Born to wait?

A study on allocation rules in booking systems

Lingbo Huang

Center for Economics Research, Shandong University.

Address: 27 Shanda South Road, Licheng District, Jinan, China, 250100

Email: lingbo.huang@outlook.com.

Tracy Xiao Liu

Department of Economics, School of Economics and Management, Tsinghua University.

Address: No 30, Shuangqing road, Haidian District, Beijing, China, 100872

Email: liuxiao@sem.tsinghua.edu.cn.

Jun Zhang

Institute for Social and Economic Research, Nanjing Audit University.

Address: 86 Yushan West Road, Pukou District, Nanjing, China, 211815

Email: zhangjun404@gmail.com.

December 23, 2024

Abstract

Queue-based rules for allocating scarce goods are widely utilized in booking systems due to their perceived efficiency. However, empirical investigations into the externalities and opportunity costs of queuing in multitasking scenarios are limited. This paper reports on two laboratory experiments that compare a queue-based rule with a lottery-based rule by quantifying their respective efficiency losses. Our findings indicate that while the queue-based rule demonstrates superior allocative efficiency, it incurs significant losses in productive efficiency attributed to opportunity costs of time. In contrast, the lottery-based rule exhibits improved overall efficiency with minimal time spent on the booking system. Additionally, under the queue-based rule, participants display bimodal behavior, either engaging fully or abstaining from the booking system, influenced by their time valuations. Further, while providing queue length information facilitates more efficient coordination, it also leads to more frequent task-switching

behavior that negates any productive efficiency gain from improved coordination. This research underscores the crucial need to reevaluate allocation mechanisms in booking systems, taking into account their externalities.

Keywords: market design; booking system; queue; lottery; opportunity cost of time

Received: February, 2024; Accepted: December 2024 by Gary Bolton after two revisions.

JEL Classification: C92, D47

The first parent lined up at 4 a.m. on a Sunday, when the only other people around were out just long enough to stumble from warm taxis through sobering 19-degree air into their homes.

Twenty minutes later, other parents showed up and a line began to form down Atlantic Avenue in Brooklyn. One father kept a list so that anyone searching for a thawing hot coffee could do so without losing a place in the line. He abandoned that project as more and more people trickled in and the end of the line was no longer visible from the front.

[...] It was too dark to read, so they chatted about things like schools or children, and they poked fun at one another for being there. Every few minutes, someone would check his watch and express the hope that Carmelo the Science Fellow would open his doors early for his annual summer camp registration.

— "Born to wait", New York Times, February 22th, 2013

1 Introduction

Queue rules based on a first-come, first-served basis are commonly used to manage the allocation of scarce resources. Extensive research has been conducted in various academic fields, including computer science, operations research, and economics, to analyze the performance of queue rules. In some situations, people put their names on a waiting list for goods that arrive over time (e.g., public housing, daycare spots, and deceased donor organs), without spending time queuing. But in many other situations, people must physically wait in line. For instance, the epigraph quoted from a New York Times article describes how New York City parents wait in line for hours in advance at the entrance of the institute for science camp registration. Similar situations arise when consumers line up in front of Apple Stores during the release of new iPhone models or when shoppers wait for hours to gain entry into retail stores during Black Friday sales.

While previous studies have provided valuable insights into the optimal design of a queue rule that restricts attention to an allocation problem,[1] little attention has been paid to the externality of a queue rule on other parallel activities in which people participate.[2] In particular, the first-come, first-served rule may result in individuals devoting significant time and effort to unproductive activities like queuing, thereby reducing the time available for other productive activities. According to a New York Times article, Americans spend

[1]For instance, Platz and Østerdal (2017); Bloch and Cantala (2017); Che and Tercieux (2024); Schummer (2021); Leshno (2022), among many others.

[2]The literature on transportation economics has emphasized the time lost due to traffic congestion as one of the largest externalities associated with the use of automobiles (see Naor (1969); Parry, Walls and Harrington (2007); Heller et al. (2019), among others). But, as far as we know, such externalities due to time lost have not attracted enough attention in other literature.

approximately 37 billion hours annually waiting in line, highlighting the magnitude of this issue.[3] A survey conducted in 2014 indicated that U.S. businesses lose around $130 billion in employee productivity every year ($900 per employee) due to the time wasted on service inefficiencies during working hours; 40% of surveyees reported spending at least one hour waiting in line or on a telephone queue to resolve service-related issues.[4]

In this paper, we study a multi-tasking situation in which people can obtain scarce goods (e.g., time slots, event tickets) via a booking system while also having the chance to spend time on a productive activity. An example is the science camp registration quoted above where parents could have spent the queuing time on other productive work. We investigate the extent of externalities generated by different allocation rules within the booking system on the parallel productive activity. In particular, we compare a queue rule with an alternative rule based on lottery. A queue rule could potentially enhance *allocative efficiency* since the queuing time could signal people's valuation. We, however, highlight the time cost incurred by queuing. Since the time people spend queuing could have been utilized for the parallel productive activity, the queue rule could lead to lower *productive efficiency*. On the other hand, a lottery rule eliminates the necessity of competition via spending time on the booking system, although it could hamper allocative efficiency.[5]

Learning about individuals' opportunity costs of time is critical to quantifying the aforementioned two types of efficiency, that is, allocative efficiency in the booking system and productive efficiency in the productive activity. First, to quantify the productive efficiency, we must compute individuals' forgone payoffs from the productive activity due to the time they spend on the booking system. Second, individuals' opportunity costs of time also affect the time they spend in a booking queue, which complicates the determination of allocative efficiency. As Holt and Sherman (1982) and the follow-up literature have theoretically shown,[6] if individuals' opportunity costs of time are heterogeneous, the queue rule does not necessarily produce a more efficient allocation of goods than a random allocation.[7] Using field data to quantify these types of

[3] "Why Waiting is Torture," *New York Times*, August 18th, 2012.

[4] See https://www.huffpost.com/entry/waiting-in-line-is-bad-bu_b_12523316; last accessed on May 26, 2022.

[5] Lottery is widely used in market design environments, including public school choices (Abdulkadiroğlu and Sönmez, 2003), on-campus housing placements (Chen and Sönmez, 2002), allocation of vehicle licenses (Li, 2018), and allocation of nonimmigrant work permits in the U.S. (Pathak, Rees-Jones and Sönmez, 2022). The market design literature advocates lottery for fairness, while we highlight its advantage in saving people's time costs.

[6] The economics of rationing and queuing has been studied by Tobin (1952); Nichols, Smolensky and Tideman (1971); Barzel (1974); Holt and Sherman (1982); Suen (1989); Taylor, Tsui and Zhu (2003), among others.

[7] In theory, a participant's queuing time is a function of her valuation of goods and her opportunity cost of time. If a participant with a high valuation also has a high opportunity cost of time, she may actually spend *less* time queuing compared to a participant with a low valuation and a low opportunity cost of time.

efficiency is challenging given the difficulty in obtaining individual-level data on opportunity costs of time. Therefore, we introduce an experimental framework that allows for a tight control of individual opportunity cost of time, enabling quantification of the different sources of efficiency.

To quantify and compare the different sources of efficiency losses under different allocation rules, we design an experiment in a dual-tasking environment in which each participant faces two simultaneous tasks in four minutes: an appointment booking task and a production task. In the booking task, each participant needs to book one slot, for which their valuation is private and randomly generated. Participants compete for these slots by queueing or entering a lottery, depending on the treatments. The slots are allocated at the end of the fourth minute. In the production task, participants earn a flat payoff for every second spent on the task screen, which is also generated privately and independently for each participant. This payoff reflects the opportunity cost of the time spent on the booking task. Participants can freely switch between the two tasks at any time, but cannot work on both simultaneously. So, they face a time allocation problem between the two tasks.

Using a between-subjects design, we compare two booking rules: the queue rule and the lottery rule. Under the lottery rule, when participants visit the booking system, they can apply to enter a lottery by pressing a button on the screen at any time in a round, and slots are randomly allocated to applicants at the end of the round. Under the queue rule, participants can enter the queue in the booking system at any time in a round and remain in the queue. But if a participant switches to the production task and later returns to the booking system, she must go to the back of the queue. Slots are allocated according to participants' ranks in the queue at the end of the round.

Under the queue rule, we further vary whether participants can observe the current queue length and their ranks upon entering the queue. In some real-life queues, participants can see where they are and use that information to infer their winning probabilities. In other queuing situations, participants may be uncertain about their winning probabilities, especially when queues are long or when the supply of goods is uncertain. Therefore, we design two treatments in which queuing participants either have precise knowledge of their ranks in queues or have no such information at all. We want to investigate whether providing such ranking information can help improve the overall efficiency of queuing systems. For example, those who realize they have entered the queue too late to secure a slot might choose to leave before the round ends, potentially

reducing productive inefficiency.

We distinguish between two sources of efficiency loss in our theoretical framework: the inefficient allocation of booking slots (allocative efficiency loss) and the total of participants' opportunity cost of time spent on the booking task (productive efficiency loss). Consistent with our theoretical predictions, our experimental results show that queue participants spend substantial amounts of time on the booking task while lottery participants spend only a few seconds submitting their lottery entry and the remainder of their time on the production task. Although allocative efficiency is higher under the queue rule, the productive efficiency loss associated with this rule far exceeds the allocative efficiency improvement, resulting in significantly lower overall efficiency compared to the lottery rule. Additionally, we observe bimodal behavior under the queue rule, which is significantly correlated with participants' time valuations: those with high time valuations (i.e., the ratio of monetary slot valuation and opportunity cost of time) tend to spend nearly all of their time queuing, while those with low time valuations tend to spend little to no time queuing. Moreover, we find that providing ranking information to queuing participants does not influence their overall queuing time. Although observable queues promote more efficient coordination, they simultaneously lead to much more frequent task-switching. Consequently, productive efficiency loss attributable to these switches negates any efficiency gain through improved coordination. Furthermore, while most switches can be considered rational when queues are too short or too long, a small fraction of plausibly irrational switching behavior (around 10%) has a disproportionate impact on productive efficiency loss (around 40%).

We examine the robustness of our main findings in alternative settings. The experimental manipulation is summarized in Section 5. We find that regardless of the degree of market competitiveness, the nature of the task (abstract vs. real-effort), the complexity of the booking system (single vs. two-stage; solo vs. dual-track), queuing consistently induces lower overall efficiency compared to lottery. At the individual level, similar bimodal behavior is also observed across these settings, although it is not significantly associated with participants' time valuations. This lack of association is likely due to the endogenous nature of productivity in the real effort task, which makes it difficult for participants to accurately evaluate their opportunity cost of time.

Our paper is positioned within the broad experimental literature on matching markets (see Roth (2021) and Hakimov and Kübler (2021) for recent surveys). However, we differ from this literature in our introduc-

tion of a new experimental framework for evaluating various forms of efficiency loss that arise during the matching process. In a related study, Hakimov et al. (2021) observe scalping in booking systems based on first-come, first-served rules, and propose a lottery-based batch system that periodically collects applications and then draws lotteries to allocate slots. This system eliminates the importance of speed in the allocation process and thus deters scalpers from entering the market. Our paper could complement their study by demonstrating another advantage of lottery-based booking systems, that is, eliminating the efficiency loss due to the opportunity cost of time spent queuing in (offline) booking systems.

Our multitasking experimental environment parallels previous studies in which agents can choose between work tasks and leisure activities. The leisure utility is represented by either an abstract flat wage (Li, Lariviere and Bassamboo, 2024; Beer, Qi and Ríos, 2024; Dutcher, Salmon and Saral, 2024), similar to our experiment, or a tangible activity such as Internet browsing (Corgnet, Hernán-González and Schniter, 2015; Corgnet, Gómez-Minambres and Hernán-González, 2015; Corgnet and Hernán-González, 2019). However, these studies address different sets of research questions, including inter-team dynamics and principal-agent incentive problems. In contrast, the purpose of our research is to quantify efficiency losses arising from externalities in various allocation systems.

Our study also contributes to the burgeoning literature on behavioral queues, specifically the behavioral impacts of queuing systems on the individual behaviors of both customers and service workers. For example, Allon and Hanany (2012) theoretically investigates how social norms and community enforcement can rationalize the phenomenon of cutting in line and the rejection of such intrusions. Buell (2021) demonstrates that last-place aversion can lead to inefficient switching and abandonment behavior. Shunko, Niederhoff and Rosokha (2018) identifies that both parallel queues and queue-length visibility exert behavioral impacts on service worker productivity. Estrada Rodriguez, Ibrahim and Zhan (2024) finds that lying aversion may limit customers' attempts to reduce their waiting times through misreporting their private information in unobservable queues. Wang and Zhou (2018) finds in a natural field experiment that shared queues, as opposed to dedicated queues, slow down service times due to the social loafing effect. In contrast, our experiment demonstrates that customers exhibit bimodal behavior in queues and are more likely to do so in unobservable queues, as opposed to observable ones.

Finally, it is important to note that there are two types of allocation systems also called queues in real life but

different from the one we study. One type is a queue system where a facility continuously provides services to people who arrive over time in which situation an important reason for people to arrive earlier is to be served earlier.[8] For instance, at airports, passengers are checked in based on the order of their arrival. In our queue system, slots on booking systems are released at a predetermined time, so an earlier arrival does not result in an earlier assignment. The other type is a waiting list, where people enter their names on a list but do not physically queue. In this situation, there is no opportunity cost of time as we study, but there may be other types of waiting costs for individuals who are delay sensitive.[9]

The rest of the paper is organized as follows. Section 2 presents the theoretical framework. Section 3 describes the experimental design and outline hypotheses. Section 4 reports our experimental results. Section 5 summarizes main findings of the robustness experiment. Section 6 provides concluding remarks.

2 Theoretical Framework

This section presents the theoretical framework that guides our experimental design. In our model, n participants face a time allocation problem between the booking and a production task. There are m identical slots in the booking task, and $n > m \geq 1$. Denote the set of participants by $I = \{1, 2, \ldots, n\}$ and denote the set of slots by $S = \{s_1, s_2, \ldots, s_m\}$. Each participant i demands one slot and values each slot at $v_i \in \mathbf{R}_+$. Each v_i is independently drawn from a commonly known uniform distribution on an interval $[\underline{v}, \overline{v}] \subset \mathbf{R}_+$. Each i knows her valuation v_i but does not know the other participants' valuations except for the underlying distribution. Each i also has constant productivity denoted by $w_i \in \mathbf{R}_+$ on the production task, which means that i will obtain a payoff w_i by spending one unit of time on the production task. Each w_i is independently drawn from a commonly known uniform distribution on an interval $[\underline{w}, \overline{w}] \subset \mathbf{R}_+$, and it is also independent of v_i. Each i knows her productivity w_i but does not know the other participants' productivity except for the underlying distribution. We call $y_i = v_i / w_i$ the *time valuation* of slots for participant i, which is the valuation of slots measured in time units. So, y_i is distributed on $[\underline{y}, \overline{y}]$, where $\underline{y} = \underline{v}/\overline{w}$ and $\overline{y} = \overline{v}/\underline{w}$. Let F denote the cumulative distribution function of y_i, which is not a uniform distribution. As will be shown shortly,

[8] Numerous studies in economics and management science have been devoted to this type of queue (e.g., Naor, 1969; Platz and Østerdal, 2017; Che and Tercieux, 2024).

[9] Some papers have studied the trade-off between quick matching to cut down waiting costs and slow matching to generate higher match surplus (e.g., Akbarpour, Li and Gharan (2020); Baccara, Lee and Yariv (2020); Schummer (2021); Leshno (2022)). Other papers take agents' waiting time as endogenous choices and design mechanisms to encourage truthful reports (e.g., Schummer and Abizada (2017); Dimakopoulos and Heller (2019)).

in equilibrium, participants' strategies are determined by their time valuation of slots. We assume that all participants are risk-neutral. Each participant has $T \in \mathbf{R}_+$ units of time to allocate between the two tasks.

In the booking task, slots are assigned using either the queue rule or the lottery rule. Under the *lottery rule*, participants do not need to spend time on the booking task. They only need to show up once in the booking task to become an applicant and the rule assigns slots to applicants uniformly at random. If applicants are no more than the number of slots, every applicant is assigned a slot. Therefore, under the lottery rule, participants can spend all of their time on the production task.

In contrast, under the *queue rule*, slots are assigned to participants based on their queuing time in a line at the end of the game. If there are more participants in the queue than the number of slots, only the first m participants in the queue are assigned a slot. Otherwise, every participant in the queue is assigned a slot. Participants face a time allocation problem between queuing for the booking task and engaging in the production task.

To understand how participants allocate their time between the two tasks under the queue rule, following the literature (e.g., Holt and Sherman, 1982; Suen, 1989; Taylor, Tsui and Zhu, 2003), we model the queuing game as an all-pay auction in which participants compete for slots by bidding their amount of queuing time. The symmetric Bayesian Nash equilibrium is characterized by an increasing function $t(y)$, which determines a participant's queuing time when her time valuation is y. It is worth emphasizing that the queuing auction we study here differs from the standard auction in which participants' bidding strategies are determined by their valuation of slots. Here, because participants may have heterogeneous productivity, a participant with a high valuation of slots and an even higher productivity may spend less time queuing than another participant with a low valuation of slots and an even lower productivity. In our analysis, the bid cap T is ignored because our experimental parameters are carefully chosen to ensure that the cap is never binding.

Formally, let H denote the cumulative distribution function of the m-th order statistics among $n-1$ independent draws from the time valuation distribution, F. Then, $H(y_i)$ is the probability for any participant i with time valuation y_i to win a slot in equilibrium. Proposition 1 derives the symmetric Bayesian Nash equilibrium. The proof is provided in Section A of the E-Companion.

Proposition 1. *Under the queue rule in the booking task, in the symmetric Bayesian Nash equilibrium, every*

participant with time valuation y_i spends $t(y_i)$ units of time in the queue, where $t(y_i) = y_i H(y_i) - \int_{\underline{y}}^{y_i} H(s)ds$.

The above analysis assumes that all players queue until slots are allocated. However, some real-life situations correspond to an alternative queue rule in which participants arriving too late to secure a slot are informed of this when they arrive, thus avoiding unnecessary waiting. Although this alternative queue rule would appear to conserve participants' queuing time, Holt and Sherman (1982) have shown that modeling it as a winner-pay auction results in an equilibrium effect that encourages participants to arrive earlier. As a consequence, the equilibrium expected queuing time for each participant remains unchanged.[10] This observation is reminiscent of the revenue equivalence theorem in the standard auction theory (Myerson, 1981; Riley and Samuelson, 1981). Therefore, using either setup of the queue rule does not change our comparison between the two allocation rules in theory.

An immediate corollary to Proposition 1 is that participants spend more time queuing when they have higher time valuations.

Corollary 1. *Under the queue rule in the booking task, participants with higher time valuations spend more time queuing.*

Next, we utilize the equilibrium outcome to compare the efficiency under the two allocation rules. Let a function $\mu : S \to I \cup \{0\}$ denote an allocation of slots where for every slot $s \in S$, $\mu(s)$ denotes the participant who obtains s, and if $\mu(s) = 0$, it means that s is unassigned. Let μ denote the allocation of slots and let $t_i \in [0, T]$ denote the units of time that every participant i spends on the booking task in the allocation process. We identify two types of potential efficiency loss for each rule. The first type is the efficiency loss in the (mis)allocation of slots, which we call *allocative efficiency loss*. In the most efficient allocation, slots should be allocated to those who value them the most. Given participants' valuations of slots, let $v_{(\ell)}$ denote the ℓ-th highest valuation among the n participants. Then, taking the most efficient allocation as the benchmark, we define the (expected) allocative efficiency loss of the rule as follows:

$$\text{Allocative efficiency loss} = \textstyle\sum_{\ell=1}^{m} \mathbb{E}[v_{(\ell)}] - \sum_{\ell=1}^{m} \mathbb{E}[v_{\mu(s_\ell)}].^{11}$$

[10] Specifically, under the alternative queue rule, every participant with time valuation y_i will bid the amount of queuing time $t'(y_i) = y_i - \frac{\int_{\underline{y}}^{y_i} H(s)ds}{H(y_i)}$. Because such a participant has a probability $H(y_i)$ of winning a slot in the equilibrium, his expected queuing time is $H(y_i)t'(y_i)$, which equals $t(y_i)$ in Proposition 1.

[11] If a slot is unassigned, we let $v_0 = 0$.

The second type is the efficiency loss stemming from the opportunity costs associated with participants not working on the production task but on the booking task, which we call *productive efficiency loss*:

$$\text{Productive efficiency loss} = \sum_{i=1}^{n} \mathbb{E}[t_i w_i].$$

In our theoretical framework, participants' valuations of slots are positively correlated with their time valuations, though they are not fully aligned. Consequently, our analysis predicts that both types of efficiency loss will be positive under the queue rule, while only the allocative efficiency loss will be positive under the lottery rule. Additionally, the allocative efficiency loss under the queue rule is expected to be smaller than that under the lottery rule. When considering both types of efficiency loss, the overall efficiency of each rule will depend on the specific parameters of our model. We will provide more precise predictions following a presentation of our experimental design and parameters.

3 Design of Experiment 1

Motivated by our theoretical framework, we implement a dual-tasking experimental environment. Participants are randomly matched into groups of five for each round of the experiment. Each session consists of eight rounds, with each round lasting four minutes. After each round, participants are randomly rematched to simulate a one-shot setting. In each round, a participant engages in two tasks displayed on separate screens: an appointment booking task and an abstract-effort production task designed to impose an exogenous opportunity cost of queuing. At the beginning of each round, each participant chooses which task to display first. During each round, participants can freely switch between the two tasks at any time and as often as they wish.

In the booking task, three slots are available for each group of five participants, with each participant allowed to acquire at most one slot per round. At the beginning of each round, participants are privately informed of their valuation for a slot, drawn independently from a uniform distribution ranging from 400 to 600 Experimental Currency Units (ECUs). During the round, participants compete for slots either by queuing or entering a lottery, depending on their treatment condition. All slots are allocated at the end of a round, and any unassigned slots are wasted.

In the production task, participants receive a flat payoff for every second they spend on the task screen.

The payoff per second for every participant is privately and independently drawn from 1.50 to 2.50 ECUs (accurate to two decimal places).[12] To mitigate the potential psychological costs of idleness, we also offer participants the option to engage in an unpaid simple counting task. We choose this abstract setting rather than a real-effort production environment since it allows for tighter control over participants' productivity, thereby facilitating a cleaner test of the theoretical predictions.

3.1 The booking task

In the booking task, slots are assigned using either the queue rule or the lottery rule. Under the *queue rule*, participants can enter the booking system at any time each round and immediately begin queuing by pressing a button on the screen. Those who enter the booking task earlier occupy a higher position in the queue. However, if a participant switches to the production task and later returns to the booking system, she must go to the back of the queue.

The *lottery rule* collects participants' applications and assigns slots randomly to applicants at the end of each round. Participants can enter the booking system at any time during the round and apply for entry into the lottery by pressing a button on the screen. All applications are gathered into a virtual urn. When the round ends, applications are randomly drawn from the urn one by one until all available slots have been allocated.

3.2 Treatments

We implement three treatments in Experiment 1: one for the lottery rule, Lottery5, and two for the queue rule, Queue5 and Queue5_rank. In the two queue treatments, we manipulate participants' awareness of the queue length and their ranking positions upon entering the queue. In Queue5, participants receive no feedback about their position in the queue, while in Queue5_rank, they are informed of their ranking position and the queue length. As discussed in the introduction, the primary reason for studying the observable queue treatment is the intuition that it may enable participants to make more efficient queuing decisions. However, recall that the theoretical model predicts that an observable queue should have no impact on productive efficiency under the queue rule. Table 1 summarizes the main features of our experimental design.

[12] If participants dedicate the entire four minutes of each round to the production task, their payoffs will range from 360 to 600 ECUs, comparable to the valuation of a slot.

Table 1: Design of Experiment 1

Treatments	Allocation rule	Ranking info	# of participants	# of sessions
Queue5	Queue	No	60	6
Queue5_rank	Queue	Yes	60	6
Lottery5	Lottery	N/A	60	6

3.3 Procedure

The experiment was conducted at the Nanjing Audit University Economics Experimental Lab with a total of 180 university students, using the software z-Tree (Fischbacher, 2007). Each session has 10 participants who are randomly re-matched in each round. After every round, all participants receive feedback about whether they were allocated a slot and their payoffs from the booking and production tasks. At the end of a session, one round is privately and randomly chosen for each participant and the participant receives her payoff from that round.

During the experiment, as participants arrived, they were randomly seated at a partitioned computer terminal. The experimental instructions were given to participants in printed form and were also read aloud by the experimenter. Participants then completed a comprehension quiz before proceeding. At the end of the experiment, they completed a questionnaire concerning their demographics and a number of psychological measures. For every 10 ECUs, participants earned 1 RMB. A typical session lasted about one hour with average earnings of 67.1 RMB, including a show-up fee of 15 RMB. All instructions for Experiment 1 are provided in the E-Companion.

3.4 Hypotheses

Here, we apply the theoretical analysis from Section 2 using our experimental parameters to derive a set of testable hypotheses regarding participants' strategies and efficiency outcomes across the different treatments.

It is clear that in the lottery treatment, participants do not need to spend time on the booking task except for submitting their application to enter the lottery. Therefore, we expect that participants will minimize the time spent on the booking task: they will likely visit the booking task only once and stay for just a few seconds to submit their application. All slots will be assigned randomly at the end of the round.

In contrast, in the queue treatments, participants need to compete for slots by queuing. In our experiments,

five participants are competing for three booking slots, so $n = 5$ and $m = 3$. The slot valuation v follows the uniform distribution on $[400, 600]$, while the productivity w (per minute) follows the uniform distribution on $[90, 150]$.[13] The expected efficiency losses and expected queuing time under the two rules are summarized in Table 2.

Table 2: Comparisons of efficiency and time spent in the booking system under the queue and lottery rules

	Queue	**Lottery**
Allocative efficiency loss	$\approx$ 35 ECU	100 ECU
Productive efficiency loss	$\approx$ 1267 ECU	0 ECU
Time spent on booking task	$\approx$ 139 seconds per subject ($\approx$ 57.8% of 4 minutes)	$\approx$ 0 second

While the queue rule is anticipated to incur a smaller allocative efficiency loss compared to the lottery rule, its productive efficiency loss is expected to be significantly greater. Overall, the lottery rule outperforms the queue rule in terms of overall efficiency under our experimental parameters.

Now we formally state our hypotheses as follows. Our first hypothesis is about the overall time spent on the booking system.

Hypothesis 1. *(Lottery vs. Queue on Time Spent on Booking) Participants spend more time on the booking system in Queue5 and Queue5_rank than in Lottery5. i.e.,* $\forall\, y_i,\ t_{lottery} < t_{queue}(y_i) = t_{rank}(y_i)$

Our next hypothesis is about the probability of obtaining a slot under the two rules.

Hypothesis 2. *In Lottery5, every participant have an equal chance of winning a slot. In Queue5 and Queue5_rank, participants with higher time valuations spend more time queuing and consequently have a greater chance of winning a slot.*

Moreover, the calculations regarding the productive and allocative efficiency losses under each rule immediately lead to the following hypothesis:

Hypothesis 3. *(Lottery vs. Queue on Efficiency) (a) Productive efficiency loss is higher in Queue5 and Queue5_rank than in Lottery5. (b) Allocative efficiency loss is higher in Lottery5 than in Queue5 and Queue5_rank.*

[13] We omit the detail that both parameters are discrete integers in our experiment.

It is worth noting that participants in our experiment are allowed to freely switch back and forth between the two task screens. While our theoretical framework is based on a sealed-bid all-pay auction that does not explicitly account for their switching behavior, it should be clear that in the Queue5 treatment, where the queue length is unobservable, a rational participant should enter the queue at some time $t \in [0, T]$ and remain there thereafter. Therefore, to characterize a participant's strategy in Queue5, it suffices to focus on their length of queuing time, as detailed in Proposition 1. For the Queue5_rank treatment, as mentioned in Section 2, the theoretical literature typically models this queue rule as a sealed-bid winner-pay auction in which losers do not incur costs for their bids. The intuition is that losers are latecomers to the queue who would drop out immediately after finding out they have no chance of winning, effectively not paying for their bids. We apply this winner-pay auction framework to provide a theoretical prediction for participants' queuing length in the Queue5_rank treatment. As discussed in Section 2, given their time valuations, participants' expected queuing time and overall efficiency are not affected by the availability of the ranking information regarding the queue.[14]

Hypothesis 4. *(Effect of Ranking Information) Queue5 and Queue5_rank do not differ in terms of (a) participants' expected queuing time and (b) each type of efficiency loss.*

4 Results of Experiment 1

We first present aggregated and individual-level results to test Hypotheses 1, 2, and 4(a), and then quantify and compare the different sources of efficiency loss across treatments to test Hypotheses 3 and 4(b).

4.1 Treatment Effect on Time Spent on Booking

We first examine how participants allocate their time between the booking and production tasks. We find strong support for Hypothesis 1. As shown in Figure 1, participants in the Queue5 treatment spend approximately 50% of their time queuing, which is slightly lower than the predicted level ($p = 0.063$, Wilcoxon signed-rank test).[15] Further, ranking information appears to have little impact on the average queuing time

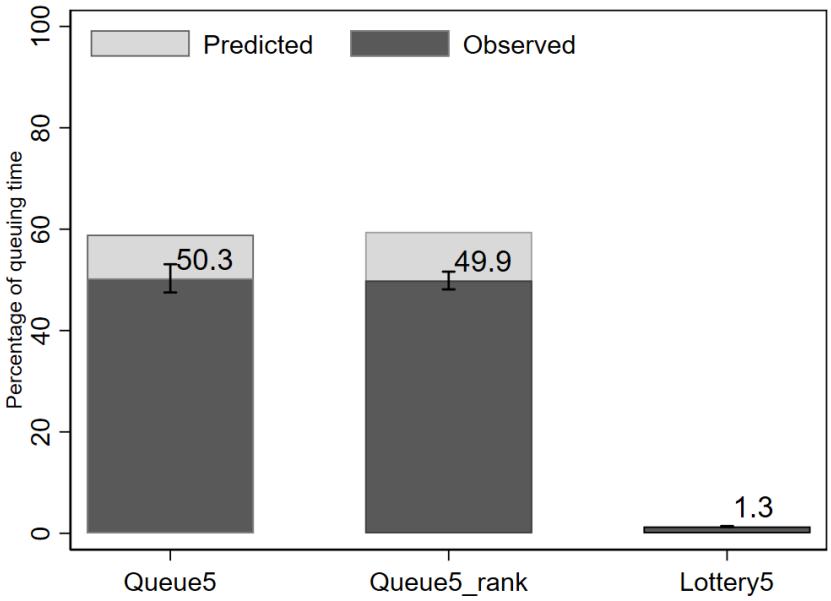
[14] Intuitively, when the queue length is observable, participants may adopt more complex strategies and exhibit increased switching behavior. For example, a participant who arrives and finds themselves alone in the queue might perceive it as beneficial to switch to the production task for a short period before returning to the queue. However, characterizing participants' equilibrium strategies in such a dynamic setting is challenging and beyond the scope of this paper. Therefore, we will mainly explore experimental data to investigate such behavior.

[15] Unless otherwise stated, we treat each session as a unit of observation in all reported statistics.

($p = 0.818$, Wilcoxon rank-sum test), supporting Hypothesis 4(a). In contrast, participants in the lottery treatment spend only a few seconds on the booking task. This evidence strongly suggests that, compared to the lottery rule, the queue rule leads to a substantial productive efficiency loss in terms of the opportunity cost of time.[16]

Result 1. *Participants spend significantly more time on the booking system in the queue treatments than in the lottery treatment.*

Figure 1: Percentage of time spent on the booking system



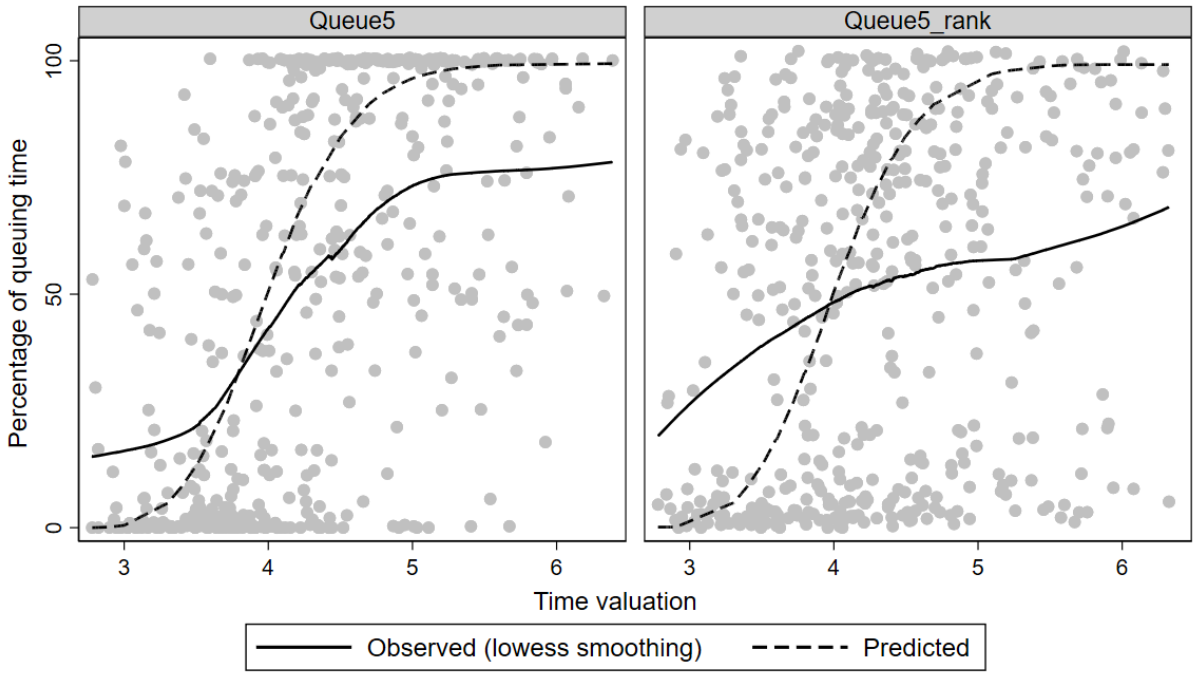
Notes: Error bars represent one standard error of means clustered at the session level.

Next, we find strong support for Hypothesis 2. Figure 2 illustrates a positive relationship between time valuation and total queuing time in both queue treatments. The solid line represents the Lowess smoothing curve, while the dashed line indicates the predicted proportion of timing spent in the queue. Each scatter data point represents the percentage of time spent queuing by participants with varying time valuations. In Queue5, the observed relationship shows a similar monotonic pattern to the theoretically predicted one. However, participants with low time valuations tend to overspend time queuing, whereas those with high time valuations tend to underspend. In Queue5_rank, the positive relationship persists but appears weaker.

To further quantify the relationship between time valuation and queuing time, we conduct random effects

[16]Figure B1 in the E-Companion shows the time allocation behavior over rounds, indicating that the overall pattern is generally stable over time.

Figure 2: Relationship between time valuation and percentage of queuing time



Notes: The observed relationship between time valuation and percentage of queuing time is produced using lowess smoothing.

regressions where the dependent variable is the percentage of time spent in the booking system for each treatment, and the independent variables include time valuation, slot valuation and time cost per minute. Table 3 reports the estimates from these regressions. We find statistical evidence supporting a positive relationship between time valuation and queuing time in both queue treatments. Notably, the estimated coefficient for time valuation is significantly higher in Queue5 compared to Queue5_rank ($p < 0.001$).[17] In contrast, there is no significant relationship in the lottery treatment. Additionally, both a higher slot valuation and a lower time cost are associated with increased queuing time, indicating that these two factors influencing time valuation significantly contribute to the observed results.

Furthermore, we conduct random effects probit regressions, changing the dependent variable to an indicator variable of winning a slot. Table 4 reports the average marginal effect estimates, indicating that increased queuing time effectively translates to a higher likelihood of obtaining a slot, with this effect being significantly larger in Queue5 ($p < 0.001$). Additionally, in the lottery treatment, the assignment of slots is uncorrelated with time valuations, suggesting that it is effectively random.

[17] The p-value is produced by estimating a specification with an interaction term between time valuation and a treatment indicator.

Result 2. *A higher time valuation is associated with more time spent queuing, which in turn increases the likelihood of obtaining a slot. Furthermore, this effect is more pronounced in the treatment with an unobservable queue.*

Table 3: Random effects regressions on time spent in booking

	Queue5		Queue5_rank		Lottery5	
Time valuation	28.903***		11.306***		-0.018	
	(2.073)		(1.095)		(0.059)	
Slot valuation		25.710***		13.272***		0.064
		(4.371)		(1.095)		(0.108)
Time cost per minute		-107.687***		-35.230***		0.299
		(6.712)		(7.113)		(0.369)
Constant	-73.616***	49.969**	1.284	25.198*	1.416***	0.658
	(7.688)	(23.668)	(4.351)	(13.484)	(0.236)	(0.812)
Clusters	6	6	6	6	6	6
N	480	480	480	480	480	480

Notes: Standard errors clustered at the session level are in parentheses. We rescale the slot valuation and time cost per minute by dividing them by 100. The time valuation is the ratio of the slot valuation and the time cost per minute.

Table 4: Random effects probit regressions on the likelihood of obtaining a slot

	Average marginal effects					
	Queue5		Queue5_rank		Lottery5	
Time valuation	0.279***		0.112***		-0.025	
	(0.021)		(0.017)		(0.036)	
Slot valuation		0.237***		0.143***		-0.035
		(0.031)		(0.023)		(0.042)
Time cost per minute		-1.011***		-0.314***		0.060
		(0.079)		(0.068)		(0.169)
Clusters	6	6	6	6	6	6
N	480	480	480	480	480	480

Notes: Standard errors clustered at the session level are in parentheses. We rescale the slot valuation and time cost per minute by dividing them by 100. The time valuation is the ratio of the slot valuation and the time cost per minute.

Though our results at the aggregate level are consistent with theoretical predictions, Figure 2 suggests a tendency for bimodal behavior in the queue treatments, which contradicts the equilibrium behavior where there is no mass at either extreme of never queuing or queuing all the time. To further validate this observation, we plot the cumulative probability function (CDF) of total queuing time in Figure 3. We find a substantial proportion of observations at the extremes (either fewer than 5 seconds or more than 235 seconds), with this proportion being significantly higher in Queue5 compared to Queue5_rank (46.0% vs. 18.3%, $p = 0.002$, Wilcoxon rank-sum test).[18] Moreover, Table 5 reports estimates from random effects probit regressions,

[18] A similar bimodal behavior is also observed in an experimental all-pay auction with incomplete information about individual

where the dependent variable is whether the participant spend almost no time queuing (columns 1 and 3) or whether they spend almost all the time queuing (columns 2 and 4). As expected, participants with higher time valuations are significantly more likely to spend nearly all their time queuing and less likely to drop out altogether.

Figure 3: CDF of percentage of total queuing time

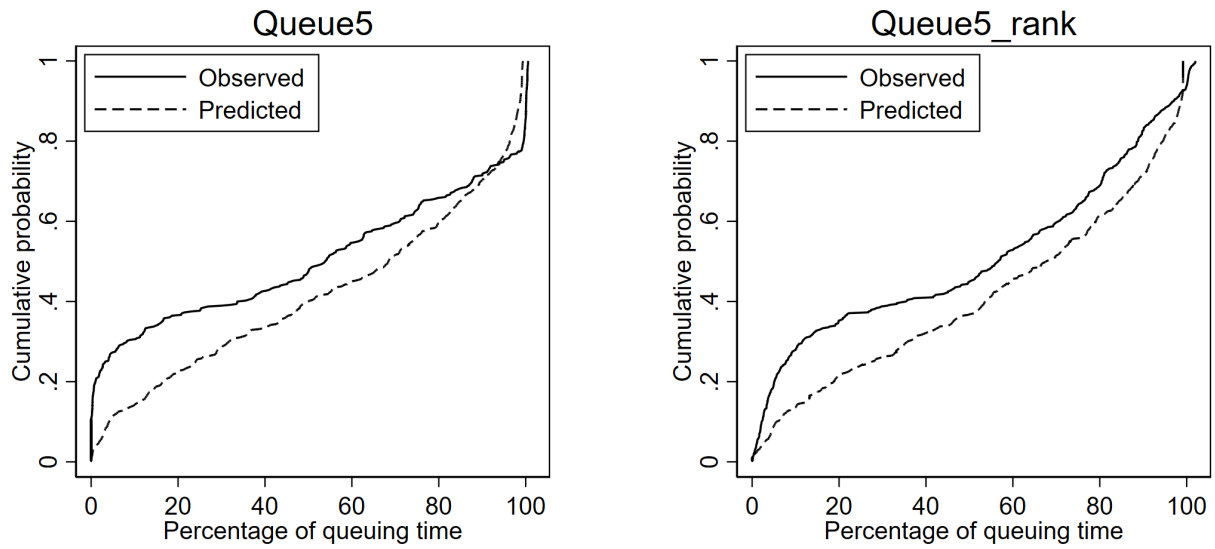


Table 5: Random effects probit regressions on the likelihood of never queuing or queuing all the time

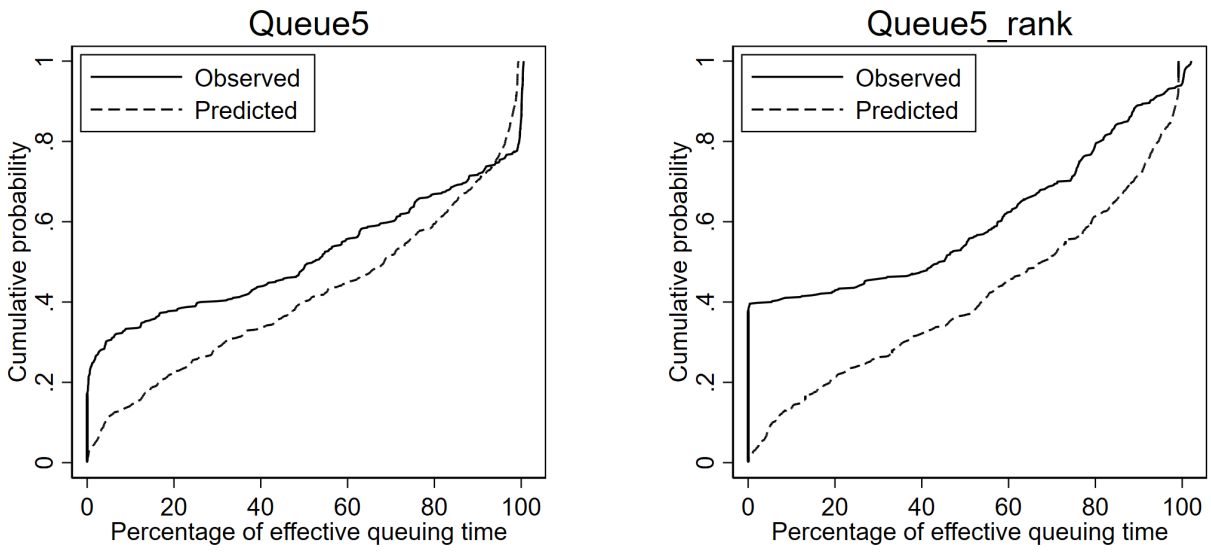
	Average marginal effects			
	Queue5		Queue5_rank	
	Always queue	Never queue	Always queue	Never queue
Time valuation	0.204***	-0.252***	0.042**	-0.062**
	(0.033)	(0.038)	(0.019)	(0.026)
Clusters	6	6	6	6
N	480	480	480	480

Notes: Standard errors clustered at the session level are in parentheses. We rescale the slot valuation and time cost per minute by dividing them by 100. The time valuation is the ratio of the slot valuation and the time cost per minute.

This finding is initially counterintuitive because, in the presence of an observable queue, we would expect more successful coordination, meaning that precisely two participants in each group should spend very little time queuing. They would switch to the booking task only to find that the queue length exceeds the number of available slots, leading them to drop out of the queue immediately. However, in practice, participants' actions are far from ideal, as switching back and forth incurs a non-negligible amount of unproductive time. Conversely, in the case of an unobservable queue, participants base their decisions primarily on their

marginal costs of bidding (Müller and Schotter, 2010).

Figure 4: CDF of percentage of effective queuing time



own time valuations, which likely leads to more bimodal behavior as a decision heuristic. To validate this conjecture, we examine the frequency of switching in the two queue treatments. On average, participants switch only once per round in Queue5, while they switch 21.4 times in Queue5_rank ($p = 0.002$, Wilcoxon rank-sum test). It remains unclear a priori whether participants with high or low time valuations are more likely to switch. On the one hand, those with high time valuations may switch early to find a short queue, then return to the production task for a while before rejoining the queue. On the other hand, participants with low time valuations might arrive at the queue too late; however, they do not simply drop out but instead switch back and forth in search of a better opportunity. Table B1 in the E-Companion reports estimates from a random effects regression where the dependent variable is the participant's switching frequency per round in Queue5_rank. We find no significant association between time valuation and switching frequency.

Due to the non-negligible switching behavior, we compare the total queuing time with the "effective" queuing time, defined as the amount of time participants spend queuing uninterruptedly until the end of a round. We find that while the "ineffective" queuing time (calculated by subtracting the "effective" queuing time from the total queuing time) is merely 2.9 seconds in Queue5, it increases significantly to 22.3 seconds in Queue5_rank ($p = 0.002$, Wilcoxon rank-sum test).[19] This suggests that 18% of productive efficiency loss in Queue5_rank can be attributed to switching behavior. In Figure 4, we plot the CDF of effective queuing

[19]Figure B2 in the E-Companion shows the percentage of effective and ineffective queuing time over rounds, indicating a relatively stable pattern over time.

time, revealing that 40% of participants in Queue5_rank spend no time queuing at all. This means that, on average, two out of five participants in each group drop out of the queue once we remove the ineffective queuing time. Additionally, we verify that the significantly positive association between queuing time and time valuation reported in Table 3 remains robust when replacing the dependent variable of total queuing time with effective queuing time (see Table B2 in the E-Companion). Overall, these findings indicate that although observable queues do not negatively impact the overall efficiency of the queue rule at the aggregate level, participants engage in switching behavior much more frequently and waste more unproductive time in the process. This also implies that observable queues lead to significantly less effective queuing time, which is the only information relevant to the final allocation of slots. Therefore, in the sense of reducing effective queuing time and that exactly three participants per group stayed in the queue till the end, making queues observable does make coordination more efficient. However, the presence of significant ineffective queuing time due to frequent switching negates this efficiency gain.

Finally, we have collected data on the queue length and the participants' positions in the queue at the moment when they switched from the booking task to the production task. This information allows us to further explore the reasons for participants' switching behavior. Overall, in the Queue5_rank treatment, participants' switching behavior is largely rational. 40.0% of switches occur when the queue is too short (i.e., when the queue length is shorter than three), while 50.4% of switches take place when the queue is too long and the participant is not in a good position to secure the slot (i.e., when both the queue length and the participant's position are greater than three). Conversely, only 9.6% of switches can be categorized as presumably irrational or at least riskier than the first two types; these switches occur when the participant is favorably positioned in a long queue (i.e., when the queue length is equal to or greater than three and their queuing position is equal to or fewer than three). Furthermore, we examine the cumulative ineffective queuing time associated with each type of switching. We find that the last type of switching, which is presumably irrational, has a disproportionately significant impact on overall efficiency: 41.1% of ineffective queuing time can be attributed to this type. In contrast, switching due to a short queue accounts for 33.4% of ineffective queuing time, whereas switching related to a long queue constitutes the remaining 25.4%. This pattern appears to remain consistent in the later rounds of the experiment. Thus, a small fraction of plausibly irrational switching behavior (around 10%) results in a disproportionate impact on ineffective queuing time (around 40%).

Result 3. *In both queue treatments, participants with higher time valuations are significantly more likely to spend nearly all their time queuing and less likely to drop out altogether. Observable queues generate less bimodal behavior and induce significantly more switching behavior. However, since observable queues simultaneously decrease effective queuing time and increase ineffective queuing time attributable to task-switching, they do not impact the total time spent in the booking system.*

4.2 Quantifying different sources of efficiency loss

The previous subsection shows that the lottery rule is superior to the queue rule in terms of productive efficiency. In this subsection, we compare the different types of efficiency losses at the group level across the two allocation rules. Table 6 reports the quantified efficiency loss (in ECUs) for each treatment, including the total efficiency loss, which is the sum of productive efficiency loss and allocative efficiency loss. Overall, each type of efficiency loss is reasonably close to the predicted level in each treatment. More importantly, we find strong support for Hypotheses 3 and 4(b). While allocative efficiency loss is significantly higher under the lottery rule than the queue rule (Queue5 vs. Lottery5: $p = 0.078$; Queue5_rank vs. Lottery5: $p = 0.010$, Wilcoxon rank-sum test), productive efficiency loss is significantly higher under the queue rule than the lottery rule ($p = 0.004$ in both comparisons) and exceeds allocative efficiency loss by orders of magnitude. As suggested in the previous subsection, Queue5_rank results in more ineffective queuing time and less effective queuing time than Queue5, with the two effects canceling each other out. Consequently, making queuing information transparent to participants does not help mitigate overall productive inefficiency.

Table 6: The Breakdown of Efficiency Loss

	Queue5	Queue5_rank		Lottery5	
	Observed		Predicted	Observed	Predicted
Allocative efficiency loss	74.990	74.490	35	105.375	100
	(13.036)	(7.061)		(6.835)	
Productive efficiency loss	1142.742	1169.229	1267	32.545	0
	(42.173)	(25.980)		(2.041)	
Total efficiency loss	1217.731	1243.719	1302	137.921	100
	(40.676)	(27.027)		(6.747)	
Obs. (group $\times$ round)	96	96		96	

Notes: Standard errors are in parentheses.

Result 4. *Compared to the lottery rule, the queue rule results in significant losses in productive efficiency, which outweigh its advantages in allocative efficiency, leading to a considerably greater overall efficiency*

loss. Ranking information about queues does not impact either type of efficiency.

5 Experiment 2: Robustness Check in More Complex Environments

In this section, we briefly report results from a robustness experiment which shares a similar research purpose as Experiment 1 but uses a more complex experimental design that can perhaps represent some real-world situations more closely. The overall takeaway message, however, is similar in both experiments: the productive efficiency loss under the queue rule is overwhelmingly large compared to any potential gain in the allocative efficiency. A detailed report on Experiment 2 is provided in the E-Companion.

Experiment 2 has two major differences from Experiment 1. First, the abstract-effort production task is replaced by a real-effort production task. A participant's payoff in this task is determined by the number of correctly solved problems. On average a participant who spends a longer time on the production task will receive a higher payoff, but now her opportunity cost of time becomes endogenous and can vary across time. Second, Experiment 2 has two stages. The first stage is similar to that of Experiment 1. The second stage models real-life situations in which some participants who fail to book slots in stage 1 may still visit the booking system to search for any remaining or canceled slots. Therefore, we let any unassigned slots in stage 1 be available at the beginning of stage 2. Furthermore, exactly one of the slots allocated in stage 1 is canceled in Stage 2, but the cancellation timing is randomly determined. Participants were informed of the cancellation rule in the experimental instructions. In stage 2, only those who have not obtained a slot can request one either on a first-come-first-served basis or through entering another lottery. In addition to the booking task, participants may also work on the real-effort production task.

Using a between-subjects design, we first compare two *solo-track* booking systems that use either the queue rule or the lottery rule exclusively in both stages. We also vary the degree of market competitiveness to test for the robustness of our results. In such settings with real-effort production task, we distinguish between three sources of efficiency loss: inefficient allocation of booking slots (allocative efficiency loss), the opportunity cost of time spent on the booking task (productive efficiency loss), and changes in on-the-job productivity due to distraction of the booking task (behavioral efficiency loss). Consistent with Experiment 1, results from Experiment 2 also show that queue participants spend substantial amounts of time on the booking task in both stages while lottery participants spend only a few seconds submitting their applications

and the remainder of their time on the production task. The productive efficiency loss under the queue rule outweighs the other two sources by a large margin, leading to a much lower overall efficiency under the queue rule than the lottery rule. We further observe that allocative efficiency is actually not higher under the queue rule, either. The reason is that most participants exhibit bimodal behavior under the queue rule just like in Experiment 1: they spend either a few seconds or almost all of their time on the booking task. However, unlike our finding in Experiment 1, this bimodal behavior is largely uncorrelated to their time valuations, perhaps because it is much harder for participants to evaluate their opportunity cost of time due to the endogenous nature of their productivity in the real-effort task.

In addition to the solo-track systems, we also design a novel *dual-track* booking system in which slots are provided in two tracks, each implementing one of the two allocation rules, and each participant can freely choose which track to enter at the beginning (but she cannot choose both).[20] Our designed dual-track system can serve two purposes. First, when the queue rule and the lottery rule have their distinct advantages—the former may achieve higher allocative efficiency while the latter may achieve higher productive efficiency—the designer may use the hybrid system to achieve a balance between their respective advantages. Second, the designer may consider a transition from a queue rule to a more efficient lottery rule but worry that an abrupt transition is not practical.[21] The hybrid system can be used to help participants build familiarity with both rules to facilitate the final transition. In our experiment, we are primarily interested in observing whether participants are more likely to choose the lottery rule over the queue rule when both rules are available and offer the same ex-ante chance of obtaining a slot. Specifically, in our dual-track system, slots are split evenly between the queue track and the lottery track in stage 1. Stage 2 does not implement the dual-track; it implements either the queue rule or the lottery rule depending on the treatment. We find that participants are more likely to choose the lottery track over the queue track in stage 1 . We further find that participant behavior under each track is similar to the corresponding solo-track system. Consequently, those who choose the lottery track earn a higher payoff than those who choose the queue track, offsetting their lower probability of obtaining a slot. The efficiency loss due to opportunity costs of time in the queue track remains substantial. But the total efficiency loss is lower than that in the solo-track queue system due

[20] Similar dual-track or hybrid systems where individuals can choose between lottery and auction systems are observed in the real world. One example is the assignment of vehicle licenses in major cities of China (Li, 2018; Huang and Wen, 2019). Beijing uses lotteries exclusively; Shanghai uses only auctions; Guangzhou, Shenzhen, Tianjin, and Hangzhou use a dual-track system.

[21] For example, participants may have concerns about the transparency of lottery draws, especially those for high-stakes goods or services; they may also be concerned about the inability of a lottery to distinguish participants with greater needs.

to a lower number of participants choosing the queue track in the dual-track setting. Finally, the dual-track system reduces allocative efficiency loss by channeling some participants with high valuations to compete for slots in the queue track.

6 Concluding Remarks

When scarce resources are provided for free or under price control, how to ration resources becomes a design problem. To the best of our knowledge, our paper is the first to systemically evaluate efficiency across various allocation systems in a multi-tasking environment, with a particular focus on the externality of an allocation system on parallel production tasks. One commonly used rule, the queuing system, is criticized for efficiency losses due to the opportunity cost of time spent on the queuing process.

Specifically, we develop a flexible dual-tasking experimental framework to compare the performance of a queue rule based on a first-come, first-served principle with that of a lottery rule that relies on a random selection process to allocate slots on a booking system when participants can also participate in a parallel production task. Our experimental results show that the lottery rule yields superior efficiency. Under the queue rule, the opportunity cost of queuing time is substantial enough to overwhelm other efficiency sources, leading to lower participant welfare. Further, our findings indicate that while providing ranking information to queuing participants reduces their effective queuing time, it simultaneously results in much more frequent switching between the two tasks, leading to significant ineffective queuing time. Consequently, the loss in productive efficiency attributed to these switches undermines the efficiency gains achieved through improved coordination. Therefore, although enhancing the observability of queues may facilitate more efficient coordination, overall efficiency would be significantly higher if the time wasted due to task switching could be minimized.

While our experimental results strongly support the superiority of the lottery rule to the queue rule, we acknowledge that the precise magnitude of each source of efficiency loss depends on the valuation of the appointment slot and the opportunity cost of time chosen in our experiment. Therefore, to what extent a real-life queuing system is inefficient requires careful calibration of these two theoretical parameters. In situations where the valuation of a slot is much higher than the opportunity cost of time, such as applying for an immigrant visa, the concern for allocative efficiency dominates and could justify the use of the queue rule

rather than the lottery. Nevertheless, for such situations, our finding of bimodal behavior suggests that the presumed greater allocative efficiency of the queue rule cannot be taken for granted because, in high-stakes booking systems, people are more likely to spend all their time queueing, resulting in what is effectively a random allocation. In some applications, the lottery system may also entail some hidden costs due to its low participation cost, which attracts more applicants than otherwise desirable and then reduces every applicant's expected payoff in the lottery system. To what extent this issue reduces the attractiveness of the lottery system is another direction for future research.

Our experimental framework is versatile enough to be applied to numerous settings. It can form the basis for more complex booking scenarios, such as when participants have preferences over different slots. For example, some patients visiting hospitals may prefer morning slots over afternoon ones. One potential solution to this issue is to borrow process steps from school choice matching algorithms. Participants begin by submitting rank-order lists of slots to reveal their preferences. The process then uses a matching algorithm (which could involve lotteries to break ties) to find an allocation of slots. This system can avoid the competition via queuing as the lottery rule.

Finally, it is worth mentioning that although this paper primarily discusses offline booking systems, where individuals physically wait in line, the externalities of queuing may also manifest in online booking systems, wherein individuals wait in front of electronic devices or in telephone queues. Although such queues ostensibly allow individuals to engage in other activities while waiting, in highly competitive scenarios, they may become distracted or may focus solely on the booking system, expending time or energy being "glued to their device" until slots are allocated. Such experiences are prevalent in contemporary society. The extent to which these online booking systems generate externalities and productive inefficiencies is likely contingent upon specific contexts and is an important avenue for future research.

Acknowledgment

All authors contributed equally to this work and are corresponding authors. We thank Bo Chen, Philipp Heller, Dorothea Kübler, Sen Geng, and participants from several seminars and conferences for helpful comments. Lingbo Huang gratefully acknowledges financial support by NSFC (Grant 72192842, 72203099, 72422018). Tracy Xiao Liu gratefully acknowledges financial support by NSFC (2222005, 72342032). Jun

Zhang gratefully acknowledges financial support by NSFC (72122009,72394391,72033004) and the Wu Jiapei Foundation of the China Information Economics Society (E21103567).

References

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *The American Economic Review*, 93(3): 729–747.

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan. 2020. "Thickness and Information in Dynamic Matching Markets." *Journal of Political Economy*, 128(3): 783–815.

Allon, Gad, and Eran Hanany. 2012. "Cutting in Line: Social Norms in Queues." *Management Science*, 58(3): 493–506.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv. 2020. "Optimal Dynamic Matching." *Theoretical Economics*, 15(3): 1221–1278.

Barzel, Yoram. 1974. "A Theory of Rationing by Waiting." *The Journal of Law and Economics*, 17(1): 73–95.

Beer, Ruth, Anyan Qi, and Ignacio Ríos. 2024. "Behavioral Externalities of Process Automation." *Management Science*, forthcoming.

Bloch, Francis, and David Cantala. 2017. "Dynamic Assignment of Objects to Queuing Agents." *American Economic Journal: Microeconomics*, 9(1): 88–122.

Buell, Ryan W. 2021. "Last-Place Aversion in Queues." *Management Science*, 67(3): 1430–1452.

Chen, Yan, and Tayfun Sönmez. 2002. "Improving Efficiency of On-Campus Housing: An Experimental Study." *The American Economic Review*, 92(5): 1669–1686.

Che, Yeon-Koo, and Olivier Tercieux. 2024. "Optimal Queue Design." *Journal of Political Economy*, forthcoming.

Corgnet, Brice, and Roberto Hernán-González. 2019. "Revisiting the Trade-Off Between Risk and Incentives: The Shocking Effect of Random Shocks?" *Management Science*, 65(3): 1096–1114.

Corgnet, Brice, Joaquin Gómez-Minambres, and Roberto Hernán-González. 2015. "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Science*, 61(12): 2926–2944.

Corgnet, Brice, Roberto Hernán-González, and Eric Schniter. 2015. "Why Real Leisure Really Matters: Incentive Effects on Real Effort in the Laboratory." *Experimental Economics*, 18(2): 284–301.

Dimakopoulos, Philipp D, and C-Philipp Heller. 2019. "Matching with Waiting Times: The German Entry-Level Labor Market for Lawyers." *Games and Economic Behavior*, 115: 289–313.

Dutcher, E. Glenn, Timothy C. Salmon, and Krista J. Saral. 2024. "Is "Real" Effort More Real?" *Experimental Economics*, 1–32.

Estrada Rodriguez, Arturo, Rouba Ibrahim, and Dongyuan Zhan. 2024. "On Customer (Dis-)Honesty in Unobservable Queues: The Role of Lying Aversion." *Management Science*, forthcoming.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics*, 10(2): 171–178.

Hakimov, Rustamdjan, and Dorothea Kübler. 2021. "Experiments on Centralized School Choice and College Admissions: A Survey." *Experimental Economics*, 24(2): 434–488.

Hakimov, Rustamdjan, C.-Philipp Heller, Dorothea Kübler, and Morimitsu Kurino. 2021. "How to Avoid Black Markets for Appointments with Online Booking Systems." *The American Economic Review*, 111(7): 2127–51.

Heller, C, Johannes Johnen, Sebastian Schmitz, et al. 2019. "Congestion Pricing: A Mechanism Design Approach." *Journal of Transport Economics and Policy (JTEP)*, 53(1): 74–98.

Holt, Charles A, and Roger Sherman. 1982. "Waiting-Line Auctions." *Journal of Political Economy*, 90(2): 280–294.

Huang, Yangguang, and Quan Wen. 2019. "Auction–Lottery Hybrid Mechanisms: Structural Model and Empirical Analysis." *International Economic Review*, 60(1): 355–385.

Leshno, Jacob D. 2022. "Dynamic Matching in Overloaded Waiting Lists." *The American Economic Review*, 112(12): 3876–3910.

Li, Shanjun. 2018. "Better Lucky than Rich? Welfare Analysis of Automobile Licence Allocations in Beijing and Shanghai." *The Review of Economic Studies*, 85(4): 2389–2428.

Li, Simin, Martin A. Lariviere, and Achal Bassamboo. 2024. "Is Full Price the Full Story When Consumers Have Time and Budget Constraints?" *Manufacturing & Service Operations Management*, 26(1): 370–388.

Müller, Wieland, and Andrew Schotter. 2010. "Workaholics and Dropouts in Organizations." *Journal of the European Economic Association*, 8(4): 717–743.

Myerson, Roger B. 1981. "Optimal Auction Design." *Mathematics of Operations Research*, 6(1): 58–73.

Naor, Pinhas. 1969. "The Regulation of Queue Size by Levying Tolls." *Econometrica*, 15–24.

Nichols, Donald, Eugene Smolensky, and T Nicolaus Tideman. 1971. "Discrimination by Waiting Time in Merit Goods." *The American Economic Review*, 61(3): 312–323.

Parry, Ian W H, Margaret Walls, and Winston Harrington. 2007. "Automobile Externalities and Policies." *Journal of Economic Literature*, 45(2): 373–399.

Pathak, Parag A, Alex Rees-Jones, and Tayfun Sönmez. 2022. "Immigration Lottery Design: Engineered and Coincidental Consequences of H-1B Reforms." *Review of Economics and Statistics*, 1–43.

Platz, Trine Tornøe, and Lars Peter Østerdal. 2017. "The Curse of the First-In–First-Out Queue Discipline." *Games and Economic Behavior*, 104: 165–176.

Riley, John G, and William F Samuelson. 1981. "Optimal Auctions." *The American Economic Review*, 71(3): 381–392.

Roth, Alvin E. 2021. "Experiments in Market Design." In *Handbook of Experimental Economics*. Vol. 2, , ed. John H Kagel and Alvin E Roth, 290–346. Princeton, NJ:Princeton University Press.

Schummer, James. 2021. "Influencing Waiting Lists." *Journal of Economic Theory*, 195: 105263.

Schummer, James, and Azar Abizada. 2017. "Incentives in Landing Slot Problems." *Journal of Economic Theory*, 170: 29–55.

Shunko, Masha, Julie Niederhoff, and Yaroslav Rosokha. 2018. "Humans Are Not Machines: The Behavioral Impact of Queueing Design on Service Time." *Management Science*, 64(1): 453–473.

Suen, Wing. 1989. "Rationing and Rent Dissipation in the Presence of Heterogeneous Individuals." *Journal of Political Economy*, 97(6): 1384–1394.

Taylor, Grant A, Kevin KK Tsui, and Lijing Zhu. 2003. "Lottery or Waiting-Line Auction?" *Journal of Public Economics*, 87(5-6): 1313–1334.

Tobin, James. 1952. "A Survey of the Theory of Rationing." *Econometrica*, 521–553.

Wang, Jingqi, and Yong-Pin Zhou. 2018. "Impact of Queue Configuration on Service Time: Evidence from a Supermarket." *Management Science*, 64(7): 3055–3075.