantify the susceptibility. Indeed, customers' sensitivity to sunk cost could be different under different service contexts. One could potentially utilize the equilibrium solution in Theorem 2 to run a field experiment in a real business setting. In particular, if we make sure that the system is under-loaded in the sense that $\min\{\lambda_f, \lambda_0\} = \lambda_0$, then by comparing two joining probabilities in the second stage associated with two different levels of sunk cost, we can analytically solve for θ. The quality of the estimation then depends on the sample size. Although implementing repeated interactions of this nature in a real-effort online experiment is challenging (Allon and Kremer 2018), this paper provides evidence about the importance of sunk cost in mitigating service incompletion. Consequently, businesses may consider conducting field experiments to assess their customers' sensitivity to sunk costs. In this section, we consider the case when the service provider cannot accurately estimate θ and explore the impact of estimation error on the system performance. For the rest of this section, we will assume that μ, R, c, λ_0 are fixed constant. With a little abuse of notation, we write the equilibrium arrival rate following the second stage decision as $\lambda_{e2}^{\gamma}(s, \theta)$ to mark its dependence on s and θ explicitly.

First, we study the suboptimality gap when θ is overestimated or underestimated by a deviation parameter δ. We denote the true sensitivity to sunk cost as $\theta^\star$. Define

$$\begin{aligned} M_{\delta^+} &= \bigcup_{\theta \in \Theta^+} \left\{ (s,\gamma) \mid (s,\gamma) = \arg\max_{(s',\gamma')} \lambda_{e2}^{\gamma'}(s', \theta) \right\} \\ M_{\delta^-} &= \bigcup_{\theta \in \Theta^-} \left\{ (s,\gamma) \mid (s,\gamma) = \arg\max_{(s',\gamma')} \lambda_{e2}^{\gamma'}(s', \theta) \right\} \end{aligned} \tag{7}$$

where $\Theta^+ = [\theta^\star, (1+\delta)\theta^\star]$ and $\Theta^- = [(1-\delta)\theta^\star, \theta^\star]$. Define the suboptimality gap when overestimating $\theta^\star$ by a deviation of δ as

$$\frac{\max_{(s,\gamma)} \lambda_{e2}^{\gamma}(s, \theta^\star) - \min_{(s,\gamma) \in M_{\delta^+}} \lambda_{e2}^{\gamma}(s, \theta^\star)}{\max_{(s,\gamma)} \lambda_{e2}^{\gamma}(s, \theta^\star)}.$$

The suboptimality gap when underestimating $\theta^\star$ by a deviation of δ is similarly defined by replacing M_{δ^+} with M_{δ^-}. In other words, we compare the worst-case throughput when applying the policy derived based on the wrongly estimated θ against the optimal throughput when θ^* is known. Figure 10 shows the suboptimality gaps when overestimating θ with a deviation of 1%, i.e., $\delta = 1\%$. Despite the small deviation, we note that the suboptimality gap could be substantial (i.e., as large as 25%).

On the other hand, Figure 11 shows the suboptimality gaps when underestimating θ by a deviation of 1%, 5%, and 20% respectively. When θ is 5% underestimated, the suboptimality gap is less than 5% (see, column (b) in Figure 11). Even when θ is 20% underestimated, the suboptimality gap is no larger than 15% (see, column (c) in Figure 11). This asymmetry between over- and under-estimated θ mainly comes from the discontinuity in s of the equilibrium strategy of informed customers in the second stage, i.e., q_2^I. In particular, q_2^I jumps upward at s_k. The throughput could be substantially smaller when using $s_k - \Delta$ for a very small $\Delta > 0$, because we miss the jump upward at s_k. See, for example, the jump at s_1 in Figure 5. Since s_k decreases as θ increases (see (5)) and the optimal sunk cost is one of the s_k's, the service provider uses a smaller-than-optimal sunk cost when overestimating θ, leading to a sharp decrease in the throughput.

We also test other values of system parameters (see, Figures 23 and 24 in Appendix E for different values of c). Overall, if the service provider is not able to accurately estimate the customer's sensitivity to sunk cost, it is recommended that they use a lower bound of the estimate.

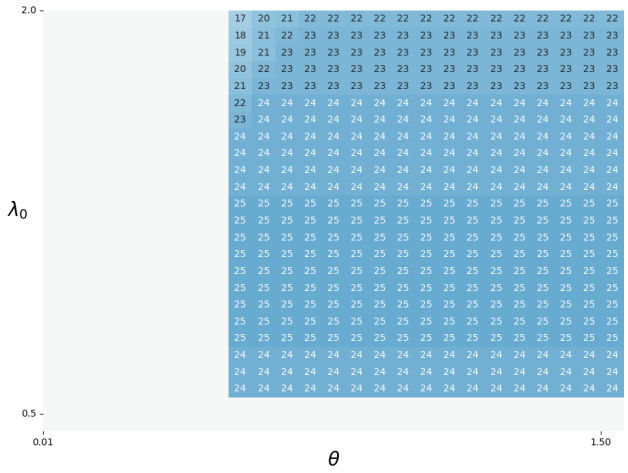
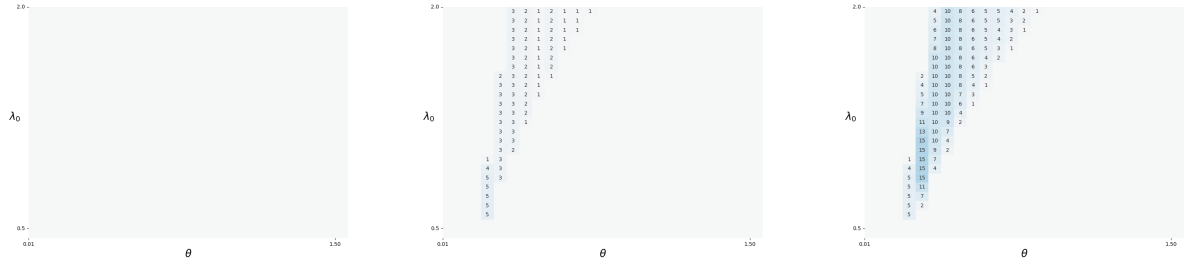


Figure 10 **Worst-case suboptimality gap of using policies derived from overestimated θ (by as large as 1% deviation) when $R = 1$ and $c = 1.1$. Numbers are in percentage scale.**



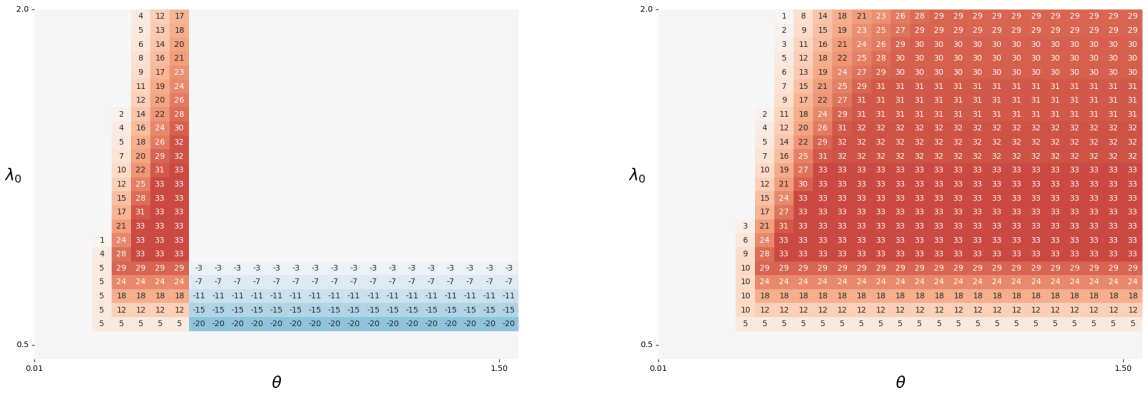
(a) 1% deviation (b) 5% deviation (c) 20% deviation

Figure 11 **Worst-case suboptimality gap of using policies derived from underestimated θ (by as large as 1%, 5%, and 20% deviations, respectively) when $R = 1$ and $c = 1.1$. Numbers are in percentage scale.**

Next, we compare to using a zero sunk cost when θ cannot be estimated accurately. Define the worst-case improvement when overestimating $\theta^\star$ by a deviation of δ as

$$\frac{\min_{(s,\gamma)\in M_{\delta+}} \lambda_{e2}^{\gamma}(s,\theta^\star) - \max_{(s,\gamma)} \lambda_{e2}^{\gamma}(0,\theta^\star)}{\max_{(s,\gamma)} \lambda_{e2}^{\gamma}(0,\theta^\star)},$$

where $M_{\delta+}$ is defined in (7). The worst-case improvement when underestimating $\theta^\star$ by a deviation of δ is similarly defined by replacing $M_{\delta+}$ with $M_{\delta-}$. In other words, we are comparing the worst-case throughput when θ is wrongly estimated against the throughput when using a zero sunk cost (i.e., either **None** policy or **Info Only** policy). The first plot in Figure 12 shows the worst-case improvement when overestimating θ by 1%. In the figure, red represents the region where the worst-case throughput outperforms the throughput when using zero sunk cost. Blue represents the region where the worst-case throughput underperforms the throughput when using a zero sunk cost. We observe that when θ is slightly overestimated, the worst-case throughput when using the wrong policy could be worse than not using any sunk cost by as much as 20%. Such a situation happens when the system is less congested (i.e., $\lambda_0 = 0.62$) and customers are sensitive to sunk cost. The second plot in Figure 12 shows the worst-case improvement when underestimating θ by 20%. We observe that even when θ is underestimated by as much as 20%, using a non-zero sunk cost still outperforms using a zero sunk cost. Overall, we recommend that the service provider use a conservative underestimate of θ when θ cannot be accurately estimated. Figures 25 in Appendix E shows the improvement of the worst-case throughput based on the wrong estimates of θ for $c = 0.7$ and $c = 1.5$ respectively, and the conclusions are similar.



(a) overestimated by 1% (b) underestimated by 20%

Figure 12 Worst-case improvement to the system throughput of using policies derived from overestimated θ (by as large as 1%) or underestimated θ (by as large as 20%) against the policy using zero sunk cost when $R = 1$ and $c = 1.1$. Blue indicates that the worst-case policy underperforms the baseline suboptimal policy. Numbers are in percentage scale.

9. Conclusions

In this work, we study the sunk cost effect in a service system setting and show that a service provider can optimally utilize this effect to mitigate service incompletion and achieve a higher system throughput.

We run a controlled experiment and show evidence that a higher sunk cost leads to a lower likelihood of abandonment. To study the optimal sunk cost strategy in real service systems, where customers can form a reasonable expectation about the equilibrium waiting time through multiple interactions with the system, and reject high sunk cost before incurring it, e.g., by seeking service elsewhere, we study a game-theoretic queueing model with sunk cost. We derive customers' equilibrium behavior and find that the proportion of customers balking the system in the second stage, i.e., after incurring the sunk cost, decreases with the sunk cost. However, the proportion of customers joining the system in the first stage, i.e., before incurring the sunk cost, decreases with the sunk cost. When the real-time waiting information is not provided, the combined effect in the two stages is negative, i.e., the system throughput decreases with the sunk cost. When real-time waiting information is available, a non-zero sunk cost can lead to a higher throughput than no sunk cost when the system is not too overloaded. We then study the service provider's optimal policy. We show that a non-zero sunk cost combined with delay announcement can substantially improve the system throughput when the system is moderately loaded and customers' sensitivity to sunk cost is relatively high. In addition, when the sensitivity to sunk cost cannot be accurately estimated, it is recommended to underestimate the parameter rather than overestimate it for a more robust performance improvement.

Our work has some limitations that warrant future study. First, our online experiment is not able to fully capture the dynamics of the service system that we model. In particular, the experiment is a one-shot interaction, rather than multiple interactions between the participants and the system. In addition, the participants do not interact with each other, i.e., there are no queueing dynamics involved. As such, while our experiment is meaningful in showing the sunk cost effect in a novel time domain, it is not able to capture the queueing dynamics and participants' equilibrium behavior. It is in general challenging to implement a real-effort experiment in a repeated game (Allon and Kremer 2018). Second, although we provide some insights for service providers to estimate the customers' sensitivity parameter to sunk cost, further empirical investigations are needed to accurately quantify the sunk cost effect. Lastly, while our stylized queueing model captures some core trade-offs in a congested service system setting, extensions of the model that capture various complications in real service systems should be explored. Examples include more general customer heterogeneity, time-varying demand, and wait-dependent service requirements, etc.

References

Akşin Z, Ata B, Emadi SM, Su CL (2017) Impact of delay announcements in call centers: An empirical approach. *Operations Research* 65(1):242–265.

Allon G, Bassamboo A, Gurvich I (2011) "we will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations research* 59(6):1382–1394.

Allon G, Kremer M (2018) Behavioral foundations of queueing systems. *The handbook of behavioral operations* 9325:325–366.

Althenayyan A, Cui S, Ulku S, Yang L (2022) Not all lines are skipped equally: an experimental investigation of line-sitting and express lines. *Georgetown McDonough School of Business Research Paper* (4179751).

Anderson-Bergman C (2017) icenreg: regression models for interval censored data in r. *Journal of Statistical Software* 81:1–23.

Arkes HR, Blumer C (1985) The psychology of sunk cost. *Organizational behavior and human decision processes* 35(1):124–140.

Armony M, Maglaras C (2004) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research* 52(2):271–292.

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Operations Research* 57(1):66–81.

Baliga S, Ely JC (2011) Mnemonomics: The sunk cost fallacy as a memory kludge. *American Economic Journal: Microeconomics* 3(4):35–67.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

Buell RW (2021) Last-place aversion in queues. *Management Science* 67(3):1430–1452.

Chen H, Frank M (2004) Monopoly pricing when customers queue. *IIE Transactions* 36(6):569–581.

Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–202.

Cui S, Su X, Veeraraghavan S (2019) A model of rational retrials in queues. *Operations Research* 67(6):1699–1718.

Dick AS, Lord KR (1998) The impact of membership fees on consumer attitude and choice. *Psychology & Marketing* 15(1):41–58.

Dube L, Renaghan LM (1994) Measuring customer satisfaction for strategic management: For financial success, a restaurant's management must make the connection between service attributes and return patronage. here's a way to establish that connection. *Cornell Hotel and Restaurant Administration Quarterly* 35(1):39–47.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* 16(1):13–39.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* 53(6):962–970.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.

Hassin R (2016) *Rational queueing* (CRC press).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Hassin R, Roet-Green R (2020) On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* 23(4):989–1004.

Hathaway BA, Kagan E, Dada M (2023) The gatekeeper's dilemma:"when should i transfer this customer?". *Operations Research* 71(3):843–859.

He YT, Down DG (2009) On accommodating customer flexibility in service systems. *INFOR: Information Systems and Operational Research* 47(4):289–295.

Ho TH, Png IP, Reza S (2018) Sunk cost fallacy in driving the world's costliest cars. *Management Science* 64(4):1761–1778.

Hong F, Huang W, Zhao X (2019) Sunk cost as a self-management device. *Management Science* 65(5):2216–2230.

Hu M, Li Y, Wang J (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Science* 64(6):2650–2671.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing & Service Operations Management* 15(2):263–279.

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing* 60(2):81–90.

Ibrahim R (2018a) Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management* 27(2):234–250.

Ibrahim R (2018b) Sharing delay information in service systems: a literature survey. *Queueing Systems* 89(1):49–79.

Ibrahim R, Armony M, Bassamboo A (2017) Does the past predict the future? the case of delay announcements in service systems. *Management Science* 63(6):1762–1780.

Ibrahim R, Estrada Rodriguez A, Zhan D (2024) On customer (dis) honesty in priority queues: The role of lying aversion. *Management Science* .

Jain S, Chen H (2023) Sunk cost bias and time inconsistency: A strategic analysis of pricing decisions. *Management Science* 69(4):2383–2400.

Jouini O, Akşin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* 13(4):534–548.

Kanodia C, Bushman R, Dickhaut J (1989) Escalation errors and the sunk cost effect: An explanation based on reputation and information asymmetries. *Journal of Accounting research* 27(1):59–77.

Kim SH, Tong J, Peden C (2020) Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Management Science* 66(11):5151–5170.

Kremer M, de Véricourt F (2023) Mismanaging diagnostic accuracy under congestion. *Operations Research* 71(3):895–916.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Science* 62(10):3023–3038.

Kruger J, Dunning D (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77(6):1121.

Larson RC (1987) Or forum—perspectives on queues: Social justice and the psychology of queueing. *Operations research* 35(6):895–905.

Lee C, Ward AR (2019) Pricing and capacity sizing of a service facility: Customer abandonment effects. *Production and Operations Management* 28(8):2031–2043.

Li DR, Brennan JJ, Kreshak AA, Castillo EM, Vilke GM (2019) Patients who leave the emergency department without being seen and their follow-up behavior: a retrospective descriptive analysis. *The Journal of emergency medicine* 57(1):106–113.

Lingenbrink D, Iyer K (2019) Optimal signaling mechanisms in unobservable queues. *Operations research* 67(5):1397–1416.

Liu N, van Jaarsveld W, Wang S, Xiao G (2023) Managing outpatient service with strategic walk-ins. *Management Science* 69(10):5904–5922.

Luo J, Valdés L, Linardi S (2022) Experienced and prospective wait in queues: A behavioral investigation. *Available at SSRN 4169028* .

Mandelbaum A, Shimkin N (2000) A model for rational abandonments from invisible queues. *Queueing Systems* 36:141–173.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society* 15–24.

Navarro AD, Fantino E (2009) The sunk-time effect: An exploration. *Journal of behavioral decision making* 22(3):252–270.

Nozick R (1994) The nature of rationality .

Plambeck EL, Wang Q (2013) Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Science* 59(8):1927–1946.

Pronin E (2007) Perception and misperception of bias in human judgment. *Trends in cognitive sciences* 11(1):37–43.

Pronin E, Lin DY, Ross L (2002) The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin* 28(3):369–381.

Ren H, Huang T (2018) Modeling customer bounded rationality in operations management: A review and research opportunities. *Computers & Operations Research* 91:48–58.

Schectman JM, Schorling JB, Voss JD (2008) Appointment adherence and disparities in outcomes among patients with diabetes. *Journal of general internal medicine* 23:1685–1687.

Schulz AKD, Cheng MM (2002) Persistence in capital budgeting reinvestment decisions–personal responsibility antecedent and information asymmetry moderator: A note. *Accounting & Finance* 42(1):73–86.

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: The behavioral impact of queueing design on service time. *Management Science* 64(1):453–473.

Soman D (2001) The mental accounting of sunk time costs: Why time is not like money. *Journal of behavioral decision making* 14(3):169–185.

Soman D, Cheema A (2001) The effect of windfall gains on the sunk-cost effect. *Marketing Letters* 12:51–62.

Staw BM (1976) Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational behavior and human performance* 16(1):27–44.

Staw BM (1981) The escalation of commitment to a course of action. *Academy of management Review* 6(4):577–587.

Thaler R (1980) Toward a positive theory of consumer choice. *Journal of economic behavior & organization* 1(1):39–60.

Ülkü S, Hydock C, Cui S (2020) Making the wait worthwhile: Experiments on the effect of queueing on consumption. *Management Science* 66(3):1149–1171.

Wang J, Hu M (2020) Efficient inaccuracy: User-generated information sharing in a queue. *Management Science* 66(10):4648–4666.

Yang L, Debo LG, Gupta V (2019) Search among queues under quality differentiation. *Management Science* 65(8):3605–3623.

Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? an empirical study. *Management Science* 63(1):1–20.

Yu Q, Allon G, Bassamboo A (2021) The reference effect of delay announcements: A field experiment. *Management Science* 67(12):7417–7437.

Yu Q, Allon G, Bassamboo A, Iravani S (2018) Managing customer expectations and priorities in service systems. *Management Science* 64(8):3942–3970.

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production and Operations Management* 23(5):788–801.

Zhang X, Iyer G, Xu X, Chong JK (2023) Sunk cost effect, self-control, and contract design. *Journal of Marketing Research* 00222437231196824.

Appendix A: Experiment Details

In this section, we present more details about the online experiment we conducted on Amazon Mechanical Turk. We advertised the experiment as image classification tasks. Figure 13 shows the instructions to the subjects before they officially agree to participate. The first stage is the qualification tasks and a typical task is shown in Figure 14. Participants are randomly assigned to a low sunk cost group and a high sunk cost group in this stage. The second stage is the manipulated waiting period. Participants will wait on the page shown in Figure 15, assuming that they are being verified about their eligibility to do the main task. The third stage is the attention checking buffer, that is, after the manipulated waiting, participants will have to proceed within one minute to the main task. In this stage, participants will see a countdown clock as shown in Figure 16. Finally, in the fourth stage, participants are asked to tell between chihuahuas and muffins, shown in Figure 17.

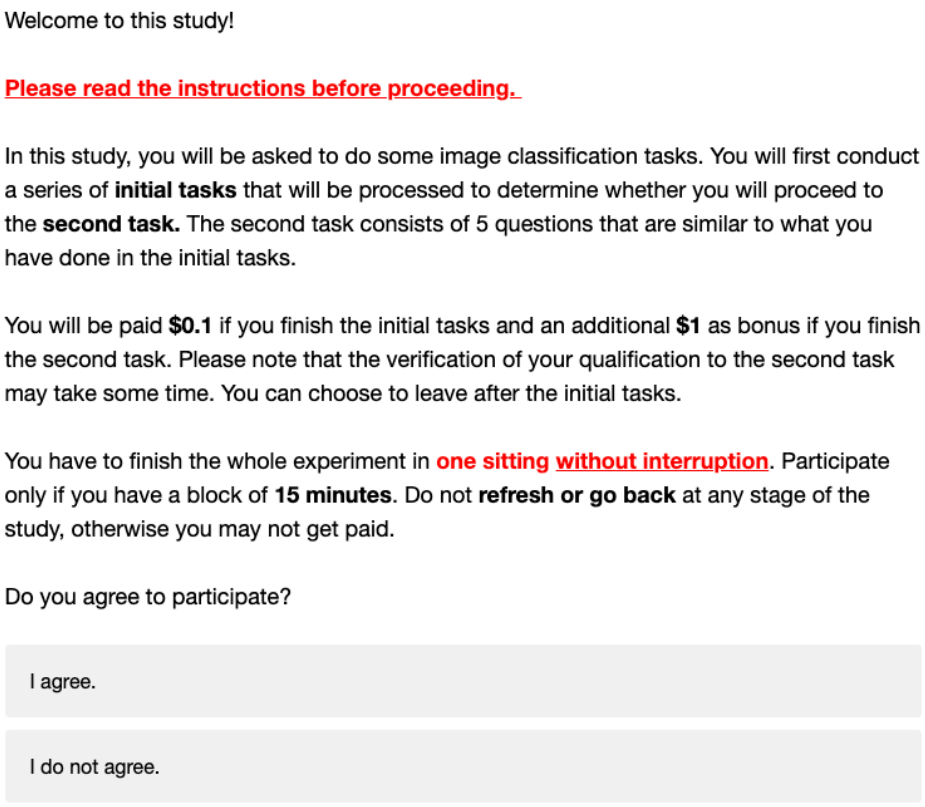


Figure 13 Introduction page of the online experiment.

We next conduct a sensitivity analysis to check the robustness of our estimation results. In our experiments, some participants quit during the qualification task. In our main analysis in Section 3.2, we exclude these participants. However, there is a concern that the inherently more patient participants (e.g., participants with a lower cost of waiting) are more likely to finish the longer qualification task, and they are also more likely to wait for the main reward. Thus, there can be a sample selection bias. To assess the effect of this potential bias, we consider a "worst-case" hypothetical scenario where we assume all participants who quit

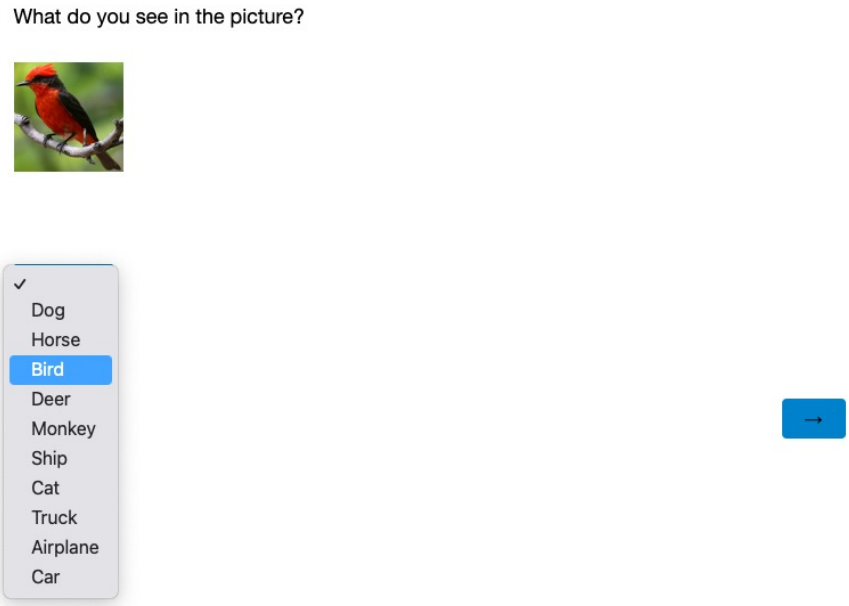


Figure 14 **The qualification task (first stage) of the online experiment.**

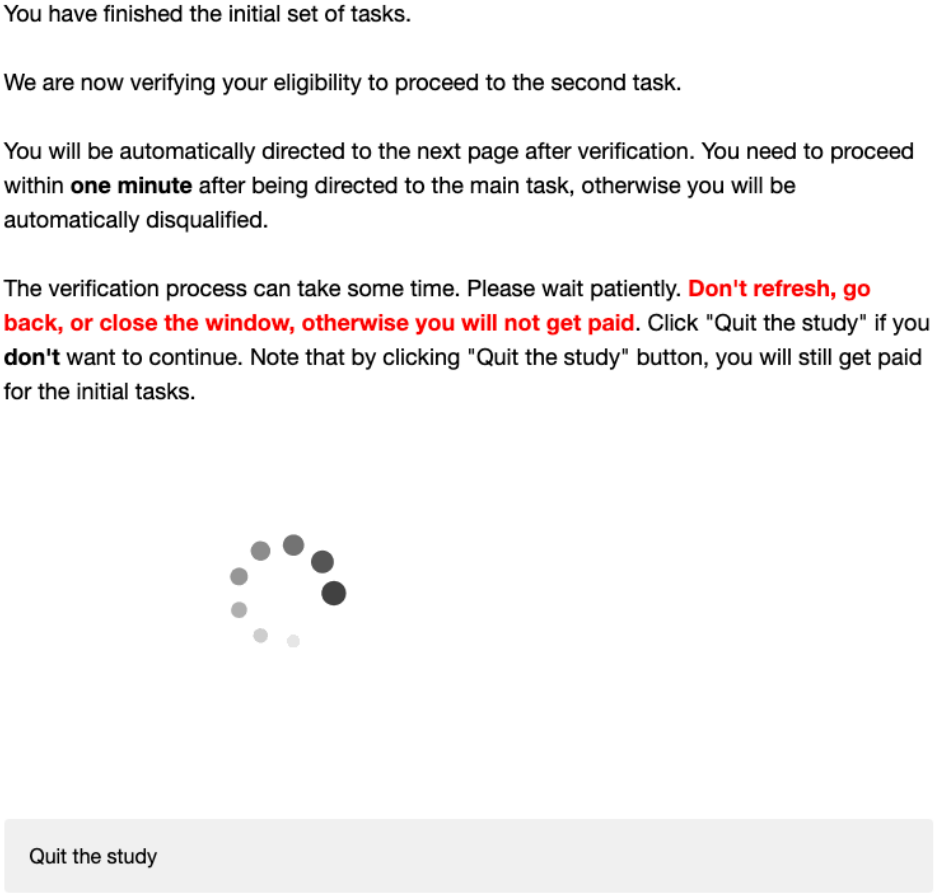


Figure 15 **The manipulated waiting period (second stage) of the online experiment.**

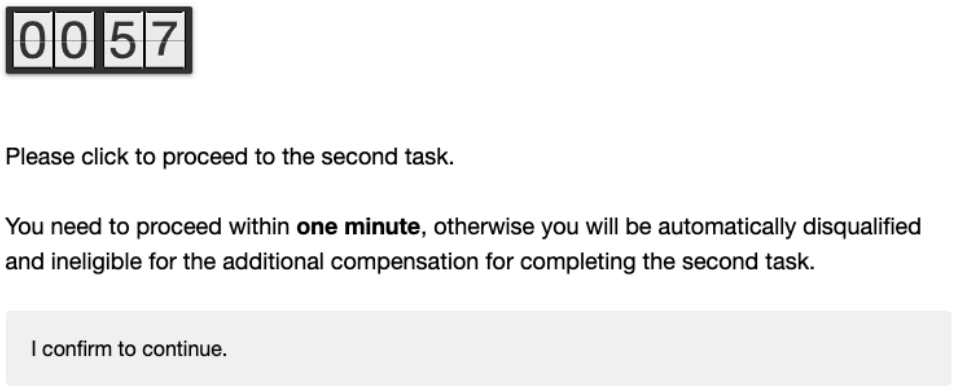


Figure 16 **The waiting buffer (third stage) page of the online experiment.**

Figure 17 **The "main" task (fourth stage) of the online experiment. The pictures are presented one by one to the participants, who are asked if they see a "chihuahua" or "muffin".**

the qualification task in the low sunk cost group (there are 5 of them) incurred a low sunk cost and waited for the main reward. On the other hand, we assume all participants who quit the qualification task in the high sunk cost group (there are 13 of them) incurred a high sunk cost but did not wait for the main reward, i.e., they abandoned while waiting. In this case, 56.1% (119/212) of the participants who incurred a low sunk cost abandoned while waiting, which is smaller than the 57.5% abandonment rate in our main analysis, and 45.5% (97/213) participants who incurred a high sunk cost abandoned while waiting, which is higher than the 42% abandonment rate in our main analysis. The two-proportion Z-test of this conservative imputation still shows that injecting a higher sunk cost significantly decreases the abandonment rate at a 5% significance level (p-value: 0.037).

Appendix B: Proofs of the Results in Section 5

In this section, since we focus on either the uninformed or informed system in each proof, for the brevity of notations, we suppress the superscript "U" or "I" in λ's and q's when there is no ambiguity.

Proof of Lemma 1. For an $M/M/1$ queue with service rate $\mu = 1$, let $W(\lambda)$ denote the steady-state average waiting time when the arrival rate is λ. Then, for $\lambda < 1$,

$$W(\lambda) = \frac{\lambda}{1-\lambda}.$$

Recall that λ_{e1} is the equilibrium arrival rate after the first stage decision. We first note that if $R - cW(\lambda_{e1}) \geq -\theta s$, then $q_2 = 1$, i.e., all uninformed customers who incur the sunk cost will wait in the second stage. The condition $R - cW(\lambda_{e1}) \geq -\theta s$ is equivalent to

$$\lambda_{e1} \leq \frac{R+\theta s}{c+R+\theta s}.$$

On the other hand, if

$$\lambda_{e1} > \frac{R+\theta s}{c+R+\theta s},$$

only a fraction of uninformed customers who incur the sunk cost will wait in the second stage. In this case, the equilibrium arrival rate after the second stage decision, $\lambda_{e2} = \lambda_{e1} q_2$, solves:

$$R - cW(\lambda_{e2}) = -\theta s \Leftrightarrow \lambda_{e2} = \frac{R+\theta s}{c+R+\theta s}.$$

Putting this all together, given λ_{e1},

$$\begin{aligned}
\lambda_{e2} &= \lambda_{e1} \mathbb{I}\left\{\lambda_{e1} \leq \frac{R+\theta s}{c+R+\theta s}\right\} + \frac{R+\theta s}{c+R+\theta s}\mathbb{I}\left\{\lambda_{e1} > \frac{R+\theta s}{c+R+\theta s}\right\} = \min\left\{\lambda_{e1}, \frac{R+\theta s}{c+R+\theta s}\right\}, \\
q_2 &= \mathbb{I}\left\{\lambda_{e1} \leq \frac{R+\theta s}{c+R+\theta s}\right\} + \frac{R+\theta s}{\lambda_{e1}(c+R+\theta s)}\mathbb{I}\left\{\lambda_{e1} > \frac{R+\theta s}{c+R+\theta s}\right\} = \min\left\{1, \frac{R+\theta s}{\lambda_{e1}(c+R+\theta s)}\right\}, \\
\bar{W}^U &= \frac{\lambda_{e1}}{1-\lambda_{e1}}\mathbb{I}\left\{\lambda_{e1} \leq \frac{R+\theta s}{c+R+\theta s}\right\} + \frac{R+\theta s}{c}\mathbb{I}\left\{\lambda_{e1} > \frac{R+\theta s}{c+R+\theta s}\right\} = \min\left\{\frac{\lambda_{e1}}{1-\lambda_{e1}}, \frac{R+\theta s}{c}\right\}.
\end{aligned}$$

□

Proof of Lemma 2. In the first stage, the uninformed customers decide whether to incur the sunk cost by comparing the expected utility of joining the system, $-s + \hat{U}_2^U$, to zero, i.e.,

$$U_1^U = \max_{q \in [0,1]} q(-s + \hat{U}_2^U),$$

where $\hat{U}_2^U = \max_{\hat{q} \in [0,1]} \hat{q}(R - c\bar{W}^U)$ is customers' *belief* about the expected utility in the second stage. Note that we assume customers learn the steady-state average waiting time through repeated interaction with the system. Thus, $\bar{W}^U$ is the average waiting time characterized in Lemma 1. On the other hand, customers believe they will be rational in the second stage when making the first-stage decision. Thus, they assume they will wait in the second stage if the expected utility of waiting, $R - c\bar{W}^U$, is no less than 0.

Given the equilibrium arrival rate after the first stage decision λ_{e1}, we note from Lemma 1 that if $\lambda_{e1} \leq R/(c+R)$, $R - c\bar{W}^U \geq 0$; otherwise $R - c\bar{W}^U < 0$. Thus, we have the following characterization of an uninformed customer's *belief* in the first stage about the probability of continuing to wait in the second stage

$$\hat{q}_2 = \mathbb{I}\left\{\lambda_{e1} \leq \frac{R}{c+R}\right\}.$$

Then,

$$\hat{U}_2^U = \begin{cases} R - c\frac{\lambda_{e1}}{1-\lambda_{e1}} & \text{if } \lambda_{e1} \leq \frac{R}{c+R} \\ 0 & \text{if } \lambda_{e1} > \frac{R}{c+R}. \end{cases}$$

Note that if $\hat{U}_2^U = 0$, no customer will join the system in the first stage.

Next if $-s + R - c\frac{\lambda_0}{1-\lambda_0} \geq 0$, i.e.,

$$\lambda_0 \leq \frac{R-s}{c+R-s},$$

$q_1 = 1$, i.e., all customers will join the system in the first stage. Otherwise, the equilibrium arrival rate after the first stage decision, $\lambda_{e1} = \lambda_0 q_1$, solves

$$-s + R - c\frac{\lambda_{e1}}{1-\lambda_{e1}} = 0,$$

which leads to

$$\lambda_{e1} = \frac{R-s}{c+R-s} \text{ and } q_1 = \frac{R-s}{\lambda_0(c+R-s)}.$$

□

Proof of Theorem 1. For all possible values of λ_{e1} in Lemma 2, we have $\lambda_{e1} \leq (R-s)/(c+R-s)$. This implies that

$$\lambda_{e1} \leq \frac{R+\theta s}{c+R+\theta s}.$$

Then, by Lemma 1, we have $q_2 = 1$.

Note that there is some ambiguity when $s = 0$. In this case, the first- and second-stage decisions collapse into a single decision. Without loss of generality, when $s = 0$, we set $q_1 = \frac{R}{\lambda_0(c+R)}$ and $q_2 = 1$.

Combining the above analysis with the results in in Lemma 2, we have if $\lambda_0 \leq (R-s)/(c+R-s)$, the unique equilibrium is $q_1 = 1$ and $q_2 = 1$. If $\lambda_0 > (R-s)/(c+R-s)$, the unique equilibrium is

$$q_1 = \frac{R-s}{\lambda_0(c+R-s)} \text{ and } q_2 = 1.$$

□

Proof of Lemma 3. The informed customers will wait in the second stage as long as $R - cN \geq 0$, where N is the observed queue length (number of customers in the system) upon the arrival of the focal customer at the second stage. Equivalently, customers will wait as long as $N < n_s = \lfloor \frac{R+\theta s}{c} \rfloor + 1$. In this case, the system evolves as an $M/M/1/n_s$ queue, i.e., a single server queue with a finite waiting room. The balance equations of the steady-state queue length are

$$\begin{aligned}
\lambda_{e1}\pi_0 &= \pi_1 \\
\lambda_{e1}\pi_0 + \pi_2 &= (\lambda_{e1}+1)\pi_1 \\
&\cdots \\
\lambda_{e1}\pi_{n_s-2} + \pi_{n_s} &= (\lambda_{e1}+1)\pi_{n_s-1} \\
\lambda_{e1}\pi_{n_s-1} &= \pi_{n_s}.
\end{aligned}$$

Solving the balance equations gives

$$\pi_i = (\lambda_{e1})^i \frac{1-\lambda_{e1}}{1-(\lambda_{e1})^{n_s+1}} \text{ for } 0 \leq i \leq n_s \text{ when } \lambda_{e1} \neq 1$$

and

$$\pi_i = \frac{1}{n_s+1} \text{ for } 0 \leq i \leq n_s \text{ when } \lambda_{e1} = 1.$$

The proportion of informed customers who continue to wait in the second stage is thus

$$q_2 = 1 - \pi_{n_s} = \frac{1-(\lambda_{e1})^{n_s}}{1-(\lambda_{e1})^{n_s+1}} \text{ when } \lambda_{e1} \neq 1$$

and

$$q_2 = 1 - \pi_{n_s} = \frac{n_s}{n_s+1} \text{ when } \lambda_{e1} = 1.$$

□

Lemma 6. *For*

$$f(\lambda; s) := -s + \mathbb{E}[\hat{U}_2^I] = -s + \sum_{i=0}^{n_0-1} (R - ci)\pi_i$$

where

$$\pi_i = \begin{cases} (\lambda)^i \frac{1-\lambda}{1-(\lambda)^{n_s+1}} & \text{for } i = 0, 1, \cdots, n_s \quad \text{if } \lambda \neq 1 \\ \frac{1}{n_s+1} & \text{if } \lambda = 1 \end{cases}, \quad n_s = \left\lfloor \frac{R+\theta s}{c} \right\rfloor + 1, \quad n_0 = \left\lfloor \frac{R}{c} \right\rfloor + 1,$$

when $s > 0$, the equation $f(\lambda; s) = 0$ has a unique solution $\lambda_f \in (0, \infty)$ and

$$f(\lambda; s) \begin{cases} > 0 & \text{if } \lambda < \lambda_f \\ < 0 & \text{if } \lambda > \lambda_f. \end{cases}$$

When $s = 0$, $f(\lambda; s) \geq 0$ for all $\lambda > 0$.

Proof. For abbreviation of notations, we write n_0 as n in the following proof. When $\lambda = 1$,

$$\begin{aligned} f(\lambda; s) &= -s + \sum_{i=0}^{n-1} (R - ci) \frac{1}{n_s+1} \\ &= -s + \frac{n}{n_s+1} R - \frac{c}{n_s+1} \frac{c(n-1)}{2} \\ &= \frac{-s(n_s+1) + nR - \frac{cn(n-1)}{2}}{n_s+1}. \end{aligned}$$

Consider $k(s) = s(n_s+1)$, which is monotonically increasing in s. Since $k(0) = 0$ and $\lim_{s\to\infty} k(s) = \infty$, $k(s) = nR - cn(n-1)/2$ has a unique solution which we denote as $\hat{s}$. For $s > \hat{s}$,

$$\begin{aligned} f(1; s) &= -s + \frac{n}{n_s+1} R - \frac{n}{n_s+1} \frac{c(n-1)}{2} \\ &= -s + \frac{n}{n_s+1} (R - \frac{c(n-1)}{2}) \\ &< -s + \frac{s(n_s+1)}{n_s+1} \\ &= 0 \end{aligned}$$

Similarly, for $s < \hat{s}$, $f(1; s) > 0$.

For $s \neq \hat{s}$ and $s > 0$, when $\lambda \neq 1$,

$$\begin{aligned} f(\lambda; s) &= -s + \sum_{i=0}^{n-1} (R - ci)\pi_i \\ &= -s + \frac{R(1-\lambda^n) - c[\frac{\lambda - \lambda^n}{1-\lambda} - (n-1)\lambda^n]}{1-\lambda^{n_s+1}} \\ &= \frac{-s\lambda^{n_s+2} + s\lambda^{n_s+1} + (R - c(n-1))\lambda^{n+1} + (cn - R)\lambda^n + (s - R - c)\lambda + (R - s)}{(1-\lambda)(1-\lambda^{n_s+1})}. \end{aligned}$$

We first show that the function

$$g(x) = -sx^{n_s+2} + sx^{n_s+1} + (R - c(n-1))x^{n+1} + (cn - R)x^n + (s - R - c)x + (R - s)$$

has a unique root in $(0,\infty)\backslash\{1\}$.

We take the first and second-order derivatives

$$g'(x) = -s(n_s+2)x^{n_s+1} + s(n_s+1)x^{n_s} + (n+1)(R - c(n-1))x^n + n(cn - R)x^{n-1} + (s - R - c)$$
$$g''(x) = x^{n-2}[-s(n_s+2)(n_s+1)x^2x^{n_s-n} + s(n_s+1)n_sxx^{n_s-n} + (n+1)n(R - c(n-1))x + n(n-1)(cn - R)]$$

Define the function $h(x)$ and take its first and second-order derivatives as follows

$$h(x) = -s(n_s+2)(n_s+1)x^2x^{n_s-n} + s(n_s+1)n_sxx^{n_s-n} + (n+1)n(R - c(n-1))x + n(n-1)(cn - R)$$
$$h'(x) = -s(n_s+2)(n_s+1)(n_s - n + 2)x^{n_s-n+1} + s(n_s+1)n_s(n_s - n + 1)x^{n_s-n} + (n+1)n(R - (n-1)c)$$
$$h''(x) = x^{n_s-n-1}s(n_s - n + 1)[-(n_s+2)(n_s+1)(n_s - n + 2)x + (n_s+1)n_s(n_s - n)]$$

Note that

$$h''(x) > 0 \text{ for } x < \frac{n_s(n_s-n)}{(n_s+2)(n_s-n+2)} \text{ and } h''(x) < 0 \text{ for } x > \frac{n_s(n_s-n)}{(n_s+2)(n_s-n+2)}.$$

Thus, depending on whether $n_s = n$ or $n_s > n$, $h'(x)$ is either monotonically decreasing in x or first increasing then decreasing in x for $x > 0$. We then note that

$$h'(0) = (n+1)n(R - (n-1)c) \geq 0$$

and

$$h'\left(\frac{n_s(n_s-n)}{(n_s+2)(n_s-n+2)}\right) = \left(\frac{n_s(n_s-n)}{(n_s+2)(n_s-n+2)}\right)^{n_s-n} sn_s(n_s+1) + n(n+1)(R - (n-1)c)$$
$$\geq 0$$

Thus, $h(x)$ is first increasing and then decreasing on $x > 0$. Since $h(0) \geq 0$, $g''(x)$ is first increasing then decreasing on $x > 0$. Since $g''(0) = 0$, $g'(x)$ is first increasing then decreasing on $x > 0$. We further note that

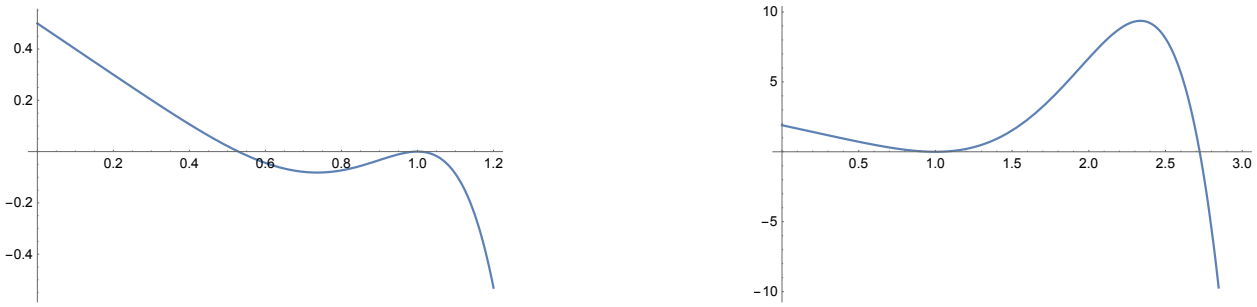
$$g'(0) < 0 \text{ and } g'(1) = 0.$$

We distinguish between two cases. Case I: $g''(1) < 0$. The transition of $g'(x)$ from an increasing to a decreasing function happens prior to $x = 1$, i.e., there exists $0 < x_- < 1$ such that $g(x)$ decreases when $x < x_-$, increases when $x_- < x < 1$, and decreases when $x > 1$. Case II: $g''(1) > 0$. The transition of $g'(x)$ from an increasing to a decreasing function happens after $x = 1$, i.e., there exists $x_+ > 1$ such that $g(x)$ decreases when $x < 1$, increases when $1 < x < x_+$, and decreases when $x > x_+$. Since

$$g(0) = R - s > 0 \text{ and } g(1) = 0,$$

for either case above, $g(x)$ has a unique root $\tilde{x} \in (0,\infty)\backslash\{1\}$ such that

$$g(x) \begin{cases} > 0 & \text{if } x < \tilde{x} \text{ and } x \neq 1 \\ < 0 & \text{if } x > \tilde{x} \text{ and } x \neq 1. \end{cases}$$



Case I. **Case II.**

Figure 18 **Plot showing the two possible cases of the $g(x)$ function. In each case, $g(x)=0$ has a unique distinct solution other from $x=1$.**

For $s=\hat{s}$, from previous analysis, we know that $g'(x)$ is first monotonically increasing and then monotonically decreasing. In addition, when $s=\hat{s}$,

$$\begin{aligned} g''(1) &= -s(n_s+2)(n_s+1)+s(n_s+1)n_s+(n+1)n(R-c(n-1))+n(n-1)(cn-R) \\ &= -2s(n_s+1)+2nR+cn(1-n) \\ &= -2(nR-\frac{cn(n-1)}{2})+2nR+cn(1-n) \\ &= 0. \end{aligned}$$

Thus, $g'(x)\le 0$ for $x>0$, i.e., $g(x)$ is decreasing on $x>0$. Since $g(0)>0$ and $g(1)=0$, we have when $s=\hat{s}$,

$$g(x)\begin{cases} >0 & \text{if } x<1 \\ <0 & \text{if } x>1. \end{cases}$$

Thus,

$$f(\lambda;\hat{s})\begin{cases} >0 & \text{if } \lambda<1 \\ =0 & \text{if } \lambda=1 \\ <0 & \text{if } \lambda>1. \end{cases}$$

Above all, for fixed R, c, θ, and $s>0$, the equation $f(\lambda;s)=0$ has a unique solution $\lambda_f\in(0,\infty)$ and

$$f(\lambda;s)\begin{cases} >0 & \text{if } \lambda<\lambda_f \\ <0 & \text{if } \lambda>\lambda_f. \end{cases}$$

In the special case when $s=0$, we note that $h''(x)=0$ and $h'(x)=(n+1)n(R-(n-1)c)\ge 0$. Thus, $h(x)$ and $g''(x)$ is non-decreasing when $x>0$. Since $g''(0)=0$, we have $g'(x)$ non-decreasing when $x>0$. Since $g'(0)<0$ and $g'(1)=0$, we have $g(x)$ decreases when $0<x<1$ and increases when $x>1$. Since $g(1)=0$, we have $g(x)\ge 0$ when $x>0$. □

Proof of Lemma 5 and Theorem 2. For $f(\lambda;s):=-s+\mathbb{E}[\hat{U}_2^I]$, if $f(\lambda_0;s)\ge 0$, $\lambda_{e1}=\lambda_0$ and $q_1=1$, i.e., all customers will join the system in the first stage. If $f(\lambda_0;s)<0$, then the equilibrium λ_{e1} solves $f(\lambda_{e1};s)=0$, i.e., $\lambda_{e1}=\lambda_f$. We note from Lemma 6 that $f(\lambda;s)<0$ for $\lambda>\lambda_f$ and $f(\lambda;s)>0$ for $\lambda<\lambda_f$. Thus, if $f(\lambda_0;s)<0$, $\lambda_0>\lambda_f$. Then, the equilibrium arrival rate following the first stage decision is

$$\lambda_{e1}=\lambda_0\mathbb{I}\{\lambda_0\le\lambda_f\}+\lambda_f\mathbb{I}\{\lambda_0>\lambda_f\}=\min\{\lambda_0,\lambda_f\}.$$

Together with Lemma 3, we have

$$(q_1, q_2) = \left(\min\left\{1, \frac{\lambda_f}{\lambda_0}\right\}, \frac{1-\min\{\lambda_f,\lambda_0\}^{n_s}}{1-\min\{\lambda_f,\lambda_0\}^{n_s+1}} \right).$$

□

Proof of Lemma 4. Define

$$g(x;s) = -sx^{n_s+2} + sx^{n_s+1} + (R - c(n-1))x^{n+1} + (cn - R)x^n + (s - R - c)x + (R - s),$$

where $n = n_0$. From the proof of Lemma 6, we have $g(x;s) = (1-x)(1-x^{n_s+1})f(x;s)$.

In the proof of Lemma 6, we show that for $s > 0$, there exists a unique solution $\tilde{x}(s) > 0$ for $g(x;s) = 0$ such that

$$g(x;s) > 0 \text{ when } x < \tilde{x}(s) \text{ and } x \neq 1$$
$$g(x;s) < 0 \text{ when } x > \tilde{x}(s) \text{ and } x \neq 1.$$

Then, it is sufficient to show that

$$g(\tilde{x}(s); s + \Delta) < 0$$

for any $\Delta > 0$. In particular, this implies that $\tilde{x}(s+\Delta) \leq \tilde{x}(s)$.

For simplicity of notation, we write $\tilde{x}(s)$ as $\tilde{x}$ below.

$$\begin{aligned} g(\tilde{x}; s+\Delta) =& -(s+\Delta)\tilde{x}^{n_{s+\Delta}+2} + (s+\Delta)\tilde{x}^{n_{s+\Delta}+1} \\ &+ (R-(n-1)c)\tilde{x}^{n+1} + (cn-R)\tilde{x}^n + (-R-c+s+\Delta)\tilde{x} + R - s - \Delta \\ =& s(\tilde{x}-1)\tilde{x}^{n_s+1} + s(1-\tilde{x})\tilde{x}^{n_{s+\Delta}+1} + \Delta(1-\tilde{x})\tilde{x}^{n_{s+\Delta}+1} + \Delta(\tilde{x}-1) \\ =& s(1-\tilde{x})(\tilde{x}^{n_{s+\Delta}+1} - \tilde{x}^{n_s+1}) + \Delta(1-\tilde{x})(\tilde{x}^{n_{s+\Delta}+1} - 1) \end{aligned}$$

If $\tilde{x} < 1$, then

$$\begin{aligned} 1 - \tilde{x} &> 0, \\ \tilde{x}^{n_{s+\Delta}+1} - \tilde{x}^{n_s+1} &< 0, \\ \tilde{x}^{n_{s+\Delta}+1} - 1 &< 0. \end{aligned}$$

In this case, $g(\tilde{x}; s+\Delta) < 0$. If $\tilde{x} > 1$, then

$$\begin{aligned} 1 - \tilde{x} &< 0, \\ \tilde{x}^{n_{s+\Delta}+1} - \tilde{x}^{n_s+1} &> 0, \\ \tilde{x}^{n_{s+\Delta}+1} - 1 &> 0. \end{aligned}$$

In this case, we also have $g(\tilde{x}; s+\Delta) < 0$.

□

Appendix C: Proofs of the Results in Section 6

Proof of Proposition 1 By Theorem 1, the throughput or the equilibrium arrival rate after the second stage decision is

$$\lambda_{e2}^U(s,\lambda_0)=\begin{cases}\lambda_0 & \text{if } \lambda_0\le\frac{R-s}{c+R-s}\\ \frac{R-s}{c+R-s} & \text{if } \lambda_0>\frac{R-s}{c+R-s}.\end{cases}$$

For any $s>0$, if $\lambda_0\le\frac{R-s}{c+R-s}<\frac{R}{c+R}$,

$$\lambda_{e2}^U(s,\lambda_0)=\lambda_0=\lambda_{e2}^U(0,\lambda_0).$$

If $\frac{R-s}{c+R-s}<\lambda_0\le\frac{R}{c+R}$,

$$\lambda_{e2}^U(s,\lambda_0)=\frac{R-s}{c+R-s}<\lambda_0=\lambda_{e2}^U(0,\lambda_0).$$

If $\lambda_0>\frac{R}{c+R}$, then

$$\lambda_{e2}^U(s,\lambda_0)=\frac{R-s}{c+R-s}<\frac{R}{c+R}=\lambda_{e2}^U(0,\lambda_0).$$

Thus, for any fixed λ_0,

$$\max_{s\ge0}\lambda_{e2}^U(s,\lambda_0)=\lambda_{e2}^U(0,\lambda_0).$$

□

Before we prove Proposition 2, we first introduce two auxiliary lemmas.

Lemma 7. *For any fixed n, consider the function for $\lambda>0$*

$$h_n(\lambda)=\begin{cases}\lambda\frac{1-\lambda^n}{1-\lambda^{n+1}} & \text{if } \lambda\neq1\\ \frac{n}{n+1} & \text{if } \lambda=1.\end{cases}$$

$h_n(\lambda)$ is monotonically increasing and continuous in λ for $\lambda>0$.

Proof. When $\lambda\neq1$,

$$h_n(\lambda)=\lambda\frac{1-\lambda^n}{1-\lambda^{n+1}}.$$
$$h_n'(\lambda)=\frac{1+n\lambda^{n+1}-n\lambda^n-\lambda^n}{(1-\lambda^{n+1})^2}.$$

Next, consider the function

$$g_n(\lambda)=1+nx^{n+1}-n\lambda^n-\lambda^n.$$
$$g_n'(\lambda)=\lambda^{n-1}(n(n+1)\lambda-n^2-n)=n(n+1)(\lambda-1)\lambda^{n-1}.$$

For $\lambda>0$, $g_n(\lambda)$ achieves its minimum at $g_n(1)=0$. $g_n(\lambda)$ is positive elsewhere on $\lambda>0$ and so is $h_n'(\lambda)$. Thus, $h_n(\lambda)$ is monotonically increasing on $0<\lambda<1$ and $\lambda>1$. Since $\lim_{\lambda\to1-}h_n(\lambda)=\lim_{\lambda\to1+}h_n(\lambda)=\frac{n}{n+1}$, $h_n(\lambda)$ is monotonically increasing and continuous. □

Lemma 8. *For fixed R, c, θ, and λ_0, the equilibrium arrival rate after the second stage decision $\lambda_{e2}^I(s,\lambda_0)$ is non-increasing in s for $s\in[s_k,s_{k+1})$.*

Proof. When $s\in[s_k,s_{k+1})$, the equilibrium arrival rate after the second stage decision can be written as a function of the initial arrival rate and sunk cost:

$$\lambda_{e2}^I(s,\lambda_0)=\lambda_0q_1^I(s,\lambda_0)\frac{1-(\lambda_0q_1^I(s,\lambda_0))^{n_0+k}}{1-(\lambda_0q_1^I(s,\lambda_0))^{n_0+k+1}}.$$

We first note that since $\lambda_f(s)$ is decreasing in s by Lemma 4, $\lambda_0 q_1^I(s,\lambda_0) = \min\{\lambda_0, \lambda_f(s)\}$ is non-increasing in s. Next, since

$$x\frac{1-x^{n_0+k}}{1-x^{n_0+k+1}}$$

is non-decreasing in x by Lemma 7, we have $\lambda_{e2}^I(s,\lambda_0)$ is non-increasing in s. □

Proof of Proposition 2. We first consider the case where $\lambda_0 \le \lambda_f(s_1)$. In this case, $q_1^I(s_1) = 1$ and $q_1^I(0) = 1$. Then,

$$\lambda_{e2}^I(s_1,\lambda_0) = \lambda_0\frac{1-\lambda_0^{n_1}}{1-\lambda_0^{n_1+1}} > \lambda_0\frac{1-\lambda_0^{n_0}}{1-\lambda_0^{n_0+1}} = \lambda_{e2}^I(0,\lambda_0),$$

i.e., $s = s_1$ leads to a larger throughput than than $s = 0$. This implies that when $\lambda_0 \le \lambda_f(s_1)$, there always exists a non-zero sunk cost that increases the throughput.

Next, we consider the case where $\lambda_0 > \lambda_f(s_1)$. Recall that

$$T^I = \sup\left\{\lambda : \max_{k\in\mathbb{N}^+} \lambda_{e2}^I(s_k,\lambda) > \lambda_{e2}^I(0,\lambda)\right\}.$$

If $\lambda_0 > T^I$, by definition,

$$\max_{k\in\mathbb{N}^+} \lambda_{e2}^I(s_k,\lambda_0) \le \lambda_{e2}^I(0,\lambda_0).$$

From Lemma 8, $\lambda_{e2}^I(s,\lambda_0)$ is non-increasing in s for $s_k \le s < s_{k+1}$, we have

$$\lambda_{e2}^I(s_k,\lambda_0) \ge \lambda_{e2}^I(s,\lambda_0) \quad \text{for } s_k < s < s_{k+1} \quad \forall k \in \mathbb{N}^+.$$

Thus, for any $s > 0$, $\lambda_{e2}^I(s;\lambda_0) \le \lambda_{e2}^I(0;\lambda_0)$.

If $\lambda_f(s_1) < \lambda_0 \le T^I$, define

$$\bar{s} = \arg\max_{s_k} \lambda_{e2}^I(s_k,\lambda_0).$$

Note that $\bar{s} \ge s_1$. Then, $\lambda_0 > \lambda_f(s_1) \ge \lambda_f(\bar{s})$. This implies that

$$\begin{aligned}
\lambda_{e2}^I(\bar{s},\lambda_0) &= \lambda_f(\bar{s})\frac{1-\lambda_f(\bar{s})^{n_{\bar{s}}}}{1-\lambda_f(\bar{s})^{n_{\bar{s}}+1}} \\
&= \lambda_{e2}^I(\bar{s},T^I) \quad \text{since } \lambda_f(\bar{s}) < \lambda_0 \le T^I \\
&\ge \lambda_{e2}^I(0,T^I) \quad \text{by definition of } T^I \\
&= T^I\frac{1-(T^I)^{n_0}}{1-(T^I)^{n_0+1}} > \lambda_0\frac{1-\lambda_0^{n_0}}{1-\lambda_0^{n_0+1}} = \lambda_{e2}^I(0,\lambda_0).
\end{aligned}$$

□

Before we prove Proposition 3, we first introduce two auxiliary lemmas.

Lemma 9. *The function*

$$f_0^{UI}(\lambda) = \begin{cases} \frac{R}{c+R} - \frac{\lambda-\lambda^{n_0+1}}{1-\lambda^{n_0+1}} & \text{if } \lambda \ne 1 \\ \frac{R}{c+R} - \frac{n_0}{n_0+1} & \text{if } \lambda = 1 \end{cases}$$

has a unique root.

Proof. Note that $f_0^{UI}(\lambda) = R/(c+R) - h_n(\lambda)$, where $h_n(\lambda)$ is defined in Lemma 7. From Lemma 7, $h_n(\lambda)$ is monotonically increasing on $\lambda > 0$ and is continuous. Since $h_n(0) = 0$ and $\lim_{\lambda\to\infty} h_n(\lambda) = 1$, $h_n(\lambda) - \frac{R}{c+R} = 0$ has a unique solution. □

Lemma 10. *For the unique solution of $f_0^{UI}(\lambda) = 0$, $\bar{T}$, $\frac{R}{c+R} < \bar{T} < 1$.*

Proof. For simplicity, we write n_0 as n. $\bar{T}$ is such that

$$\begin{aligned} & \frac{R}{c+R} = \frac{\bar{T} - (\bar{T})^{n+1}}{1-(\bar{T})^{n+1}} \\ \Leftrightarrow \quad & \frac{R}{c+R} - \frac{R}{c+R}\bar{T} = \bar{T} - (\bar{T})^{n+1} \\ \Leftrightarrow \quad & \bar{T} - \frac{R}{c+R} = (\bar{T})^{n+1} - \frac{R}{c+R}(\bar{T})^{n+1} > 0. \end{aligned}$$

Thus, $\bar{T} > \frac{R}{c+R}$.

Next,

$$\begin{aligned} & \frac{R}{c+R} = \frac{\bar{T} - (\bar{T})^{n+1}}{1-(\bar{T})^{n+1}} \\ \Leftrightarrow \quad & c(\bar{T})^{n+1} - (c+R)\bar{T} + R = 0 \end{aligned}$$

For any fixed n, consider the function and its first and second-order derivatives:

$$\begin{aligned} g(x) &= cx^{n+1} - (c+R)x + R \\ g'(x) &= c(n+1)x^n - (c+R) \\ g''(x) &= c(n+1)nx^{n-1} > 0. \end{aligned}$$

Thus, $g'(x)$ is monotonically increasing in x. Since

$$g'(0) < 0 \ \text{ and } g'(1) = cn - R \geq 0$$

and

$$g(0) = R > 0 \ \text{ and } g(1) = 0,$$

we have $\bar{T} < 1$. □

Proof of Proposition 3. First, from Lemma 10, we have $\bar{T} > \frac{R}{c+R}$.

Next, recall that

$$\begin{aligned} \lambda_{e2}^I(0,\lambda_0) &= \lambda_0 \frac{1-\lambda_0^{n_0}}{1-\lambda_0^{n_0+1}}, \\ \lambda_{e2}^U(0,\lambda_0) &= \lambda_0 \mathbb{I}\left\{\lambda_0 \leq \frac{R}{c+R}\right\} + \frac{R}{c+R}\mathbb{I}\left\{\lambda_0 > \frac{R}{c+R}\right\}. \end{aligned}$$

When $\lambda_0 \leq \frac{R}{c+R} < \bar{T}$,

$$\lambda_{e2}^I(0,\lambda_0) = \lambda_0 \frac{1-\lambda_0^{n_0}}{1-\lambda_0^{n_0+1}} < \lambda_0 = \lambda_{e2}^U(0,\lambda_0).$$

When $\frac{R}{c+R} < \lambda_0 < \bar{T}$,

$$\lambda_{e2}^I(0,\lambda_0) = \lambda_0 \frac{1-\lambda_0^{n_0}}{1-\lambda_0^{n_0+1}} < \frac{R}{c+R} = \lambda_{e2}^U(0,\lambda_0).$$

When $\lambda_0 > \bar{T}$,

$$\lambda_{e2}^I(0,\lambda_0) = \lambda_0 \frac{1-\lambda_0^{n_0}}{1-\lambda_0^{n_0+1}} > \frac{R}{c+R} = \lambda_{e2}^U(0,\lambda_0).$$

Lastly, when $\lambda_0 = \bar{T}$,

$$\lambda_{e2}^I(0,\lambda_0) = \bar{T} \frac{1-(\bar{T})^{n_0}}{1-(\bar{T})^{n_0+1}} = \frac{R}{c+R} = \lambda_{e2}^U(0,\lambda_0).$$

□

Before we prove Theorem 3, we first introduce an auxiliary lemma.

Lemma 11. *For fixed R, c, and θ, $\lambda_{e2}^I(s^\star(\lambda_0),\lambda_0)$ is non-decreasing in λ_0.*

Proof. Recall that $s^\star(\lambda_0)$ is the optimal level of sunk cost when the initial arrival rate is λ_0. For any $\epsilon>0$, consider an arrival rate $\lambda_0+\epsilon$.

$$\begin{aligned}\lambda_{e2}^I(s^\star(\lambda_0+\epsilon),\lambda_0+\epsilon) &\geq \lambda_{e2}^I(s^\star(\lambda_0),\lambda_0+\epsilon)\\ &= \min\left\{\lambda_f(s^\star(\lambda_0)),\lambda_0+\epsilon\right\}\frac{1-(\min\{\lambda_f(s^\star(\lambda_0)),\lambda_0+\epsilon\})^{n_{s^\star(\lambda_0)}}}{1-(\min\{\lambda_f(s^\star(\lambda_0)),\lambda_0+\epsilon\})^{n_{s^\star(\lambda_0)+1}}}\\ &\geq \min\{\lambda_f(s^\star(\lambda_0)),\lambda_0\}\frac{1-(\min\{\lambda_f(s^\star(\lambda_0)),\lambda_0\})^{n_{s^\star(\lambda_0)}}}{1-(\min\{\lambda_f(s^\star(\lambda_0)),\lambda_0\})^{n_{s^\star(\lambda_0)+1}}}\\ &= \lambda_{e2}^I(s^\star(\lambda_0),\lambda_0).\end{aligned}$$

□

Proof of Theorem 3. We first compare an uninformed system without sunk cost with an informed system i.e, $\lambda_{e2}^U(0,\lambda_0)$ versus $\max_{s\geq 0}\lambda_{e2}^I(s,\lambda_0)$. Since

$$\lambda_{e2}^U(0,\lambda_0)=\lambda_0\mathbb{I}\left\{\lambda_0\leq\frac{R}{c+R}\right\}+\frac{R}{c+R}\mathbb{I}\left\{\lambda_0>\frac{R}{c+R}\right\},$$

when $\lambda_0\leq\frac{R}{c+R}$,

$$\max_{s\geq 0}\lambda_{e2}^I(s,\lambda_0)\leq\lambda_0=\lambda_{e2}^U(0,\lambda_0).$$

By Proposition 3, for any $\lambda_0\geq\bar{T}$,

$$\max_{s\geq 0}\lambda_{e2}^I(s,\lambda_0)\geq\lambda_{e2}^I(0,\lambda_0)\geq\lambda_{e2}^U(0,\lambda_0).$$

When $\frac{R}{c+R}\leq\lambda_0\leq\bar{T}$, since $\lambda_{e2}^U(0,\lambda_0)=R/(c+R)$ and by Lemma 11, $\max_{s\geq 0}\lambda_{e2}^I(s,\lambda_0)$ is non-decreasing in λ_0, there exists a unique critical value $T^{UI}\in[\frac{R}{c+R},\bar{T}]$ such that if $\lambda_0>T^{UI}$,

$$\max_{s\geq 0}\lambda_{e2}^I(s,\lambda_0)>\lambda_{e2}^U(0,\lambda_0);$$

if $\lambda_0\leq T^{UI}$,

$$\max_{s\geq 0}\lambda_{e2}^I(s,\lambda_0)\leq\lambda_{e2}^U(0,\lambda_0).$$

Combined with Propositions 1 – 3, we can conclude with two possible scenarios.

In the first scenario, $T^I\leq T^{UI}$. In this case, $T^{UI}=\bar{T}$. If $\lambda_0\leq T^{UI}$, then

$$\max_{s\geq 0,\gamma\in\{I,U\}}\lambda_{e2}^\gamma(s)=\lambda_{e2}^U(0).$$

If $\lambda_0>T^{UI}$, then

$$\max_{s\geq 0,\gamma\in\{I,U\}}\lambda_{e2}^\gamma(s)=\lambda_{e2}^I(0).$$

In the second scenario, $T^I>T^{UI}$. In this case, if $\lambda_0\leq T^{UI}$, then

$$\max_{s\geq 0,\gamma\in\{I,U\}}\lambda_{e2}^\gamma(s)=\lambda_{e2}^U(0).$$

If $\lambda_0\geq T^I$, then

$$\max_{s\geq 0,\gamma\in\{I,U\}}\lambda_{e2}^\gamma(s)=\lambda_{e2}^I(0).$$

If $T^{UI}<\lambda_0<T^I$, then there exists $s^\star>0$ such that

$$\max_{s\geq 0,\gamma\in\{I,U\}}\lambda_{e2}^\gamma(s)=\lambda_{e2}^I(s^\star).$$

□

Appendix D: Proofs of the Results in Section 7

We refer to high (low) reward customers as H-class (L-class) customers in the proof below.

Lemma 12. *For fixed R^H, R^L, c, θ, and s, given the equilibrium arrival rate after the first stage decision λ_{e1H} and λ_{e1L}, for H-class customers, the equilibrium arrival rate following the second stage decision is*

$$\lambda_{e2H} = \min\left\{\lambda_{e1H}, \frac{R^H+\theta s}{c+R^H+\theta s}\right\};$$

and the proportion of uninformed customers that wait in the second stage is

$$q_{2H} = \min\left\{1, \frac{R^H+\theta s}{\lambda_{e1H}(c+R^H+\theta s)}\right\}.$$

For L-class customers, the equilibrium arrival rate following the second stage decision is

$$\lambda_{e2L} = \lambda_{e1L}\mathbb{I}\left\{\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}\right\} + \max\left\{0, \frac{R^L+\theta s}{c+R^L+\theta s} - \lambda_{e1H}\right\}\mathbb{I}\left\{\lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s}\right\};$$

and the proportion of uninformed customers that wait in the second stage is

$$q_{2L} = \mathbb{I}\left\{\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}\right\} + \max\left\{0, \frac{\frac{R^L+\theta s}{c+R^L+\theta s} - \lambda_{e1H}}{\lambda_{e1L}}\right\}\mathbb{I}\left\{\lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s}\right\}.$$

The expected waiting time in the second stage is

$$\begin{aligned}\bar{W}^U = {} & \frac{\lambda_{e1}}{1-\lambda_{e1}}\mathbb{I}\left\{\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}\right\} + \frac{R^L+\theta s}{c}\mathbb{I}\left\{\lambda_{e1H} \le \frac{R^L+\theta s}{c+R^L+\theta s}, \lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s}\right\} \\ & + \frac{\lambda_{e1H}}{1-\lambda_{e1H}}\mathbb{I}\left\{\frac{R^L+\theta s}{c+R^L+\theta s} < \lambda_{e1H} \le \frac{R^H+\theta s}{c+R^H+\theta s}\right\} \\ & + \frac{R^H+\theta s}{c}\mathbb{I}\left\{\lambda_{e1H} > \frac{R^H+\theta s}{c+R^H+\theta s}\right\}.\end{aligned}$$

Proof of Lemma 12. For an $M/M/1$ queue with service rate $\mu = 1$, let $W(\lambda)$ denote the steady-state average waiting time when the total arrival rate is λ. Then, for $\lambda < 1$,

$$W(\lambda) = \frac{\lambda}{1-\lambda}.$$

Recall that $\lambda_{e1} = \lambda_{e1H} + \lambda_{e1L}$ is the total equilibrium arrival rate after the first stage decision. We first note that if $R^L - cW(\lambda_{e1}) \ge -\theta s$, then $q_{2H} = q_{2L} = 1$, i.e., all uninformed customers who incur the sunk cost will wait in the second stage. The condition $R^L - cW(\lambda_{e1}) \ge -\theta s$ is equivalent to

$$\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}.$$

In the second case, if

$$\frac{R^L+\theta s}{c+R^L+\theta s} < \lambda_{e1} \le \frac{R^H+\theta s}{c+R^H+\theta s},$$

all H-class customers will wait in the second stage but only a fraction of L-class customers will wait in the second stage because $R^H - cW(\lambda_{e1}) \le -\theta s < R^L - cW(\lambda_{e1})$. In this case, the equilibrium arrival rate after the second stage decision, $\lambda_{e2L} = \lambda_{e1L} q_{2L}$, solves:

$$R^L - cW(\lambda_{e2L} + \lambda_{e1H}) = -\theta s \Leftrightarrow \lambda_{e2L} = \frac{R^L+\theta s}{c+R^L+\theta s} - \lambda_{e1H}.$$

In this case, if

$$\lambda_{e1H} \geq \frac{R^L + \theta s}{c + R^L + \theta s},$$

$q_{2L} = 0$. If

$$\lambda_{e1H} < \frac{R^L + \theta s}{c + R^L + \theta s},$$

$$q_{2L} = \frac{\frac{R^L + \theta s}{c + R^L + \theta s} - \lambda_{e1H}}{\lambda_{e1L}}.$$

In the last case, if

$$\lambda_{e1} > \frac{R^H + \theta s}{c + R^H + \theta s},$$

only a fraction of H-class and L-class customers will wait in the second stage. In this case, we consider three possible scenarios.

1. If

$$\lambda_{e1H} \geq \frac{R^H + \theta s}{c + R^H + \theta s},$$

then

$$q_{2H} = \frac{R^H + \theta s}{\lambda_{e1H}(c + R^H + \theta s)} \quad \text{and } q_{2L} = 0$$

is the unique equilibrium, for the reasons below. For any positive q_{2L}, the equilibrium arrival rate after the second stage decision, $\lambda_{e2L} = \lambda_{e1L} q_{2L}$, solves

$$R^L - cW(\lambda_{e2L} + \lambda_{e2H}) = -\theta s \Leftrightarrow \lambda_{e2L} + \lambda_{e2H} = \frac{R^L + \theta s}{c + R^L + \theta s}.$$

However, for the H-class, since $\lambda_{e1H} \geq (R^H + \theta s)/(c + R^H + \theta s)$, we can increase λ_{e2H} such that $R^H - cW(\lambda_{e2L} + \lambda_{e2H}) = -\theta s$. As such, any positive q_{2L} is not an equilibrium. Since $q_{2L} = 0$, the equilibrium arrival rate after the second stage decision, $\lambda_{e2H} = \lambda_{e1H} q_{2H}$, solves

$$R^H - cW(\lambda_{e2H}) = -\theta s \Leftrightarrow \lambda_{e2H} = \frac{R^H + \theta s}{c + R^H + \theta s}.$$

2. If

$$\frac{R^L + \theta s}{c + R^L + \theta s} \leq \lambda_{e1H} < \frac{R^H + \theta s}{c + R^H + \theta s},$$

$$q_{2H} = 1 \quad \text{and } q_{2L} = 0$$

is the unique equilibrium for similar arguments. In this case, all H-class customers who incur the sunk cost will wait in the second stage because $R^H - cW(\lambda_{e1H}) > -\theta s$.

3. If

$$\lambda_{e1H} < \frac{R^L + \theta s}{c + R^L + \theta s},$$

$$q_{2H} = 1 \quad \text{and } q_{2L} = \frac{\frac{R^L + \theta s}{c + R^L + \theta s} - \lambda_{e1H}}{\lambda_{e1L}}$$

is the unique equilibrium for similar arguments. In this case, all H-class customers will wait in the second stage but only a fraction of L-class customers will wait.

Putting this all together, given λ_{e1H} and λ_{e1L}, for H-class customers, the equilibrium arrival rate following the second stage decision is

$$\lambda_{e2H} = \min\left\{\lambda_{e1H}, \frac{R^H+\theta s}{c+R^H+\theta s}\right\};$$

and the proportion of uninformed customers that wait in the second stage is

$$q_{2H} = \min\left\{1, \frac{R^H+\theta s}{\lambda_{e1H}(c+R^H+\theta s)}\right\}.$$

For L-class customers, the equilibrium arrival rate following the second stage decision is

$$\lambda_{e2L} = \lambda_{e1L}\mathbb{I}\left\{\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}\right\} + \max\left\{0, \frac{R^L+\theta s}{c+R^L+\theta s} - \lambda_{e1H}\right\}\mathbb{I}\left\{\lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s,}\right\};$$

and the proportion of uninformed customers that wait in the second stage is

$$q_{2L} = \mathbb{I}\left\{\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}\right\} + \max\left\{0, \frac{\frac{R^L+\theta s}{c+R^L+\theta s} - \lambda_{e1H}}{\lambda_{e1L}}\right\}\mathbb{I}\left\{\lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s}\right\}.$$

The expected waiting time in the second stage is

$$\begin{aligned}\bar{W}^U = {} & \frac{\lambda_{e1}}{1-\lambda_{e1}}\mathbb{I}\left\{\lambda_{e1} \le \frac{R^L+\theta s}{c+R^L+\theta s}\right\} + \frac{R^L+\theta s}{c}\mathbb{I}\left\{\lambda_{e1H} \le \frac{R^L+\theta s}{c+R^L+\theta s}, \lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s}\right\} \\ & + \frac{\lambda_{e1H}}{1-\lambda_{e1H}}\mathbb{I}\left\{\frac{R^L+\theta s}{c+R^L+\theta s} < \lambda_{e1H} \le \frac{R^H+\theta s}{c+R^H+\theta s}\right\} \\ & + \frac{R^H+\theta s}{c}\mathbb{I}\left\{\lambda_{e1H} > \frac{R^H+\theta s}{c+R^H+\theta s}\right\}.\end{aligned}$$

□

Lemma 13. *For fixed R^H, R^L, c, θ, and s, given the initial arrival rate to the first stage λ_{0H} and λ_{0L}, in the unique equilibrium of an uninformed system, for H-class customers, the equilibrium arrival rate following the first stage decision is*

$$\lambda_{e1H} = \min\left\{\lambda_0, \frac{R^H-s}{c+R^H-s}\right\};$$

the proportion of uninformed customers who choose to incur the sunk cost in the first stage is

$$q_{1H} = \min\left\{1, \frac{R^H-s}{\lambda_0(c+R^H-s)}\right\}.$$

For L-class customers, the equilibrium arrival rate following the first stage decision is

$$\lambda_{e1L} = \lambda_{0L}\mathbb{I}\left\{\lambda_0 \le \frac{R^L-s}{c+R^L-s}\right\} + \max\left\{0, \frac{R^L-s}{c+R^L-s} - \lambda_{0H}\right\}\mathbb{I}\left\{\lambda_{0H} < \frac{R^H-s}{c+R^H-s}\right\};$$

the proportion of uninformed customers who choose to incur the sunk cost in the first stage is

$$q_{1L} = \mathbb{I}\left\{\lambda_0 \le \frac{R^L-s}{c+R^L-s}\right\} + \max\left\{1 - \frac{\lambda_0 - \frac{R^L-s}{c+R^L-s}}{\lambda_{0L}}, 0\right\}\mathbb{I}\left\{\lambda_{0H} < \frac{R^H-s}{c+R^H-s}\right\}.$$

Proof of Lemma 13. In the first stage, the H-class (and L-class) customers decide whether to incur the sunk cost by comparing the expected utility of joining the system, $-s+\hat{U}^U_{2H}$ (and $-s+\hat{U}^U_{2L}$), to zero, i.e.,

$$U^U_{1H} = \max_{q\in[0,1]} q(-s+\hat{U}^U_{2H}) \text{ and } U^U_{1L} = \max_{q\in[0,1]} q(-s+\hat{U}^U_{2L}),$$

where $\hat{U}^U_{2H} = \max_{\hat{q}\in[0,1]} \hat{q}(R^H - c\bar{W}^U)$ (and $\hat{U}^U_{2L} = \max_{\hat{q}\in[0,1]} \hat{q}(R^L - c\bar{W}^U)$) is H-class (and L-class) customers' *belief* about the expected utility in the second stage. Note that we assume customers learn the steady-state

average waiting time through repeated interaction with the system. Thus, $\bar{W}^U$ is the average waiting time characterized in Lemma 12. On the other hand, customers believe they will be rational in the second stage when making the first-stage decision. Thus, H-class (and L-class) customers assume they will wait in the second stage if the expected utility of waiting, $R^H - c\bar{W}^U$ (and $R^L - c\bar{W}^U$), is no less than 0.

We characterize uninformed customer's *belief* in the first stage about the probability of continuing to wait in the second stage for H-class and L-class separately. For H-class, given the equilibrium arrival rate after the first stage decision λ_{e1H} and λ_{e1L}, by the characterization of $\bar{W}^U$ in Lemma 12, we consider the following cases:

1. If
$$\lambda_{e1H} > \frac{R^H + \theta s}{c + R^H + \theta s},$$
then
$$R^H - c\bar{W}^U = -\theta s < 0.$$
2. If
$$\frac{R^L + \theta s}{c + R^L + \theta s} < \lambda_{e1H} \leq \frac{R^H + \theta s}{c + R^H + \theta s},$$
then
$$R^H - c\bar{W}^U = R^H - c\frac{\lambda_{e1H}}{1 - \lambda_{e1H}} \geq 0 \Leftrightarrow \lambda_{e1H} \leq \frac{R^H}{c + R^H}.$$
   (a) If $R^H > R^L + \theta s$, then $R^H - c\bar{W}^U \geq 0$ when $\lambda_{e1H} \leq R^H/(c + R^H)$.
   (b) If $R^H \leq R^L + \theta s$, then $R^H - c\bar{W}^U < 0$.
3. If
$$\lambda_{e1H} \leq \frac{R^L + \theta s}{c + R^L + \theta s} \text{ and } \lambda_{e1} > \frac{R^L + \theta s}{c + R^L + \theta s},$$
then
$$R^H - c\bar{W}^U = R^H - R^L - \theta s.$$
   (a) If $R^H > R^L + \theta s$, then $R^H - c\bar{W}^U \geq 0$.
   (b) If $R^H \leq R^L + \theta s$, then $R^H - c\bar{W}^U < 0$.
4. If
$$\lambda_{e1H} \leq \frac{R^L + \theta s}{c + R^L + \theta s} \text{ and } \lambda_{e1} \leq \frac{R^L + \theta s}{c + R^L + \theta s},$$
then
$$R^H - c\bar{W}^U = R^H - c\frac{\lambda_{e1}}{1 - \lambda_{e1}} \geq 0 \Leftrightarrow \lambda_{e1} \leq \frac{R^H}{c + R^H}.$$
   (a) If $R^H > R^L + \theta s$, then $R^H - c\bar{W}^U \geq 0$.
   (b) If $R^H \leq R^L + \theta s$, then $R^H - c\bar{W}^U \geq 0$ when $\lambda_{e1} \leq \frac{R^H}{c + R^H}$.

Thus, for H-class, we can summarize from above that if $R^H > R^L + \theta s$,

$$\hat{q}_{2H} = \mathbb{I}\left\{\lambda_{e1H} \leq \frac{R^H}{c + R^H}\right\}$$

$$\hat{U}^U_{2H} = \begin{cases} R^H - c\frac{\lambda_{e1H}}{1-\lambda_{e1H}} & \text{if } \frac{R^L+\theta s}{c+R^L+\theta s} < \lambda_{e1H} \leq \frac{R^H}{c+R^H} \\ R^H - (R^L + \theta s) & \text{if } \lambda_{e1H} \leq \frac{R^L+\theta s}{c+R^L+\theta s} \text{ and } \lambda_{e1} > \frac{R^L+\theta s}{c+R^L+\theta s} \\ R^H - c\frac{\lambda_{e1H}}{1-\lambda_{e1H}} & \text{if } \lambda_{e1H} \leq \frac{R^L+\theta s}{c+R^L+\theta s} \text{ and } \lambda_{e1} \leq \frac{R^L+\theta s}{c+R^L+\theta s} \\ 0 & \text{if } \lambda_{e1H} > \frac{R^H}{c+R^H}. \end{cases}$$

If $R^H \le R^L + \theta s$,

$$\hat{q}_{2H} = \mathbb{I}\left\{\lambda_{e1} \le \frac{R^H}{c+R^H}\right\}$$
$$\hat{U}^U_{2H} = \begin{cases} R^H - c\frac{\lambda_{e1}}{1-\lambda_{e1}} & \text{if } \lambda_{e1} \le \frac{R^H}{c+R^H} \\ 0 & \text{if } \lambda_{e1} > \frac{R^H}{c+R^H}. \end{cases}$$

For L-class, given the equilibrium arrival rate after the first stage decision λ_{e1H} and λ_{e1L}, by the characterization of $\bar{W}^U$ in Lemma 12, we consider the following cases:

1. If
$$\lambda_{e1H} > \frac{R^H + \theta s}{c + R^H + \theta s},$$
then
$$R^L - c\bar{W}^U = R^L - R^H - \theta s < 0.$$

2. If
$$\frac{R^L + \theta s}{c + R^L + \theta s} < \lambda_{e1H} \le \frac{R^H + \theta s}{c + R^H + \theta s},$$
then
$$R^L - c\bar{W}^U = R^L - c\frac{\lambda_{e1H}}{1-\lambda_{e1H}} \ge 0 \Leftrightarrow \lambda_{e1H} \le \frac{R^L}{c+R^L}.$$
This cannot happen because $\lambda_{e1H} > (R^L + \theta s)/(c + R^L + \theta s) \ge R^L/(c + R^L)$.

3. If
$$\lambda_{e1H} \le \frac{R^L + \theta s}{c + R^L + \theta s} \text{ and } \lambda_{e1} > \frac{R^L + \theta s}{c + R^L + \theta s},$$
then
$$R^L - c\bar{W}^U = -\theta s < 0.$$

4. If
$$\lambda_{e1H} \le \frac{R^L + \theta s}{c + R^L + \theta s} \text{ and } \lambda_{e1} \le \frac{R^L + \theta s}{c + R^L + \theta s},$$
then
$$R^L - c\bar{W}^U = R^L - c\frac{\lambda_{e1}}{1-\lambda_{e1}} \ge 0 \Leftrightarrow \lambda_{e1} \le \frac{R^L}{c+R^L}.$$

Thus, for L-class, we can summarize from above that

$$\hat{q}_{2L} = \mathbb{I}\left\{\lambda_{e1} \le \frac{R^L}{c+R^L}\right\}$$
$$\hat{U}^U_{2L} = \begin{cases} R^L - c\frac{\lambda_{e1}}{1-\lambda_{e1}} & \text{if } \lambda_{e1} \le \frac{R^L}{c+R^L} \\ 0 & \text{if } \lambda_{e1} > \frac{R^L}{c+R^L}. \end{cases}$$

Note that if $\hat{U}^U_2 = 0$, no customer will join the system in the first stage.

Below, we characterize the equilibrium joining probabilities of H-class and L-class, given the initial arrival rate to the first stage λ_{0H} and λ_{0L}. First, we consider the case where $R^H \le R^L + \theta s$.

1. If
$$\lambda_0 \le \frac{R^L - s}{c + R^L - s},$$
$$-s + R^L - c\frac{\lambda_0}{1-\lambda_0} \ge 0 \text{ and } -s + R^H - c\frac{\lambda_0}{1-\lambda_0} > 0.$$
$q_{1H} = q_{1L} = 1$, i.e., all customers will join the system in the first stage.

2. If

$$\frac{R^L - s}{c + R^L - s} < \lambda_0 \leq \frac{R^H - s}{c + R^H - s},$$

$$-s + R^L - c\frac{\lambda_0}{1-\lambda_0} < 0 \text{ and } -s + R^H - c\frac{\lambda_0}{1-\lambda_0} \geq 0.$$

The equilibrium arrival rate after the first stage decision solves

$$-s + R^L - c\frac{\lambda_{e1}}{1-\lambda_{e1}} = 0,$$

which leads to

$$\lambda_{e1} = \frac{R^L - s}{c + R^L - s} \text{ and } q_{1H} = 1, q_{1L} = \max\left\{0, 1 - \frac{\lambda_0 - \frac{R^L - s}{c + R^L - s}}{\lambda_{0L}}\right\}.$$

3. If

$$\lambda_0 > \frac{R^H - s}{c + R^H - s} \text{ and } \lambda_{0H} > \frac{R^H - s}{c + R^H - s},$$

$$q_{1L} = 0 \text{ and } q_{1H} = \frac{R^H - s}{\lambda_{0H}(c + R^H - s)}$$

is the unique equilibrium, for the reasons below. For any positive q_{1L}, the equilibrium arrival rate after the first stage decision, $\lambda_{e1L} = \lambda_{0L} q_{1L}$, solves

$$-s + R^L - c\frac{\lambda_{e1L} + \lambda_{e1H}}{1 - (\lambda_{e1L} + \lambda_{e1H})} = 0 \Leftrightarrow \lambda_{e1L} + \lambda_{e1H} = \frac{R^L - s}{c + R^L - s}.$$

However, for the H-class, since $\lambda_{0H} \geq (R^H - s)/(c + R^H - s)$, we can increase λ_{e1H} such that $\lambda_{e1L} + \lambda_{e1H} = (R^H - s)/(c + R^H - s)$. As such, any positive q_{1L} is not an equilibrium. Since $q_{1L} = 0$, the equilibrium arrival rate after the second stage decision, $\lambda_{e1H} = \lambda_{0H} q_{1H}$, solves

$$-s + R^H - c\frac{\lambda_{e1H}}{1 - \lambda_{e1H}} = 0 \Leftrightarrow \lambda_{e1H} = \frac{R^H - s}{c + R^H - s}.$$

4. If

$$\lambda_0 > \frac{R^H - s}{c + R^H - s} \text{ and } \frac{R^L - s}{c + R^L - s} \leq \lambda_{0H} < \frac{R^H - s}{c + R^H - s},$$

$$q_{1H} = 1 \text{ and } q_{1L} = \max\left\{0, 1 - \frac{\lambda_0 - \frac{R^L - s}{c + R^L - s}}{\lambda_{0L}}\right\}$$

is the unique equilibrium for similar arguments. In this case, all H-class customers will join in the first stage because $-s + R^H - c\lambda_0/(1 - \lambda_0) > 0$.

5. If

$$\lambda_0 > \frac{R^H - s}{c + R^H - s} \text{ and } \lambda_{0H} < \frac{R^L - s}{c + R^L - s},$$

$$q_{1H} = 1 \text{ and } q_{1L} = 1 - \frac{\lambda_0 - \frac{R^L - s}{c + R^L - s}}{\lambda_{0L}}$$

is the unique equilibrium for similar arguments. In this case, all H-class customers will join in the first stage but only a fraction of L-class customers will join.

Putting this all together, given λ_{0H} and λ_{0L}, for H-class customers, the equilibrium arrival rate following the first stage decision is

$$\lambda_{e1H} = \min\left\{\lambda_{0H}, \frac{R^H - s}{c + R^H - s}\right\};$$

and the proportion of uninformed customers that join in the first stage is

$$q_{1H} = \min\left\{1, \frac{R^H - s}{\lambda_{0H}(c + R^H - s)}\right\}.$$

For L-class customers, the equilibrium arrival rate following the first stage decision is

$$\lambda_{e1L} = \lambda_{0L}\mathbb{I}\left\{\lambda_0 \le \frac{R^L - s}{c + R^L - s}\right\} + \max\left\{0, \frac{R^L - s}{c + R^L - s} - \lambda_{0H}\right\}\mathbb{I}\left\{\lambda_0 > \frac{R^L - s}{c + R^L - s}, \lambda_{0H} \le \frac{R^H - s}{c + R^H - s}\right\};$$

and the proportion of uninformed customers that join in the first stage is

$$q_{1L} = \mathbb{I}\left\{\lambda_{e1} \le \frac{R^L - s}{c + R^L - s}\right\} + \max\left\{0, \frac{\frac{R^L - s}{c + R^L - s} - \lambda_{0H}}{\lambda_{0L}}\right\}\mathbb{I}\left\{\lambda_0 > \frac{R^L - s}{c + R^L - s}, \lambda_{0H} \le \frac{R^H - s}{c + R^H - s}\right\}.$$

Similarly, for the case where $R^H < R^L + \theta s$, we can show that the equilibrium behavior is the same as what we have derived above. □

Theorem 5. *(Equilibrium Strategy of Uninformed Heterogeneous Customers) For fixed R^H, R^L, c, θ, and s, given the initial arrival rate to the first stage λ_{0H} and λ_{0L}, in the unique global equilibrium of an uninformed system, the proportions of uninformed customers that join in the first stage and wait in the second stage are*

$$(q_{1H}, q_{1L}, q_{2H}, q_{2L}) = \begin{cases} (1,1,1,1) & \text{if } \lambda_0 \le \frac{R^L - s}{c + R^L - s} \\ \left(1, \max\left\{1 - \frac{\lambda_0 - \frac{R^L - s}{c + R^L - s}}{\lambda_{0L}}, 0\right\}, 1, 1\right) & \text{if } \lambda_0 > \frac{R^L - s}{c + R^L - s} \text{ and } \lambda_{0H} \le \frac{R^H - s}{c + R^H - s} \\ \left(\frac{R^H - s}{\lambda_{0H}(c + R^H - s)}, 0, 1, 1\right) & \text{if } \lambda_{0H} > \frac{R^H - s}{c + R^H - s} \end{cases}$$

Proof of Theorem 5. For H-class customers, for all possible values of λ_{e1} in Lemma 13, we have $\lambda_{e1} \le (R^H - s)/(c + R^H - s)$. This implies that

$$\lambda_{e1} \le \frac{R^H + \theta s}{c + R^H + \theta s}.$$

Then, by Lemma 12, we have $q_{2H} = 1$. Similarly, for L-class customers, for all possible values of λ_{e1} in Lemma 13, we have $\lambda_{e1} \le (R^L - s)/(c + R^L - s)$. This implies that

$$\lambda_{e1} \le \frac{R^L + \theta s}{c + R^L + \theta s}.$$

Again, by Lemma 12, we have $q_{2L} = 1$.

Combining the above analysis with the results in in Lemma 13, we have if $\lambda_0 \le (R^L - s)/(c + R^L - s)$, the unique equilibrium is $q_{1H} = 1$, $q_{1L} = 1$, $q_{2H} = 1$, and $q_{2L} = 1$. If $\lambda_0 > (R^L - s)/(c + R^L - s)$ and $\lambda_{0H} \le (R^H - s)/(c + R^H - s)$, the unique equilibrium is

$$q_{1H} = 1, \quad q_{1L} = \max\left\{1 - \frac{\lambda_0 - \frac{R^L - s}{c + R^L - s}}{\lambda_{0L}}, 0\right\}, \quad q_{2H} = 1, \quad \text{and } q_{2L} = 1.$$

If $\lambda_{0H} > (R^H - s)/(c + R^H - s)$, the unique equilibrium is

$$q_{1H} = \frac{R^H - s}{\lambda_{0H}(c + R^H - s)}, \quad q_{1L} = 0, \quad q_{2H} = 1, \quad \text{and } q_{2L} = 1.$$

Note that there is some ambiguity when $s=0$. In this case, the first- and second-stage decisions collapse into a single decision. Without loss of generality, when $s=0$, we set

$$q_{1L}=\max\left\{1-\frac{\lambda_0-\frac{R^L}{c+R^L}}{\lambda_{0L}},0\right\}$$

when $\lambda_0>(R^L-s)/(c+R^L-s)$ and $\lambda_{0H}\le(R^H-s)/(c+R^H-s)$ and

$$q_{1H}=\frac{R^H}{\lambda_{0H}(c+R^H)}$$

when $\lambda_{0H}>(R^H-s)/(c+R^H-s)$. □

Proposition 4. *(Optimal Sunk Cost for Uninformed Heterogeneous Customers) For fixed λ_0, R, c, and θ, the throughput is the largest when $s=0$ for an uninformed system, i.e.,*

$$\max_{s\ge0}\lambda_{e2}^U(s)=\lambda_{e2}^U(0).$$

Proof of Proposition 4 By Theorem 5, the throughput or the equilibrium arrival rate after the second stage decision is

$$\lambda_{e2}^U(s,\lambda_0)=\begin{cases}\lambda_0 & \text{if } \lambda_0\le\frac{R^L-s}{c+R^L-s}\\ \lambda_{0H}+\lambda_{0L}\max\left\{1-\frac{\lambda_0-\frac{R^L-s}{c+R^L-s}}{\lambda_{0L}},0\right\} & \text{if } \lambda_0>\frac{R^L-s}{c+R^L-s} \text{ and } \lambda_{0H}\le\frac{R^H-s}{c+R^H-s}\\ \frac{R^H-s}{c+R^H-s} & \text{if } \lambda_{0H}>\frac{R^H-s}{c+R^H-s}.\end{cases}$$

For any $s>0$, if $\lambda_0\le\frac{R^L-s}{c+R^L-s}<\frac{R^L}{c+R^L}$,

$$\lambda_{e2}^U(s,\lambda_0)=\lambda_0=\lambda_{e2}^U(0,\lambda_0).$$

If $\frac{R^L-s}{c+R^L-s}<\lambda_0\le\frac{R^L}{c+R^L}$ and $\lambda_{0H}\le\frac{R^H-s}{c+R^H-s}$,

$$\lambda_{e2}^U(s,\lambda_0)=\lambda_{0H}+\lambda_{0L}\max\left\{1-\frac{\lambda_0-\frac{R^L-s}{c+R^L-s}}{\lambda_{0L}},0\right\}<\lambda_0=\lambda_{e2}^U(0,\lambda_0).$$

If $\frac{R^L-s}{c+R^L-s}<\lambda_0\le\frac{R^L}{c+R^L}$ and $\lambda_{0H}>\frac{R^H-s}{c+R^H-s}$,

$$\lambda_{e2}^U(s,\lambda_0)=\frac{R^H-s}{c+R^H-s}<\lambda_0=\lambda_{e2}^U(0,\lambda_0).$$

If $\lambda_0>\frac{R^L}{c+R^L}$ and $\lambda_{0H}\le\frac{R^H-s}{c+R^H-s}$,

$$\lambda_{e2}^U(s,\lambda_0)=\lambda_{0H}+\lambda_{0L}\max\left\{1-\frac{\lambda_0-\frac{R^L-s}{c+R^L-s}}{\lambda_{0L}},0\right\}<\lambda_{0H}+\lambda_{0L}\max\left\{1-\frac{\lambda_0-\frac{R^L}{c+R^L}}{\lambda_{0L}},0\right\}=\lambda_{e2}^U(0,\lambda_0).$$

If $\lambda_0>\frac{R^L}{c+R^L}$ and $\frac{R^H-s}{c+R^H-s}<\lambda_{0H}\le\frac{R^H}{c+R^H}$,

$$\lambda_{e2}^U(s,\lambda_0)=\frac{R^H-s}{c+R^H-s}<\lambda_{0H}<\lambda_{0H}+\lambda_{0L}\max\left\{1-\frac{\lambda_0-\frac{R^L}{c+R^L}}{\lambda_{0L}},0\right\}=\lambda_{e2}^U(0,\lambda_0).$$

If $\lambda_{0H}>\frac{R^H}{c+R^H}$, then

$$\lambda_{e2}^U(s,\lambda_0)=\frac{R^H-s}{c+R^H-s}<\frac{R^H}{c+R^H}=\lambda_{e2}^U(0,\lambda_0).$$

Thus, for any fixed λ_0,

$$\max_{s\ge0}\lambda_{e2}^U(s,\lambda_0)=\lambda_{e2}^U(0,\lambda_0).$$

□

We denote $n_s^H = \lfloor \frac{R^H+\theta s}{c} \rfloor + 1$ and $n_s^L = \lfloor \frac{R^L+\theta s}{c} \rfloor + 1$.

Lemma 14. *For fixed R^H, R^L, c, θ, and s, given the equilibrium arrival rate from the first stage to the second stage λ_{e1H} and λ_{e1L}, when $\lambda_{e1} \neq 1$ and $\lambda_{e1H} \neq 1$, the steady-state distribution of the queue length in the second stage is*

$$\begin{aligned}
\pi_0 &= \frac{(1-\lambda_{e1})(1-\lambda_{e1H})}{1-\lambda_{e1H}+\lambda_{e1}^{n_s^L}(\lambda_{e1H}-\lambda_{e1}+\lambda_{e1H}^{n_s^H-n_s^L+1}(\lambda_{e1}-1))} \\
\pi_i &= (\lambda_{e1})^i \pi_0 \quad \textit{for} \quad i = 1, \cdots, n_s^L \\
\pi_i &= (\lambda_{e1H})^{i-n_s^L}(\lambda_{e1})^{n_s^L}\pi_0 \quad \textit{for} \quad i = n_s^L+1, \cdots, n_s^H
\end{aligned}$$

When $\lambda_{e1} = 1$ and $\lambda_{e1H} = 1$, the steady-state distribution of the queue length in the second stage is

$$\pi_i = \frac{1}{n_s^H+1} \quad \textit{for} \quad i = 0, \cdots, n_s^H$$

When $\lambda_{e1} = 1$ and $\lambda_{e1H} < 1$, the steady-state distribution of the queue length in the second stage is

$$\begin{aligned}
\pi_0 &= \frac{1-\lambda_{e1H}}{n_s^L(1-\lambda_{e1H})+\lambda_{e1H}-\lambda_{e1H}^{n_s^H-n_s^L+1}} \\
\pi_i &= \pi_0 \quad \textit{for} \quad i = 1, \cdots, n_s^L \\
\pi_i &= (\lambda_{e1H})^{i-n_s^L}\pi_0 \quad \textit{for} \quad i = n_s^L+1, \cdots, n_s^H
\end{aligned}$$

When $\lambda_{e1} > 1$ and $\lambda_{e1H} = 1$, the steady-state distribution of the queue length in the second stage is

$$\begin{aligned}
\pi_0 &= \frac{1-\lambda_{e1}}{\lambda_{e1}^{n_s^L}(n_s^H-n_s^L)(1-\lambda_{e1})+\lambda_{e1}-\lambda_{e1}^{n_s^L+1}} \\
\pi_i &= (\lambda_{e1})^i \pi_0 \quad \textit{for} \quad i = 1, \cdots, n_s^L \\
\pi_i &= (\lambda_{e1})^{n_s^L}\pi_0 \quad \textit{for} \quad i = n_s^L+1, \cdots, n_s^H
\end{aligned}$$

Proof of Lemma 14. The H-class (L-class) informed customers will wait in the second stage as long as $R^H - cN \geq 0$ ($R^L - cN \geq 0$), where N is the observed queue length (number of customers in the system) upon the arrival of the focal customer at the second stage. Equivalently, H-class (L-class) customers will wait as long as $N < n_s^H = \lfloor \frac{R^H+\theta s}{c} \rfloor + 1$ ($N < n_s^L = \lfloor \frac{R^L+\theta s}{c} \rfloor + 1$). In this case, the system evolves as a multi-class loss queue with different balking thresholds for each class. The balance equations of the steady-state queue length are

$$\begin{aligned}
\lambda_{e1}\pi_0 &= \pi_1 \\
\lambda_{e1}\pi_0 + \pi_2 &= (\lambda_{e1}+1)\pi_1 \\
&\cdots \\
\lambda_{e1}\pi_{n_s^L-2} + \pi_{n_s^L} &= (\lambda_{e1}+1)\pi_{n_s^L-1} \\
\lambda_{e1}\pi_{n_s^L-1} + \pi_{n_s^L+1} &= (\lambda_{e1H}+1)\pi_{n_s^L} \\
\lambda_{e1H}\pi_{n_s^L} + \pi_{n_s^L+2} &= (\lambda_{e1H}+1)\pi_{n_s^L+1} \\
&\cdots \\
\lambda_{e1H}\pi_{n_s^H-2} + \pi_{n_s^H} &= (\lambda_{e1H}+1)\pi_{n_s^H-1} \\
\lambda_{e1H}\pi_{n_s^H-1} &= \pi_{n_s^H}.
\end{aligned}$$

When $\lambda_{e1} \neq 1$ and $\lambda_{e1H} \neq 1$, solving the balance equations gives

$$\pi_0 = \frac{(1-\lambda_{e1})(1-\lambda_{e1H})}{1-\lambda_{e1H}+\lambda_{e1}^{n_s^L}(\lambda_{e1H}-\lambda_{e1}+\lambda_{e1H}^{n_s^H-n_s^L+1}(\lambda_{e1}-1))}$$
$$\pi_i = (\lambda_{e1})^i \pi_0 \quad \text{for} \quad i = 1, \cdots, n_s^L$$
$$\pi_i = (\lambda_{e1H})^{i-n_s^L}(\lambda_{e1})^{n_s^L}\pi_0 \quad \text{for} \quad i = n_s^L+1, \cdots, n_s^H$$

When $\lambda_{e1} = 1$ and $\lambda_{e1H} = 1$, solving the balance equations gives

$$\pi_i = \frac{1}{n_s^H+1} \quad \text{for} \quad i = 0, \cdots, n_s^H$$

When $\lambda_{e1} = 1$ and $\lambda_{e1H} < 1$, solving the balance equations gives

$$\pi_0 = \frac{1-\lambda_{e1H}}{n_s^L(1-\lambda_{e1H})+\lambda_{e1H}-\lambda_{e1H}^{n_s^H-n_s^L+1}}$$
$$\pi_i = \pi_0 \quad \text{for} \quad i = 1, \cdots, n_s^L$$
$$\pi_i = (\lambda_{e1H})^{i-n_s^L}\pi_0 \quad \text{for} \quad i = n_s^L+1, \cdots, n_s^H$$

When $\lambda_{e1} > 1$ and $\lambda_{e1H} = 1$, solving the balance equations gives

$$\pi_0 = \frac{1-\lambda_{e1}}{\lambda_{e1}^{n_s^L}(n_s^H-n_s^L)(1-\lambda_{e1})+\lambda_{e1}-\lambda_{e1}^{n_s^L+1}}$$
$$\pi_i = (\lambda_{e1})^i \pi_0 \quad \text{for} \quad i = 1, \cdots, n_s^L$$
$$\pi_i = (\lambda_{e1})^{n_s^L}\pi_0 \quad \text{for} \quad i = n_s^L+1, \cdots, n_s^H$$

□

Proof of Theorem 4. We begin with the first part of the theorem. If

$$\lambda_0 \leq \frac{R^L}{c+R^L},$$

by Theorem 5,

$$\lambda_{e2}^U(0,\lambda_0) = \lambda_0 \geq \max_{s\geq 0} \lambda_{e2}^I(s,\lambda_0).$$

Now we turn to the second part of the theorem. By Proposition 4 and Theorem 5, we know that the throughput is the largest when $s = 0$ for an uninformed system and

$$\max_{s\geq 0} \lambda_{e2}^U(s,\lambda_0) = \lambda_{e2}^U(0,\lambda_0) = \frac{R^H}{c+R^H}.$$

In other words, regardless of λ_0, the equilibrium throughput after the second stage decision is capped at $(R^H)/(c+R^H)$. Define

$$f_H(\lambda_H,\lambda_L;s) = -s + \mathbb{E}[\hat{U}_{2H}] = -s + \sum_{i=0}^{n_0^H-1}(R^H-ci)\pi_i$$

and

$$f_L(\lambda_H,\lambda_L;s) = -s + \mathbb{E}[\hat{U}_{2L}] = -s + \sum_{i=0}^{n_0^L-1}(R^L-ci)\pi_i,$$

where π_i depends on λ_H and λ_L as characterized in Lemma 14 with $\lambda_{e1H} = \lambda_H$ and $\lambda_{e1L} = \lambda_L$. $f_H(\lambda_H,\lambda_L;s)$ ($f_H(\lambda_H,\lambda_L;s)$) can be interpreted as the H-class (L-class) customer's utility of joining in the first stage when $\lambda_{e1H} = \lambda_H$ and $\lambda_{e1L} = \lambda_L$.

Next, we show that when $\lambda_{0H} \geq 1$ and $s = 0$, the equilibrium arrival rate after first stage decision being $\lambda_{e1H} = 1$ and $\lambda_{e1L} = 0$ is incentive compatible, i.e.,

$$f_H(1,0;0) \geq 0 \text{ and } f_L(1,0;0) \geq 0.$$

By Lemma 14,

$$\begin{aligned}
f_H(1,0;0) &= \sum_{i=0}^{n_0^H - 1} (R^H - ci)\pi_i \\
&= R^H \frac{n_0^H}{n_s^H + 1} - c\frac{1}{n_s^H + 1}\frac{(n_0^H - 1)n_0^H}{2} \\
&= \frac{n_0^H}{n_s^H + 1}(R^H - \frac{c(n_0^H - 1)}{2}) \\
&\geq \frac{n_0^H}{n_s^H + 1}(R^H - \frac{c(\frac{R^H}{c} + 1 - 1)}{2}) \\
&= \frac{n_0^H R^H}{2(n_s^H + 1)} \\
&> 0.
\end{aligned}$$

Similarly we can show that $f_L(1,0;0) > 0$. When $\lambda_{e1H} = 1$ and $\lambda_{e1L} = 0$, by Lemma 14, the proportion of customers that wait in the second stage is

$$1 - \pi_{n_s^H} = 1 - \frac{1}{n_s^H + 1} = \frac{n_s^H}{n_s^H + 1}.$$

It suffices to show that

$$\frac{n_s^H}{n_s^H + 1} \geq \frac{R^H}{c + R^H}$$

where $(R^H)/(c + R^H)$ is the equilibrium throughput after the second stage decision when no information is provided.

$$\begin{aligned}
n_s^H(c + R^H) - (n_s^H + 1)R^H &= cn_s^H - R^H \\
&\geq c\frac{R^H + \theta s}{c} - R^H \\
&= \theta s.
\end{aligned}$$

We can conclude that if $\lambda_0 \leq \frac{R^L}{c + R^L}$

$$\max_{s \geq 0, \gamma \in \{I,U\}} \lambda_{e2}^{\gamma}(s) = \lambda_{e2}^{U}(0).$$

If $\lambda_{0H} \geq 1$,

$$\max_{s \geq 0, \gamma \in \{I,U\}} \lambda_{e2}^{\gamma}(s) = \max_{s \geq 0} \lambda_{e2}^{I}(s).$$

□

Appendix E: Supplementary Numerical Experiment

In this section, we provide additional numerical results to complement those presented in Section 8. Figure 19 shows the optimal policy and optimal level of sunk cost for extended ranges of the sensitivity to sunk cost parameter θ and the initial arrival rate λ_0. Figure 20 shows the optimal policy and optimal level of sunk cost under different waiting costs. Figure 21 shows the improvement to the system throughput for extended regimes of the sensitivity to sunk cost parameter θ and the initial arrival rate λ_0. Figure 22 shows the improvement to the system throughput under different waiting costs. Figure 23 shows the worst-case suboptimality gap of using policies derived from *overestimated* θ's under different waiting costs. Figure 24 shows the worst-case suboptimality gap of using policies derived from *underestimated* θ's under different waiting costs. Figure 25 shows the worst-case improvement of the system throughput when using policies derived from overestimated or underestimated θ's compared to the zero sunk cost policy under different waiting costs.

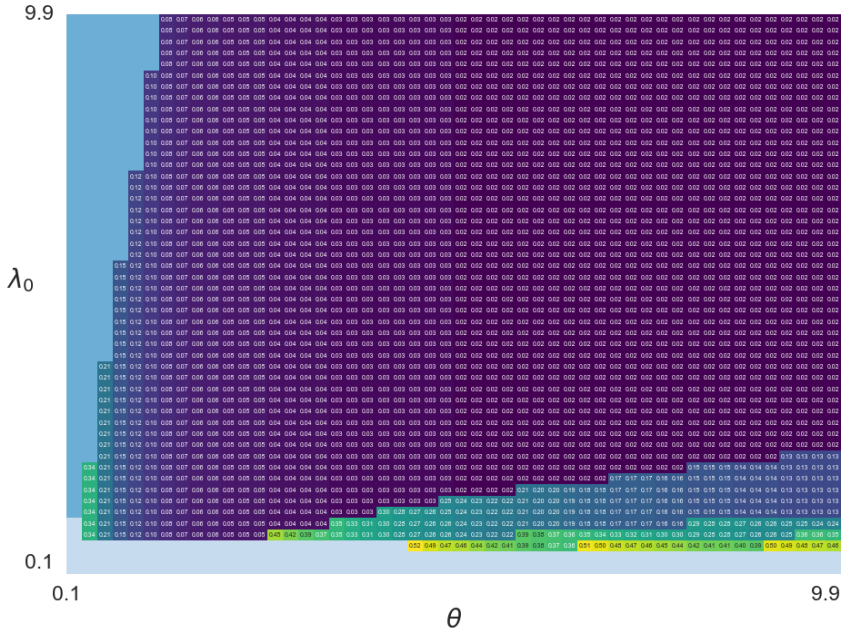
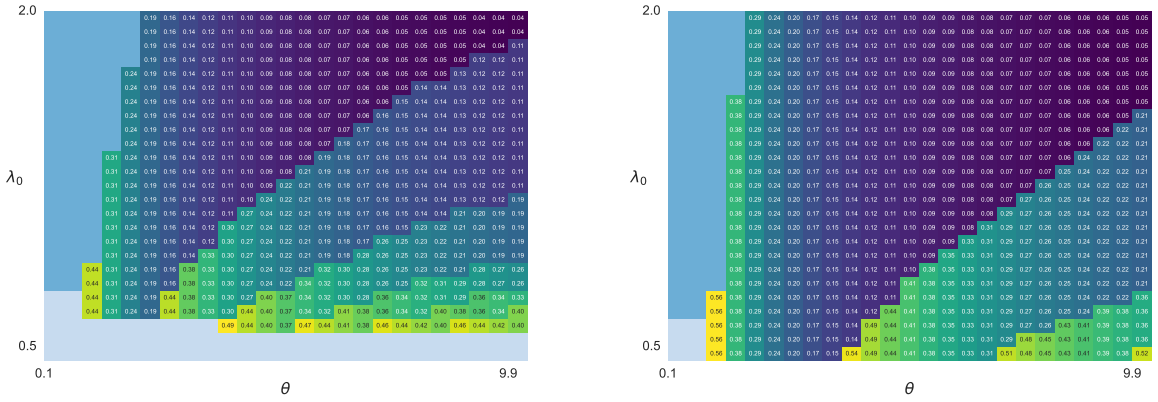


Figure 19 Optimal policy for the service provider when $R = 1$ and $c = 1.1$. The sensitivity to sunk cost parameter θ ranges from 0.1 to 9.9. The initial arrival rate λ_0 ranges from 0.1 to 9.9. Lighter blue without numbers indicates that the optimal policy is None. Darker blue without numbers indicates that the optimal policy is Info Only. The region with numbers indicates that the optimal policy is Both and the numbers indicate the optimal sunk cost.



(a) $c = 0.7$ (b) $c = 1.5$

Figure 20 **Optimal policy for the service provider when $R = 1$ and $c = 0.7$ (column (a)) or $c = 1.5$ (column (b)). The sensitivity to sunk cost parameter θ ranges from 0.1 to 9.9. The initial arrival rate λ_0 ranges from 0.5 to 2.0. Lighter blue without numbers indicates that the optimal policy is None. Darker blue without numbers indicates that the optimal policy is Info Only. The region with numbers indicates that the optimal policy is Both and the numbers indicate the optimal sunk cost.**

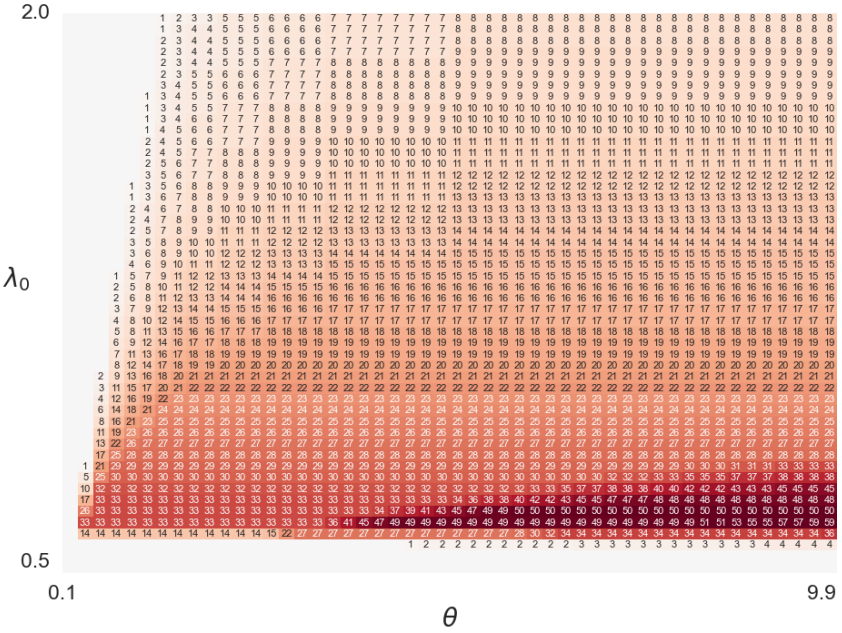
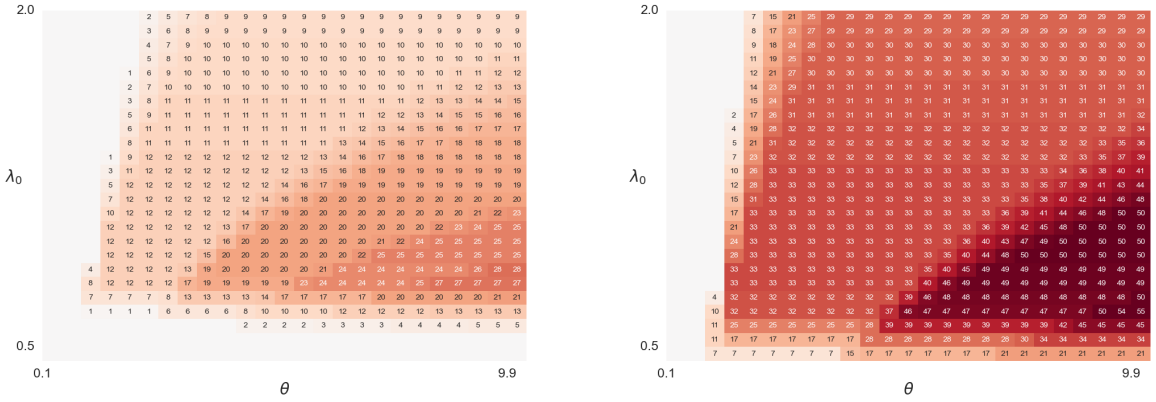
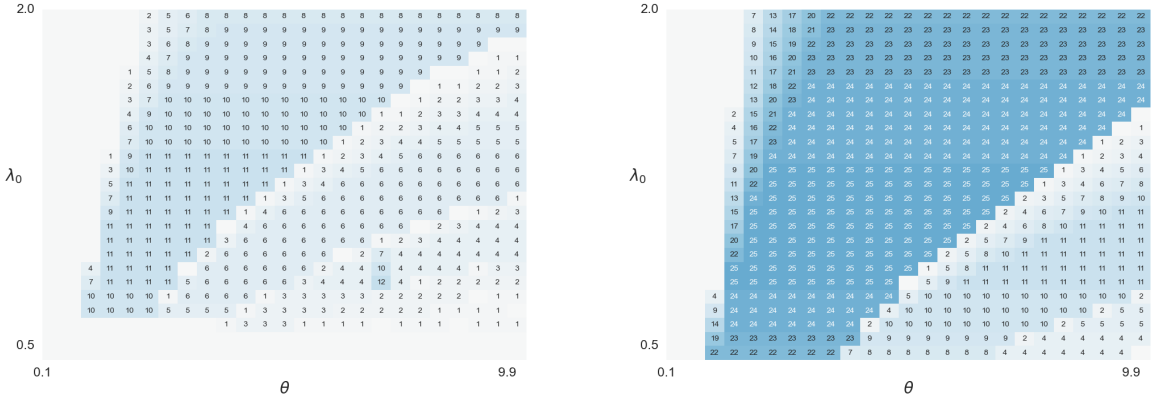


Figure 21 **Improvement to the system throughput of the optimal policy against the suboptimal policy using zero sunk cost when $R = 1$ and $c = 1.1$. Numbers are in percentage scale.**



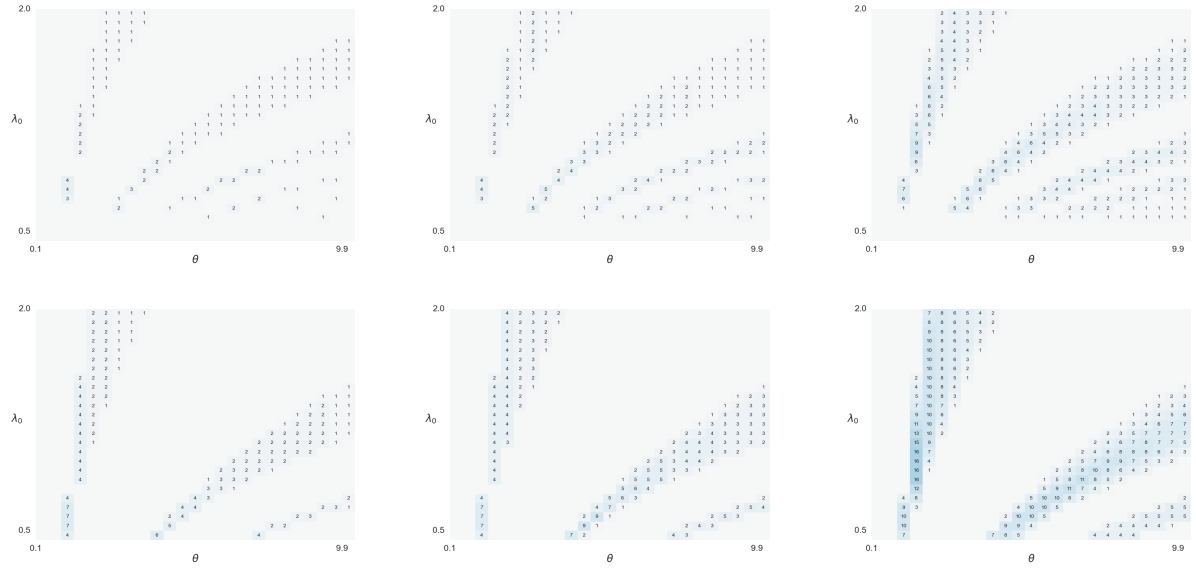
(a) $c = 0.7$ (b) $c = 1.5$

Figure 22 **Improvement to the system throughput of the optimal policy against the suboptimal policy using zero sunk cost. $R = 1$ and $c = 0.7$ or $c = 1.5$. Numbers are in percentage scale.**



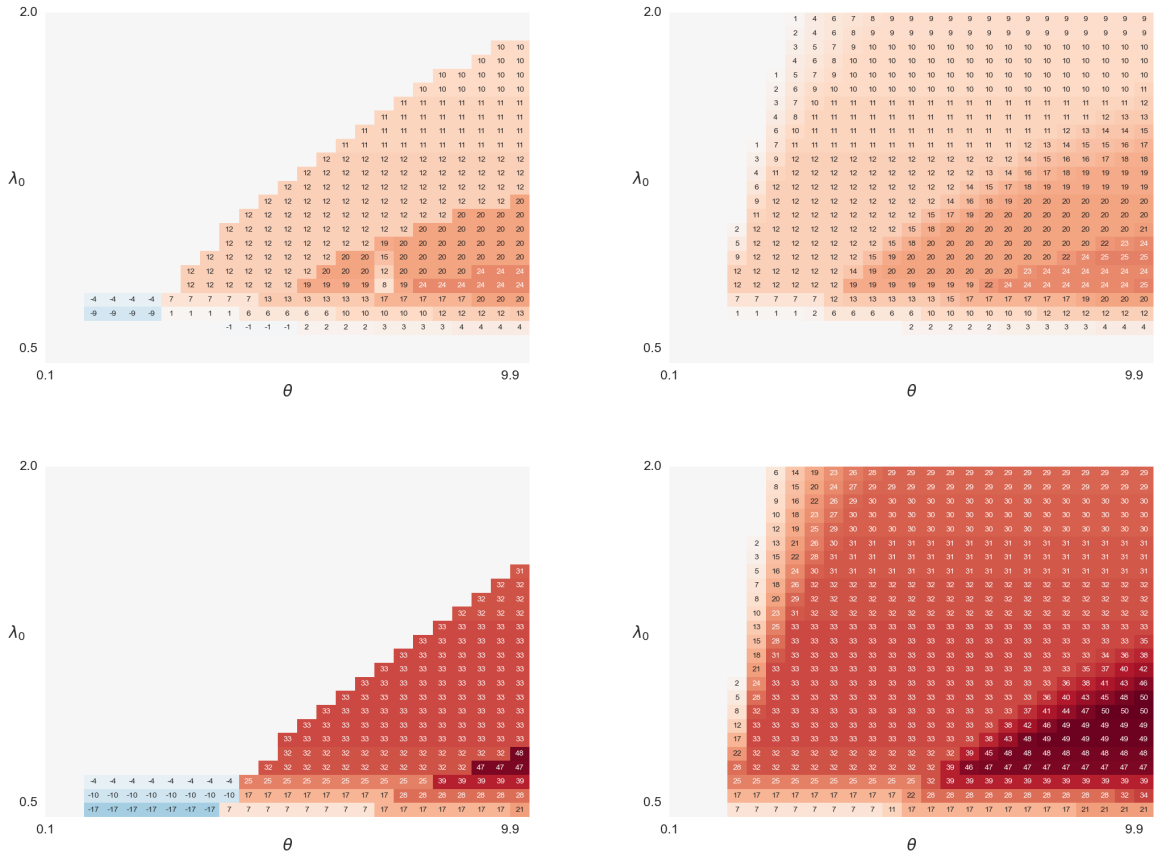
(a) $c = 0.7$ (b) $c = 1.5$

Figure 23 **Worst-case suboptimality gap of using policies derived from overestimated θ by as large as 1% deviation, i.e., $\delta = 1\%$. $R = 1$ and $c = 0.7$ or $c = 1.5$. Numbers are in percentage scale.**



(a) 5% deviation (b) 10% deviation (c) 20% deviation

Figure 24 Worst-case suboptimality gap of using policies derived from underestimated θ by as large as 5% (left panel), 10% (middle panel), and 20% (right panel) deviations. $R=1$ and $c=0.7$ (first row) or $c=1.5$ (second row). Numbers are in percentage scale.



(a) overestimated by 1% (b) underestimated by 20%

Figure 25 Worst-case improvement to the system throughput of using policies derived from overestimated θ by as large as 1% (left panel) or underestimated θ by as large as 20% (right panel) against the policy using zero sunk cost. $R = 1$ and $c = 0.7$ (first row) or $c = 1.5$ (second row). Blue indicates that the worst-case policy underperforms the baseline suboptimal policy. Numbers are in percentage scale.