Redistributing Priority Payments in Service Systems

Andrew E. Frazelle, Elena Katok
Jindal School of Management, The University of Texas at Dallas
andrew.frazelle@utdallas.edu, ekatok@utdallas.edu

Problem Definition: We study a service system in which each customer has a random waiting cost and must choose between two queues: regular (no cost) or priority (for a fee). Social preferences may make customers more willing to purchase priority if the payments are redistributed to compensate inconvenienced customers. **Methodology/Results:** We study the impact of the recipient of the priority payment—the service provider or the inconvenienced customers in the regular queue—on customer priority-purchasing decisions. To determine the waiting-cost-dependent optimal strategies, we first establish structural results at a sample-path level and then prove that they generalize. With no redistribution, we show that the equilibrium obeys a cost-dependent, increasing-threshold structure. We then prove that the same structure prevails with redistribution. We also prove that redistribution results in fewer priority purchases because compensating regular-queue customers makes priority relatively less attractive. We then study both settings in the lab. Subjects were not biased to purchase priority more often in the compensation treatment, failing to display a social preference. Instead, the aggregate number of priority purchases aligned closely with theory in both treatments, i.e., more priority purchases occurred in the no-compensation treatment. When subjects made mistakes, across both treatments they were more likely to incorrectly choose the regular queue than they were to incorrectly choose the priority queue. **Managerial Implications:** Our theoretical results can help service providers predict priority purchasing patterns for better capacity planning. That our experiments align closely with the theory reinforces the value of these predictions. Moreover, for systems (like city congestion tolls) whose implicit goal is to achieve fewer purchases, our results imply that a system planner can disincentivize purchases merely by redistributing proceeds. Finally, subject tendency to under-purchase priority relative to the theory suggests that customers may have an intrinsic preference for the "free" option (the regular queue).

Key words: priority queues, behavioral operations, strategic queueing, social norms, behavioral queueing
This version: August 18, 2021.

1. Introduction

Paid priority programs operate in various industries including theme parks (e.g., Six Flags Flash Pass, see Levine 2019); air travel (e.g., expedited screening, see Transportation Security Administration 2021 and Clear 2021); and food delivery (e.g., Papa Priority, see Papa John's 2021), as well as in the public sector (e.g., expedited U.S. passport renewal, see U.S. Department of State 2021b, and express lanes on highways, see Texas Department of Transportation 2021). The service provider—the restaurant, theme park, screening contractor, or government agency—often keeps the priority fee. However, customers may bristle at being overtaken for no compensation, especially as the service provider incurs little to no cost to operate a priority program: all that is required is to process customers in a different order from first-come, first-serve (FCFS). Moreover, by purchasing priority, a customer risks being seen as spoiled or entitled. Particularly in the airline industry, there

is growing dissent between regular customers and those with elite status, who receive various forms of priority (Lipsey 2016).

The extant literature focuses mainly on programs with priority payments retained by the service provider. However, those who actually experience the externality are the customers without priority who are overtaken when another customer purchases priority. As pointed out in Yang et al. (2016), customer priority purchasing decisions, as well as perceived fairness, are likely to alter dramatically if the priority payments are used to compensate the non-priority customers, rather than line the service provider's pockets. Oberholzer-Gee (2006) studies a related setting behaviorally by having experimenters in the field make different monetary offers to move in front of someone in line. Subjects tended to accept such offers, but they frequently declined the monetary payment, suggesting a behavioral sensitivity to others' experiences that is relevant to our work.

The focus of our study is the effect of the recipient of the priority payment on customers' priority purchasing decisions in a service system. This effect is relevant for service providers who operate such programs, whether to optimize revenue or some other objective function like social welfare. For example, the U.S. Department of State offers regular and expedited passport renewal processing. As of July 2021, regular processing is estimated to take 18 weeks, while expedited processing is estimated to take 12 weeks, and COVID-19-related issues have created additional uncertainty in processing times (U.S. Department of State 2021a). By paying for expedited processing, a citizen moves her application ahead of regular applicants and mitigates the risk of a longer delay. As a government entity interested in improving the lives of its citizens, the State Department might consider funneling payments for expedited service to the non-expedited applicants whose renewals will take longer because they will be displaced by expedited applicants. Redistribution would improve fairness, as those who experience longer waiting times would be compensated by those who overtake them. Also, expedited applicants likely have higher waiting costs, so the reordering of applicants and subsequent redistribution of proceeds could be an efficient exchange. Similar fairness reasons could motivate redistributing tolls from highway express lanes (e.g., Texas Department of Transportation 2021) or congestion charges in crowded cities (e.g., Transport for London 2021); hurried commuters facing a traffic jam can pay to bypass the regular lanes and reduce their travel time (or to drive into the city, avoiding slower public transportation). Commuters who do not pay, who could have shorter travel times if the express lane was converted to a regular lane (or if they could drive into the city with no congestion charge instead of using other transport modes), may justifiably perceive this system as unfairly favoring the wealthy. However, they might feel—and behave—differently if they were compensated with a share of the tolls.

Service providers may wish to predict how customer decisions will change if non-priority customers are compensated. For example, city congestion charges aim to alleviate congestion and reduce emissions, both of which aims are served by fewer commuters choosing to pay the charge. Thus, city planners should be influenced in their decision to redistribute tolls by whether more commuters pay the charge in the equilibrium with vs. without compensation of commuters who opt for other transportation modes. On the other hand, a firm like Papa John's, which operates Papa Priority (customers can pay extra for their pizza to be made sooner: see Papa John's 2021), may hesitate to share proceeds because the service exists to generate revenue. Still, to engender goodwill among customers, it might choose to share a (perhaps small) fraction of the payments. Also, behaviorally, more customers might purchase priority if they knew that their payments would compensate inconvenienced non-priority customers. Service providers must thus consider both the strategic and behavioral effects of redistribution.

To investigate these issues, we propose a model of a service system with a single server and two customer classes—regular and priority—and we analyze equilibrium priority purchasing decisions both with and without compensation of regular customers (throughout the paper, compensation is only redistributive and comes from no source other than priority payments). We also conduct experiments on a special case of our model to test the behavioral response of subjects in this setting. In our model, a customer observes the regular and priority queues and chooses between joining at the end of the regular queue at no cost and joining at the end of the priority queue for a fee. While more complicated auction-based systems could allow the service provider to exert more influence over the outcome by effectively permitting an arbitrary number of priority classes, in practice there are usually far fewer priority classes than there are customers. Perhaps the most common setup is to have only two classes, as in our model (e.g., Papa Priority is a two-class system, as are U.S. passport renewal and many highway express lanes). We model disutility from waiting as a linear function of waiting time, with cost coefficients independent and identically distributed (IID) for each customer. A customer knows only her own waiting cost when making her priority purchase decision. In our finite and sequential setup, the random waiting costs significantly complicate the equilibrium analysis. To determine the equilibrium strategy for a focal customer requires computing the strategies for every combination of waiting-cost realizations for the customers after her, finding the waiting time for the focal customer for both queue choices in each such combination, and finally taking expectation.

We first study the *base* model in which the service provider keeps all priority proceeds. To navigate the difficulty created by the many possible combinations of equilibrium strategies for different waiting-cost realizations for those after a focal customer, we analyze individual sample paths. We

prove that for a given waiting cost and *any* threshold strategies for the customers after a focal customer, the customer optimally also uses a threshold strategy. Taking expectation over the later customers' strategies and performing an induction over the queue lengths and the customers establishes that all customers use *cost-dependent threshold strategies* in the perfect Bayesian equilibrium (PBE), i.e., they purchase priority if the priority queue is below a threshold, with possibly different thresholds for different waiting-cost realizations. The sample path approach allows us to simplify an extremely difficult problem—that of computing the equilibrium for arbitrary waiting-cost distributions, in which each customer must account for a potentially huge cross product of the other customers' optimal strategies, which depend on their realizations—into a still challenging but tractable one with unknown but fixed threshold strategies for the other customers.

Under cost-dependent threshold strategies, a customer purchases priority if and only if the priority queue is *short* enough. With a short priority queue (and thus a longer regular queue), a given customer will overtake more customers by purchasing priority. For a long priority queue, however, there are few regular-queue customers to overtake; above a threshold priority queue length, which depends on her realized waiting cost, a customer should not purchase priority because the reduction in expected waiting time that she achieves will not offset the priority fee. Even for the same waiting cost, the equilibrium thresholds differ among customers. We show that the thresholds are increasing, i.e., the first customer will have a lower threshold than the second, etc. Later customers have more regular-queue customers to overtake for a given priority queue length, and this makes priority more attractive than it would be to an earlier customer. To prove this, we employ a similar sample-path approach to compare the regular-queue waiting times of successive customers.

On top of these structural results, we obtain additional insights about the equilibrium, which are simplest to illustrate for the special case of deterministic waiting costs. Intuitively, we might expect that earlier customers should join the regular queue because there are few customers to overtake. However, if the priority fee is low, then the equilibrium path can begin with a string of consecutive priority purchases, followed by alternation. In other words, the *first* several customers purchase priority, even though they do not overtake anyone by doing so. The reason for these priority purchases is preemptive: knowing that later customers to arrive may purchase priority and overtake them, the early customers purchase priority to protect their privileged positions. This phenomenon is somewhat reminiscent of findings in Arlotto et al. (2019) for two-station service networks in which customers must visit both stations but can choose the order. In their model, counterintuitively, a customer who observes one busy station and one idle station optimally chooses to start with the busy station; this choice delays the time until her first service begins, but it can reduce her overall waiting time by protecting her position at the busy station. Choosing the busy

station is analogous to purchasing priority in our model with an empty or short regular queue: the immediate consequences appear negative, but there is a net benefit because one is protected from being overtaken later. However, unlike Arlotto et al. (2019), in our model we can also observe the opposite phenomenon. If priority is expensive, then customers will not pay merely to protect their position in line. Instead, the regular queue must build up to the point that priority greatly reduces the wait: in this case, customers choose the regular queue until a tipping point is reached at which priority becomes worth the fee.

We then study the *compensation* model, in which a fraction of the priority proceeds is shared with the regular-queue customers. The dynamics are even more complex because both compensation and wait time depend on others' decisions, and it is not intuitively clear whether threshold strategies should be optimal. Nevertheless, we prove that the equilibrium with compensation also has a cost-dependent, increasing-threshold structure. Importantly, priority is relatively less valuable with compensation because a customer can expect a payment if she chooses the regular queue. Accordingly, we prove that the equilibrium thresholds in the compensation model are *lower* than those in the base model because redistribution makes priority relatively less attractive. This implies that fewer customers purchase priority in the equilibrium in the compensation model than the base model. Additionally, for a given waiting cost and with fixed strategies of those after her, a focal customer's equilibrium threshold is *decreasing* in the compensation fraction. That is, the more that the service provider shares with regular customers, the less inclined each customer will be to purchase priority, at least strategically. These findings imply a two-pronged revenue reduction for service providers. First, by sharing revenue at all, the provider receives less for each priority purchase. Second, by redistributing proceeds, the provider dampens the value of priority, which reduces the equilibrium number of priority purchases. Service providers should account for both effects when deciding whether and how to operate a priority program.

All the same, the findings in Oberholzer-Gee (2006) demonstrate that humans are sympathetic to the urgency of others, which suggests that they may be hesitant to purchase priority and cause inconvenience in the base model. With redistribution, by contrast, knowing that the priority payment is used to compensate those overtaken could reduce this hesitation and mitigate the reduced value of priority with redistribution. Hence, the behavioral and strategic factors are at odds: the regular queue is more attractive strategically when regular-queue customers are compensated, but the priority queue may also become more attractive under compensation for behavioral reasons. The extent of this phenomenon depends on the strength of individuals' social preferences.

To test our theoretical predictions and to gauge the importance of fairness considerations in a priority service system, we conducted laboratory experiments that required subjects to choose

between a priority queue and a regular queue. To ensure that social pressures were not diminished by the anonymity of a software-based experiment, the experiments were conducted "live" (before the COVID-19 pandemic). That is, subjects were in the same room, and decisions were made in full view of the others. Subjects physically stood in the queue that they chose and waited to be served, and they were paid based on their queue choice and waiting time.

Using a between-subjects design, we tested the base model and the compensation model. Subjects were initially sequenced randomly, and according to this order, each subject successively chose between the regular and priority queue. We did not observe a "guilt"-related phenomenon: there was *not* a reversal of the theoretical prediction that more customers would purchase priority in the base model than with compensation. However, in all sessions for both treatments, fewer subjects purchased priority than the equilibrium predicted. Moreover, comparing subject decisions to the theoretical best-response functions reveals that they were much more likely *not* to purchase priority when they should have than to purchase priority when they should have chosen the regular queue.

Our key behavioral findings are as follows. First, the experiments hewed reasonably close to the theoretical prediction. In line with the equilibrium for the tested parameters, subjects early in the selection order indeed tended to realize that they needed to purchase priority preemptively to protect their favorable position. Not all did, however: some subjects early in the order chose the regular queue and earned a very long wait because later subjects purchased priority and overtook them. Second, as mentioned, fairness considerations did not appear to play a large role, at least in our experiments. Finally, in all treatments, subjects were conservative when it came to purchasing priority: they frequently did not purchase priority when it would have been rational to do so. We make no firm conclusions about the origin of this phenomenon, but one possible explanation is that a subject who is uncertain about the right decision may have an intrinsic preference for the free option—the regular queue—merely because it is free: in her decision-making process, she may place too much weight on the priority fee and not enough on the waiting-cost considerations.

That subject decisions closely matched the theory suggests that any fairness concerns did not outweigh the strategic imperative to value priority less in the compensation case. Thus, both strategically and behaviorally, we observe that redistribution tends to reduce the number of priority purchases. The implications for this finding are different depending on the goal of the priority system. Revenue-maximizing firms may be less likely to share priority proceeds given our findings. However, although the revenue impact of sharing priority proceeds may be negative, a desire for a better customer-facing image could motivate such sharing (perhaps with a low compensation fraction), especially for firms like airlines with tenuous customer relations. For a social planner

like a government who is less concerned about revenue, e.g., the State Department, redistributing waiting time and priority payments creates a fairer, more efficient system. Moreover, for entities implementing a congestion charge or a related system, we have identified another lever to help reduce congestion. Our findings imply that fewer commuters can be expected to pay the charge to drive into the city if they know that they will be compensated for not doing so. The commuters who do pay would then be subsidizing and thus incentivizing the "good behavior" of those who use other modes, and what is more, the subsidy is self-funding. The city can decide how much of the proceeds to redistribute by trading off the predicted congestion and emissions benefits of redistribution against its own revenue needs. Overall, we believe that both our theoretical and behavioral results can be useful for providers operating priority systems, especially those considering redistribution.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the theoretical model. In Section 4, we derive structural results for the base model and illustrate the nature of the equilibrium. Section 5 extends our model to the case with compensation, deriving additional structural results and contrasting the equilibrium with the base model. We discuss our behavioral experiments in Section 6, and we conclude in Section 7.

2. Literature Review

Theoretically, strategic customer behavior in priority service systems has been studied extensively. One of the earliest works in this stream is Kleinrock (1967), in which customers place bids for priority and are served in decreasing order of their bids. This system is equivalent to an infinite number of priority classes. In Adiri and Yechiali (1974), the service provider administers a finite number of priority classes, with a fixed price for each class. Both of these papers study queueing systems in steady state. Additional work in this stream includes Hassin and Haviv (1997), which studies priority purchasing in an observable $M/M/1$ queue, and references therein, as well as Erlichman and Hassin (2015), which we discuss below. We refer the reader to Yang et al. (2016) for an excellent review of the theoretical priority-pricing literature. Hassin and Haviv (2003) and Hassin (2016) thoroughly survey the strategic queueing literature.

In Yang et al. (2016), the focus is on a marketplace in which customers may bid for favorable positions in line, considering various transfer mechanisms. An appropriately designed auction mechanism, in which customers submit bids to overtake others based on their value of time, results in customers being served efficiently. Customers' bids increase in their waiting costs and therefore upon joining the system, a customer will overtake only customers with lower waiting costs than herself. An earlier work that also studies trading positions is Rosenblum (1992), which shows that in a $G/M/S$ system, customers can achieve a socially efficient equilibrium by paying to trade

positions; importantly, each customer's waiting cost is common knowledge to the other customers. Wang et al. (2019) studies a model of priority-purchasing for both unobservable and observable queues. In the observable setting, which is closer to our work, in equilibrium customers should purchase priority only when the queue is long. The analysis is in steady state and only symmetric equilibria are considered. Haviv and Winter (2020) derives revenue-optimal mechanisms for a two-class, steady-state queueing system. Their mechanisms result in different customers paying different fees to join the same priority class, while we implement a single priority price. A rare analytical study on social preferences in queueing is Allon and Hanany (2012), which studies queue-jumping behavior. They note that in practice customers in line may allow a new arrival to "jump" the queue if they have a very urgent request, or if they can be served quickly. Reputation effects can result in the socially optimal $c\mu$ rule being implemented in equilibrium. None of the above papers consider redistribution of priority payments.

Our study differs from most of the papers above in that we study a sequential game in a finite system, as opposed to a queueing network in steady state. Such a model affords us two important benefits. First, it allows us to drill down into the detailed dynamics of customer decisions at different positions in the queue. Second, it is well suited for behavioral experiments. While subjects may not understand stationary distributions or the dynamics of an unobservable queue, a fixed number of customers and observable moves as in our model affords them full queue-state information with no need to reason about steady state. Eliminating this cognitive challenge allows us to focus on the key incentives surrounding redistribution as well as isolate any behavioral biases. In this sense, our modeling approach is similar to Kremer and Debo (2016), but with different motivation. Like ours, their model is of a finite, observable queue in which customers decide sequentially; their focus is on whether customers learn about product quality through the actions of others who may be privately informed. Also like us, Kremer and Debo (2016) first studies the model analytically and then tests the theory via experiments, finding that uninformed subjects indeed learn from informed ones.

Another study involving finite queues is the "static" model in Erlichman and Hassin (2015); in that model, instead of paying a fixed fee to access a priority queue, each customer may pay a certain amount per overtaken customer; the unique equilibrium involves no overtaking. By contrast, in our model, priority purchases are common in equilibrium: two key differences between their framework and ours are that (i) in their model, customers choose how many others they wish to overtake and pay *per customer*, and (ii) in their model, a customer who overtakes others has no ability to avoid being overtaken herself by later customers. Curiel et al. (1989) is another example of a finite queueing game (they use the term "sequencing game," related to the scheduling literature); they

use cooperative game theory to investigate what sequence customers will self-select. For more on the scheduling literature, see Pinedo (2012) and references therein.

On the behavioral side, Larson (1987) offers many examples of the importance of individual perceptions of social justice in queueing, as well as conjectures based on these examples. Allon and Kremer (2018) gives an overview of the current state of behavioral queueing research, specifically noting the need for more research that studies social preferences. In a laboratory setting, El Haji and Onderstal (2019) studies different auction mechanisms to determine service order, both of which award all payments to other customers. They consider a server-initiated auction, in which customers bid to be served next and the winning bid is split among the losing customers, and a customer-initiated auction in which customers can pay each other to switch places. Like Haviv and Winter (2020), these mechanisms can result in different customers paying different amounts, and they are much more complicated than a two-class priority system. Buell (2021) studies last-place aversion, a special disutility from being at the very back of a queue. It is found in a field study and several lab experiments that from last place, customers are more likely to switch queues or abandon the system altogether, independent of the length of the queue in front of them. This phenomenon reflects a different side of social preferences: customers who are in last place cannot make a "downward social comparison," which reduces their satisfaction.

Dold and Khadjavi (2017) conducts laboratory experiments in which one subject can bribe another to reduce her own waiting time. Their subjects displayed strong social preferences; some were even willing to lengthen their own waiting times merely to punish another subject who paid to receive a better position, even though that subject had imposed no externality on others. The individual nature of the transactions in Dold and Khadjavi (2017) could provoke a stronger response to a norm violation than we observe in our experiments. In our setup, the payment is to the service provider—i.e., the experimenter—rather than between subjects, and we do not observe evidence of such strong social preferences.

Previous theoretical studies have explored customer and firm decisions in priority service environments with priority payments kept either by the service provider or traded among customers. Yang et al. (2016) also incorporates a combination of these, such that customers pay a fee to participate in trading and then engage in transfer payments. Previous behavioral studies have focused on one or the other of these settings: either the service provider keeps the payments, or the customers pay each other to trade positions. We believe that we are the first to study both theoretically and behaviorally the impact of the recipient of the priority payment on priority-purchase decisions.

3. The Model

Consider a service system with a single server and two queues, "regular" and "priority". There are N customers in the system who choose queues sequentially according to some order (e.g., random), and the server begins processing only after all customers have made their decisions. An example of a real-life setting in which customers are present before service starts is airplane boarding, in which passengers are present together in the gate area when the boarding process starts.

In our model, all N customers must be processed by the server. For ease of exposition, we assume that each service lasts exactly one unit of time. We note, however, that all of our results would hold unchanged for general independent and identically distributed service times because each customer makes her decision based on her expected waiting time and decisions are made before any service begins. Customers $i \in \{1, \ldots, N\}$ value the service at $V > 0$, which is deterministic and homogeneous. They incur a waiting disutility proportional to their total waiting time (including the time in service), with cost coefficients $C_i > 0$. These waiting costs are independent and identically distributed across customers, drawn from a distribution with cumulative distribution function (CDF) F with nonnegative and bounded support. Each customer's waiting-cost realization is her private information.

Upon arrival, each customer observes both the regular and priority queues and has a one-shot, irrevocable opportunity to either (i) pay a price p to join at the end of the priority queue, or (ii) join at the end of the regular queue at no cost. For simplicity, we assume that a customer joins the priority queue if indifferent between the two options. Also, let $\bar{c}$ be an upper bound on the waiting cost random variables and $\underline{c}$ a lower bound. To ensure that no customer has an incentive to balk and not receive service, and also that priority is at least inexpensive enough that it would be worth purchasing to move from last place to first place, we assume that $V \geq \bar{c}N$ and $p \leq \underline{c}(N-1)$, respectively. The setup of the game and the system parameters—including the price p, the number of customers N, the waiting cost distribution F, and the valuation V—are common knowledge. Importantly, a customer who purchases priority will not be overtaken and therefore knows exactly how many people will be served before her. On the other hand, a customer who joins the regular queue may be overtaken by a customer behind her who purchases priority, so she must take this possibility into account when making her decision.

Owing to the random waiting costs, our equilibrium concept is perfect Bayesian equilibrium (PBE). In a PBE, each customer i must use Bayes' rule to update her beliefs about the waiting costs of customers $1, \ldots, i-1$ after observing their decisions. However, the waiting costs for these customers are not payoff-relevant for customer i because only their actions, which are observable,

affect her waiting time. Hence, while customer i should of course use Bayes' rule to update them, her beliefs about these customers do not affect her decision or those of others. Regarding the customers after her, there is no information that customer i can use to update the prior distribution F for their waiting costs until after she has already made her decision. So, in the PBE, each customer determines her optimal action for each state by calculating her expected net utility after inferring the waiting-cost-dependent optimal strategies of the customers after her and taking expectation over the waiting costs of these customers.

For $i \in \{1, \ldots, N\}$, let x_i be the number of customers that purchase priority, up to and including the i-th customer to make her decision. By convention we take $x_0 = 0$. We will refer to the i-th customer as customer i. The quantity x_i defines the state of the priority and regular queues that is observed by customer $i+1$. For $i \leq i'$, by definition we have $x_i \leq x_{i'}$. Denote by $\sigma_i(x_{i-1}, C_i)$ a strategy function of customer i; that is, $\sigma_i : \mathbb{Z}_+ \times \text{supp}(C_i) \to \{0, 1\}$ maps from the number of customers x_{i-1} that customer i observes in the priority queue and her waiting cost C_i to a decision of either the regular queue (encoded as 0) or the priority queue (encoded as 1). Let $\boldsymbol{\sigma_{i,j}}$ represent a vector of strategy functions $(\sigma_i, \ldots, \sigma_j)$ for customers $i, \ldots, j$. Conditional on her own waiting-cost realization c_i, let $U_{R,i}(x_{i-1}; \boldsymbol{\sigma_{i+1,N}})$ be customer i's net utility from joining the regular queue if customers $i+1, \ldots, N$ follow the strategies $\boldsymbol{\sigma_{i+1,N}}$ and the number of customers in the priority queue that customer i observes is x_{i-1}. This utility is random even though customer i knows her own waiting cost because the strategies $\boldsymbol{\sigma_{i+1,N}}$ are functions of the later customers' random waiting costs, which are not known to her. Similarly, let $U_{P,i}(x_{i-1})$ be customer i's net utility from joining the priority queue, given x_{i-1}. Unlike the regular queue, customer i's utility from joining the priority queue is deterministic given her waiting-cost realization and does not even depend on the strategies of later arrivals because they cannot overtake her if she purchases priority.

4. Base Model

We start with the base model, in which all payments for priority are kept by the service provider, and we perform backward induction to characterize the equilibrium priority purchasing strategies.

4.1. Equilibrium Analysis and Structural Results

When customer N makes her decision, she observes x_{N-1} customers in the priority queue, and $N-1-x_{N-1}$ customers in the regular queue. As the last customer, her utility from each decision is fully determined by x_{N-1}. For waiting-cost realization c_N, her utility from the regular queue is $U_{R,N}(x_{N-1}) = V - c_N N$. Similarly, her utility from the priority queue is $U_{P,N}(x_{N-1}) = V - p - c_N(x_{N-1}+1)$. Optimally, she will purchase priority if and only if $U_{R,N}(x_{N-1}) \leq U_{P,N}(x_{N-1})$, i.e.,

$$V - c_N N \leq V - p - c_N(x_{N-1}+1) \iff x_{N-1} \leq N - 1 - \frac{p}{c_N}. \tag{1}$$

So, customer N purchases priority only if x_{N-1} is *small* enough, i.e., only if not too many customers have purchased priority (we take the floor of the quantity on the far right-hand side of equation (1) to get the integer-valued threshold). The longer the priority queue, the smaller the difference in customer N's wait time between the two queues. If almost everyone has purchased priority, then it is in customer N's best interest to join at the back of the regular queue; then, she saves p and still waits only a little bit longer than if she had joined at the back of the priority queue.

Observe that customer N's optimal threshold depends on her waiting-cost realization c_N. That is, customer N uses a *cost-dependent threshold strategy.*

Definition 1 (Cost-Dependent Threshold Strategy). A *cost-dependent threshold strategy* for customer i is a strategy function σ_i such that $\sigma_i(k+1, c_i) \leq \sigma_i(k, c_i)$ for all c_i in the support of C_i and $k \in \{0, \ldots, i-2\}$. For such a strategy, we define the shorthand notation $\bar{x}_{i-1}(C_i) := \max\{k : \sigma_i(k, C_i) = 1\}$.

In words, a cost-dependent threshold strategy is a strategy function such that the customer purchases priority if and only if the priority queue that she observes is below a threshold length, which threshold is a function of her random waiting cost (i.e., $\bar{x}_{i-1}(c_i)$ can be different for different realizations c_i). The case in which $\sigma_i(x_{i-1}, c_i) = 0$ for all possible x_{i-1} can be expressed by the threshold strategy $\bar{x}_{i-1}(c_i) = -1$ for the realization c_i; we adopt this convention.

We will see that customers $1, \ldots, N-1$ optimally also use cost-dependent threshold strategies, but to prove this requires us to establish an important property of the system under such strategies.

Lemma 1 **(Effect of One Additional Priority-Queue Customer)**. *Consider a customer $i \in \{1, \ldots, N-1\}$, and suppose that customers $j \in \{i+1, \ldots, N\}$ use cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$. Given these strategies, let L_i^k be the random number of services (including her own) that customer i will wait through if $x_{i-1} = k$ and she chooses the regular queue. We have*

$$0 \leq \mathsf{E}[L_i^k] - \mathsf{E}[L_i^{k+1}] \leq 1.$$

All proofs can be found in the appendix. Lemma 1 reveals a qualitative feature of cost-dependent threshold strategies that facilitates comparison of a customer's waiting time in the regular queue for different queue states that she observes. Namely, if every customer behind a focal customer uses such strategies, then the difference is at most 1 between the expected numbers of services that the focal customer must wait through after choosing the regular queue upon observing k versus $k+1$ priority-queue customers in front of her. This lemma exemplifies the sample-path proof approach that we use repeatedly to navigate the randomness in each customer's waiting cost. Ex ante,

customers earlier in the order must account for an enormous number of possible strategy functions for the later customers, mapping from each waiting-cost realization and queue state to an action. However, for a fixed sample path of waiting-cost realizations, each customer will use a pure strategy. We exploit the pure strategies to analyze the outcomes on each sample path, circumventing the combinatorial problem described above to reveal the structure of the equilibrium.

Suppose that the customers after a focal customer use deterministic threshold strategies. A focal customer i who chooses the regular queue will wait through at least i services because customers $1,\ldots,i-1$ will be served before her regardless of their and her queue choices. However, with fixed threshold strategies for the customers after customer i, more priority purchases in front of her may *reduce* the number of priority purchases after her because some customers whose thresholds are not exceeded if $x_i = k$ will be exceeded if $x_i \geq k+1$. In Lemma 1, we prove that on a given waiting-cost sample path, if $x_i = k+1$, among customers $i+1,\ldots,N$ there is either the same number of priority purchases or one less, compared to the case with $x_i = k$. Thus, if $x_{i-1} = k+1$ and customer i chooses the regular queue, then she will wait through either exactly the same number of services or one less than if $x_{i-1} = k$ and she chose the regular queue. This result holds for *any* threshold strategies among customers $i+1,\ldots,N$, so if these customers use cost-dependent threshold strategies, then it holds for every sample path and is preserved by expectation, hence the lemma.

With Lemma 1 in hand, we can characterize the structure of the PBE.

THEOREM 1 **(Base Model: Cost-Dependent Threshold Strategies)**. *In the unique PBE of the base model, all customers use cost-dependent threshold strategies.*

Because the sample-path approach used in Lemma 1 does not depend on the form of the waiting-cost distribution, Theorem 1 holds in full generality. Intuitively, Lemma 1 implies that the value of priority for customer i is less when observing $k+1$ priority customers in front of her than when observing k priority customers; in the regular queue her expected waiting time will be the same or better with $x_{i-1} = k+1$, but in the priority queue she will wait longer if $x_{i-1} = k+1$, with one more priority customer served before her than if $x_{i-1} = k$. Consequently, if priority is not worth the fee with k priority customers before customer i, then neither is it worth it with $k+1$, so a threshold strategy is optimal. The proof of Theorem 1 formalizes this reasoning with an induction.

Because all customers use cost-dependent threshold strategies in equilibrium, we might expect the equilibrium path to be "smooth" in that consecutive customers all make the same decision up to a point, after which a switch occurs and the rest of the customers make the opposite decision. However, this is not the case because the optimal thresholds differ among customers, even for the

same waiting-cost realization. If customer thresholds oscillate, then the equilibrium path might reflect an almost arbitrary sequence of customer decisions. To better understand the equilibrium, we establish the following theorem about the relationship among the optimal thresholds.

THEOREM 2 **(Base Model: Increasing Thresholds)**. *Let* $\bar{\boldsymbol{x}}^* = (\bar{x}_0^*(C_1), \ldots, \bar{x}_{N-1}^*(C_N))$ *be the vector of equilibrium threshold functions. For a constant* c *in the support of* F*, we have*

$$\bar{x}_{i-1}^*(c) \leq \bar{x}_i^*(c) \text{ for } i = 1, \ldots, N-1. \tag{2}$$

Moreover, for $c < c'$*, we have* $\bar{x}_i^*(c) < \bar{x}_i^*(c')$.

The preceding theorem demonstrates that the equilibrium thresholds are increasing. That is, a customer earlier in the order will have a lower threshold than one later in the order, conditional on the same waiting-cost realization. Intuitively, higher thresholds for later customers makes sense because a given priority queue length k implies one more regular queue customer for customer $i+1$ than for customer i. A longer regular queue should increase the value of priority because there are more customers to overtake and thus a greater wait time reduction from purchasing priority. However, the outcome also depends on the actions of the later customers. We use a similar sample-path approach to that used in Lemma 1 to compare the difference in expected regular-queue waiting time for customers i and $i+1$ if both observe the same priority queue length. We prove that in this case customer $i+1$ has a longer expected waiting time in the regular queue than customer i. Thus, a customer further back in the order finds priority more attractive for a given priority queue length, and hence the equilibrium thresholds are increasing for a given waiting cost. Also, as we would expect, customer i's threshold is increasing in her waiting cost.

4.2. Numerical Examples and Equilibrium Computation

Despite our structural results, calculating the PBE thresholds is still computationally difficult, even for simple waiting-cost distributions, because the optimal thresholds may be different for different realizations. Customers early in the order may then need to take expectation over a huge number of possible vectors of thresholds for the later customers. To build intuition, we first give some numerical examples for a degenerate waiting-cost distribution, in which case the PBE reduces to a subgame perfect Nash equilibrium (SPNE). Then, we give a general algorithm to compute the equilibrium, which we apply to a two-point distribution to determine the PBE threshold functions.

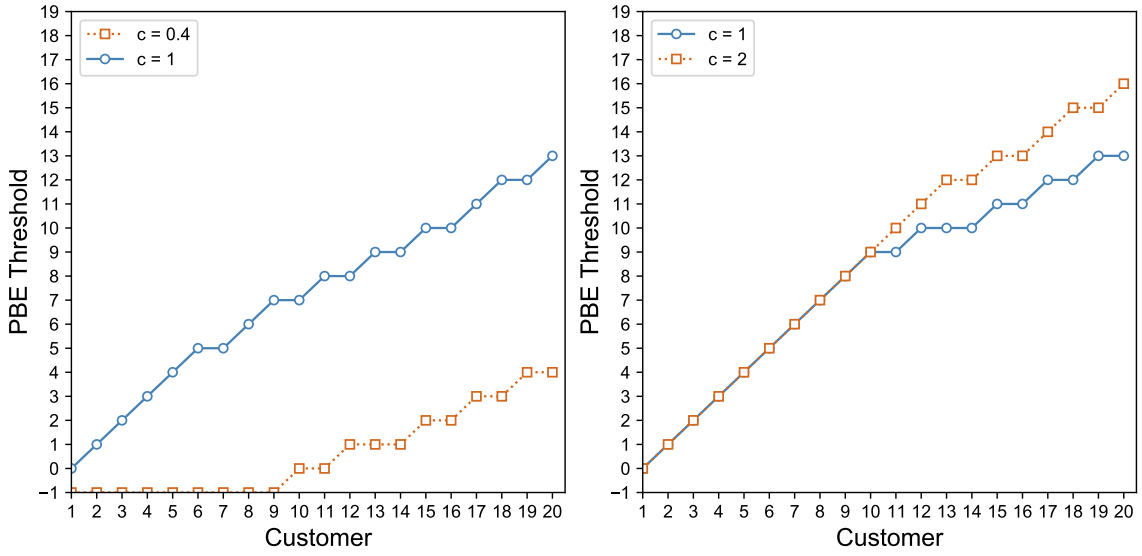
Table 1 reports the equilibrium path and number of priority purchases for degenerate waiting-cost distributions equal to some c with probability 1. For a given waiting cost, the total number of priority purchases decreases as the priority fee increases. Similarly, for a given priority fee, the number of priority purchases increases as the waiting cost increases. In extreme cases (in the table,

N	V	c	p	x_N	Equilibrium Path (0 = Regular, 1 = Priority)
20	35	0.2	3	5	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
20	35	0.2	6	0	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
20	35	0.2	9	0	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
20	35	0.6	3	15	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
20	35	0.6	6	10	(1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
20	35	0.6	9	5	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
20	35	1	3	17	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0)
20	35	1	6	14	(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
20	35	1	9	11	(1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)

Table 1 Subgame perfect equilibrium path for deterministic waiting cost c.

when $c = 0.2$ and $p = 6$ or 9), there are zero priority purchases because the fee is just too high relative to the waiting cost and the number of customers. We observe two other qualitatively different types of equilibria. In the last three rows of the table, for example, the first several customers purchase priority, and the later customers alternate between the two queues. These cases reflect a low priority fee relative to the waiting cost (these three rows have the same three priority fees considered in the rest of the table, but with a higher waiting cost than the rest of the table). When priority is cheap, the first customers to arrive have an incentive to purchase it, not to overtake other customers but instead to protect their privileged position in the queue. By contrast, in the first and fifth rows of the table, we see an almost opposite phenomenon. The early customers do *not* purchase priority, and only once the regular queue reaches a certain length do customers begin to purchase. In these cases, priority is expensive relative to the waiting cost, so it is not worth it to the early customers to purchase merely to protect their position. Instead, customers choose the regular queue until this queue is so long that it is worth the high fee to overtake these customers, after which the remaining customers alternate between the priority and regular queue.

In Appendix F, we provide an algorithm to compute the PBE threshold functions for an arbitrary bounded waiting-cost distribution. This algorithm codifies the backward induction process in which each customer must take expectation over all possible equilibrium strategies of the customers after her. For each customer i, starting from customer N, we must determine the probability of each threshold vector among customers $i+1,\ldots,N$ that occurs with positive probability (this process is trivial for customer N). The result is a finite probability distribution over vectors of integers, but it can have as many as $\prod_{j=i+1}^{N}(j+1)$ mass points, and $N-1$ such distributions must be determined. With this probability distribution, customer i can calculate her expected utility from each action as a function of her waiting cost and choose her threshold function accordingly. Importantly, it would *not* help our purpose to consider every vector of thresholds and determine



(a) Waiting Cost $C_i \in \{0.4, 1\}$ (b) Waiting Cost $C_i \in \{1, 2\}$

Figure 1 PBE thresholds for two-point waiting-cost distributions ($N = 20$, $V = 35$, and $p = 6$).

the subgame perfect equilibrium strategies for each vector; this would indeed reveal an equilibrium (one for each vector), but that of a different game. The reason is that each customer knows only her own waiting cost, so when determining her optimal actions, the uncertainty in the others' strategies must be maintained. This point reveals an important difference between calculating the PBE threshold functions and the sample path method that we use to prove the cost-dependent threshold structure. The numerical value of the later customers' thresholds matters when determining each customer's optimal strategy. By contrast, to prove our structural results, the values of other customers' thresholds are irrelevant; all that matters is that the strategies have a threshold structure for all possible waiting-cost realizations.

We apply our algorithm to compute the PBE thresholds for two-point distributions with equal probabilities. In contrast to the deterministic case, here the thresholds and hence also the equilibrium path depend on the waiting-cost realizations, so in principle there could be as many possible realized threshold vectors in the PBE as there are waiting-cost vectors in the support of $(C_1, \ldots, C_N)$. We plot the PBE thresholds for two different distributions in Figure 1. As implied by our structural results, for each possible waiting cost, the thresholds increase as we move back in the order. Moreover, in each case, the thresholds are higher for the higher realization. In the left panel of the figure, we observe both "types" of equilibria that we saw in the deterministic case, each for a different waiting-cost realization. The thresholds for customers early in the order who draw a waiting cost of 1 are equal to the maximum possible priority queue length that they can observe (the first customer has a threshold of 0, the second a threshold of 1, etc.), meaning that these

customers will purchase priority no matter what state they observe. This is a "protective" strategy: customers purchase priority to avoid being overtaken by others. On the other hand, customers early in the order who draw a waiting cost of 0.4 take the opposite approach. Their thresholds are -1, meaning that they will choose the regular queue in any state. For these customers, priority is expensive relative to their lower waiting cost, so they will only purchase priority if there are many regular-queue customers to overtake, which can only happen further back in the order. Because the sequences of strategies for the two possible waiting costs are so different, there can be almost arbitrary sequences of decisions depending on the sample path of realizations. Hence, the possible equilibrium paths do not follow a simple pattern, even though the equilibrium strategies can be neatly characterized.

Comparing the two plots highlights the impact of the randomness in the waiting costs. In both the left and right panel, one of the possible waiting-cost realizations is $c_i = 1$ (the thresholds for this realization are plotted in the same color and style in both panels for ease of comparison). At first thought, we might expect the PBE thresholds to be the same for a customer with $c_i = 1$ in a given position in the queue in both plots, but this is not the case. The reason is that the equilibrium strategies of customers after a focal customer will be different if the other realization is different, which changes the expected utilities of each action. For a waiting cost of 1, it is more important in the right panel to purchase priority to protect her position than in the left panel: preemptive purchases continue only through customer 6 in the left panel (other realization of 0.4), while they go through customer 10 in the right panel (other realization of 2) because the waiting cost and thus the thresholds for the other realization are higher in the right panel. These two plots both have two-point distributions with equal probabilities, with different supports. Similarly, for a customer with a waiting cost of 1 in a given position, compared to her threshold for the $\{1, 2\}$ equal probability case, her threshold could even be different for another distribution with the same support of $\{1, 2\}$ but different probabilities. There are thus many possible equilibria (let alone equilibrium paths) even among two-point waiting cost distributions with a given support, and the possibilities increase for more complicated distributions.

From our results, a service provider knows that in equilibrium customers use cost-dependent threshold strategies, and he can use our algorithm to calculate the equilibrium, as well as to perform sensitivity analysis to evaluate the effect of parameter changes. For instance, the provider may have a good estimate of the range of customers' waiting costs but not their distribution. He can use our results and algorithm to calculate the equilibrium for different possible distributions over this range, to assess what range of outcomes to plan for. Having now a sound understanding of the base model, we next move to extend our model to incorporate compensation, such that a fraction of the proceeds from priority is redistributed to the customers that choose the regular queue.

5. Compensation Model

A service provider can keep the priority payments, redistribute them fully among inconvenienced regular-queue customers, or keep a portion of the proceeds and redistribute the rest. The base model treated the case in which the service provider keeps all payments. In this section, we extend our model to a setting in which the service provider redistributes a fraction $\gamma \in (0,1]$ of the proceeds from priority to the regular-queue customers. The extreme of $\gamma = 1$ corresponds to the service provider fully redistributing the priority proceeds to regular-queue customers.

5.1. Structural Results

The "compensation" setting has more complicated dynamics because of the redistribution, and it is not obvious ex ante what form the equilibrium strategies should take, i.e., whether they can be shown to have a similar structure to those in the base model. If all customers purchase priority, then the payments are deemed to be forfeited. However, as in the base model, it is clearly sub-optimal for customer N to purchase priority if the first $N-1$ customers have all purchased priority because she would not improve her wait but would forfeit compensation. Therefore, the set of PBE would be the same under most reasonable assumptions about the priority payments in this outcome, e.g., if each customer's payment were returned to her.

Formally, if x_N customers purchase priority, leaving $N - x_N$ customers in the regular queue, then each regular-queue customer receives a payment of

$$\gamma\Big(\frac{px_N}{N-x_N}\Big). \tag{3}$$

For customer N with realized waiting cost c_N who arrives to find x_{N-1} customers in the priority queue, her utility from purchasing priority is the same as in Section 4, namely $U_{P,N}(x_{N-1}) = V - p - c_N(x_{N-1}+1)$. However, her utility from joining the regular queue is different because the compensation must be added to her utility, which (noting that $x_N = x_{N-1}$ in this case) yields

$$U_{R,N}(x_{N-1}) = V + \gamma\Big(\frac{px_{N-1}}{N-x_{N-1}}\Big) - c_N N,$$

Customer N will purchase priority if and only if $U_{R,N}(x_{N-1}) \le U_{P,N}(x_{N-1})$, which for $\gamma = 1$ is equivalent to

$$\begin{aligned} V + \frac{px_{N-1}}{N-x_{N-1}} - c_N N &\le V - p - c_N(x_{N-1}+1) \\ \iff x_{N-1} &\le N - \frac{1}{2}\Big(1 + \sqrt{1 + \frac{4pN}{c_N}}\Big). \end{aligned} \tag{4}$$

For $\gamma < 1$, the threshold takes a similar form but with a significantly more complicated expression under the radical, and the strategies of earlier customers are still more complex because of the

required equilibrium inference. To characterize the overall equilibrium structure requires us to understand the impact of the compensation payments on the equilibrium strategies. By doing so, we can prove that if the customers after a focal customer use cost-dependent threshold strategies, then her expected compensation payment is weakly increasing in the priority queue length.

LEMMA 2 **(Compensation Effect of One Additional Regular-Queue Customer)**. *Consider a customer $i \in \{1,\ldots,N-1\}$, and suppose that customers $j \in \{i+1,\ldots,N\}$ use cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$. Let $g_i^\gamma(k)$ be customer i's compensation after choosing the regular queue if $x_{i-1}=k$, for compensation fraction γ. For $k \in \{0,\ldots,i-2\}$, we have*

$$\mathsf{E}[g_i^\gamma(k)] \leq \mathsf{E}[g_i^\gamma(k+1)].$$

This result builds on Lemma 1. If the number of priority purchases in front of a focal customer increases by one, then Lemma 1 shows for a sample path that there will be either the same number or one less priority purchase after her. Thus, the total number of priority purchases is either the same or one more. So, either the compensation payment is the same or, if there is one more priority purchase, then it is more because the total proceeds from priority are higher but they are split among fewer regular-queue customers. Therefore, the expected compensation from choosing the regular queue after observing priority queue length k is weakly less than with priority queue length $k+1$. The direction of the effect on compensation is the same as that of the waiting-time effect of a longer priority queue length, so the overall effect is monotonic. Specifically, more customers in the priority queue means both a diminished waiting-cost gain from choosing priority (Lemma 1) and more compensation in the regular queue (Lemma 2), so for long enough priority queue lengths it is optimal to choose the regular queue. Accordingly, we can show that cost-dependent threshold strategies are optimal in the compensation model just as they were in the base model.

THEOREM 3 **(Compensation: Cost-Dependent Threshold Strategies)**. *In the unique PBE of the compensation model, all customers use cost-dependent threshold strategies.*

Although the equilibrium involves cost-dependent threshold strategies both with and without compensation, the values of the thresholds are different with compensation. Moreover, the dynamics are more complex because the compensation is a non-linear function of the number of priority purchases, and customers must account for the reactions of the others to this incentive. Nonetheless, we can still show that the equilibrium obeys a similar structure to the base model.

THEOREM 4 **(Compensation: Increasing Thresholds)**. *Let $\bar{\boldsymbol{x}}^* = (\bar{x}_0^*(C_1),\ldots,\bar{x}_{N-1}^*(C_N))$ be the vector of equilibrium threshold functions in the compensation model. For a constant c in the support of F, we have*

$$\bar{x}_{i-1}^*(c) \leq \bar{x}_i^*(c) \text{ for } i=1,\ldots,N-1. \tag{5}$$

Moreover, for $c < c'$, we have $\bar{x}_i^*(c) < \bar{x}_i^*(c')$.

Theorem 4 uses a related argument to Lemma 2, showing that the expected compensation for customer $i+1$ if $x_i = k$ is lower than customer i's expected compensation if $x_{i-1} = k$. Combined with the effect on waiting time from being one spot farther back in the queue (see the proof of Theorem 2), the reduced compensation implies that customer $i+1$ finds priority relatively more valuable than would customer i. Additionally, in a given state, if priority is worth the fee for a customer with a lower waiting cost, then it is clearly also worth the fee with a higher waiting cost.

Theorems 3 and 4 demonstrate that the underlying structure of the equilibrium is similar both with and without compensation. To wit, for a given waiting-cost realization, customers farther back in the queue are willing to purchase priority when facing longer priority queues. However, the equilibria are different in the two models because redistribution makes priority less valuable by improving the regular-queue outcome. The next theorem formalizes this intuition and highlights the side effect of redistributing priority proceeds, which service providers should consider carefully.

THEOREM 5 **(Thresholds Lower in Compensation Model)**. *For the base model, let $\bar{x}_{i-1}^*(C_i)$ be the equilibrium threshold function for customer $i \in \{1, \ldots, N\}$. Similarly, in the compensation model with compensation fraction $0 < \gamma \leq 1$, let $\bar{x}_{i-1,\gamma}^*(C_i)$ be the equilibrium threshold function for customer i. For $i \in \{1, \ldots, N\}$ and all c_i in the support of C_i, we have*

$$\bar{x}_{i-1,\gamma}^*(c_i) \leq \bar{x}_{i-1}^*(c_i),$$

i.e., the thresholds with compensation are lower than the corresponding thresholds in the base model. Furthermore, for fixed cost-dependent threshold strategies for customers $j \in \{i+1, \ldots, N\}$, customer i's optimal threshold with waiting-cost realization c_i decreases with γ.

Despite both involving cost-dependent threshold strategies, the equilibria differ between the two models because the effect of one additional priority customer is magnified with compensation. Theorem 5 establishes an unambiguous relationship between the equilibrium thresholds in the compensation model and those in the base model (without compensation), solidifying our intuition that redistribution makes priority less valuable. Namely, the thresholds are lower with compensation than without, meaning that there will be fewer equilibrium priority purchases under compensation. So, a firm that redistributes a share of priority proceeds to customers takes a double hit to its priority revenue, if customers are rational and ignore others' payoffs. First, the revenue is directly reduced because the firm keeps only a $1-\gamma$ fraction of the priority payments. Second, the revenue reduces indirectly because the equilibrium thresholds decrease, meaning that fewer customers

	Base Model		Compensation Model ($\gamma = 1$)	
p	x_N	$\bar{\boldsymbol{x}}^*$	x_N	$\bar{\boldsymbol{x}}^*$
3	17	(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,14,15,15,16,16)	12	(0,0,1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9)
6	14	(0,1,2,3,4,5,6,7,8,8,9,9,10,10,11,11,12,12,13,13)	9	(-1,-1,0,0,1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8)
9	11	(0,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10)	7	(-1,-1,-1,-1,-1,-1,-1,0,0,1,1,2,2,3,3,4,4,5,5,6)

Table 2 Equilibrium thresholds and total purchases with deterministic waiting costs ($N = 20$, $V = 35$, $c = 1$).

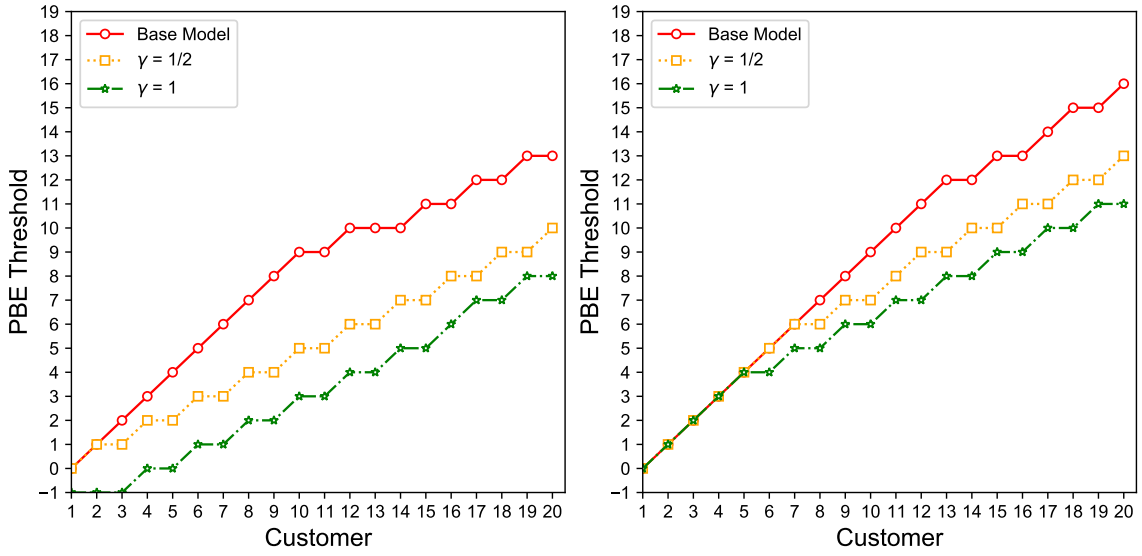
purchase priority. Thus, a firm considering redistributing priority payments to engender customer goodwill should beware of the dual hit to its revenue, at least for rational customers. The question remains whether behavioral factors will result in this gap being closed, and we study this question in the laboratory in Section 6.

Among different positive compensation fractions, it is difficult to directly compare the equilibrium thresholds because if the thresholds behind a focal customer decrease with the compensation fraction, then a higher compensation fraction means a shorter wait time but a worse compensation in the regular queue, with an ambiguous net effect on the utility. Still, if we fix the strategies of those after a focal customer, then we show in Theorem 5 that the focal customer's optimal threshold is decreasing in the compensation fraction γ. Numerically, we observe that the equilibrium thresholds also decrease with γ. The examples below demonstrate the different equilibria across the two models and for different compensation fractions.

5.2. Numerical Examples

As with the base model, we first look at the special case of deterministic waiting costs. Table 2 compares the PBE thresholds and number of priority purchases from the base model with those in the compensation model with $\gamma = 1$. Recall that a threshold of -1 means that a customer will join the regular queue no matter what. For both models, the thresholds are naturally decreasing in the priority fee p. Also, as implied by Theorem 5, the equilibrium thresholds are lower with compensation, and the number of priority purchases is thus also lower. This reflects the fact that when priority proceeds are shared with regular-queue customers, priority becomes relatively less valuable. For brevity, we do not include results for intermediate compensation fractions, but as an example, for the prices and waiting cost in the table, the thresholds for $\gamma = 1/2$ are indeed between those for the base model and those for the compensation model with $\gamma = 1$.

In Appendix F, we also provide an algorithm to compute the PBE thresholds for arbitrary distributions in the compensation model. We compute the equilibrium and plot the changing thresholds for a two-point waiting-cost distribution with equal probabilities for two different compensation fractions in Figure 1. The thresholds labeled "Base Model" are the PBE thresholds from the right panel of Figure 1 for the respective waiting-cost realization. The figures match our intuition from Theorem 5 and Table 2: the PBE threshold is decreasing in the compensation fraction γ, and indeed



(a) Waiting Cost Realization $c_i = 1$ (b) Waiting Cost Realization $c_i = 2$

Figure 2 PBE thresholds with/out compensation for waiting costs drawn from $\{1, 2\}$ ($N = 20$, $V = 35$, $p = 6$).

the difference can be significant. For example, for customer 20 with waiting-cost realization 1, the threshold of 13 in the base model is more than 60% higher than the threshold of 8 for full compensation (i.e., for $\gamma = 1$). So, the effect of compensation has more than just an incremental impact on the equilibrium. Indeed, the number of priority purchases can decrease significantly with redistribution. The case of $\gamma = 1/2$ is also plotted, with predictably intermediate results: thresholds are lower than in the base model but higher than with full compensation. These findings reinforce the analytical insight from Theorem 5. Service providers who share priority proceeds with customers should do so with clear eyes and account in their planning for the impact of compensation on priority purchases. If they wish to promote a benevolent image by sharing priority payments, then because of the double dampening effect of compensation (smaller share and fewer purchases), it may be prudent to share a relatively small fraction. At least, this is the case with fully rational customers who do not adjust their behavior out of fairness considerations.

Behaviorally, however, customers may not play the equilibrium. In the compensation case, for instance, there could be less "guilt" for purchasing priority and bypassing other customers because those customers are compensated for the extra time that they must spend in the queue. To the extent that this phenomenon influences customer decisions, we could observe more priority purchases in the compensation setting than in the base model, contrary to the theoretical predictions.

6. Behavioral Experiments

We tested subjects' priority purchase decisions in two settings: the base model and the compensation model. All currency calculations were in units of actual U.S. dollars. Subjects were paid a

$5 participation fee, with any additional payment determined from how well they performed on the task. Subjects were endowed with $30 each on top of the participation fee, and their final payments were determined by deducting their waiting cost and any priority payment from their endowment, then adding the participation fee. The instructions given to subjects can be found in Appendix G (H) for the base model (compensation model). These instructions also give illustrations of how subject payments are determined in each setting: calculations are based on the respective payoff functions from the analyses in Sections 4-5. Experiments were conducted through a behavioral laboratory at a large public university in the southern United States. Subjects were a mix of graduate and undergraduate students, recruited through an online recruiting platform.

6.1. Experimental Setup

We conducted two sessions for each treatment, for four sessions in total. Each session included 20 subjects ($N = 20$), with a different group of subjects for each session. The main purpose of this behavioral experiment was to identify any systematic deviations from the models' predictions, especially any that may be due to social preferences. With this goal in mind, we conducted the sessions in person, rather than on the computer. This deviation from the standard method of conducting laboratory experiments was intentional because the lack of anonymity highlighted potential feelings of fairness that may arise in a real situation in which customers pay for priority, effectively "cutting in line" in front of other customers. The sterility and anonymity of a computer interface would have served to minimize these effects, and our goal was to stress test this aspect of the model. As a first step to understand the behavioral effects of priority payment redistribution, we tested the special case of deterministic and homogeneous waiting costs, with a compensation fraction $\gamma = 1$ in the compensation model treatment. In this special case, the PBE reduces to a subgame perfect Nash equilibrium (SPNE). The combined endowment and participation fee imply a valuation $V = \$35$. We set the priority fee p at $6, and the waiting cost c at $1 *per service*, including the subject's own. Indexing the waiting cost to the number of services rather than time serves two purposes. First, it allows flexibility in the implementation: this choice eliminated the need to ensure that each service took exactly the same amount of time, which would have been more difficult in the live setting than in a computerized setup. Second, since we imposed a specific waiting cost on the subjects, indexing to the number of services instead of time further divorced subject decisions from their own intrinsic waiting costs. For similar reasons, we ensured that all subjects were released from the experiment at the same time; that is, their decisions only affected their monetary payoffs, not their total time commitment to the experiment.

Upon entering the room, subjects received a random "decision sequence" number which determined the order in which they made their choice between the two queues. When a subject's turn

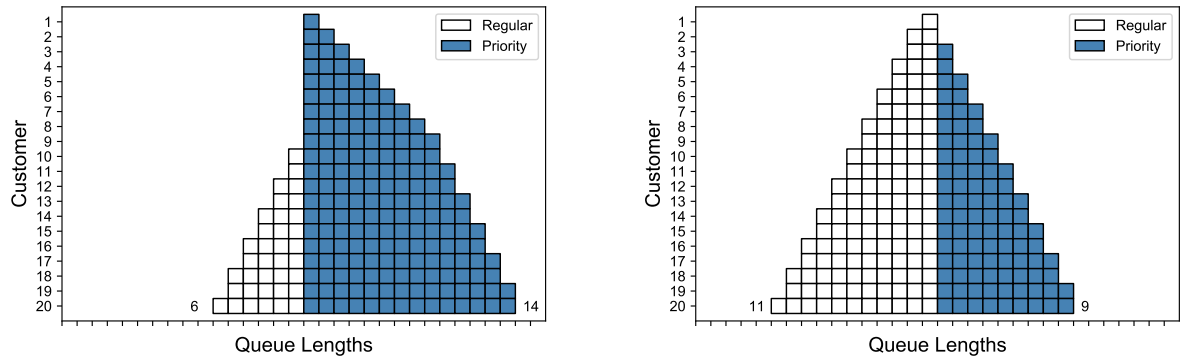
Treatment	x_N	Profile (0 = Regular, 1 = Priority)
Base Model	14	(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
Compensation Model ($\gamma = 1$)	9	(0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)

Table 3 Equilibrium path for experimental treatments ($N = 20$, $V = 35$, $c = 1$, $p = 6$).

arrived, they could observe the decisions made by those earlier in the sequence before making their own choice. After announcing which queue they wished to join, the subject physically stood in that queue and waited to be served. The choices resulted in a service sequence which differed from the decision sequence. After all subjects had made their decisions and were standing in their chosen queues, they came forward to be served. The service consisted of recording the subject's decision sequence number and service sequence number, calculating the payoff, and paying the subject. Subjects in the priority queue were served first, with FCFS service within the priority queue; once all priority subjects had been served, the regular queue subjects were served, also FCFS within the regular queue.

6.2. Experimental Results

Figure 3 depicts the SPNE paths for the base model (left panel) and the compensation model (right panel), for the parameter values used in the experiments. We also report these paths as vectors in Table 3. In Figure 3, play proceeds from the top down: the change in the queues from one row to the next indicates the decision made by the current subject. An unshaded square (shaded square) represents a customer in the regular queue (priority queue). For instance, in Figure 3a, the first 9 customers should purchase priority in equilibrium, and the remaining 11 should alternate between the regular and priority queues, i.e., this equilibrium involves protective priority purchases, as discussed in Section 4.2. So, the number of shaded squares (priority queue) increases by 1 starting at the top and moving down each of the first 9 rows, after which the number of unshaded squares (regular queue) increases by 1 in the tenth row, followed by each queue increasing in an alternating fashion for the remaining rows. The total number of subjects who should choose each queue is listed in the bottom row of the chart and matches the number of squares of the given type in that row. The differing nature of the equilibria in the base model versus the compensation case comes across sharply in the figure: in the left panel depicting the base model, as mentioned play is "smoother:" customers join the priority queue until it reaches a certain length, after which they alternate between the queues. By contrast, in the compensation model customers alternate back and forth almost from the very beginning, giving the chart more of a "Christmas tree" appearance. In the end, 9 total customers should purchase priority in the SPNE for the compensation case, fewer than the base model because, as discussed, priority is relatively less valuable when regular customers are compensated.



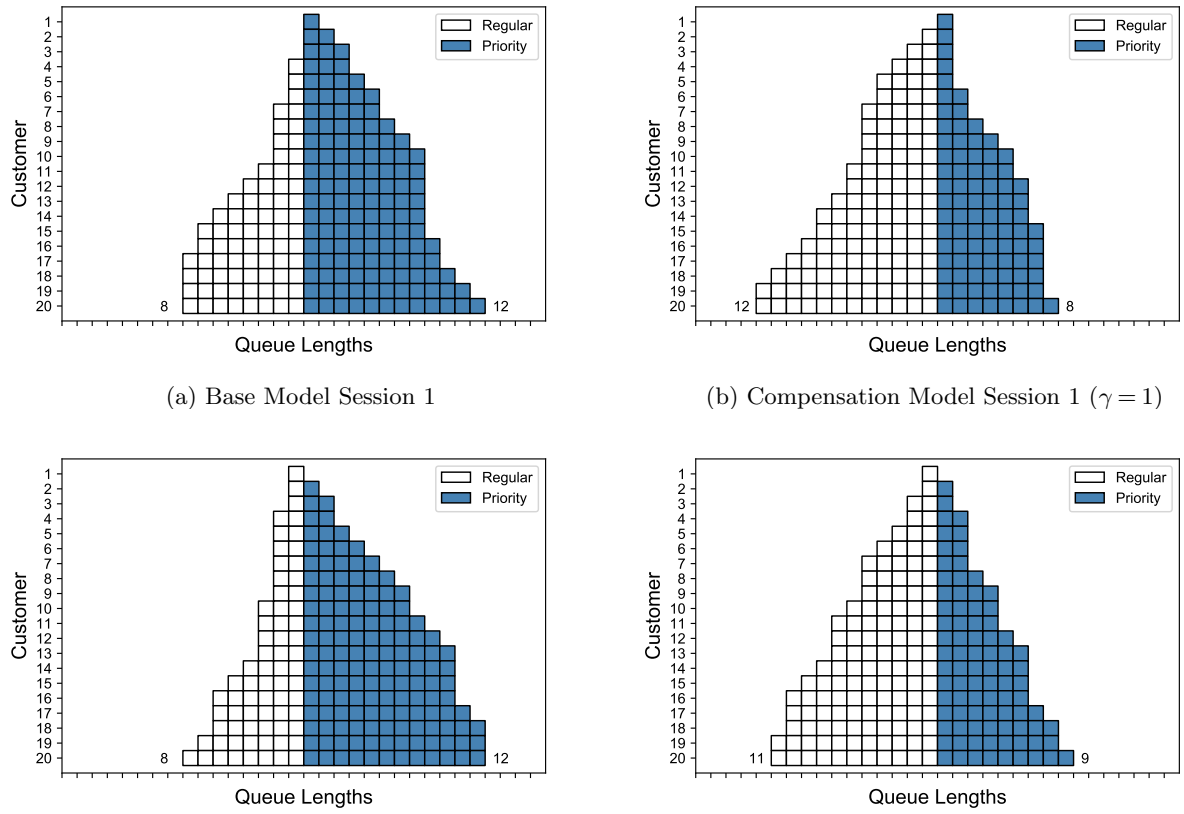
(a) Base Model (b) Compensation Model ($\gamma = 1$)

Figure 3 Subgame perfect equilibrium paths for both treatments with $N = 20$, $V = 35$, $c = 1$, $p = 6$.

The left column of plots in Figure 4 gives the sequences of subject decisions in the base model experimental sessions. In both such sessions, 12 subjects in total joined the priority queue, and the remaining 8 subjects joined the regular queue. While the realized paths do not exactly match the theoretical equilibrium, they are remarkably close to it. In both sessions there was an initial flurry of priority purchases, as indicated by the number of shaded squares growing faster than the number of unshaded squares, roughly for the first 10 subjects. This pattern suggests that many subjects indeed internalized the possibility of being overtaken by later arrivals, and they correctly purchased priority to preclude this possibility. At the same time, in both sessions there were subjects much earlier than predicted who chose to stay in the regular queue, and they suffered for it—particularly the subject in the second session who was first to choose, remained in the regular queue, and in the end was thirteenth to be served.

The right column of plots in Figure 4 depicts results from the compensation sessions. In the first (second) session, 8 (9) subjects purchased priority and the remaining 12 (11) joined the regular queue. These totals are quite close (exactly equal in the second session) to the theoretical prediction, although the exact sequence of decisions does not match the theory. The average number of subjects who joined the priority queue in the compensation treatment is 8.5, which is significantly lower than the average of 12 in the base treatment (t-test p-value = 0.0198, using session average as the unit of analysis). This difference is qualitatively consistent with the SPNE.

Overall, our experiments exhibit the same phenomenon found in the theoretical analysis: fewer subjects purchased priority in the compensation treatment than in the base model treatment. From a pecuniary standpoint, this outcome is not surprising. However, it lends little support to the notion that "guilt" for purchasing priority is less in the compensation treatment. We do observe fewer priority purchases in the base model sessions than in the theoretical equilibrium—12 versus 14—although the origin of this difference is not obvious. We next conduct a brief analysis, comparing each subject's decision to the SPNE move given the queue lengths that the subject observed.



(a) Base Model Session 1

(b) Compensation Model Session 1 ($\gamma = 1$)

(c) Base Model Session 2

(d) Compensation Model Session 2 ($\gamma = 1$)

Figure 4 **Subject decisions for treatments with $N = 20$, $V = 35$, $c = 1$, $p = 6$.**

	Base			Compensation			Aggregate			
Queue Type	Chosen	Not Chosen	Best Reply	Chosen	Not Chosen	Best Reply	Chosen	Not Chosen	Best Reply	Error Rate
Priority	24	16	40	12	10	22	36	26	62	41.9%
Regular	0	0	0	13	5	18	13	5	18	27.8%

Table 4 **Analysis of subjects' decisions.**

6.3. Analysis of Individual Decisions

Because the game unfolds sequentially, we can analyze each subject's decision relative to what backward induction would have prescribed (we will call this the *best reply*). Each subject observed the lengths of the two queues when making her decision, and these lengths may or may not have matched the SPNE path to that point (most likely they did not). However, it is possible to use backward induction, and assuming that the rest of the subjects would do the same, compute the decision that would be the best reply given the current state of the two queues. Of course, even if a subject possessed the necessary sophistication to find the optimal strategy, she might choose differently if she did not believe that the remaining subjects would act optimally. We believe that it is nonetheless useful to compare observed decisions to the best reply to better understand how closely subjects' decisions match theoretical predictions at a more granular level.

Independent Variable	Model (1)	Model (2)	Model (3)	Model (4)
Best Reply	1.281** (0.586)			
Threshold - Priority Queue Length		0.510** (0.208)	0.874*** (0.267)	1.045*** (0.369)
Priority Queue Length			-0.231** (0.091)	-0.234** (0.092)
Compensation Treatment				0.0504 (0.0728)
Constant	-0.956* (0.526)	-0.194 (0.251)	0.810* (0.469)	0.488 (0.653)
Log Likelihood	-52.80	-52.11	-48.52	-48.28
Observations (Groups)	80(4)			

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 5 Logit models of the likelihood of selecting priority queue.

Table 4 presents the number of decisions in each treatment that fall into each of four categories: (1) priority queue is the best reply and is chosen, (2) priority queue is the best reply and is not chosen, (3) regular queue is the best reply and is chosen, and (4) regular queue is the best reply and is not chosen. The columns labeled "Chosen" report the number of subjects who *correctly* chose a given queue ("correctly" in the sense that their choices aligned with backward induction), while the columns labeled "Not Chosen" report the number of subjects who incorrectly did *not* choose that queue when they should have chosen it. The columns labeled "Best Reply" give the total number of either priority or regular queue choices that should have been made given the state of the queues that each subject observed. We give results for each treatment separately in the left two sections, and then aggregate them in the far right part of the table. The column labeled "Error Rate" represents the number of not-chosen decisions for each queue divided by the total number of times that this queue should have been chosen, had subjects followed the best reply strategy.

First, we note that in the base model treatments, not once did a subject correctly choose the regular queue or incorrectly choose the priority queue. The reason is that no subject's SPNE threshold was ever reached, so a scenario in which one of these two outcomes would be possible was never observed. That is, for every subject i, we had $x_{i-1} \leq \bar{x}^*_{i-1}$, so if the subjects were each performing backward induction, then not one of them should have chosen the regular queue. Another observation that we can make from Table 4 is that although the lengths of the two queues at the end of the session, in the two treatments, are close to what is predicted by the SPNE, this is not because individual subjects closely follow the equilibrium strategies. They do make both types of errors (fail to choose a type of queue when they should), but the errors end up canceling out, so the final outcome closely resembles SPNE. Nevertheless, we observe that subjects are much more

likely to fail to choose the priority queue when they should (41.9% of decisions) than they are to fail to choose the regular queue when they should (27.8% of decisions). Note that this behavior is not consistent with risk aversion, because priority queue guarantees a serving sequence, while the regular queue does not. But this behavior may be consistent with loss aversion, because the priority queue is costly, while the regular queue is free, so if a subject is uncertain, she may feel that it is safer to err on the side of the free option.

We next fit several logit models, presented in Table 5, with the subject's decision as the dependent variable (which we coded as 1 when the priority queue was chosen and 0 when the regular queue was chosen). Since each treatment includes two sessions, we include a random effect for session. In Model (1) we use the best reply as the independent variable. The Best Reply variable is set to 1 if the best reply is to choose the priority queue and 0 otherwise. The coefficient for the best reply is positive and significant, indicating that subjects are more likely to select the priority queue when that is the optimal decision. We get a slightly better fit with Model (2) if we replace the Best Reply variable with the difference between the threshold and the priority queue length. This variable conveys the same information as the Best Reply variable, but the difference variable also reflects the magnitude of the benefit of choosing the priority queue. Naturally, the likelihood of selecting the priority queue increases with the difference between the threshold and the queue length. When we add a variable for the priority queue length in Model (3), the fit improves substantially. The length variable is negative and significant, so the likelihood of selecting the priority queue decreases with the length of the queue, even beyond the decrease that is captured by the difference between the priority queue length and the threshold. This observation may shed some light on why subjects err more often on the side of failing to choose the priority queue when they should: when the priority queue is long, subjects may underestimate the benefit of the priority status. In Model (4) we add the indicator variable for the compensation treatment. This variable is not significant, which suggests that the treatment effect we observe is fully explained by the analytical model and the priority queue length that subjects observe.

7. Conclusion

We have studied a priority service system with two classes (priority and regular) and random waiting costs. We considered two variations of our model reflecting the recipient of the priority payments. These we call the base model and the compensation model, in the latter of which the service provider redistributes a fraction of the priority proceeds to the regular customers.

In the base model, despite the analytical challenge of a sequential game with randomness at every stage, we derived structural properties of the perfect Bayesian equilibrium. We used a sample-path

approach to prove that the equilibrium has a cost-dependent, increasing threshold structure: for the same realized waiting cost, customers further back in the decision sequence are willing to purchase priority given a longer priority queue than those earlier in the sequence. However, although the strategies are nicely structured, the equilibrium path is not, even in the special case of deterministic and homogeneous waiting costs. Alternation can occur between the regular queue and the priority queue. Moreover, the equilibria can take different forms: if priority is cheap for a given waiting-cost realization, then customers early in the order purchase priority to protect their position, while if priority is very expensive, then only customers later in the order will purchase priority because only for them will there be enough regular-queue customers to overtake to justify the fee.

In the compensation model, we proved that the equilibrium strategies also obey a cost-dependent threshold structure. However, there is an important difference from the base model. Namely, in the compensation model, customers choosing between queues must internalize the compensatory payment received by regular-queue customers. This payment makes the regular queue relatively more desirable, and we proved that this results in lower thresholds and fewer priority purchases in equilibrium. We also found that the higher the compensation fraction shared with customers, the *less* valuable is priority.

We hope that our theoretical findings can help service providers understand and predict customer behavior in priority service systems. That priority becomes less valuable with more compensation is particularly relevant for service providers considering redistributing priority proceeds. The provider gives up a piece of the pie by sharing proceeds, and the pie also shrinks because redistribution improves the regular-queue outcome. This revenue reduction could be a deal-breaker for some firms. Others, however, may still wish to gain customer goodwill by promoting fairness and being seen as less greedy. To mitigate the negative revenue impact, such a firm might choose to share a smaller fraction of the proceeds, both to give away less of the pie and to keep it from shrinking too much. Alternatively, a social planner like a government (as in our motivating example of expedited and non-expedited passport renewals) may deem the redistributive benefits of sharing priority proceeds to be worth a reduction in revenue. Indeed, in some systems reducing the number of priority purchases could even be a goal, as with city congestion charges. Our findings suggest a new lever for managing congestion in such systems, given that redistribution tends to reduce priority purchases. Namely, planners can potentially reduce congestion (measured by the number of people who pay the toll) merely by sharing a higher fraction of the proceeds with those who do not purchase.

We also studied the behavior of human decision makers in a controlled laboratory environment. Although we anticipated that social preferences might result in subjects over-purchasing priority

in the compensation model relative to theory (and relative to the base model), we did not observe such a phenomenon. We did find in both treatments that human decision makers purchased priority less frequently than predicted by the theoretical equilibrium. Moreover, an analysis of individual decisions revealed that subjects were substantially more likely to stay in the regular queue when they should have purchased priority than they were to purchase priority when they should have stayed in the regular queue. This may be because the likelihood of purchasing priority decreases with the length of the priority queue that subjects observe, beyond the effect driven by their threshold for priority purchase that is implied by the equilibrium. Apart from the slight overall reduction in the number of priority purchases, subject decisions in aggregate aligned reasonably well with the theoretical equilibria in both settings. We also found that most (though not all) subjects early in the order correctly internalized the need to preemptively purchase priority to protect their position, which matched the theoretical equilibrium for the tested parameters.

Redistributing priority payments results in different theoretical equilibria. It also results in different purchasing decisions by experimental subjects in our study that are largely consistent with equilibrium behavior, at least in aggregate. We did not find evidence of social preferences or the desire not to inconvenience others. It is possible that even though we conducted the experiments in person, our experimental setup was still not ideal for inducing social preferences. In our treatments, subjects were symmetric; this was intentional. Subjects knew that the decision sequence was determined randomly, so everyone had an equal chance to be early in the decision sequence. In other words, subjects had no reason to feel any property right to their place in line. The fact that the cost of waiting was also symmetric as well as deterministic (we tested this special case of our more general theoretical model, which includes stochastic waiting costs) only strengthened the sense of symmetry. This obvious symmetry may have resulted in subjects being largely unencumbered by feelings of guilt. It may be possible to change the design of the experiment in a way that preserves theoretical properties of the game, while making social preferences more salient. This would require a design that would instill in subjects a sense of entitlement, or a property right, to their spot in the queue.

Another finding that could lead to future research is that subjects seem to be more likely to incorrectly pick the regular queue than to incorrectly pick the priority queue. For service providers, it would be valuable to understand the origin of this bias. If it emanates from something related to loss aversion, then the cost of the priority queue could be framed to make the cost seem less salient. If it emanates from the fact that the regular queue is viewed as the status quo, then service providers could reframe the decision as, for example, providing a discount for forgoing priority, which would make the priority queue the status quo (there are some examples of this attempt

at reframing, such as the "basic economy" airline class, or an option Amazon gives to its Prime customers to accept a later delivery in exchange for a small reward). If merely observing the length of the priority queue deters priority purchases, then service providers could hide this information from customers. Different underlying causes could suggest different interventions. We do believe that social preferences play some role in real queues, even if not in our experiment, because for example, inequity in the airline industry has been raising increasing concerns as the number of status classes has increased and customers without status have been pushed further and further towards the back of the line. So we observe that plenty of customers "vote with their feet" and either purchase priority or exert significant effort to obtain elite status, or switch to airlines, such as Southwest, with a more straightforward loyalty program. Of course, what we observe in the real world may be due to inherent asymmetry in customers, which did not exist in our experiment, but is covered by our analytical model and results. Understanding when customers are more or less sensitive to the fairness of priority queues is a ripe area of future study.

References

Adiri I, Yechiali U (1974) Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research* 22(5):1051–1066.

Allon G, Hanany E (2012) Cutting in line: Social norms in queues. *Management Science* 58(3):493–506.

Allon G, Kremer M (2018) Behavioral foundations of queueing systems. Donohue K, Katok E, Leider S, eds., *The Handbook of Behavioral Operations*, 325–366 (John Wiley & Sons).

Arlotto A, Frazelle AE, Wei Y (2019) Strategic open routing in service networks. *Management Science* 65(2):735–750.

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

Clear (2021) *How It Works: The Clear Travel Experience*. `https://www.clearme.com/how-it-works`, Accessed 08/11/2021.

Curiel I, Pederzoli G, Tijs S (1989) Sequencing games. *European Journal of Operational Research* 40(3):344–351.

Dold M, Khadjavi M (2017) Jumping the queue: An experiment on procedural preferences. *Games and Economic Behavior* 102:127–137.

El Haji A, Onderstal S (2019) Trading places: An experimental comparison of reallocation mechanisms for priority queuing. *Journal of Economics & Management Strategy* 28(4):670–686.

Erlichman J, Hassin R (2015) Strategic overtaking in a monopolistic m/m/1 queue. *IEEE Transactions on Automatic Control* 60(8):2189–2194.

Hassin R (2016) *Rational Queueing* (CRC Press).

Hassin R, Haviv M (1997) Equilibrium threshold strategies: The case of queues with priorities. *Operations Research* 45(6):966–973.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Springer Science & Business Media).

Haviv M, Winter E (2020) An optimal mechanism charging for priority in a queue. *Operations Research Letters* 48(3):304–308.

Kleinrock L (1967) Optimum bribing for queue position. *Operations Research* 15(2):304–318.

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

Larson RC (1987) OR Forum–Perspectives on queues: Social justice and the psychology of queueing. *Operations Research* 35(6):895–905.

Levine A (2019) *Is Six Flags' Flash Pass worth the cost?* https://www.tripsavvy.com/six-flags-flash-pass-3224566, Accessed 1/20/2020.

Lipsey S (2016) *Are airlines creating a class of spoiled, entitled brats?* https://www.yahoo.com/lifestyle/are-airlines-creating-a-class-1336347816460342.html, Accessed 1/20/2020.

Oberholzer-Gee F (2006) A market for time fairness and efficiency in waiting lines. *Kyklos* 59(3):427–440.

Papa John's (2021) *Terms & conditions: Papa Priority.* https://www.papajohns.com/papa-priority, Accessed 7/21/2021.

Pinedo ML (2012) *Scheduling: Theory, Algorithms, and Systems* (Springer Science & Business Media).

Rosenblum DM (1992) Allocation of waiting time by trading in position on a $G/M/s$ queue. *Operations Research* 40(3-supplement-2):S338–S342.

Texas Department of Transportation (2021) *TEXpress Lanes.* https://www.txdot.gov/driver/managed-lanes/texpress.html, Accessed 08/09/2021.

Transport for London (2021) *Congestion Charge/ULEZ Zone.* https://tfl.gov.uk/modes/driving/congestion-charge/congestion-charge-zone?intcmp=2055, Accessed 08/11/2021.

Transportation Security Administration (2021) *TSA PreCheck®.* https://www.tsa.gov/precheck, Accessed 08/16/2021.

US Department of State (2021a) *Passport Operations in Response to COVID-19.* https://travel.state.gov/content/travel/en/traveladvisories/ea/passport-covid-19.html, Accessed 08/16/2021.

US Department of State (2021b) *Renew my Passport.* https://travel.state.gov/content/travel/en/passports/have-passport/renew.html, Accessed 7/21/2021.

Wang J, Cui S, Wang Z (2019) Equilibrium strategies in $M/M/1$ priority queues with balking. *Production and Operations Management* 28(1):43–62.

Yang L, Debo L, Gupta V (2016) Trading time in a congested environment. *Management Science* 63(7):2377–2395.

Appendix. Proofs, Algorithms, and Laboratory Instructions

This appendix contains a section for each theorem. In each section, we first state and/or prove a necessary lemma, then prove the theorem. After these sections we give algorithms for computing the PBE thresholds in the base model and compensation model, followed by the laboratory instructions provided to the subjects in our experiments.

A. Proofs of Lemma 1 and Theorem 1

Proof of Lemma 1. We take a sample path approach. Consider a particular vector of realized waiting costs $(c_1,\ldots,c_N)$ and a focal customer i. Under the threshold strategy $\bar{x}_{j-1}(c_j)$, customer $j\in\{i+1,\ldots,N\}$ purchases priority if and only if $x_{j-1}\le\bar{x}_{j-1}(c_j)$. Given the fixed (but arbitrary) threshold strategies on the sample path, L_i^k is no longer random: we use ℓ_i^k to denote its realization corresponding to the realized waiting costs $(c_1,\ldots,c_N)$. Denote by x_j^k the length of the priority queue that is observed by customer $j+1$, given that $x_{i-1}=k$, customer i chooses the regular queue, and each customer $j\in\{i+1,\ldots,N\}$ uses the threshold strategy $\bar{x}_{j-1}(c_j)$. We proceed by cases.

Case 1: $\bar{x}_{j-1}(c_j)\neq x_{j-1}^k$ for all $j\in\{i+1,\ldots,N\}$. In this case, if customer i chooses the regular queue, then all customers $j\in\{i+1,\ldots,N\}$ will take the same actions whether $x_{i-1}=k$ or $x_{i-1}=k+1$. To see this, consider customer $i+1$. We have $x_i^{k+1}=x_i^k+1$. Because $x_i^k\neq\bar{x}_i(c_{i+1})$, we either have $\bar{x}_i(c_{i+1})\ge x_i^{k+1}=x_i^k+1>x_i^k$, or $\bar{x}_i(c_{i+1})<x_i^k<x_i^k+1=x_i^{k+1}$. Either way, customer $i+1$ will make the same decision with $x_i=x_{i-1}=k$ as with $x_i=x_{i-1}=k+1$, and by induction, so will customers $j\in\{i+2,\ldots,N\}$.

Hence, after choosing the regular queue, customer i will wait through the same number of services whether $x_{i-1}=k$ or $x_{i-1}=k+1$. Denoting by α the number of priority purchases among customers $j\in\{i+1,\ldots,N\}$, we then have

$$\ell_i^k=i+\alpha=\ell_i^{k+1}. \tag{6}$$

Case 2: $\bar{x}_{j-1}(c_j)=x_{j-1}^k$ for some $j\in\{i+1,\ldots,N\}$. In this case, define j' by

$$j':=\min\left\{j\in\{i+1,\ldots,N\}:\bar{x}_{j-1}(c_j)=x_{j-1}^k\right\}.$$

By the same argument as in Case 1, if customer i chooses the regular queue, then whether $x_{i-1}=k$ or $x_{i-1}=k+1$, customers $j\in\{i+1,\ldots,j'-1\}$ will take the same actions in either case because $\bar{x}_{j-1}(c_j)\neq x_{j-1}^k$ for all $j\in\{i+1,\ldots,j'-1\}$ (if $j'=i+1$, then this interval of customers is vacuous and thus trivially there is no difference in this empty set of customers between the cases with $x_{i-1}=k$ and $x_{i-1}=k+1$). By the definition of j', we have $x_{j'-1}^k=\bar{x}_{j'-1}(c_j)$, so if $x_{i-1}=k$, then customer j' will purchase priority because the priority queue length will be exactly at her threshold. Also, because all customers $j\in\{i+1,\ldots,j'-1\}$ take the same actions with $x_{i-1}=k+1$ as with $x_{i-1}=k$, we have $x_{j'-1}^{k+1}=x_{j'-1}^k+1=\bar{x}_{j'-1}(c_j)+1$. So, if $x_{i-1}=k+1$, then customer j' will *not* purchase priority because her threshold will be exceeded by one. Consequently, we have $x_{j'}^{k+1}=x_{j'-1}^k+1=x_{j'}^k$, meaning that if customer i chooses the regular queue, then $x_{j'}$ is the same whether $x_{i-1}=k$ or $x_{i-1}=k+1$.

Therefore, all customers $j \in \{j'+1,\ldots,N\}$ take the same action whether $x_{i-1}=k$ or $x_{i-1}=k+1$ because $x^k_{j'}=x^{k+1}_{j'}$ implies that $x^k_{j-1}=x^{k+1}_{j-1}$ for all $j \in \{j'+1,\ldots,N\}$ (if $j'=N$, then again this empty interval of customers has no effect on ℓ^k_i or ℓ^{k+1}_i). So, conditional on customer i choosing the regular queue, the total number of customers to purchase priority among customers $j \in \{i+1,\ldots,j'-1,j'+1,\ldots,N\}$ is the same whether $x_{i-1}=k$ or $x_{i-1}=k+1$. Denoting this number by β, we can write

$$\ell^k_i = i+\beta+1 = \ell^{k+1}_i + 1, \tag{7}$$

where the difference of 1 between ℓ^k_i and ℓ^{k+1}_i is due to customer j' purchasing priority if $x_{i-1}=k$ (because in this case $x_{j'-1}=\bar{x}_{j'-1}(c_{j'})$)—and accordingly being served before customer i—but choosing the regular queue if $x_{i-1}=k+1$ (because in this case $x_{j'-1}=\bar{x}_{j'-1}(c_{j'})+1$).

Equations (6) and (7) imply the bounds $0 \le \ell^k_i - \ell^{k+1}_i \le 1$. Taking expectation over the waiting costs (and by extension, over the other customers' cost-dependent thresholds) yields the lemma. □

Proof of Theorem 1. The proof is by a double induction. Consider a customer $i \in \{1,\ldots,N-1\}$, and suppose that all customers $j \in \{i+1,\ldots,N\}$ use some cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$. That is, customer j purchases priority if and only if $x_{j-1} \le \bar{x}_{j-1}(C_j)$.

For a given waiting-cost realization c_i, consider customer i's optimal strategy as a function of x_{i-1}. Given the cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$ for customers $j \in \{i+1,\ldots,N\}$, let $0 \le k \le i-1$ be the smallest integer such that, if $x_{i-1}=k$, then it is optimal for customer i to stay in the regular queue. We note that, by definition, it is optimal for customer i to purchase priority if $x_{i-1}<k$. If it is never optimal for customer i to choose the regular queue, then the optimal strategy for customer i is the threshold strategy $\bar{x}_{i-1}=i-1$, and by convention we say that $k=i$ in this case. Similarly, if $k=i-1$, then the optimal strategy for customer i is the threshold strategy $\bar{x}_{i-1}=i-2$. The remainder of the argument establishes that a threshold strategy is also optimal if $k \le i-2$.

As in Lemma 1, let L^k_i denote the random number of services (including her own) that customer i will wait through if $x_{i-1}=k$ and she chooses the regular queue; the exact value of L^k_i will depend on the thresholds used by the later customers, which in turn depend on their waiting-cost realizations. Let $\bar{\boldsymbol{x}}_{\boldsymbol{i,j-1}}$ be a vector of cost-dependent threshold strategies $(\bar{x}_i(C_{i+1}),\ldots,\bar{x}_{j-1}(C_j))$ for customers $i+1,\ldots,j$. For the case in which customers $i+1,\ldots,N$ use the cost-dependent threshold strategies $\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}}$, we represent customer i's net utilities from choosing the regular or priority queue by $U_{R,i}(x_{i-1};\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})$ and $U_{P,i}(x_{i-1})$, respectively. Note that the utility from the regular queue is a random variable because the strategies of the later customers depend on their realized waiting costs. Taking expectation over the remaining customers' waiting costs (the earlier customers' waiting costs are irrelevant because their decisions have already been observed), the assumption that the regular queue is optimal for customer i if $x_{i-1}=k$ implies

$$\mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V - c_i\mathsf{E}[L^k_i] > V - p - c_i(k+1) = U_{P,i}(k). \tag{8}$$

By Lemma 1, we then have

$$\begin{aligned}\mathsf{E}[U_{R,i}(k+1\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V - c_i\mathsf{E}[L_i^{k+1}] &\geq V - c_i\mathsf{E}[L_i^k] \\ &> V - p - c_i(k+1) \\ &> V - p - c_i(k+2) \\ &= U_{P,i}(k+1),\end{aligned} \tag{9}$$

where the inequality on the second line holds by equation (8). We conclude that if it is optimal for customer i to choose the regular queue when $x_{i-1} = k$, then it is also optimal for her to choose the regular queue when $x_{i-1} = k+1$, and therefore by induction for any $k \leq x_{i-1} \leq i-1$. Because by the definition of k it is optimal for customer i to join the priority queue if $x_{i-1} < k$, we conclude that customer i's optimal strategy for waiting-cost realization c_i is the threshold strategy $\bar{x}^*_{i-1}(c_i) = k-1$. The above derivation holds for any realization of the waiting cost, so the overall optimal strategy for customer i is a cost-dependent threshold strategy $\bar{x}^*_{i-1}(C_i)$.

The outer induction hypothesis is verified in equilibrium for customer $N-1$ by equation (1): customer N optimally uses the cost-dependent threshold strategy $\bar{x}^*_{N-1}(C_N) = \lfloor N-1-p/C_N \rfloor$. The above then implies that it is also optimal for customers $i \in \{1,\ldots,N-1\}$ to use cost-dependent threshold strategies. □

B. Supplementary Result and Proof for Theorem 2

LEMMA 3. *Consider a customer $i \in \{1,\ldots,N-1\}$, and suppose that each customer $j \in \{i+1,\ldots,N\}$ uses a cost-dependent threshold strategy $\bar{x}_{j-1}(C_j)$. Given these strategies, for $k \in \{0,\ldots,i-1\}$, let L_i^k (L_{i+1}^k) be the random variable for the number of services (including her own) that customer i ($i+1$) will wait through if $x_{i-1} = k$ ($x_i = k$) and customer i ($i+1$) chooses the regular queue. We have*

$$0 \leq \mathsf{E}[L_{i+1}^k] - \mathsf{E}[L_i^k] \leq 1.$$

Proof. Consider a particular vector of realized waiting costs $(c_1,\ldots,c_N)$, and again let ℓ_i^k (ℓ_{i+1}^k) denote the realization of L_i^k (L_{i+1}^k) for these waiting costs and the corresponding thresholds. If $x_{i-1} = k$ and at least one of customers i and $i+1$ chooses the regular queue, then we will have $x_{i+1} \in \{k, k+1\}$.

Case 1: $\bar{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{c}) < \boldsymbol{k}$. In this case, if $x_{i-1} = k$ and customer i chooses the regular queue, then we have $x_i = k$, and customer $i+1$ will not purchase priority because her threshold is exceeded. Let the number of priority purchases among customers $i+2,\ldots,N$ be denoted by α in this case. We have $\ell_i^k = i+\alpha$. If $x_i = k$, and if customer $i+1$ chooses the regular queue, then the number of priority purchases among customers $i+2,\ldots,N$ will also be α, so we have $\ell_{i+1}^k = i+1+\alpha$, and therefore

$$\ell_{i+1}^k = \ell_i^k + 1. \tag{10}$$

Case 2: $\bar{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{c}) \geq \boldsymbol{k}$. In this case, if $x_{i-1} = k$ and customer i chooses the regular queue, then we again have $x_i = k$, but now customer $i+1$'s strategy will prescribe priority for her because her threshold is at least k. Denote by $x_{j,m}^k$ the length of the priority queue that is observed by customer $j+1$, given that $x_m = k$, customer m chooses the regular queue, and each customer $j \in \{m+1,\ldots,N\}$ uses the threshold strategy

$\bar{x}_{j-1}(c_j)$. Suppose first that $\bar{x}_j(c_{j+1}) \neq x^k_{j,i+1}$ for all $j \in \{i+2,\ldots,N\}$. By arguments analogous to Case 1 of the proof of Lemma 1, in this case the number of priority purchases among customers $i+2,\ldots,N$ will be the same with $x_{i+1}=k$ and with $x_{i+1}=k+1$. Denoting this number by α, and for ℓ^k_{i+1} letting customer $i+1$ contemplate choosing the regular queue even though the strategy $\bar{x}_i(c_i)$ prescribes priority, we have

$$\ell^k_i = i+1+\alpha = \ell^k_{i+1}, \tag{11}$$

where $\ell^k_i = i+1+\alpha$ because customer i anticipates that customer $i+1$ will purchase priority, and then there will be an additional α priority purchases among customers $i+2,\ldots,N$.

If instead $\bar{x}_j(c_{j+1}) = x^k_{j,i+1}$ for at least one $j \in \{i+2,\ldots,N\}$, then an analogous argument to that in Case 2 of the proof of Lemma 1 implies that there will be one less priority purchase among customers $i+2,\ldots,N$ with $x_{i+1}=k+1$ than with $x_{i+1}=k$. Let these numbers be denoted $\alpha-1$ and α, respectively. We then have $\ell^k_i = i+1+(\alpha-1) = i+\alpha$ and $\ell^k_{i+1} = i+1+\alpha$, which implies

$$\ell^k_{i+1} = \ell^k_i + 1. \tag{12}$$

Combining equations (10), (11), and (12) gives

$$0 \leq \ell^k_{i+1} - \ell^k_i \leq 1,$$

and taking expectation over the waiting costs completes the proof. □

Proof of Theorem 2. For a given constant c, suppose that the equilibrium threshold for customer i is $\bar{x}^*_{i-1}(c) \geq k$, so if $x_{i-1}=k$, then in equilibrium customer i will purchase priority. We must then have

$$U_{P,i}(k) = V - p - c(k+1) \geq V - c\mathsf{E}[L^k_i] = \mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^*_{\boldsymbol{i,N-1}})]. \tag{13}$$

Lemma 3 and equation (13) then imply that

$$\begin{aligned} U_{P,i+1}(k) = V - p - c(k+1) &\geq V - c\mathsf{E}[L^k_i] \\ &\geq V - c\mathsf{E}[L^k_{i+1}] \\ &= \mathsf{E}[U_{R,i+1}(k\,;\,\bar{\boldsymbol{x}}^*_{\boldsymbol{i+1,N-1}})], \end{aligned} \tag{14}$$

where customer $i+1$'s comparisons are made assuming the same waiting-cost realization c. Thus, for a given k, if in equilibrium customer i purchases priority upon observing $x_{i-1}=k$, then customer $i+1$ must also purchase priority if she observes $x_i=k$. We conclude that customer $i+1$'s equilibrium threshold is at least as large as that for customer i, which in turn implies that $\bar{x}^*_i(c) \leq \bar{x}^*_j(c)$ for $i<j$.

Finally, consider a given customer i and two waiting-cost realizations c and c', with $c<c'$. Suppose that $\bar{x}^*_{i-1}(c) \geq k$. Upon observing $x_{i-1}=k$, then, customer i with waiting-cost realization c will purchase priority, which implies $\mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^*_{\boldsymbol{i,N-1}}\,;c)] - U_{P,i}(k;c) \leq 0$. We then have

$$\begin{aligned} \mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^*_{\boldsymbol{i,N-1}}\,;c')] - U_{P,i}(k;c') &= p - c'\big(\mathsf{E}[L^k_i] - (k+1)\big) \\ &< p - c\big(\mathsf{E}[L^k_i] - (k+1)\big) \\ &= \mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^*_{\boldsymbol{i,N-1}}\,;c)] - U_{P,i}(k;c) \\ &\leq 0. \end{aligned}$$

Therefore, for customer i, for any priority queue length such that with waiting cost c she will purchase priority, she will also purchase priority with waiting cost $c'>c$ for the same queue length. We conclude that the corresponding thresholds must satisfy $\bar{x}^*_{i-1}(c) < \bar{x}^*_{i-1}(c')$. □

C. Proofs of Lemma 2 and Theorem 3

First, it is important to note that Lemma 1 applies to the compensation model as well as the base model because it holds for any cost-dependent threshold strategies for customers $j \in \{i+1,\dots,N\}$, independent of how these thresholds were determined. Theorem 3 also depends on Lemma 2.

Proof of Lemma 2. Under the cost-dependent threshold strategies $\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}}$, let A^k denote the random number of priority purchases among customers $j \in \{i+1,\dots N\}$ if $x_{i-1} = k$ and customer i chooses the regular queue. By equation (3), we have

$$g_i^\gamma(k) = \gamma \frac{p(k+A^k)}{N-(k+A^k)} \quad \text{and} \quad g_i^\gamma(k+1) = \gamma \frac{p(k+1+A^{k+1})}{N-(k+1+A^{k+1})}.$$

Let α^k denote a realization of the random variable A^k for a given vector of realized waiting costs. It follows from the proof of Lemma 1, for any vector $(c_1,\dots,c_N)$ of waiting-cost realizations, we have $\alpha^{k+1} \geq \alpha^k - 1$, which implies

$$\gamma \frac{p(k+\alpha^k)}{N-(k+\alpha^k)} \leq \gamma \frac{p(k+1+\alpha^{k+1})}{N-(k+1+\alpha^{k+1})}.$$

Taking expectation over the waiting costs gives $\mathsf{E}[g_i^\gamma(k)] \leq \mathsf{E}[g_i^\gamma(k+1)]$, as desired. □

Proof of Theorem 3. The proof uses a similar approach to that of Theorem 1. Consider a customer $i \in \{1,\dots,N-1\}$, and suppose that all customers $j \in \{i+1,\dots,N\}$ use some cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$. Fix a waiting-cost realization c_i for customer i. Given the cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$ for customers $j \in \{i+1,\dots,N\}$, let $0 \leq k \leq i-1$ be the smallest integer such that, if $x_{i-1} = k$, then it is optimal for customer i to stay in the regular queue. We note that, by definition, it is optimal for customer i to purchase priority if $x_{i-1} < k$. The cases with $k = i$ and $k = i-1$ trivially imply a threshold strategy, as in the proof of Theorem 1. We proceed to the case with $k \leq i-2$.

As in Theorem 1, let L_i^k denote the random number of services (including her own) that customer i will wait through if $x_{i-1} = k$ and she chooses the regular queue. Taking expectation over the remaining customers' waiting costs, the assumption that the regular queue is optimal for customer i if $x_{i-1} = k$ implies

$$\mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V + \mathsf{E}[g_i^\gamma(k)] - c_i\mathsf{E}[L_i^k] > V - p - c_i(k+1) = U_{P,i}(k). \tag{15}$$

Lemmas 1 and 2 then imply

$$\begin{aligned} \mathsf{E}[U_{R,i}(k+1\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V + \mathsf{E}[g_i^\gamma(k+1)] - c_i\mathsf{E}[L_i^{k+1}] &\geq V + \mathsf{E}[g_i^\gamma(k)] - c_i\mathsf{E}[L_i^k] \\ &> V - p - c_i(k+1) \\ &> V - p - c_i(k+2) \\ &= U_{P,i}(k+1), \end{aligned} \tag{16}$$

where the inequality on the second line holds by equation (15).

The same logic as in Theorem 1—with the outer induction hypothesis verified for $i = N-1$ by equation (4)—then implies that it is optimal for all customers to use cost-dependent threshold strategies. □

D. Supplementary Result and Proof for Theorem 4

As with Lemma 1, we note that Lemma 3 also applies to the compensation model because it does not depend on how the threshold strategies are determined. We also need an additional lemma for Theorem 4.

LEMMA 4. *Consider a customer $i \in \{1,\dots,N-1\}$, and suppose that each customer $j \in \{i+1,\dots,N\}$ uses a cost-dependent threshold strategy $\bar{x}_{j-1}(C_j)$. Under these strategies for the other customers, let A_i^k (A_{i+1}^k) denote the random number of priority purchases among customers $j \in \{i+2,\dots N\}$ if customer i ($i+1$) observes $x_{i-1}=k$ ($x_i=k$) and chooses the regular queue. Also, let $g_i^\gamma(k)$ ($g_{i+1}^\gamma(k)$) be the compensation that customer i ($i+1$) receives by choosing the regular queue after observing $x_{i-1}=k$ ($x_i=k$), for compensation fraction γ. For $k \in \{0,\dots,i-1\}$, we have*

$$\mathsf{E}[g_{i+1}^\gamma(k)] \le \mathsf{E}[g_i^\gamma(k)].$$

Proof. Let α_i^k (α_{i+1}^k) denote a realization of the random variable A_i^k (A_{i+1}^k) for a given vector of realized waiting costs (and note that we are using A_i^k and A_{i+1}^k to both cover the same customers $i+2,\dots,N$, different from A^k in the proof of Lemma 2). From the proof of Lemma 3, for any vector $(c_1,\dots,c_N)$ of waiting-cost realizations, we have $\alpha_i^k \ge \alpha_{i+1}^k - 1$.

Case 1: $\bar{x}_i(c_{i+1}) < k$. In this case, customer $i+1$ will not purchase priority if $x_i=k$, we will have $\alpha_i^k = \alpha_{i+1}^k$, and both customers will receive the same compensation in the respective scenario, i.e., we have

$$g_{i+1}^\gamma(k) = \gamma\frac{p(k+\alpha_{i+1}^k)}{N-(k+\alpha_{i+1}^k)} = \gamma\frac{p(k+\alpha_i^k)}{N-(k+\alpha_i^k)} = g_i^\gamma(k).$$

Case 2: $\bar{x}_i(c_{i+1}) \ge k$. In this case, customer $i+1$ will purchase priority upon observing $x_i=k$. By arguments in the proof of Lemma 3, we will either have $\alpha_{i+1}^k = \alpha_i^k$, or $\alpha_{i+1}^k = \alpha_i^k + 1$. If $\alpha_i^k = \alpha_{i+1}^k$, then because customer $i+1$'s strategy prescribes priority if $x_i=k$, customer i will receive one more customer's worth of compensation from choosing regular with $x_{i-1}=k$ than would customer $i+1$ from choosing regular with $x_i=k$, so we have

$$g_{i+1}^\gamma(k) = \frac{p(k+\alpha_{i+1}^k)}{N-(k+\alpha_{i+1}^k)} < \frac{p(k+1+\alpha_{i+1}^k)}{N-(k+1+\alpha_{i+1}^k)} = \frac{p(k+1+\alpha_i^k)}{N-(k+1+\alpha_i^k)} = g_i^\gamma(k).$$

If instead $\alpha_{i+1}^k = \alpha_i^k + 1$, then we have

$$g_{i+1}^\gamma(k) = \frac{p(k+\alpha_{i+1}^k)}{N-(k+\alpha_{i+1}^k)} = \frac{p(k+1+\alpha_i^k)}{N-(k+1+\alpha_i^k)} = g_i^\gamma(k),$$

where the last equality holds because for customer i's calculations, customer $i+1$ will purchase priority if $x_i=k$ by the assumption of this case, so after customer i there will be α_i^k+1 priority purchases in total.

We conclude that for any waiting-cost realizations and their corresponding thresholds, we have $g_{i+1}^\gamma(k) \le g_i^\gamma(k)$. Taking expectation over the waiting costs completes the proof. □

Proof of Theorem 4. For a given constant c and compensation fraction γ, suppose that the equilibrium threshold for customer $i+1$ is $\bar{x}_i^*(c) < k \le i-1$, so if $x_i=k$, then in equilibrium customer $i+1$ will *not* purchase priority. We must then have

$$U_{P,i+1}(k) = V - p - c(k+1) < V + \mathsf{E}[g_{i+1}^\gamma(k)] - c\mathsf{E}[L_{i+1}^k] = \mathsf{E}[U_{R,i+1}(k\,;\,\bar{\boldsymbol{x}}^*_{\boldsymbol{i+1,N-1}})]. \tag{17}$$

Lemmas 3 and 4 and equation (17) then imply that

$$\begin{aligned} U_{P,i}(k) = V - p - c(k+1) &< V + \mathsf{E}[g_{i+1}^{\gamma}(k)] - c\mathsf{E}[L_{i+1}^{k}] \\ &\leq V + \mathsf{E}[g_{i}^{\gamma}(k)] - c\mathsf{E}[L_{i}^{k}] \\ &= \mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^{*}_{\boldsymbol{i,N-1}})], \end{aligned}$$

where customer $i+1$'s comparisons are made assuming the same waiting-cost realization c. Thus, if in equilibrium customer $i+1$ chooses the regular queue upon observing $x_i = k$, then it must also be that customer i chooses the regular queue if she observes $x_{i-1} = k$. Put another way, there does not exist a queue length k such that customer i will purchase priority if $x_{i-1} = k$ but customer $i+1$ will choose the regular queue if $x_i = k$, for the same waiting-cost realization. We conclude that customer $i+1$'s equilibrium threshold is at least as large as that for customer i, which in turn implies that $\bar{x}_i^*(c) \leq \bar{x}_j^*(c)$ for $i < j$.

Finally, consider a given customer i and two waiting-cost realizations c and c', with $c < c'$. Suppose that $\bar{x}_i^*(c) \geq k$. Upon observing $x_{i-1} = k$, then, customer i with waiting-cost realization c will purchase priority, which implies $\mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^{*}_{\boldsymbol{i,N-1}}\,;c)] - U_{P,i}(k;c) \leq 0$. We then have

$$\begin{aligned} \mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^{*}_{\boldsymbol{i,N-1}}\,;c')] - U_{P,i}(k;c') &= p + \mathsf{E}[g_{i}^{\gamma}(k)] - c'\big(\mathsf{E}[L_{i}^{k}] - (k+1)\big) \\ &< p + \mathsf{E}[g_{i}^{\gamma}(k)] - c\big(\mathsf{E}[L_{i}^{k}] - (k+1)\big) \\ &= \mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}^{*}_{\boldsymbol{i,N-1}}\,;c)] - U_{P,i}(k;c) \\ &\leq 0. \end{aligned}$$

Therefore, for customer i, for any priority queue length such that with waiting cost c she will purchase priority, she will also purchase priority with waiting cost $c' > c$ for the same queue length. We conclude that the equilibrium threshold functions must satisfy $\bar{x}_{i-1}^*(c) < \bar{x}_{i-1}^*(c')$. □

E. Supplementary Result and Proof for Theorem 5.

LEMMA 5. *For customers $i+1,\ldots,N$, consider two vectors of cost-dependent threshold functions, $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}$ and $\bar{\boldsymbol{x}}'_{\boldsymbol{i,N}}$, with elements $\bar{x}_j(C_j)$ and $\bar{x}'_j(C_j)$, respectively. For a sample path of realizations $(c_{i+1},\ldots,c_N)$, suppose that $\bar{x}'_j(c_j) \leq \bar{x}_j(c_j)$ for all $j \in \{i+1,\ldots,N\}$. Let α_i^k ($\tilde{\alpha}_i^k$) be the number of priority purchases among customers $i+1,\ldots,N$ if $x_{i-1} = k$, customer i chooses the regular queue, and the thresholds are $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}(c_{i+1},\ldots,c_N)$ ($\bar{\boldsymbol{x}}'_{\boldsymbol{i,N}}(c_{i+1},\ldots,c_N)$). For $k \in \{0,\ldots,i-1\}$, we have*

$$\tilde{\alpha}_i^k \leq \alpha_i^k.$$

Proof. Consider the thresholds $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}(c_{i+1},\ldots,c_N)$. Let x_{j-1}^k be the priority queue length observed by customer j if $x_{i-1} = k$ and customer i chooses the regular queue, given these thresholds. For some $j' \in \{i+1,\ldots,N\}$, consider also the vector of thresholds obtained from $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}(c_{i+1},\ldots,c_N)$ by reducing by 1 the threshold of customer j', from $\bar{x}_{j'-1}(c_{j'})$ to $\bar{x}_{j'-1}(c_{j'}) - 1$ (the other thresholds are the same as in the original vector). Under these modified thresholds, let $x_{j-1}^{k(-)}$ be the priority queue length observed by customer j if $x_{i-1} = k$ and customer i chooses the regular queue.

For customers $i+1,\ldots,j'-1$, we have $x_{j-1}^k = x_{j-1}^{k(-)}$, so these customers will take the same actions either way, and there will be the same number of priority purchases among these customers for either vector of thresholds. We thus have $x_{j'-1}^k = x_{j'-1}^{k(-)}$.

For customer j', then, if $x^{k(-)}_{j'-1} = x^k_{j'-1} \neq \bar{x}_{j'-1}(c_{j'})$, then either $\bar{x}_{j'-1}(c_{j'}) - 1 < \bar{x}_{j'-1}(c_{j'}) < x^k_{j'-1} = x^{k(-)}_{j'-1}$, or $x^{k(-)}_{j'-1} = x^k_{j'-1} \leq \bar{x}_{j'-1}(c_{j'}) - 1 < \bar{x}_{j'-1}(c_{j'})$. Customer i takes the same action in either case, meaning that she takes the same action for either vector of thresholds. In this case, we will also have $x^k_{j-1} = x^{k(-)}_{j-1}$ for $j \in \{j'+1, \ldots, N\}$, so these customers also will take the same actions under either vector of thresholds. Thus, we have

$$\alpha^k_i = \tilde{\alpha}^k_i. \tag{18}$$

If instead $x^{k(-)}_{j'-1} = x^k_{j'-1} = \bar{x}_{j'-1}(c_{j'})$, then customer j' purchases priority with her original threshold $\bar{x}_{j'-1}(c_{j'})$, but not with her modified threshold $\bar{x}_{j'-1}(c_{j'}) - 1$. There are two cases.

Case 1: $x^{k(-)}_{j-1} \neq \bar{x}_{j-1}(c_j)$ for all $j \in \{j'+1, \ldots, N\}$. In this case, because of customer i's different action, we have $x^k_{j'} = x^{k(-)}_{j'} + 1$. So, similar to the above for customer j', by the hypothesis of this case, for customer $j'+1$, we either have $x^{k(-)}_{j'} < x^k_{j'} = x^{k(-)}_{j'} + 1 \leq \bar{x}_{j'}(c_{j'+1})$, or $\bar{x}_{j'}(c_{j'+1}) < x^{k(-)}_{j'} < x^{k(-)}_{j'} + 1 = x^k_{j'}$. Hence, customer $j'+1$ will take the same action under both the original threshold vector and that with the threshold for customer j' decreased by 1. By induction, we then have $x^k_{j-1} = x^{k(-)}_{j-1} + 1$ for all $j \in \{j'+2, \ldots, N\}$. Therefore, customers $j'+2, \ldots, N$ will also take the same actions under either vector by the same reasoning as for customer $j'+1$. In total, then, there is one less priority purchase among customers $j \in \{i+1, \ldots, N\}$ when customer j' has a decreased threshold, so we have

$$\alpha^k_i = \tilde{\alpha}^k_i + 1. \tag{19}$$

Case 2: $x^{k(-)}_{j''-1} = \bar{x}_{j''-1}(c_{j''})$ for some $j'' \in \{j'+1, \ldots, N\}$. We have $x^k_{j-1} = x^{k(-)}_{j-1} + 1$ for $j \in \{j'+1, \ldots, j''\}$ by the same reasoning as in Case 1 because of customer i's different actions under the two threshold vectors. Customers $j \in \{j'+1, \ldots, j''-1\}$ will thus take the same actions under either the original or the modified threshold vectors, also by arguments in Case 1. For customer j'', we have $\bar{x}_{j''-1}(c_{j''}) = x^{k(-)}_{j''-1} < x^k_{j''-1}$, so customer j'' will purchase priority for the modified threshold vector (when customer j' has her threshold reduced by 1), but not for the original vector. Summarizing, other than customers j' and j'', all customers $j \in \{i+1, \ldots, N\}$ will take the same action under either threshold vector. Under the original vector, customer j' will purchase priority but customer j'' will choose the regular queue, while under the modified vector, customer j' will choose the regular queue but customer j'' will purchase priority. In either case, there is exactly one priority purchase among these two customers (and no change at all for the other customers), so we conclude that in this case

$$\alpha^k_i = \tilde{\alpha}^k_i. \tag{20}$$

Combining equations (18), (19), and (20) gives $\tilde{\alpha}^k_i \leq \alpha^k_i$. By induction, we can successively reduce the thresholds customer by customer and in increments of 1 until we reach $\bar{\boldsymbol{x}}'_{i+1,N}(c_{i+1}, \ldots, c_N)$. Because $\tilde{\alpha}^k_i \leq \alpha^k_i$ at every step of this process, we have the desired result. □

Proof of Theorem 5. Consider a customer $i \in \{1, \ldots, N-1\}$, and suppose that $\bar{x}^*_{j-1,\gamma}(c_j) \leq \bar{x}^*_{j-1}(c_j)$ for $j \in \{i+1, \ldots, N\}$ and all c_j in the support of C_j. For customers $i+1, \ldots, N$, consider a given sample path of waiting costs $(c_{i+1}, \ldots, c_N)$. In the base model (compensation model with compensation fraction γ), let α^k_i ($\alpha^k_{i,\gamma}$) be the number of priority purchases among customers $i+1, \ldots, N$, under the equilibrium thresholds

for the waiting-cost sample path $(c_{i+1},\dots,c_N)$ if $x_{i-1}=k$ and customer i chooses the regular queue. For $k\in\{0,\dots,i-1\}$, Lemma 5 and our hypothesis that $\bar{x}^*_{j-1,\gamma}(c_j)\le\bar{x}^*_{j-1}(c_j)$ together imply that $\alpha^k_{i,\gamma}\le\alpha^k_i$, i.e., the number of priority purchases after customer i will be weakly less with compensation than without. Let A^k_i ($A^k_{i,\gamma}$) be the random variable for the number of priority purchases after customer i in the base model (compensation model). Because $\alpha^k_{i,\gamma}\le\alpha^k_i$ on every sample path, taking expectation over the waiting costs yields

$$\mathsf{E}[A^k_{i,\gamma}]\le\mathsf{E}[A^k_i]. \tag{21}$$

Moreover, since the number of services L^k_i ($L^k_{i,\gamma}$) that customer i must wait through if $x_{i-1}=k$ and she chooses the regular queue in the base model (compensation model) is equal to i plus the number of priority purchases after her, equation (21) also implies

$$\mathsf{E}[L^k_{i,\gamma}]=i+\mathsf{E}[A^k_{i,\gamma}]\le i+\mathsf{E}[A^k_i]=\mathsf{E}[L^k_{i,\gamma}]. \tag{22}$$

Let $U_{P,i}(x_{i-1};C_i)$ ($U_{P,i,\gamma}(x_{i-1};C_i)$) be the utility from purchasing priority in the base model (compensation model), and similarly $U_{R,i}(x_{i-1};C_i)$ ($U_{R,i,\gamma}(x_{i-1};C_i)$) for the utility from the regular queue. Because priority customers are not compensated even in the compensation model, we have $U_{P,i,\gamma}(x_{i-1};C_i)=U_{P,i}(x_{i-1};C_i)$. Suppose that for waiting-cost realization c_i, if $x_{i-1}=k$, then in equilibrium in the base model, customer i chooses the regular queue. In this case, we must have $U_{P,i}(x_{i-1};c_i)<\mathsf{E}[U_{R,i}(x_{i-1};c_i)]$. In the compensation model, customer i's compensation in the regular queue is $g^\gamma_i(k)$, which is random but nonnegative. We have

$$\begin{aligned}U_{P,i,\gamma}(x_{i-1};c_i)&=U_{P,i}(x_{i-1};c_i)\\&<\mathsf{E}[U_{R,i}(x_{i-1};c_i)]\\&=V-c_i\mathsf{E}[L^k_i]\\&\le V-c_i\mathsf{E}[L^k_{i,\gamma}]+\mathsf{E}[g^\gamma_i(k)]=\mathsf{E}[U_{R,i,\gamma}(x_{i-1};c_i)].\end{aligned}$$

Therefore, for any priority queue length k such that customer i will choose the regular queue in the base model, she will also choose the regular queue in the compensation model with the same waiting-cost realization, under any compensation fraction $0<\gamma\le1$. Under our hypothesis that $\bar{x}^*_{j-1,\gamma}(c_j)\le\bar{x}^*_{j-1}(c_j)$ for customers $j\in\{i+1,\dots,N\}$ and all c_j in the support of C_j, this implies that also $\bar{x}^*_{i-1,\gamma}(c_i)\le\bar{x}^*_{i-1}(c_i)$ for customer i and all c_i in the support of C_i. For $\gamma=1$, the induction hypothesis is verified for $i=N-1$ by comparing equations (1) and (4) under our assumption that $p\le\underline{c}(N-1)$. For $\gamma<1$, the comparison requires some algebra, but it follows by the same assumption, completing the proof of the first part of the theorem.

The second part of the theorem, that a customer's threshold decreases in the compensation fraction for fixed strategies of the customers after her, follows by a related but simpler argument, which we merely sketch here for brevity. For fixed strategies of the later customers and two compensation fractions $\gamma<\gamma'$, we have $L^k_{i,\gamma}=L^k_{i,\gamma'}$. We also have $g^\gamma_i(k)\le g^{\gamma'}_i(k)$. These two relations make the regular queue more attractive as the compensation fraction increases, so the optimal threshold decreases in the compensation fraction. □

F. Algorithms to Compute PBE Threshold Functions

Here, we give algorithms to calculate the PBE threshold functions for an arbitrary continuous waiting-cost distribution in both models. The analogous algorithms for discrete distributions are obtained in the natural way. The conditions in the indicator functions in the last lines of both algorithms are equivalent to $U_{P,i}(k) \leq \mathsf{E}[U_{R,i}(k)]$ under the respective models. Finally, note that for fixed thresholds, L_i^k is deterministic and can be calculated easily by iteratively by recording the decisions prescribed for each customer given their thresholds and determining the number of priority purchases after customer i. For each customer i, the resulting threshold function is an increasing step function in the waiting-cost realization c_i.

Algorithm 1: Compute PBE cost-dependent thresholds for base model

Result: Vector $\bar{\boldsymbol{x}}^*$ of threshold functions

for $i = N, N-1, \ldots, 1$ **do**
  for $(\bar{x}_i^m, \ldots, \bar{x}_{N-1}^m) \in \left\{ \times_{j=i}^{N-1} \{-1, 0, 1, \ldots, j\} \right\}$ **do**
    $\pi_m \leftarrow \prod_{k=i+1}^{N} \int \mathbf{1}\{\bar{x}_{k-1}^*(c) = \bar{x}_{k-1}^m\} \mathrm{d}F(c)$ `// PBE probability of threshold vector` m
  end
  $\bar{x}_{i-1}^*(c_i) \leftarrow -1$ for c_i in support of C_i;
  for $k \in \{0, 1, \ldots, i-1\}$ **do**
    $\lambda_i^k \leftarrow \sum_m \pi_m L_i^k((\bar{x}_i^m, \ldots, \bar{x}_{N-1}^m))$ `// Expected services to wait through`
    $\bar{x}_{i-1}^*(c_i) \leftarrow \bar{x}_{i-1}^*(c_i) + \mathbf{1}\{c_i \geq p/(\lambda_i^k - (k+1))\}$ for c_i in support of C_i `// If priority is preferred at current` k`, increment previous threshold`
  end
end

Algorithm 2: Compute PBE cost-dependent thresholds for compensation model

Result: Vector $\bar{\boldsymbol{x}}^*$ of threshold functions

for $i = N, N-1, \ldots, 1$ **do**
  for $(\bar{x}_i^m, \ldots, \bar{x}_{N-1}^m) \in \left\{ \times_{j=i}^{N-1} \{-1, 0, 1, \ldots, j\} \right\}$ **do**
    $\pi_m \leftarrow \prod_{k=i+1}^{N} \int \mathbf{1}\{\bar{x}_{k-1}^*(c) = \bar{x}_{k-1}^m\} \mathrm{d}F(c)$ `// PBE probability of threshold vector` m
  end
  $\bar{x}_{i-1}^*(c_i) \leftarrow -1$ for c_i in support of C_i;
  for $k \in \{0, 1, \ldots, i-1\}$ **do**
    $\lambda_i^k \leftarrow \sum_m \pi_m L_i^k((\bar{x}_i^m, \ldots, \bar{x}_{N-1}^m))$ `// Expected services to wait through`
    $\rho_i^k \leftarrow \sum_m \pi_m g_i^\gamma(k)$ `// Expected compensation`
    $\bar{x}_{i-1}^*(c_i) \leftarrow \bar{x}_{i-1}^*(c_i) + \mathbf{1}\{c_i \geq (p + \rho_i^k)/(\lambda_i^k - (k+1))\}$ for c_i in support of C_i `// If priority is preferred at current` k`, increment previous threshold`
  end
end

G. Laboratory Instructions for Base Model (Service Provider Keeps Revenue)

Instructions

You are about to participate in an experiment in the economics of decision-making. If you follow these instructions carefully and make good decisions, you will earn money that will be paid to you in cash at the end of the session. If you have a question at any time, please raise your hand and the experimenter will answer it. We ask you not to talk with one another for the duration of the experiment.

Overview of the Game
You are in the role of a customer waiting to receive a service. When you entered the room you were given a slip of paper with a sequence number. The sequence numbers were generated randomly. There are 20 people in the room. Each person will start out with $30, called your endowment. Each person will be called in the order of his or her sequence number and will be asked to make a decision to either join the **Regular** Queue or to purchase a spot in the **Priority** Queue. The priority queue costs $6. The regular queue is free. After both queues have been formed, the virtual service will start. Each service will take approximately 2 minutes. The service will be performed for priority queue customers first, followed by the regular queue customers. Within each queue the service will be performed in the order of your sequence number. Each service that you wait through (including your own) costs $1. After your service has completed you will be paid your total earnings, calculated as follows.

$5 participation fee + $30 endowment - $6 if you chose priority queue (if you purchased Priority) - $1 x (the number of services you waited).

Example:
Suppose your sequence number is 7. When your turn to make the decision comes, you observe that there are 6 people in front of you, 3 in the priority queue and 3 in the regular queue. Suppose you decided to join the regular queue. Suppose that after that, of the 13 remaining people behind you, 3 joined the priority queue. This means that once the service starts, there will be 3+3 = 6 people in the priority queue, and 3 people in the regular queue in front of you. This means that you will wait for 6+3+1 = 10 services. Your waiting cost will be $10. Your total earnings will be:
$5 + $30 - $10 = $25

Now suppose that you chose to pay $6 and join priority queue. In this case your waiting cost will be 3+1 = $4 because you only have to wait for the 3 Priority people in front of you, and for your own service. Your total earnings will be:
$5 + $30 -$6 - $4 = $25

How you will be paid
As soon as your service is completed, you will be paid your earnings in cash and in private. You will remain in the room until everyone has been served. Everybody will leave the session at the same time.

H. Laboratory Instructions for Compensation Case

Instructions

You are about to participate in an experiment in the economics of decision-making. If you follow these instructions carefully and make good decisions, you will earn money that will be paid to you in cash at the end of the session. If you have a question at any time, please raise your hand and the experimenter will answer it. We ask you not to talk with one another for the duration of the experiment.

Overview of the Game
You are in the role of a customer waiting to receive a service. When you entered the room you were given a slip of paper with a sequence number. The sequence numbers were generated randomly. There are 20 people in the room. Each person will start out with $30, called your endowment. Each person will be called in the order of his or her sequence number and will be asked to make a decision to either join the **Regular** Queue or to purchase a spot in the **Priority** Queue. The priority queue costs $6. Priority Queue fees that have been collected will be added up and equally divided and paid to the Regular Queue customers. We will call this amount Compensation.

After both queues have been formed, the virtual service will start. Each service will take approximately 2 minutes. The service will be performed for priority queue customers first, followed by the regular queue customers. Within each queue the service will be performed in the order of your sequence number. Each service that you wait through (including your own) costs $1. After your service has completed you will be paid your total earnings, calculated as follows.

$5 participation fee + $30 endowment - $6 (if you purchased Priority) - $1 x (the number of services you waited) + Compensation (if you did not purchase Priority).

Example:
Suppose your sequence number is 7. When your turn to make the decision comes, you observe that there are 6 people in front of you, 3 in the priority queue and 3 in the regular queue. Suppose you decided to join the regular queue. Suppose that after that, of the 13 remaining people behind you, 2 joined the priority queue. This means that once the service starts, there will be 2+3 = 5 people in the priority queue and 15 people (including you) in the regular queue. Out of those 15 people in the regular queue, 3 are in front of you. This means that you will wait for 5+3+1 = 9 services. Your waiting cost will be $9. Your Compensation will be ($6 x 5)/15 = $2. Your total earnings will be: $5 + $30 - $9 + $2 = $28.

Now suppose that you chose to pay $6 and join priority queue. In this case your waiting cost will be 3+1 = $4 because you only have to wait for the 3 Priority people in front of you, and for your own service. Your total earnings will be:
$5 + $30 -$6 - $4 = $25

How you will be paid
As soon as your service is completed, you will be paid your earnings in cash and in private. You will remain in the room until everyone has been served. Everybody will leave the session at the same time.