Paid Priority in Service Systems: Theory and Experiments

January 2024

Forthcoming in *Manufacturing & Service Operations Management*
https://doi.org/10.1287/msom.2021.0387

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

Problem Definition: Motivated by the prevalence of paid priority programs in practice, we study a service provider operating a system in which customers have random waiting costs and choose between two queues: regular (no cost) or priority (for a fee). We also consider a mechanism by which the provider redistributes a portion of priority revenue to compensate regular-queue customers for their longer waits. **Methodology/Results:** To determine the waiting-cost-dependent equilibrium priority purchasing strategies, we establish structural results at a sample-path level and prove that they generalize. In models both with and without compensation, the equilibrium exhibits a cost-dependent, increasing-threshold structure. We also prove that compensation entails fewer priority purchases because compensating regular-queue customers makes priority less attractive. We then analyze system-wide performance. Despite the fewer priority purchases, for a fixed (low) priority fee, compensation can actually reduce equilibrium aggregate waiting cost by filtering low-waiting-cost customers out of the priority queue; however, this finding does not hold when comparing at the optimal fees. We then test our models in the laboratory. Key behavioral regularities are that low-cost subjects are over-represented (under-represented) in the priority (regular) queue compared to equilibrium, and subjects with low and high waiting costs tend to overbuy priority at high fees. **Managerial Implications:** Our theoretical and behavioral results guide service providers in managing priority service systems. First, we find that compensation does not provide short-term performance benefits. Second, our experiments reveal that sub-optimal customer decisions partially prevent efficient reordering of customers by waiting cost, leading to higher aggregate waiting cost than the equilibrium predicts, but still lower than under first-come, first-serve service. Finally, because customers tolerate higher fees than they should, a revenue-maximizing provider can set a higher priority fee and extract more revenue than it could if customers acted rationally.

Key words: priority queues, behavioral operations, strategic queueing, behavioral queueing

1. Introduction

In November 2021, Killington Ski Resorts in Vermont introduced Fast Tracks, a priority program allowing time-sensitive skiers to buy access to an "express" line for the chair lift to the top of the slopes (Killington 2021). The resort even advertises the option to purchase the access "on-mountain," e.g., after heading to the slopes only to see a long queue. Similarly, in late 2021, Walt Disney World in Florida and Disneyland in California rolled out a new paid priority program (Maehrer 2021, Lynch 2021). If a guest observes many people in line for a ride—or a high number for the posted waiting time—she can use the Disney mobile app to purchase Lightning Lane access. Guests can purchase a day pass called Genie+ to access the Lightning Lanes at multiple lower-tier attractions, but some of the most popular rides of all, such as Star Wars: Rise of the Resistance at Disneyland Park, are not included in the Genie+ pass. Instead, they are designated as Individual Lightning Lane, meaning

payment is required each time a guest wishes to access the priority queue.[1] Universal Orlando Resort offers a similar experience called Universal Express at its Islands of Adventure and Universal Studios Florida theme parks, where guests can pay to use a "separate line with a shorter wait time;" like Fast Tracks and Lightning Lane, this access can be purchased "in-park" (Universal Orlando Resort 2023).

Paid priority queues have come under fire for promoting "money-talks culture" (Baggini 2017) and "segmenting society" (Baraniuk 2019). Indeed, with Individual Lightning Lane purchases costing up to \$20 each,[2] a family of four may find itself paying \$80 for priority access to a single ride, on top of the more than \$100 (on some dates more than \$150) per person for one-day access to the park.[3] Given the soaring costs of all aspects of a Disney vacation, from lodging to food to admission (Dumas and Rumpf 2022), some see the priority program as exacerbating an already serious affordability problem for those outside the upper classes who wish to visit. Following the program's announcement, angry customers decried it as blatant corporate greed (Maehrer 2021).[4] Matters were arguably even worse for POWDR, the parent of Killington Ski Resorts, after it announced Fast Tracks. As Allon (2021) points out when discussing Fast Tracks, allowing the wealthy to pay for a priority queue for a natural resource like a ski slope (or a public service like border control processing—see Le Seur 2012) is a particularly questionable practice. Sure enough, before the program even began operation, POWDR faced such intense resistance (see, e.g., Waite 2021) that its leaders were forced to issue a community letter to "clarify" the plans and attempt to convince customers that their experience would not be spoiled by longer waits if they did not purchase priority (Martin and Sibley 2021); it even offered customers refunds if desired.

These and related examples—e.g., the paid priority line for elevators to the SkyDeck viewing area at Willis Tower (formerly Sears Tower) in Chicago (SkyDeck 2023), as well as similar programs at 30 Rockefeller Plaza in New York City (Rockefeller Center 2023) and the CN Tower in Toronto (CN Tower 2023)—illustrate the growing prevalence of paid priority programs in service systems, as well as the substantial controversy they can generate. For service providers, on one hand, a paid priority queue enables a self-directed reorganization of customers such that those with higher waiting cost are served earlier, decreasing the aggregate waiting cost. On the other hand, a priority system elicits negative reactions related to fairness, income inequality, and the "right" to a position in the queue.

[1] https://disneyland.disney.go.com/genie/lightning-lane/ (accessed 06/13/2023)

[2] ibid

[3] https://disneyland.disney.go.com/admission/tickets/dates/ (accessed 06/13/2023)

[4] Example comments on Disney's blog post (Maehrer 2021) announcing the program: "Not a fan for this at Disney World where FastPass+ was FREE before and no silly money-grab upcharge for some premium attractions!" (User Ken); "[the new program] just let's [sic] us see that more and more they are aimed at the money and the data and not looking out for the experience of the regular class kids... you mean to tell me that only the richer class can have... a magical day with their kids...?" (User Olga)

To avoid the aforementioned backlash, service providers may wish to implement a different type of priority system that is less likely to upset customers. In a related setting that involves waiting times but does not study priority mechanisms or directly address queueing phenomena, Cohen et al. (2022) has established a significant benefit to service providers from proactively compensating customers who experience long waits. They conducted randomized field experiments with a ride-hailing platform in which some customers who experienced an excessive wait for their driver or an excessive travel time were compensated with a voucher for future service. Customers who received vouchers were found to spend significantly more money with the platform, with the average additional spending exceeding the value of the voucher. Such findings demonstrate that it can be in a service provider's interest to compensate inconvenienced customers.

Effectively managing a priority service system like those operated by Disney, POWDR, and other service providers demands a clear understanding of customer decisions in such a system and their system-wide operational implications (for, e.g., revenue and aggregate waiting cost). In the present work, we achieve this understanding through complementary theoretical and behavioral analyses. Moreover, in light of the successful implementation of customer compensation for long waits described above, we also examine the operational impact of compensation in priority service systems.

For a unified treatment, we present a stylized theoretical model of a service system with a setup that can be faithfully replicated in the laboratory. We consider two variations: the *base model* and the *compensation model*. The base model is a standard priority system: customers with heterogeneous waiting costs sequentially decide whether to pay a flat fee to the service provider to obtain priority, or to join the regular queue for free. The compensation model operates similarly with one key difference: rather than the priority proceeds going to the service provider, instead a fraction is divided among the customers in the regular queue. Hence, the regular-queue customers are *compensated* for being overtaken by the priority-queue customers via a portion of the latter's payments.

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 sequential setup, the random waiting costs significantly complicate the equilibrium analysis for both the base and compensation models. 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.

To overcome the combinatorial challenge described above, we analyze individual sample paths. In the base model, we show 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. 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—into a still challenging but tractable one with unknown but fixed threshold strategies for the other customers.

We apply our sample-path results to prove that all customers use *cost-dependent threshold strategies* in equilibrium, i.e., they purchase priority if the priority queue is below a threshold that depends on their waiting cost. With a shorter priority queue, a customer will overtake more customers by purchasing priority. But for a long priority queue, there are few regular-queue customers to overtake, and above a waiting-cost-dependent threshold, the time savings is not worth the priority fee. Even for the same waiting cost, the thresholds differ within the sequence. Later customers will overtake more regular-queue customers for a given priority queue length, and a similar sample-path argument shows that their thresholds are higher: the second customer will have a higher threshold than the first, etc.

For the compensation model, the dynamics are even more complex because both compensation and waiting time depend on others' decisions. Still, we prove that the equilibrium with compensation also has a cost-dependent, increasing-threshold structure. Importantly, priority is 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; this implies that in equilibrium, fewer customers purchase priority in the compensation model.

We next define and study three system-wide performance measures: aggregate waiting cost, customer surplus, and provider revenue. We characterize the system performance in the base and compensation models and for various priority fees and supports of the waiting-cost distribution. First, unsurprisingly, customer surplus is higher with compensation than in the base model. Additionally, with a high enough compensation fraction, customer surplus exceeds that under first-come, first-serve (FCFS) service. Second, for low priority fees, the compensation model—despite entailing fewer priority purchases—achieves lower aggregate waiting cost than the base model. At low fees, low-cost customers sometimes buy priority in the base model, a socially undesirable outcome. Compensation attracts low-cost customers to the regular queue without deterring high-cost customers from buying priority, leading to a more efficient service order. Third, when the priority fee is higher, low-cost customers are unlikely to buy priority in either model, so the "filtering" benefit of compensation is negligible and the base model tends to yield lower aggregate waiting cost. Moreover, when we allow optimization of the priority fee, the base model consistently achieves lower aggregate waiting cost than the compensation model. Thus, compensation does not deliver a short-term performance benefit when the priority fee can be optimized or if the priority fee is high.

We then take our models to the laboratory. We present two studies: the All-Human Study and the One-Human Study. In the All-Human Study, we collected data from several queues in which subjects made their decisions and then physically stood in line and waited to be served. This deviation

from the standard method of conducting laboratory experiments was intentional because the lack of anonymity highlighted any potential behavioral issue that may arise in a real situation in which customers pay for priority, effectively "cutting in line" in front of others. However, the real-time and sequential nature of the All-Human Study implies slow and expensive data collection. So, we conducted additional experiments in the One-Human Study, in which individual subjects played our game repeatedly against computerized agents. Subjects in the One-Human Study exhibited qualitatively similar behavior to the All-Human Study, and the larger data set from the One-Human Study provides much higher power.

In our experiments, although subject decisions are directionally in line with equilibrium predictions, subjects exhibit behavioral regularities with managerial implications. In particular, low-cost subjects represented a greater proportion of priority purchase decisions (and a smaller proportion of regular-queue decisions) than predicted, in both the base and compensation model treatments.

We leverage the larger data set from the One-Human Study to compute system-wide performance measures. Comparing these measures against those in equilibrium and with FCFS service yields prescriptive insights for operating a priority service system. First, in terms of aggregate waiting cost, compensation fails to achieve the predicted performance gain for low fees. Thus, despite its predicted benefit in equilibrium for such fees, compensation in priority service systems appears not to have a short-term performance benefit with human decision makers. Still, we conjecture that service providers might choose to implement it to calm strong resistance to a priority program, or, if socially motivated, for performance benefits if (i) they are constrained to charge a low priority fee and (ii) they can successfully nudge customers toward rational decisions. In this case, our results provide a tool for service providers to assess how customers' priority purchasing decisions (and the resulting system performance) will change if compensation is implemented. Second, even in the base model the higher fraction of low-cost customers in the priority queue harms the social welfare relative to the equilibrium. At the socially optimal fee, the reduction in aggregate waiting cost relative to FCFS is barely a third of that predicted by the equilibrium. Low-cost customers buying priority is socially undesirable even if individually rational, but our low-cost subjects purchase priority even more than in the rational equilibrium. This behavioral regularity thwarts the priority mechanism because high-cost customers cannot overtake a low-cost customer if that customer bought priority earlier in the sequence. Third, and perhaps most important, at high priority fees, not only low-cost customers but also high-cost customers overbuy priority, i.e., *customers tolerate higher priority fees than they rationally should*. Thus, a service provider can potentially capitalize on customers' deviations from equilibrium to earn even more priority revenue. Not only do we find that the revenue-optimal priority fee is higher based on subject decisions than in equilibrium, but also the maximum revenue is higher by nearly 5%.

So, we find that behavioral deviations in priority service systems lead to worse (i.e., higher) aggregate waiting cost relative to the rational equilibrium, but they permit a revenue-maximizing service provider to set a higher priority fee and earn more revenue. The managerial prescription depends on the service provider's priorities. For performance-critical systems, our findings suggest that the service provider should attempt to influence customers toward rational decisions so that low-cost customers do not occupy slots in the priority queue. One way to accomplish this is to communicate to customers that the regular-queue experience is a good one, hopefully convincing low-cost customers that priority is a luxury and not a necessity, while also touting the time-saving benefits of priority to those who especially dislike waiting. Indeed, Killington Ski Resort's community letter explicitly tried to reassure customers that "the impact [of the priority option] on lift line wait times across our mountains is negligible," whereas Killington elsewhere employs phrases like "upgrade" and "maximize your time" (Killington 2021) to catch the attention of high-cost customers who especially dislike waiting. On the other hand, Universal Orlando Resort's practices seem more in line with our findings on revenue maximization: the price for Universal Express, which provides access to a priority queue, can be more than $200 (Universal Orlando Resort 2023), even more than the admission ticket to the park! In this spirit, our results suggest that a provider focused on revenue can set the price for priority high and let nature run its course, anticipating that although they may cry foul as documented above, plenty of customers will likely pay up anyway.

2. Literature Review

Theoretically, strategic customer behavior in priority service systems has been studied extensively. A seminal work is Kleinrock (1967), in which customers bid 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. Yang et al. (2016) gives an excellent review of the theoretical priority-pricing literature. Hassin and Haviv (2003) and Hassin (2016) provide useful surveys. Additionally, and reflecting the current interest in and importance of priority service systems, Cui et al. (2023) theoretically studies several methods for assigning priority and waiting time in queues, such as line-sitting (paying someone else to wait in line on one's behalf) and distance-based priority (giving priority to customers who have traveled farther to the system).

In Yang et al. (2016), the focus is on transfer mechanisms in a marketplace in which customers bid for favorable positions in line. An auction mechanism in which customers submit bids to overtake others based on their value of time is shown to result in efficient service. Rosenblum (1992) shows that in a $G/M/S$ system, customers can achieve a socially efficient equilibrium by paying to trade

positions; importantly, waiting costs are common knowledge. 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. Gurvich et al. (2019) study price-waiting-time menus and the number of priority classes to offer, also in steady state and with an unobservable queue. They find that a social planner and a revenue-maximizing provider offer exactly two priority classes. However, a revenue maximizer may induce too many or too few priority customers relative to the first-best. 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. Roet-Green and Shetty (2022) studies a social planner who divides service resources between expedited and regular service options, and customers choose in advance of their service needs whether to buy access to the expedited option. The customers must be pre-processed to access the expedited service, and, importantly, the service time itself is shorter for expedited customers. The resource allocation optimization is solved, and the expedited service is shown to sometimes benefit not only overall welfare but also all individual customers. In addition to endogenous service times, key differences from our work are that (i) resources are divided between the regular and expedited service, rather than shared as in our model, (ii) their analysis is in steady state, and (iii) their queues are unobservable. Finally, a rare analytical study on social preferences in queueing is Allon and Hanany (2012), which studies queue-jumping behavior. None of these works consider redistribution of priority payments.

Rather than the conventional steady-state queueing system, we study a sequential game in a finite system. 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. Subjects may not understand stationary distributions or the dynamics of an unobservable queue, but a fixed number of customers and observable moves is more approachable. Eliminating this cognitive challenge allows us to focus on the key incentives in priority queueing. 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 with finite queues is the "static" model in Erlichman and Hassin (2015); there, instead of a flat priority fee, each customer may pay a certain amount per overtaken customer; the equilibrium has 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" from 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. Allon and Kremer (2018) gives an overview of the current state of behavioral queueing research. 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. 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. Dold and Khadjavi (2017) conducts experiments in which one subject can bribe another to reduce her own waiting time. Their subjects displayed strong social preferences, perhaps due to the individual nature of the transactions, unlike our setup where the service provider handles payments.

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.

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.

Each customer at the time of her decision 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.[5] 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 also that priority is at least cheap enough that it would be worth purchasing to move from last to first place, we assume $V \geq \bar{c}N$ and $p \leq \underline{c}(N-1)$, respectively. The setup of the game and the parameters—including p, N, F, and V—are common knowledge.

Due 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, her beliefs about these customers are irrelevant. 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 their waiting costs.

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 known to her and does not depend on the strategies of later arrivals because they will not be able to overtake her.

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.

[5] To break ties, we assume that a customer joins the priority queue if indifferent between the two options.

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 the priority queue is below a threshold length that is a function of her waiting cost, i.e., she uses a *cost-dependent threshold strategy.*[6]

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, denote $\bar{x}_{i-1}(C_i) := \max\{k : \sigma_i(k, C_i) = 1\}$.

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. For $k \in \{0, \ldots, i-2\}$, we have*

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

All proofs can be found in the e-companion. 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

[6] 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.

the pure strategies to analyze the outcomes on each sample path, circumventing the combinatorial problem described above to reveal the structure of the equilibrium.

A focal customer i who chooses the regular queue will wait through at least i services because she will not overtake anyone in either queue. However, if the customers after i use fixed threshold strategies, 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 $x_i = k$ may 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 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 on the priority queue length, below which customers purchase priority.*

The sample-path approach used in Lemma 1 is agnostic to the form of the waiting-cost distribution, and therefore so is Theorem 1. 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.

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*

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

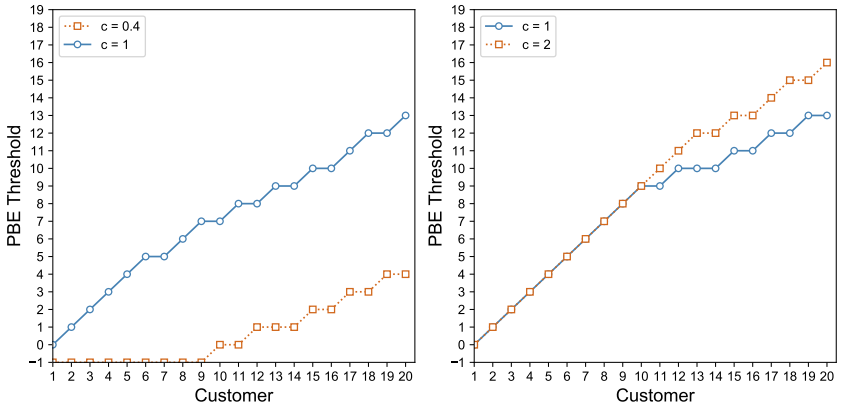
Moreover, for $c < c'$, we have $\bar{x}_i^*(c) \leq \bar{x}_i^*(c')$.

Theorem 2 implies that a customer earlier in the order will have a lower threshold than one later in the order, conditional on the same waiting-cost realization. A given priority queue length k implies one more regular queue customer for customer $i+1$ to overtake than for customer i, so intuitively, customer $i+1$ should be willing to purchase priority up to a higher priority queue length than customer i due to her greater time savings from priority for each given priority queue length. Thus, higher thresholds for later customers makes intuitive sense. 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. 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 as we move back in the order 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 equilibrium is still computationally difficult, even for simple waiting-cost distributions. Customers may need to take expectation over a huge number of possible vectors of thresholds for the later customers. In Appendix H, we provide an algorithm to compute the PBE threshold functions. 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.

We apply our algorithm to compute the PBE thresholds for two-point distributions with equal probabilities, and we plot the PBE thresholds for two different distributions in Figure 1. As expected, the thresholds increase as we move back in the order, and the thresholds are higher for the higher realization. In Figure 1(a), we observe two different "patterns" in the strategies, one for each 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



(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 (prior probability 1/2 for each waiting-cost realization, $N = 20$, $V = 35$, and $p = 6$).**

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. Notably, despite the threshold structure for given waiting cost, the random waiting costs mean that the realized path of play may not follow a simple pattern.

Comparing the two plots highlights the impact of the randomness in the waiting costs. The thresholds for a waiting cost of 1 are different in the two plots because of the different thresholds for the other realization. In the right panel, it is more important to protect her position than in the left panel: "protective" 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.

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.

5. Compensation Model

In the base model, the service provider kept all payments: we now move to a setting in which a fraction $\gamma \in (0, 1]$ of the priority revenue is redistributed to the regular-queue customers.[7]

[7] 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.

5.1. Structural Results

The compensation model has more complicated dynamics, and it is not obvious ex ante what form the equilibrium strategies should take. 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 the base model, 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 compensation on the equilibrium strategies, which we accomplish in the next lemma.

LEMMA 2 **(Compensation Effect of One Additional Regular-Queue Customer).** *Consider a customer $i \in \{1, \dots, N-1\}$, and suppose that customers $j \in \{i+1, \dots, N\}$ 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 $k \in \{0, \dots, i-2\}$, we have*

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

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 the same or one more. So, either the compensation is the same or, if there is one more priority purchase, then it is more because the (increased) priority revenue is split among fewer regular-queue customers. The expected compensation from choosing the regular queue after observing priority queue length k is thus weakly less than for $k+1$. Thus, 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.

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

Although the threshold structure is maintained, the dynamics are more complex here because of the impact of the non-linear compensation on incentives and equilibrium inference. Nonetheless, we can still show that the equilibrium thresholds follow a similar pattern 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* $\bar{x}_i^*(c) \leq \bar{x}_i^*(c')$.

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.

Despite the shared structure of cost-dependent, increasing threshold strategies, the equilibria are different in the base and compensation models because compensation 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*

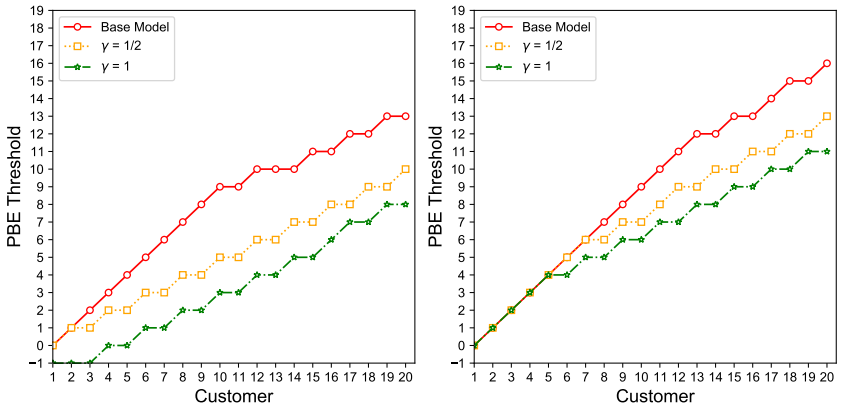
$$\bar{x}_{i-1,\gamma}^*(c_i) \leq \bar{x}_{i-1}^*(c_i),$$

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.

Compared to the base model, the effect of one additional priority customer is magnified with compensation. Theorem 5 establishes an unambiguous relationship between the equilibrium thresholds in the two models: thresholds are lower with compensation, which also implies fewer priority purchases.

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

	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)



(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 $\{1, 2\}$ (prior probability $1/2$ for each waiting-cost realization, $N = 20$, $V = 35$, $p = 6$).

Comparing the thresholds among different compensation fractions is difficult because if the thresholds behind a focal customer decrease with the compensation fraction, then a higher compensation fraction means a shorter wait time but less compensation in the regular queue. Still, if we fix the strategies behind a focal customer, then Theorem 5 shows that the focal customer's optimal threshold decreases with γ. Numerically, we observe that the equilibrium thresholds also decrease with γ.

5.2. Numerical Examples

We first look briefly at the special case of deterministic waiting costs. Table 1 compares the thresholds and number of priority purchases from the base model with those in the compensation model with $\gamma = 1$. Recall that a customer with a threshold of -1 means will join the regular queue no matter what. The thresholds in both models decrease with the priority fee. Also, as implied by Theorem 5, the equilibrium thresholds are lower with compensation, leading to fewer priority purchases.

In Appendix H, 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 in Figure 2. As expected, the threshold is decreasing in γ, and 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 $\gamma = 1$. 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 $\gamma = 1$.

6. System-Wide Performance Measures

Understanding the operation of priority service systems and our compensation mechanism in particular requires understanding not only customer strategies but also system-wide performance measures. Our measures of interest are (i) aggregate waiting cost, (ii) customer surplus, and (iii) provider net revenue from priority. We give the formal definition of each performance measure in Appendix I.

The performance measures depend on the path of play, which itself depends on the PBE thresholds and thus on the waiting-cost realizations. To compute the expected values of these measures thus requires considering each realization separately, of which there are an enormous number (2^N even for a two-point distribution). The concomitantly huge number of possible equilibrium paths makes it extremely difficult to analytically compare such measures across different parameter instances and compensation fractions. Instead, in this section, we perform two numerical studies: one that varies many parameters to cover a wide range of different scenarios, and another that considers a smaller number of scenarios in order to illuminate the effect of the priority fee, which will be important in our laboratory experiments. Throughout the section, we consider two-point waiting-cost distributions, and when making comparisons in words (e.g., "greater than"), the comparisons are in the weak sense.

First Numerical Study. For each instance in our first numerical study, we compute the PBE thresholds and then determine the equilibrium path under each of the 2^N possible vectors of waiting costs. We consider $N \in \{10, 12, 14, 16, 18, 20\}$, $p \in \{2, 4, 6\}$, $\gamma \in \{0, 1/4, 1/2, 3/4, 1\}$, $\text{supp}(C_i) \in \{\{1, 2\}, \{1, 4\}, \{1, 5\}, \{2, 4\}, \{2, 5\}, \{3, 5\}, \{4, 5\}\}$, and $\delta \in \{1/4, 1/2, 3/4\}$, where δ denotes the probability mass on the smaller element in the support of C_i (which also determines the probability of each possible waiting-cost vector). Note that the priority fees are relatively low in this study compared to the total waiting cost a customer can expect to incur. The combinations yield a total of 1,890 parameter instances, and in each instance, we set $V = \bar{c}N$, where $\bar{c}$ is the larger element in the support of C_i. For all instances, we compute the expected values of our three performance measures.

First, an important goal of any priority system is to promote social welfare; that is, a system with a priority option should ideally result in a lower aggregate waiting cost than an equivalent system without a priority option. This goal is met in 100% of the parameter instances in our study; that is, in all 1,890 instances, the expected aggregate waiting cost with a priority option is lower than that with first-come, first-serve (FCFS) service. Importantly, this finding holds both in the instances with compensation ($\gamma > 0$) and those without ($\gamma = 0$). Second, recall from Theorem 5 and surrounding that compensation reduces the expected total number of priority purchases. Since the priority option aims to promote efficiency by serving higher-cost customers earlier, we might expect that this reduction in priority purchases would harm social efficiency. However, perhaps counterintuitively, in 1,511 of 1,512 instances with $\gamma > 0$ (99.9%), the aggregate waiting cost is lower than for $\gamma = 0$ and the same

Table 2 Summary statistics for system-wide performance measures in first numerical study.

γ	Agg. Wait. Cost Less Than FCFS (Freq.)	Surplus Exceeds FCFS (Freq.)	Agg. Wait. Cost Less Than Base Model (Freq.)
0	100%	0.53%	-
1/4	100%	12.17%	99.74%
1/2	100%	29.37%	100%
3/4	100%	67.20%	100%
1	100%	99.47%	100%

other parameters. That is, despite entailing fewer priority purchases, compensation *reduces* aggregate waiting cost in this study with low priority fees.

What leads to this finding? By definition, a given amount of money is worth more time to a low-waiting-cost customer than to a high-waiting-cost customer. Thus, implementing compensation is more likely to change a low-cost customer's queue choice from priority to regular than to cause such a change for a high-cost customer. A low-cost customer whose decision switches from priority to regular will now be overtaken by any later high-cost customers who purchase priority, which reduces the aggregate waiting cost. Indeed, we observe in Figure 2 that the PBE thresholds for $\gamma = 1$ versus $\gamma = 0$ tend to be farther apart in the left panel (low cost realization) than in the right panel (high cost realization). To validate the intuition, we compare the number of priority purchases in the base model for high- and low-cost customers against the corresponding numbers for $\gamma > 0$. For $\gamma > 0$, the expected number of low-cost customers purchasing priority is 1.98 less on average than for the same parameters with $\gamma = 0$, but the expected number of high-cost customers purchasing priority is only 0.15 less; also, in 1,463 of 1,512 instances with $\gamma > 0$ (97%), the reduction in expected low-cost priority purchases is of greater magnitude compared to the reduction in expected high-cost priority purchases. That is, as conjectured, the actions of low-cost customers are more affected by compensation than those of high-cost customers. We summarize the above findings in the following observation.

Observation 1 **(Aggregate Waiting Cost for Small Priority Fee)**. *When the priority fee is relatively small: (i) in both the base model ($\gamma = 0$) and compensation model ($\gamma > 0$), the aggregate waiting cost is lower than under FCFS, and (ii) the aggregate waiting cost is lower in the compensation model ($\gamma > 0$) than in the base model ($\gamma = 0$).*

Since compensation redirects a portion of priority payments to customers, we expect it to increase customer surplus, and indeed, for all 378 combinations of N, p, supp(C_i), and δ in our study, the customer surplus is increasing in the compensation fraction γ. Next, we group the parameter instances by γ and compute the percentage of cases in which the surplus is higher than that under FCFS. We see from Table 2 that the customer surplus more frequently exceeds that of FCFS as γ increases. In the base model ($\gamma = 0$), customer surplus exceeds that of FCFS only in 2 of 378 cases. By contrast, for $\gamma = 3/4$, this happens in more than 2/3 of cases, and for $\gamma = 1$, in almost all (376 of 378). These findings validate our intuition about the value of compensation for increasing customer surplus.

OBSERVATION 2 **(Customer Surplus).** *Customer surplus is higher in the compensation model ($\gamma > 0$) than in the base model ($\gamma = 0$). Moreover, for high (low) compensation fractions, customer surplus is higher (lower) than under FCFS.*

Second Numerical Study. We now want to obtain a better understanding of the impact of the priority fee on equilibrium customer decisions and system performance. To accomplish this, we consider a smaller cross product of values for the other parameters so that we can consider a much wider range and finer grid of priority fees. We fix $\delta = 1/2$ and $N = 10$, and we consider $\gamma \in \{0, 1\}$ and $\text{supp}(C_i) \in \{\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{2,5\},\{3,5\},\{4,5\}\}$. We again let $V = \bar{c}N$, where $\bar{c}$ is the larger element in the support of C_i. An upper bound on the range of prices to consider is easily determined: if $p > \bar{c}(N-1)$, then no customer will buy priority in equilibrium because the fee is strictly more than her waiting cost savings even if she has high waiting cost and priority allows her to overtake all $N-1$ other customers. Thus, we consider prices $p \in \{1/2, 1, 3/2, \ldots, \bar{c}(N-1), \bar{c}(N-1)+1/2\}$.

Comparing base and compensation models for a fixed priority fee, the aggregate waiting cost is lower with compensation ($\gamma = 1$) than in the base model ($\gamma = 0$) in 37% of instances (205 of 547). For these, the average fee is 13.70, versus 25.19 for the 63% of cases where the base model performs better; compensation performing better with lower fees aligns with our finding in the first numerical study. This phenomenon can also be understood with similar intuition from the first study. In cases where compensation achieves lower aggregate waiting cost, there are on average 1.24 fewer low-cost priority purchases in the compensation model than in the base model; that is, compensation is filtering out low-cost customers from the priority queue. On the other hand, in cases where the base model achieves lower aggregate waiting cost, the number of low-cost priority purchases in the compensation model is almost exactly the same as in the base model (only 0.017 fewer purchases on average). The reason is that in these latter cases, the fee is usually high enough that almost no low-cost customers (0.042) purchase priority even in the base model, so the potential for further "filtering" is negligible.

It turns out that when we allow optimization of the priority fee, the base model tends to perform better. In other words, optimally choosing the priority fee acts as a substitute for implementing compensation in terms of filtering low-cost customers out of the priority queue. Indeed, for all seven supports of C_i that we consider, the lowest aggregate waiting cost in the base model is lower than the lowest aggregate waiting cost in the compensation model. We also ran the same computations for the case with $N = 14$, and again in all seven instances, the base model achieves lower aggregate waiting cost when the priority fee can be optimized. So, despite the promising feature of compensation that it prevents low-cost customers from buying priority when they otherwise would, unfortunately it does not do this job better than can be accomplished by optimally setting the priority fee in the base model. Combining the above findings leads to our third observation about system performance.

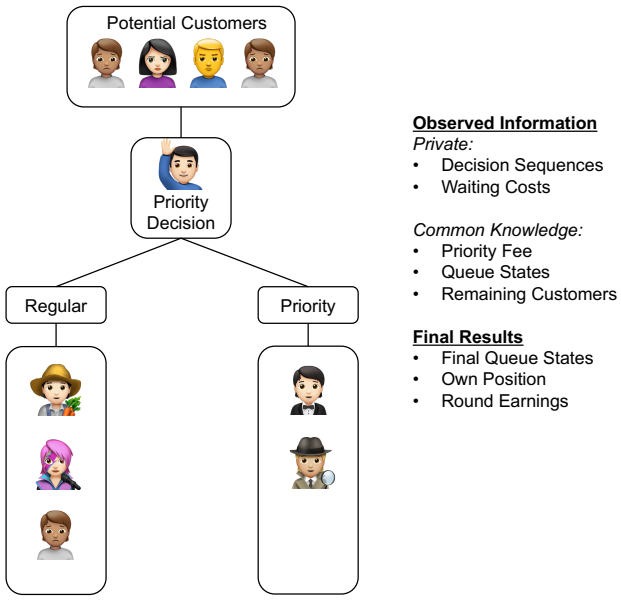


Figure 3 **All-Human Study session flow.**

Observation 3 **(Effect of Priority Fee on Aggregate Waiting Cost).** *The aggregate waiting cost tends to be lower (higher) with compensation than under the base model when the priority fee is low (high). If the fee is optimized to minimize aggregate waiting cost, then the aggregate waiting cost is lower in the base model than the compensation model.*

Armed with these findings about system performance, as well as our structural results about equilibrium customer decisions, we next test our model's predictions in the laboratory.

7. Behavioral Experiments

We conducted several controlled laboratory experiments with human subjects who were incentivized based on their decisions. The purpose of these experiments was to test the extent to which human behavior resembles the equilibrium and then use these behavioral insights to test the performance of service systems with a range of priority fees in the base model and the compensation model.

An all-human queue constitutes only a single observation, and processing an entire queue involves serving each customer, a lengthy process that is not ideal for repetition. However, in testing an equilibrium model, it is important to allow participants to learn to play the game by allowing them to play repeatedly. Also, understanding how participants react to different priority fees requires multiple observations with each fee. To tackle these challenges, we used a two-part process. We first conducted a study (the All-Human Study) collecting data from several queues consisting of 10 people ($N = 10$) for the base and compensation models. We then conducted a second study (the One-Human Study) in which one human participant interacted with nine computerized agents programmed to play the best reply. The human participants played 100 rounds of the game, each round facing a randomly-drawn priority fee, decision sequence, own waiting cost, and the waiting costs of the computerized agents.

7.1. All-Human Study

Protocol. We conducted five sessions for the base treatment and 10 sessions for the compensation treatment, for a total of 15 sessions. Each session included 10 participants ($N = 10$), with a different group of subjects for each session. Because the main purpose of this behavioral experiment is to identify any systematic deviations from the models' predictions, we conducted the sessions in person, so that participants actually waited to be served in live queues, and participants who purchased priority were served before participants who did not. This departure from the standard method of conducting laboratory experiments was intentional because the lack of anonymity highlighted any potential behavioral issue that may arise in a real situation in which customers pay for priority, effectively "cutting in line" in front of other customers. As a first step to understanding the behavioral effects of priority payment redistribution, we tested the special case of a two-type waiting cost distribution (High or Low with equal probability), with a compensation fraction $\gamma = 1$ in the compensation model treatment ($\gamma = 1$ maximizes the difference between the models and thus should also magnify differences in outcomes: the other experimental parameters were chosen with similar motivation). The combined endowment of \$15 and a participation fee of \$5 imply a valuation $V = \$20$. We set the priority fee p at \$1.50, and the waiting cost c at \$0.50 for low-type customers and \$1.50 for high-type customers 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 by eliminating 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.

Figure 3 displays the flow of experimental sessions in the All-Human Study. Upon entering the room, participants were asked to log into the experimental software on their phones. We decided to keep track of the decisions with the aid of the software in order to streamline the data collection process; this also provided a smooth transition to the One-Human Study.

Once all participants were seated, read instructions that describe the game (see Appendix J), and provided their Informed Consent, the session started and each participant observed their own decision sequence and waiting cost (either \$0.50 or \$1.50), as well as the priority cost (which was \$1.50 in the All-Human Study). They also observed the current state of both queues, which were empty at the beginning of the session, and filled up as the session progressed.

Participants were called to make the priority decision in the order of their sequence numbers. Upon making the decision, each participant entered it into the software on their smartphone, then physically

got up and stood in the chosen queue. After all participants had made their decisions and were standing in their chosen queues, they were served. The service consisted of recording the subject's earnings on a receipt, having them sign the receipt, 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. Participants left the session after all ten were served.

Results. We conducted five sessions for the base model treatment. In each session, the waiting cost corresponding to each decision sequence was generated randomly and independently. Thus, the baseline treatment consisted of five unique cost draw sets, which we call cost profiles. For the compensation treatment, we used five identical profiles. Upon examining the compensation treatment data, we decided to check whether the compensation aspect was sufficiently salient. To do this, we added to all compensation screens a blue box that computed the minimum and maximum possible compensation if the current decision-maker were to join the regular queue. The minimum compensation scenario is one in which all remaining customers join the regular queue, while the maximum compensation scenario is one in which all remaining customers join the priority queue. We repeated the identical five cost profiles for the compensation treatment with this added feature. It turned out that the saliency feature made no difference to behavior or the measures, so we pooled the data for the ten compensation treatment sessions. But findings remain the same with and without this data pooling.

Important quantities for the experiments include the five cost profiles (see Table EC.1 in Appendix A), the equilibrium paths and lengths of the priority and regular queues for both models (Table EC.2), and the observed behavior in each session of the two treatments (Table EC.3). Another measure that will be indicative of behavioral deviations from equilibrium is the percentage of low-cost customers occupying each queue. Note that the compensation treatment includes two instances of each profile—10 sessions. We first observe that in agreement with Theorem 5, priority queues are significantly shorter and regular queues significantly longer in the compensation treatment than in the base treatment. Also, in equilibrium, the regular queue should consist exclusively of low-cost customers in all five profiles for both treatments. But in our experiment, four of five sessions in the base model treatment, and eight of ten sessions in the compensation model treatment have at least one high-cost customer in the regular queue.

In Table 3, we report summary statistics for performance measures observed in the two models, and also for comparison, the same measures in equilibrium (for the same five waiting-cost profiles). As another comparison, we report the aggregate waiting cost and surplus that would have resulted in the FCFS regime. Notable behavioral regularities that we would like to emphasize are: (1) The proportion of low-cost customers observed in the priority queue is higher than in equilibrium in both treatments, and the difference is significant in the compensation treatment but not in the base treatment. (2)

Table 3 Average performance measures in All-Human Study.

	Base		Compensation		FCFS
	Equilibrium	Observed	Equilibrium	Observed	
Priority Length	7.8	7	5.4	6^{+}	-
	(0.20)	(0.32)	(0.24)	(0.15)	
Priority Proportion Low	0.52	0.62	0.31	0.62^{*}	-
	(0.10)	(0.07)	(0.12)	(0.06)	
Regular Length	2.2	3	4.6	4^{+}	-
	(0.20)	(0.32)	(0.24)	(0.15)	
Regular Proportion Low	1	0.57^{*}	1	0.63^{*}	-
		(0.12)		(0.09)	
Aggregate Waiting Cost	46.1	47.7	39.9	48.7^{*}	48.9
	(3.9)	(4.4)	(3.1)	(3.6)	(4.4)
Surplus	142.2	141.8	160.1	151.3^{*}	151.1
	(4.1)	(4.5)	(3.1)	(3.6)	(4.4)
Revenue	11.7	10.5	0	0	-
	(0.30)	(0.47)			

Note: $^{*}H_0$: Observed = Equilibrium $p < 0.05$; $^{+}H_0$: Base = Compensation $p < 0.05$

The proportion of low-cost customers observed in the regular queue is significantly lower than in equilibrium in both treatments. These behavioral deviations from equilibrium do not result in any significant difference between observed and equilibrium waiting costs or customer surplus levels in the base condition. They do, however, lead to significantly higher waiting costs (and lower surplus levels) in the compensation condition. Observed as well as equilibrium waiting costs (and surplus levels) also do not significantly differ from what they would have been under the FCFS regime in the base condition. In the compensation condition, waiting costs (and surplus levels) should be lower in equilibrium than under the FCFS, but observed averages are not different from FCFS.[8]

Next, we analyze individual decisions by fitting a sequence of logit models, where the dependent variable is the choice to buy priority, and the independent variables are in the first column of Table 4. The estimates show that the rational decision has no explanatory power. Recall that the rational decision depends on comparing the current priority queue length to a threshold that depends on the waiting cost. Model (2) shows that priority queue length does affect behavior in the predicted direction (longer length decreases the likelihood of choosing priority), but Model (3) shows that waiting cost does not affect behavior. In Model (4) we add the sequence number to the model and it turns out that the sequence number has explanatory power, being positive and significant, which indicates that customers who choose later are more likely to purchase priority. The sequence number should not affect the priority choice after controlling for the rational decision. So, participants overreact to the priority queue length, but the effect on the aggregate likelihood of choosing priority is offset by considering

[8] Since the measures in Table 3 are aggregated at the session level, the sample size is small, so the statistical power to identify differences in these measures is limited. To augment the qualitative insights from the All-Human Study, it is worthwhile to consider an alternative experimental design that permits efficient collection of a larger sample; this we accomplish with the One-Human Study in Section 7.2.

Table 4 Random effect logit model estimates for All-Human Study.

	(1)	(2)	(3)	(4)
Rational Decision	0.209	0.0786	0.185	-0.455
(1 = Priority, 0 = Regular)	(0.401)	(0.417)	(0.704)	(0.805)
Compensation	-0.325	-0.620	-0.558	-1.608*
(1 = Compensation, 0 = Base)	(0.434)	(0.463)	(0.567)	(0.729)
Priority Queue Length		-0.354***	-0.351***	-1.623***
		(0.102)	(0.104)	(0.381)
Waiting Cost			-0.118	0.118
(1 = High, 0 = Low)			(0.625)	(0.718)
Sequence Number				0.838***
				(0.236)
Constant	0.652	2.086**	2.017**	2.412**
	(0.486)	(0.663)	(0.755)	(0.874)
N (groups)	150 (15)	150 (15)	150 (15)	150 (15)
Log Likelihood	-97.71	-91.06	-91.05	-83.55
χ^2	1.7	13.09	13.18	22.63

Note: Panel variable is group ID; Standard errors in parentheses; * $p<0.05$, ** $p<0.01$, *** $p<0.001$

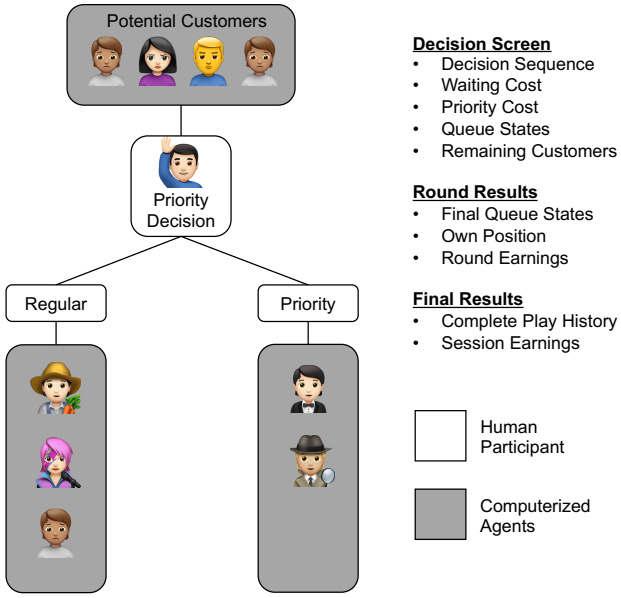


Figure 4 One-Human Study session flow.

the sequence number, with which the priority queue length is positively correlated (correlation is 0.44, which does not raise concerns about multicollinearity). The net effect is that, although the proportion of low-cost customers in the priority queue is significantly higher than equilibrium, the average total number (low-cost plus high-cost) of participants who purchase priority in the All-Human Study is not significantly different from the equilibrium prediction, as shown in Table 3. This finding, however, is for a single priority fee; when we vary the priority fee (in the One-Human Study below), the outcomes at high fees differ from the equilibrium prediction.

Table 5 Proportion of decisions to join each queue by low-cost customers in the One-Human Study.

	Base		Compensation	
Proportion Low	Equilibrium	Observed	Equilibrium	Observed
Priority	0.20	0.34*	0.11	0.38*
Regular	0.67	0.58*	0.63	0.53*

Note: H_0: Observed = Equilibrium * $p < 0.05$

7.2. One-Human Study

Protocol. Figure 4 displays the flow of experimental sessions in the One-Human Study. Besides being done entirely on the computer, the One-Human Study differed from the All-Human Study in three major ways: (1) each human participant played with nine computerized agents programmed to play the best reply (i.e., the PBE strategies for their realized waiting costs), (2) the task was repeated for 100 rounds, and (3) the priority cost was randomly chosen each round from a range between \$0.50 to \$20, each 50-cent increment equally likely. The waiting cost for a customer (human or computerized) was \$1 or \$2 per service with equal probability, and subjects started with an endowment of \$20 and received a \$5 participation fee, implying $V = 25$.[9] Subjects were paid based on their earnings averaged over all rounds in the session. We kept the decision screen identical to the decision screen in the All-Human Study. The information that participants observed was also the same. After each decision, participants saw the outcome, which included decisions made by the computerized agents who made their decisions after the subject, and the resulting waiting cost and total cost for the round. Each round all players (human and computerized) drew new waiting costs and sequence numbers, and the priority fee for the group was also determined. This information was displayed on the decision screen.

We conducted one session each, for the base treatment and the compensation treatment. Each treatment included 50 people, each using an individual cost profile. As in the All-Human Study, we matched cost profiles across the two treatments, but in this study, the sequence number of the human player and the priority cost for the round were also matched across the two sessions.

Results. Recall that in the All-Human Study, we identified two behavioral regularities, both of which we can directly measure in the One-Human Study: (1) The proportion of low-cost customers observed in the priority queue is significantly higher than in equilibrium in both treatments. (2) The proportion of low-cost customers observed in the regular queue is significantly lower than in equilibrium in both treatments. To make a qualitative comparison of this between the All-Human and the One-Human Studies, we measure the proportion of observed and equilibrium decisions to join each queue by low-cost customers and display the summary in Table 5.

We observe both behavioral regularities in the One-Human Study that we did in the All-Human Study: priority queues have too many low-cost participants, and regular queues have too few of them.

[9] We assumed for our analytical models that $p \leq \underline{c}(N-1)$. Although some of the parameter settings in the One-Human Study do not satisfy this assumption, it can be shown that all of our theoretical results continue to hold if the assumption is dropped.

Table 6 Random effect logit model estimates for the One-Human Study ($\gamma = 0$).

	(1)	(2)	(3)	(4)
Rational Decision	1.915***	1.792***	1.337***	0.977***
(1 = Priority, 0 = Regular)	(0.0710)	(0.0735)	(0.101)	(0.107)
Priority Queue Length		0.0977***	0.0290	-0.261***
		(0.0165)	(0.0220)	(0.0354)
Waiting Cost			0.428***	0.610***
(1 = High, 0 = Low)			(0.0827)	(0.0867)
Priority Fee			-0.0544***	-0.158***
			(0.0103)	(0.0148)
Sequence Number				0.211***
				(0.0200)
Constant	-1.387***	-1.498***	-1.315***	-1.134***
	(0.130)	(0.133)	(0.202)	(0.206)
N (groups)	5000 (50)	5000 (50)	5000 (50)	5000 (50)
Log Likelihood	-2691.1	-2673.8	-2653.3	-2592.9
χ^2	727.3	745.1	765.6	798.2

Note: Panel variable is subject ID; Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This link gives us the confidence to conclude that behavior in our One-Human Study is qualitatively similar to behavior in the All-Human Study, so we will compute the relevant system-wide measures based on the behavior in the One-Human Study.

Analysis of Individual Decisions. We begin by fitting a set of logit models similar to the ones that we used in analyzing behavior in the All-Human Study but separate for the base and compensation treatments, and with the additional independent variable Priority Fee. We present these estimates in Table 6 for the base case and Table 7 for the compensation case.

Comparing estimates of the full models (Model 4) for the All-Human Study in Table 4 and the One-Human Study in Tables 6 and 7, we see that behavior is qualitatively similar. The signs of the variables for the priority queue length, waiting cost, and sequence number are the same. Because the One-Human Study included a range of values for the priority fee, we can also measure its effect, which is negative and significant, as expected. Additionally, in the One-Human Study, the Rational Decision variable is positive and significant, so the behavior is qualitatively consistent with the equilibrium prediction. Also, unlike in the All-Human Study, all variables in Model (4) are statistically significant. The One-Human Study has more than 30 times the amount of data than the All-Human Study, so the power is much higher (which was one of the reasons for this design).

System-Wide Performance Measures. We are interested in comparing system performance between the base and compensation treatments, as well as comparing observed measures to equilibrium predictions and the FCFS regime. Since each decision in the One-Human data includes one human subject and nine automated customers programmed to follow the equilibrium, comparing the observed measures directly to the equilibrium would not be meaningful. Instead, we conducted a simulation study in which agents are programmed to behave according to the models we estimated in Tables 6 and 7 for the base and compensation treatments, respectively. As we already showed that behavior

Table 7 Random effect logit model estimates for the One-Human Study ($\gamma = 1$).

	(1)	(2)	(3)	(4)
Rational Decision	1.347***	1.208***	1.059***	0.586***
(1 = Priority, 0 = Regular)	(0.0756)	(0.0790)	(0.101)	(0.108)
Priority Queue Length		0.137***	0.111***	-0.238***
		(0.0232)	(0.0290)	(0.0388)
Waiting Cost			0.148	0.370***
(1 = High, 0 = Low)			(0.0808)	(0.0854)
Priority Fee			-0.0159	-0.105***
			(0.00885)	(0.0116)
Sequence Number				0.241***
				(0.0174)
Constant	-1.175***	-1.271***	-1.269***	-1.618***
	(0.170)	(0.174)	(0.226)	(0.235)
N (groups)	5000 (50)	5000 (50)	5000 (50)	5000 (50)
Log Likelihood	-2673.5	-2656.1	-2653.4	-2548.5
χ^2	317.3	344.5	348.9	483.9

Note: Panel variable is subject ID; Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

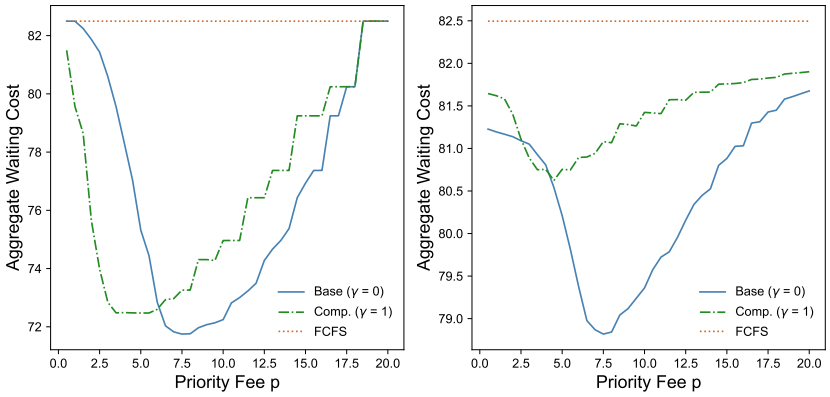
in the One-Human Study is close to the behavior in the All-Human Study, we expect the measures that we obtain from these simulations to be similar to analogous measures if we were to conduct the same experiments with all human subjects.[10] Details of the simulation procedure are provided in Appendix B; we intentionally use an extremely large sample size to yield negligible standard errors, so for comparisons, we treat sample averages as essentially equal to the true means.

The findings for customer surplus are straightforward. With simulated subjects, the surplus is higher with compensation than under FCFS or the base model; however, due to sub-optimal choices, the improvement with compensation above FCFS is smaller than in equilibrium, and the gap in surplus for the base model versus FCFS is larger than in equilibrium. Hereafter, we focus on the other two performance measures, namely aggregate waiting cost and provider revenue.

We first consult Figure 5,[11] which plots the aggregate waiting cost against the priority fee p for the base model, compensation model, and FCFS. There are similar features in the equilibrium (left panel) and with simulated subjects (right panel). In both cases, there is a region of low priority fees where compensation achieves lower aggregate waiting cost than the base model; the aggregate waiting cost follows a broadly decreasing-then-increasing pattern; and the waiting-cost-minimizing fees for the base and compensation models are roughly the same for simulated subjects as in the PBE. The first point aligns with Observation 1 from Section 6, although the range of fees where the observation holds is

[10] In the One-Human Study we have data for 100 independent observations (50 participants in two treatments make 100 decisions each, each participant representing an independent observation). If we were to collect data for 100 priority purchase decisions in all-human queues, we would have needed 1,000 participants, and because each complete queue includes 10 people, each person making a decision and waiting for the other nine, each replication would have taken 10 times longer. Therefore, each participant would have been able to make only 10 instead of 100 decisions, which would have significantly limited the range of priority costs and sequence numbers each person would have been exposed to.

[11] In order to effectively depict the respective curvatures, different vertical axis scales are used in the two panels.

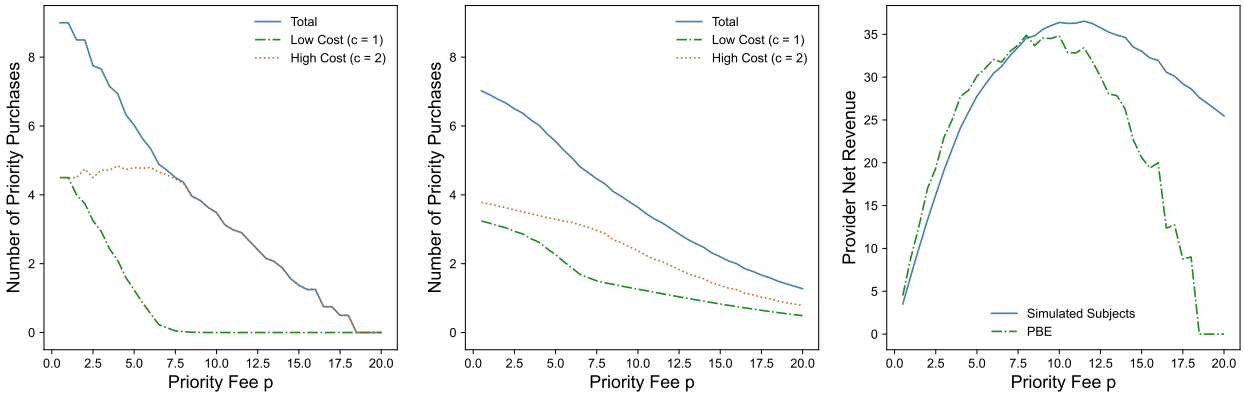


(a) Equilibrium (b) Simulated Subjects

Figure 5 Aggregate waiting costs ($N = 10$, waiting cost $C_i \in \{1, 2\}$).

much smaller with simulated subjects than in equilibrium, and the difference is also much smaller even in this range. Moreover, in line with Observation 3, with simulated subjects as in equilibrium, we find that when the priority fee can be optimized, the best-case aggregate waiting cost is lower in the base model than with compensation. Thus, overall, compensation in a priority service system appears not to offer the short-term social welfare benefit that we might hope for given its ability in equilibrium to filter low-cost customers out of the priority queue. However, we conjecture that a firm still might choose to implement it for other reasons related to customer relations, e.g., to combat the negative customer response stoked by priority queues that we document in the introduction. In this case, our results and insights can aid managers in assessing how customers' priority purchasing decisions (and the system performance) will change if compensation is implemented.

Interestingly, at the socially optimal fee and for most fees considered, the magnitude of improvement in aggregate waiting cost compared to FCFS is much smaller with simulated subjects (4.5% at optimal for base model) than in the PBE (13.0%): observe the differing scales on Figure 5's vertical axis. Why is this? As discussed (see, e.g., Table 5), low-cost customers represent a significantly greater proportion of priority purchases in our experiments than in equilibrium. Accordingly, compare Figure 6(a) (equilibrium) and (b) (simulated subjects) for the base model: as the fee increases, the number of equilibrium low-cost priority purchases quickly drops to near zero, but low-cost subjects continue to buy priority across the whole range of fees. The pattern is decreasing as expected, but at a much shallower rate, so that apart from very low fees, there are far more low-cost customers in the priority queue relative to the equilibrium. A similar phenomenon occurs in the compensation case (plots omitted for brevity). These sub-optimal decisions lead to higher aggregate waiting cost. Low-cost customers buying priority is always socially undesirable, and our subjects do this even more than in equilibrium; this behavioral regularity thwarts the priority mechanism because high-cost customers cannot overtake a low-cost customer if that customer bought priority earlier in the sequence.



(a) Purchases: Equilibrium (b) Purchases: Simulated Subjects (c) Provider Revenue

Figure 6 Priority purchases and provider revenue (both averages) for PBE vs. simulated subjects in base model.

On the other hand, customers overbuying priority at high fees has another important ramification, one that may be welcomed by some service providers. As the fee increases, the number of priority purchases decreases both with simulated subjects and in equilibrium. However, the decrease is shallower with simulated subjects, both for low-cost customers, as noted above, and also for high-cost customers (Figure 6 (a) vs. (b)).[12] Thus, for high fees, the total number of priority purchases is higher than in equilibrium, i.e., *customers tolerate higher priority fees than they rationally should*. In other words, customers' behavioral deviations allow the provider to raise the price of priority higher than it would in equilibrium and extract even more priority revenue. Turning now to Figure 6 (c), we see that not only is the revenue-optimal priority fee meaningfully higher with simulated subjects ($11.50) than in equilibrium ($8), but also the maximum revenue is nearly 5% higher.

In short, behavioral deviations lead to increased revenue but worse (i.e., higher) aggregate waiting cost. For performance-critical or publicly-operated systems where aggregate waiting cost is paramount, our results imply that service providers should attempt to nudge customers closer to the equilibrium priority-purchasing strategies, especially low-cost customers. On the other hand, customers' behavioral deviations will lead them to tolerate high priority fees and continue to buy priority. So, a service provider with revenue as its primary objective can price accordingly and let the cash roll in.

8. Conclusion

Priority service systems are increasingly prevalent, and different providers operate them differently and charge varying prices. Also, compensation of inconvenienced customers has shown promise, but its operation in priority systems is not well understood. We have illuminated how to operate a priority service system (with or without compensation) through a theoretical and behavioral lens.

[12] This finding is also consistent with actual subject decisions in the One-Human Study. For priority fees of $10.50 or more, priority was optimal (i.e., the equilibrium choice) in only 11% (568 of 5,062) of decision opportunities, but subjects purchased priority in more than 26% (1,334 of 5,062) of these (the findings are similar if we break out base vs. compensation or low vs. high cost: in each group, subjects purchase priority more frequently than in equilibrium).

We have studied a model of a priority service system designed to be tested in the lab, with two variants: one a traditional priority system (base model), and another in which priority payments are partially or fully redistributed as compensation to regular-queue customers (compensation model). Theoretically, we derived structural results for how customers make decisions and uncovered key strategic drivers. The equilibrium in both the base and compensation model follows a cost-dependent, increasing-threshold structure. Also, because compensation makes the priority queue less attractive, customers in the compensation model have lower thresholds than those in the base model, all else equal. Due to our sequential setup and the huge cross-product of possible equilibrium strategies given customers' random waiting costs, our theoretical results required careful and innovative sample-path analysis to reveal the equilibrium's structure without handling each scenario separately. Additionally, we analyzed system-wide performance measures. We found that for a low priority fee, compensation can actually reduce aggregate waiting cost by filtering low-cost customers out of the priority queue; however, this finding does not hold for higher priority fees or when the priority fee is optimized.

We then took our model and its predictions to the laboratory. Directionally, subject behavior in our experiments was in line with predictions, but some key behavioral deviations have significant consequences for service providers. We found that subjects with low waiting costs represented a greater proportion of priority purchases than the equilibrium predicted. This behavioral regularity prevents the priority mechanism from achieving its full social potential, such that fewer high-cost customers are able to overtake low-cost customers, leading to higher aggregate waiting cost than with fully rational customers. Additionally, we found that compensation, despite its promise, failed to deliver. A traditional priority system outperformed it in terms of aggregate waiting cost both in equilibrium and in simulations based on our experiments. However, we also identified a notable benefit (to service providers) of customers' behavioral deviations: at high fees, more customers purchased priority than predicted. This irrationally high willingness to pay for priority potentially allows service providers to extract more revenue from customers than they could if customers were fully rational. In deciding whether to attempt to influence customers toward rational decisions, a provider thus faces a tradeoff: the status quo (with behavioral deviations) brings more revenue but worse social efficiency. The way forward is then dictated by the relative importance ascribed to these two measures.

Future behavioral queueing research has much to investigate, including a few avenues suggested by our work and its limitations. First, all subjects in our experiments received the service, so we did not separate willingness to pay for priority (which is principally dependent on the waiting cost) from willingness to pay for the service itself. In practice, these could be statistically dependent; this could affect the propensity of customers with different waiting costs to purchase priority, and an experimental study incorporating these features could be a fruitful direction for future research. Second, our experiments imposed specific (monetary) waiting costs on subjects. This approach gives

the experimenter the most control and follows the convention in the literature (see, e.g., Section 5 of Allon and Kremer 2018), but it has its drawbacks, which present a shared challenge for most queueing-based experiments. Specifically, individuals likely possess their own intrinsic waiting costs, and imposing monetary waiting costs on subjects does not facilitate inference about these intrinsic costs. At the same time, intrinsic waiting costs are not only difficult to infer but may also be non-linear in time (and customer preferences may depend both on the notion of "cost" and on the customer's *perception* of elapsed time). Some past work has attempted to understand the functional form of waiting disutility (see Section 2.1 of Allon and Kremer 2018 for a survey); an exciting but daunting challenge for future work is to successfully harmonize and disentangle the competing goals of (i) experimentally testing a theoretical queueing model in a controlled environment and (ii) incorporating or inferring subjects' intrinsic waiting preferences. Third and finally, our experiments comprised a One-Human Study and an All-Human Study, each offering its own (dis)advantages. The One-Human Study was conducted virtually and with one human player in a given instance of the game; this facilitated a large sample while retaining the essential features of the model. However, it also entailed subjects interacting with computerized agents playing the equilibrium strategy, which could be difficult for subjects to conceptualize. To make this notion more accessible, our instructions informed subjects that the computerized agents were "programmed to make decisions in a way that would maximize their earnings." The All-Human Study, on the other hand, involved human subjects in all positions of the game and required them to physically wait in queues; this was ideal for external validity but hampered data collection due to the time commitment and expense. For future behavioral queueing experiments, all-human studies with large samples would be the ideal solution but a difficult one to bring to fruition; innovative approaches to behavioral experiments may be needed to enable this, and field experiments could also complement such work.

Acknowledgments

The authors thank department editor Guillaume Roels, as well as the anonymous associate editor and referees, for their many valuable comments that led to a significantly improved paper. Additionally, the authors are grateful to Mannat Batish and Xiwen Zhang for assistance with laboratory experiments.

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 (2021) *Priority Queues and Skiing: A Slippery Slope*. `https://gadallon.substack.com/p/priority-queues-and-skiing-a-slippery`, Accessed 06/21/2023.

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).

Baggini J (2017) It's the end of the line for queueing. What's replaced it is far worse. *The Guardian* `https://www.theguardian.com/commentisfree/2017/aug/28/line-queueing-money-priority-access-dunkirk`, Accessed 07/04/2023.

Baraniuk C (2019) The big problem with short queues. *BBC* https://www.bbc.com/worklife/article/20190624-the-big-problem-with-short-queues, Accessed 07/04/2023.

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

CN Tower (2023) *Tickets — CN Tower*. https://www.cntower.ca/plan-your-visit/tickets-and-hours/tickets, Accessed 06/21/2023.

Cohen MC, Fiszer MD, Kim BJ (2022) Frustration-based promotions: Field experiments in ride-sharing. *Management Science* 68(4):2432–2464.

Cui S, Wang Z, Yang L (2023) *Innovative Priority Mechanisms in Service Operations: Theory and Applications* (Springer Nature).

Curiel I, Pederzoli G, Tijs S (1989) Sequencing games. *EJOR* 40(3):344–351.

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

Dumas B, Rumpf S (2022) Has the cost of Disney World become unaffordable for the average American family? *Fox Business* https://www.foxbusiness.com/lifestyle/cost-disney-world-unaffordable-average-american-family, Accessed 07/28/2022.

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.

Gurvich I, Lariviere MA, Ozkan C (2019) Coverage, coarseness, and classification: Determinants of social efficiency in priority queues. *Management Science* 65(3):1061–1075.

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).

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

Killington (2021) *Fast Tracks is your all-day upgrade to Killington's new express lift lines*. https://www.killington.com/plan-your-trip/premium-experiences/fast-tracks, Accessed 07/28/2022.

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.

Le Seur A (2012) On queuing and queue jumping. *UK Constitutional Law Association* https://ukconstitutionallaw.org/2012/06/14/andrew-le-sueur-on-queuing-and-queue-jumping/, Accessed 07/04/2023.

Lynch K (2021) *Disney Genie and Disney Genie+ service coming to Disneyland Resort beginning Dec. 8*. https://disneyparks.disney.go.com/blog/2021/12/disney-genie-and-disney-genieplus-service-coming-to-disneyland-resort-beginning-dec/, Accessed 07/28/2022.

Maehrer A (2021) *Disney Genie launching Oct. 19 at Walt Disney World Resort: Create your best Disney day.* https://disneyparks.disney.go.com/blog/2021/10/disney-genie-launching-at-walt-disney-world-resort/, Accessed 07/28/2022.

Martin W, Sibley J (2021) *A letter to the community regarding Fast Tracks*. https://www.powdr.com/news/powdr-news/a-letter-to-the-community-regarding-fast-tracks, Accessed 07/28/2022.

Pinedo ML (2012) *Scheduling: Theory, Algorithms, and Systems* (Springer).

Rockefeller Center (2023) *Buy Tickets — Rockefeller Center & NYC Tourist Passes*. https://www.rockefellercenter.com/buy-tickets/, Accessed 06/21/2023.

Roet-Green R, Shetty A (2022) On designing a socially optimal expedited service and its impact on individual welfare. *Manufacturing & Service Operations Management* 24(3):1843–1858.

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.

SkyDeck (2023) *SkyDeck Chicago Tickets: Willis Tower Ledge Price*. https://theskydeck.com/tickets-shop/general-pricing, Accessed 06/21/2023.

Universal Orlando Resort (2023) *Universal Express Pass for Theme Parks*. https://www.universalorlando.com/web/en/us/tickets-packages/express-passes, Accessed 06/16/2023.

Waite A (2021) Fast track to criticism: Express-lane product at Killington garners complaints, curiosity in Capital Region. *The Daily Gazette* `https://dailygazette.com/2021/11/07/fast-track-to-criticism-express-lane-product-at-killington-garners-complaints-curiosity-in-capital-region/`, Accessed 06/21/2023.

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.

E-Companion to:

Paid Priority in Service Systems: Theory and Experiments

Andrew E. Frazelle, Elena Katok

Appendix

A. Additional Tables

Table EC.1 Cost profiles in the All-Human Study.

	Sequence									
Profile	1	2	3	4	5	6	7	8	9	10
1	H	L	H	H	H	L	L	H	L	H
2	H	H	L	L	L	L	H	L	H	H
3	H	L	L	L	L	L	H	H	H	L
4	H	L	L	L	L	L	L	H	L	L
5	L	L	H	L	L	L	L	L	L	H

Table EC.2 Equilibrium paths and summary statistics in equilibrium for the All-Human Study.

	Base Model ($\gamma = 0$):										Priority			Regular		
Profile	PBE Path (1: Pri., 0: Reg.)										Length	Low	High	Length	Low	High
1	1	1	1	1	1	0	1	1	0	1	8	2	6	2	2	0
2	1	1	1	1	1	0	1	0	1	1	8	3	5	2	2	0
3	1	1	1	1	1	0	1	1	1	0	8	4	4	2	2	0
4	1	1	1	1	1	0	1	1	0	0	7	5	2	3	3	0
5	1	1	1	1	1	0	1	0	1	1	8	6	2	2	2	0
										Average:	7.8	4	3.8	2.2	2.2	0
	Comp. Model ($\gamma = 1$):										Priority			Regular		
	PBE Path (1: Pri., 0: Reg.)										Length	Low	High	Length	Low	High
1	1	0	1	1	1	0	0	1	0	1	6	0	6	4	4	0
2	1	1	0	0	0	0	1	1	1	1	6	1	5	4	4	0
3	1	0	0	0	1	0	1	1	1	0	5	1	4	5	5	0
4	1	0	0	0	1	0	1	1	0	1	5	3	2	5	5	0
5	0	0	1	0	1	0	1	1	0	1	5	3	2	5	5	0
										Average:	5.4	1.6	3.8	4.6	4.6	0

B. Details of Simulation for One-Human Study

In the simulations, we considered the same sets of parameters tested in the One-Human Study, i.e., $N = 10$, compensation fraction $\gamma \in \{0, 1\}$, priority fee p between \$0.50 and \$20 in increments of \$0.50, and waiting costs for each customer (human or computerized) drawn IID from \$1 or \$2 per service with equal probability. In order to obtain estimates with negligible standard errors, we conducted 5 million replications for each combination of priority fee and compensation fraction. Given the huge sample size, for purposes of comparison we treat sample averages as essentially equal to the true means. For each replication, we randomly generated the vector of 10 waiting costs. As mentioned, these waiting costs were IID across customers within a replication; however, the waiting-cost vectors for a given replication were coupled across different values of the priority fee and compensation fraction; that is, the vector used in the j-th replication for $p = \$0.50$ and $\gamma = 0$ was also used in the j-th replication for all other combinations of p and γ. We also used the same 5 million randomly drawn waiting-cost vectors to compute equilibrium and FCFS performance measures.

Table EC.3 Observed paths and summary statistics in the All-Human Study.

	Base Model ($\gamma=0$):										Priority			Regular		
Profile	Observed (1: Pri., 0: Reg.)										Length	Low	High	Length	Low	High
1	1	0	0	1	1	1	1	1	1	0	7	3	4	3	1	2
2	1	1	1	1	1	0	1	1	0	0	7	4	3	3	1	2
3	1	1	0	1	1	0	1	0	1	1	7	4	3	3	2	1
4	1	1	1	0	1	1	1	0	1	1	8	7	1	2	1	1
5	1	1	1	0	0	1	1	0	0	1	6	4	2	4	4	0
										Average:	7	4.4	2.6	3	1.8	1.2
	Comp. Model ($\gamma=1$):										Priority			Regular		
Profile	Observed (1: Pri., 0: Reg.)										Length	Low	High	Length	Low	High
1	1	1	0	1	0	1	0	0	0	1	5	2	3	5	2	3
2	1	1	1	1	0	1	0	0	0	1	6	3	3	4	2	2
3	1	1	0	1	1	0	0	1	0	1	6	4	2	4	2	2
4	1	1	0	1	1	0	0	1	0	1	6	4	2	4	4	0
5	1	1	1	0	1	0	0	1	1	0	6	5	1	4	3	1
1	0	1	1	1	1	1	0	1	0	1	7	2	5	3	2	1
2	0	0	1	1	1	1	0	1	0	1	6	5	1	4	0	4
3	1	1	0	0	1	1	0	1	1	0	6	3	3	4	3	1
4	1	1	0	1	0	0	1	1	0	1	6	4	2	4	4	0
5	1	1	1	0	0	0	1	1	1	0	6	5	1	4	3	1
										Average:	6	3.7	2.3	4	2.5	1.5

We refer to a particular combination of replication number (and the associated waiting-cost vector), priority fee, and compensation fraction (e.g., fifth replication, $p=4.50$, $\gamma=1$) as an *instance*. The priority queue length facing the i-th customer in a given instance was determined from the simulated actions of the $i-1$ earlier customers in the instance, and given the priority fee and the customer's waiting cost and sequence number, the equilibrium decision was determined by comparing the priority queue length with the customer's PBE threshold. We then calculated the predicted probability of priority purchase using the logit models (4) from Tables 6 and 7 for $\gamma=0$ and 1, respectively. The customer's decision (regular or priority) was then realized from a Bernoulli distribution with the calculated probability of priority purchase.

C. 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 $(c_1,\ldots,c_N)$, a focal customer i, and $k\in\{0,\ldots,i-2\}$. 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 $j\in\{i+1,\dots,N\}$, we then have

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

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

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

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

$$\ell_i^k = i+\beta+1 = \ell_i^{k+1}+1, \tag{EC.2}$$

where the difference of 1 between ℓ_i^k and ℓ_i^{k+1} 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 (EC.1) and (EC.2) imply the bounds $0\le\ell_i^k-\ell_i^{k+1}\le 1$. 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 $i\in\{1,\dots,N-1\}$, and suppose that all customers $j\in\{i+1,\dots,N\}$ use some cost-dependent threshold strategies $\bar{x}_{j-1}(C_j)$. 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 $j\in\{i+1,\dots,N\}$, 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_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; the exact value of L_i^k 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

$$\mathsf{E}[U_{R,i}(k\,;\, \bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V - c_i \mathsf{E}[L_i^k] > V - p - c_i(k+1) = U_{P,i}(k). \tag{EC.3}$$

By Lemma 1, we then have

$$\begin{aligned} \mathsf{E}[U_{R,i}(k+1\,;\, \bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V - c_i \mathsf{E}[L_i^{k+1}] &\ge V - c_i \mathsf{E}[L_i^k] \\ &> V - p - c_i(k+1) \\ &> V - p - c_i(k+2) \\ &= U_{P,i}(k+1), \end{aligned} \tag{EC.4}$$

where the inequality on the second line holds by equation (EC.3). 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 $k \le x_{i-1} \le i-1$. 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. □

D. Supplementary Result and Proof for Theorem 2

LEMMA EC.1. *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, 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. For $k \in \{0, \ldots, i-1\}$, we have*

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

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: $\bar{x}_i(c)<k$. 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

$$\ell_{i+1}^k=\ell_i^k+1. \tag{EC.5}$$

Case 2: $\bar{x}_i(c)\geq k$. 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 $\bar{x}_j(c_{j+1})\neq x_{j,i+1}^k$ 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_{i+1}^k letting customer $i+1$ contemplate choosing the regular queue even though the strategy $\bar{x}_i(c_i)$ prescribes priority, we have

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

where $\ell_i^k=i+1+\alpha$ 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 $\bar{x}_j(c_{j+1})=x_{j,i+1}^k$ 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 $\ell_i^k=i+1+(\alpha-1)=i+\alpha$ and $\ell_{i+1}^k=i+1+\alpha$, which implies

$$\ell_{i+1}^k=\ell_i^k+1. \tag{EC.7}$$

Combining equations (EC.5), (EC.6), and (EC.7) gives

$$0\leq\ell_{i+1}^k-\ell_i^k\leq 1,$$

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

$$U_{P,i}(k)=V-p-c(k+1)\geq V-c\mathsf{E}[L_i^k]=\mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}}^*)]. \tag{EC.8}$$

Lemma EC.1 and equation (EC.8) then imply that

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

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 $\bar{x}_i^*(c)\le\bar{x}_j^*(c)$ for $i<j$.

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

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

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 $\bar{x}_{i-1}^*(c)\le\bar{x}_{i-1}^*(c')$. □

E. 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 $\alpha^{k+1}\ge\alpha^k-1$, which implies

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

Taking expectation over the waiting costs gives $\mathsf{E}[g_i^\gamma(k)]\le\mathsf{E}[g_i^\gamma(k+1)]$, 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 $0\le k\le i-1$ 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 $x_{i-1}<k$. 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 $k\le i-2$.

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

$$\mathsf{E}[U_{R,i}(k\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V + \mathsf{E}[g_i^\gamma(k)] - c_i\mathsf{E}[L_i^k] > V - p - c_i(k+1) = U_{P,i}(k). \tag{EC.10}$$

Lemmas 1 and 2 then imply

$$\begin{aligned}\mathsf{E}[U_{R,i}(k+1\,;\,\bar{\boldsymbol{x}}_{\boldsymbol{i,N-1}})] = V + \mathsf{E}[g_i^\gamma(k+1)] - c_i\mathsf{E}[L_i^{k+1}] &\geq V + \mathsf{E}[g_i^\gamma(k)] - c_i\mathsf{E}[L_i^k] \\ &> V - p - c_i(k+1) \\ &> V - p - c_i(k+2) \\ &= U_{P,i}(k+1),\end{aligned} \tag{EC.11}$$

where the inequality on the second line holds by equation (EC.10).

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

F. Supplementary Result and Proof for Theorem 4

As with Lemma 1, we note that Lemma EC.1 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 EC.2. *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*

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

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 EC.1, for any vector $(c_1,\dots,c_N)$ of waiting-cost realizations, we have $\alpha_i^k \geq \alpha_{i+1}^k - 1$.

Case 1: $\bar{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{c_{i+1}}) < \boldsymbol{k}$. 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: $\bar{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{c_{i+1}}) \geq \boldsymbol{k}$. In this case, customer $i+1$ will purchase priority upon observing $x_i=k$. By arguments in the proof of Lemma EC.1, 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

$$g^{\gamma}_{i+1}(k)=\frac{p(k+\alpha^k_{i+1})}{N-(k+\alpha^k_{i+1})}<\frac{p(k+1+\alpha^k_{i+1})}{N-(k+1+\alpha^k_{i+1})}=\frac{p(k+1+\alpha^k_{i})}{N-(k+1+\alpha^k_{i})}=g^{\gamma}_{i}(k).$$

If instead $\alpha^k_{i+1}=\alpha^k_i+1$, then we have

$$g^{\gamma}_{i+1}(k)=\frac{p(k+\alpha^k_{i+1})}{N-(k+\alpha^k_{i+1})}=\frac{p(k+1+\alpha^k_{i})}{N-(k+1+\alpha^k_{i})}=g^{\gamma}_{i}(k),$$

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^k_i+1 priority purchases in total.

We conclude that for any waiting-cost realizations and their corresponding thresholds, we have $g^{\gamma}_{i+1}(k)\le g^{\gamma}_i(k)$. 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

$$U_{P,i+1}(k)=V-p-c(k+1)<V+\mathsf{E}[g^{\gamma}_{i+1}(k)]-c\mathsf{E}[L^k_{i+1}]=\mathsf{E}[U_{R,i+1}(k\,;\,\bar{\boldsymbol{x}}^{*}_{\boldsymbol{i+1,N-1}})]. \tag{EC.12}$$

Lemmas EC.1 and EC.2 and equation (EC.12) then imply that

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

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 $\bar{x}^*_i(c)\le\bar{x}^*_j(c)$ for $i<j$.

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

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

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 $\bar{x}^*_{i-1}(c)\le\bar{x}^*_{i-1}(c')$. □

G. Supplementary Result and Proof for Theorem 5

LEMMA EC.3. *Take $i \in \{1,\dots,N-1\}$, and for customers $i+1,\dots,N$, 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 $(c_{i+1},\dots,c_N)$, suppose that $\bar{x}'_j(c_j) \le \bar{x}_j(c_j)$ for all $j \in \{i+1,\dots,N\}$. Let α_i^k ($\tilde{\alpha}_i^k$) be the number of priority purchases among customers $i+1,\dots,N$ if $x_{i-1}=k$, customer i chooses the regular queue, and the thresholds are $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}(c_{i+1},\dots,c_N)$ ($\bar{\boldsymbol{x}}'_{\boldsymbol{i,N}}(c_{i+1},\dots,c_N)$). For $k \in \{0,\dots,i-1\}$, we have*

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

Proof. Consider the thresholds $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}(c_{i+1},\dots,c_N)$. 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 $j' \in \{i+1,\dots,N\}$, consider also the vector of thresholds obtained from $\bar{\boldsymbol{x}}_{\boldsymbol{i,N}}(c_{i+1},\dots,c_N)$ 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 each customer $j \in \{i+1,\dots,j'-1\}$, 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 $x_{j'-1}^{k(-)} = x_{j'-1}^k \ne \bar{x}_{j'-1}(c_{j'})$, then either $\bar{x}_{j'-1}(c_{j'})-1 < \bar{x}_{j'-1}(c_{j'}) < x_{j'-1}^k = x_{j'-1}^{k(-)}$, or $x_{j'-1}^{k(-)} = x_{j'-1}^k \le \bar{x}_{j'-1}(c_{j'})-1 < \bar{x}_{j'-1}(c_{j'})$. In either case, customer j' takes the same action for either vector of thresholds. In this case, we will also have $x_{j-1}^k = x_{j-1}^{k(-)}$ for $j \in \{j'+1,\dots,N\}$, so these customers also will take the same actions under either vector of thresholds. Thus, we have

$$\alpha_i^k = \tilde{\alpha}_i^k. \tag{EC.13}$$

If instead $x_{j'-1}^{k(-)} = x_{j'-1}^k = \bar{x}_{j'-1}(c_{j'})$, 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: $\boldsymbol{x_{j-1}^{k(-)} \ne \bar{x}_{j-1}(c_j)}$ for all $\boldsymbol{j \in \{j'+1,\dots,N\}}$. In this case, because of customer i's different action, we have $x_{j'}^k = x_{j'}^{k(-)}+1$. So, similar to the above for customer j', by the hypothesis of this case, for customer $j'+1$, we either have $x_{j'}^{k(-)} < x_{j'}^k = x_{j'}^{k(-)}+1 \le \bar{x}_{j'}(c_{j'+1})$, or $\bar{x}_{j'}(c_{j'+1}) < x_{j'}^{k(-)} < x_{j'}^{k(-)}+1 = x_{j'}^k$. 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 $x_{j-1}^k = x_{j-1}^{k(-)}+1$ for all $j \in \{j'+2,\dots,N\}$. Therefore, customers $j'+2,\dots,N$ 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 $j \in \{i+1,\dots,N\}$ when customer j' has a decreased threshold, so we have

$$\alpha_i^k = \tilde{\alpha}_i^k + 1. \tag{EC.14}$$

Case 2: $\boldsymbol{x_{j''-1}^{k(-)} = \bar{x}_{j''-1}(c_{j''})}$ for some $\boldsymbol{j'' \in \{j'+1,\dots,N\}}$. We have $x_{j-1}^k = x_{j-1}^{k(-)}+1$ for $j \in \{j'+1,\dots,j''\}$ by the same reasoning as in Case 1 because of customer i's different actions under the two threshold vectors. Customers $j \in \{j'+1,\dots,j''-1\}$ 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 $j \in \{i+1,\dots,N\}$ 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

$$\alpha^k_i = \tilde{\alpha}^k_i. \tag{EC.15}$$

Combining equations (EC.13), (EC.14), and (EC.15) gives $\tilde{\alpha}^k_i \le \alpha^k_i$. By induction, we can successively reduce the thresholds customer by customer and in increments of 1 until we reach $\bar{\boldsymbol{x}}'_{\boldsymbol{i+1,N}}(c_{i+1},\dots,c_N)$. Because $\tilde{\alpha}^k_i \le \alpha^k_i$ at every step of this process, we have the desired result. □

Proof of Theorem 5. Consider a customer $i \in \{1,\dots,N-1\}$, and suppose that $\bar{x}^*_{j-1,\gamma}(c_j) \le \bar{x}^*_{j-1}(c_j)$ for $j \in \{i+1,\dots,N\}$ and all c_j in the support of C_j. For customers $i+1,\dots,N$, consider a given sample path of waiting costs $(c_{i+1},\dots,c_N)$. 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 $i+1,\dots,N$, 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 EC.3 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

$$\mathsf{E}[A^k_{i,\gamma}] \le \mathsf{E}[A^k_i]. \tag{EC.16}$$

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 (EC.16) also implies

$$\mathsf{E}[L^k_{i,\gamma}] = i + \mathsf{E}[A^k_{i,\gamma}] \le i + \mathsf{E}[A^k_i] = \mathsf{E}[L^k_i]. \tag{EC.17}$$

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 $0 < \gamma \leq 1$. Under our hypothesis that $\bar{x}^*_{j-1,\gamma}(c_j) \leq \bar{x}^*_{j-1}(c_j)$ for customers $j \in \{i+1, \ldots, N\}$ and all c_j in the support of C_j, this implies that also $\bar{x}^*_{i-1,\gamma}(c_i) \leq \bar{x}^*_{i-1}(c_i)$ 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 $p \leq \underline{c}(N-1)$. 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 $g^\gamma_i(k) \leq g^{\gamma'}_i(k)$. These two relations make the regular queue more attractive as the compensation fraction increases, so the optimal threshold decreases in the compensation fraction. □

H. 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 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 $(\bar{x}_i^m, \ldots, \bar{x}_{N-1}^m) \in \{ \times_{j=i}^{N-1} \{-1, 0, 1, \ldots, j\}\}$ **do**
    $\pi_m \leftarrow \prod_{k=i+1}^{N} \int \mathbb{1}\{\bar{x}_{k-1}^*(c) = \bar{x}_{k-1}^m\} \mathrm{d}F(c)$ `// 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`
    $\bar{x}_{i-1}^*(c_i) \leftarrow \bar{x}_{i-1}^*(c_i) + \mathbb{1}\{c_i \geq p/(\lambda_i^k - (k+1))\}$ 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 $(\bar{x}_i^m, \ldots, \bar{x}_{N-1}^m) \in \{ \times_{j=i}^{N-1} \{-1, 0, 1, \ldots, j\}\}$ **do**
    $\pi_m \leftarrow \prod_{k=i+1}^{N} \int \mathbb{1}\{\bar{x}_{k-1}^*(c) = \bar{x}_{k-1}^m\} \mathrm{d}F(c)$ `// 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`
    $\bar{x}_{i-1}^*(c_i) \leftarrow \bar{x}_{i-1}^*(c_i) + \mathbb{1}\{c_i \geq (p + \rho_i^k)/(\lambda_i^k - (k+1))\}$ for c_i in support of C_i `// If priority is preferred at current` k`, increment previous threshold`
  end
end

I. Performance Measure Definitions

In this appendix, we formally define each of the performance measures considered in Sections 6 and 7. As in Appendix H, the exposition is for a continuous waiting-cost distribution, and the corresponding quantities for a discrete distribution can be obtained in the natural way.

The starting point for calculating performance measures in equilibrium is the output from Algorithm 1 or 2, namely a vector $\bar{\boldsymbol{x}}^*$ of PBE threshold functions. Let L_i^* be the random variable for the number of services (including her own) that customer i waits through in the PBE. Furthermore, for waiting-cost realization vector $(c_1,\ldots,c_N)\in \mathrm{supp}(C_1,\ldots,C_N)$, let $\ell_i^*(c_1,\ldots,c_N)$ be the realization for L_i^* associated with the threshold vector $\big(\bar{x}_0^*(c_1),\ldots,\bar{x}_N^*(c_N)\big)$. Also, let A^* be the random variable for the total number of priority purchases in the PBE, and let $\alpha^*(c_1,\ldots,c_N)$ be the realization for A^* associated with the threshold vector $\big(\bar{x}_0^*(c_1),\ldots,\bar{x}_N^*(c_N)\big)$. For each $(c_1,\ldots,c_N)$, both ℓ_i^* and α^* can be determined by simple bookkeeping.

Recalling that $\gamma=0$ in the base model and $\gamma>0$ in the compensation model, we have the following definitions that apply to both models.

Definition EC.1 (Aggregate Waiting Cost). The expected aggregate waiting cost C_{Agg} is

$$C_{\mathrm{Agg}}=\mathsf{E}\Big[\sum_{i=1}^N C_iL_i^*\Big]=\int\limits_{\times_{j=1}^N\mathrm{supp}(C_j)}\cdots\int\Big(\sum_{i=1}^N c_i\ell_i^*(c_1,\ldots,c_N)\Big)\mathrm{d}F(c_1)\cdots\mathrm{d}F(c_N).$$

Definition EC.2 (Customer Surplus). The expected customer surplus S is

$$S=VN-C_{\mathrm{Agg}}-p\gamma\mathsf{E}[A^*]=VN-C_{\mathrm{Agg}}-p\gamma\int\limits_{\times_{j=1}^N\mathrm{supp}(C_j)}\cdots\int\alpha^*(c_1,\ldots,c_N)\mathrm{d}F(c_1)\cdots\mathrm{d}F(c_N).$$

Definition EC.3 (Provider Net Revenue). The expected provider net revenue Z from priority purchases (i.e., after subtracting compensation payments) is

$$Z=p(1-\gamma)\mathsf{E}[A^*]=p(1-\gamma)\int\limits_{\times_{j=1}^N\mathrm{supp}(C_j)}\cdots\int\alpha^*(c_1,\ldots,c_N)\mathrm{d}F(c_1)\cdots\mathrm{d}F(c_N).$$

The versions of these measures used in Section 7 for the experiments and logit simulations are the analog of the above for sample averages: for each instance of the game, we compute the measures based on the waiting-cost realization vector and the path of play, and we then take the average across the instances. For fair comparisons, when computing equilibrium and FCFS measures in Section 7, we also use sample averages, computed with the same set of waiting-cost realization vectors as in the experiment or simulation.

J. Laboratory Instructions

The instructions below are for the compensation treatment ($\gamma = 1$) in the All-Human Study (for the sessions where minimum and maximum compensation were displayed). The instructions for the other treatments are similar to these; they are omitted due to space constraints but are available from the authors upon request.

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 Participant code. Please use your phone to go to

https://utd.sophielabs.net

and type in your participant code to log into the software. You will see the informed consent form. Please read and sign it. Once the experimenter starts the game you will see a screen that has your sequence number. The sequence numbers were generated randomly. Also on the screen is your personal waiting cost, which is \$0.50 or \$1.50. Waiting costs were also generated randomly and \$0.50 and \$1.50 are equally likely. There are 10 people in the room. Each person will start with \$15, called your endowment. Each person will be called in the order of his or her sequence number and will be asked to decide to either join the **Regular** Queue or purchase a spot in the **Priority** Queue. The priority queue costs \$1.50. The regular queue is free. 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.

Compensation depends on how many people purchase priority, and how many people join the regular queue. The minimum amount comes about if all remaining people join the regular queue. The maximum amount comes about if all remaining people purchase priority.

For example, suppose there are currently two people in the priority queue and two people in the regular queue, and the fifth player is deciding. If this player joins the regular queue, the minimum compensation happens if the remaining 5 people also join the regular queue:

$$\frac{\$1.50 \times 2}{2+1+5} = \frac{\$3}{8} = \$0.37$$

The maximum compensation happens if the remaining 5 purchase priority:

$$\frac{\$1.50 \times (2+5)}{2+1} = \frac{\$10.50}{3} = \$3.50$$

At the start of each round, we will show you the possible minimum and maximum compensation amounts, given the current composition of the two queues.

You will record your decision on your phone and stand to join your chosen queue. After both queues have been formed, the virtual service will start. Each service will take approximately 1 minute. 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 your waiting cost (either \$0.50 or \$1.50). After your service has been completed you will be paid your total earnings, calculated as follows.

\$5 participation fee + \$15 endowment - \$1.50 if you purchased Priority – (your waiting cost) x (the number of services you waited) + Compensation if you did not purchase Priority.

Figure EC.1 Instructions for compensation treatment in the All-Human Study (page 1)

Example:
Suppose your sequence number is 7 and your waiting cost is $1.50. 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, of the 3 remaining people behind you, 2 joined the priority queue. This means that once the service starts, there will be 3+2 = 5 people in the priority queue, and 5 people (including you) in the regular queue. Out of those 5 people in the regular queue, 3 are in front of you. This means that you will wait for 5+3+1 = 9 services. Your total waiting cost will be 9 x $1.50 = $13.50. Your Compensation will be ($1.50 x 5)/5 = $1.5. Your total earnings will be: $5 + $15 - $13.50 + $1.50 = $8.00

Now suppose that you chose to pay $1.50 and join the priority queue. In this case, your total waiting cost will be $1.50 x (3+1) = $6 because you only have to wait for the 3 Priority people in front of you, and your own service. Your total earnings will be: $5 + $15 -$1.50 - $6 = $12.50

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.

Decision Screen: **Final Screen:**

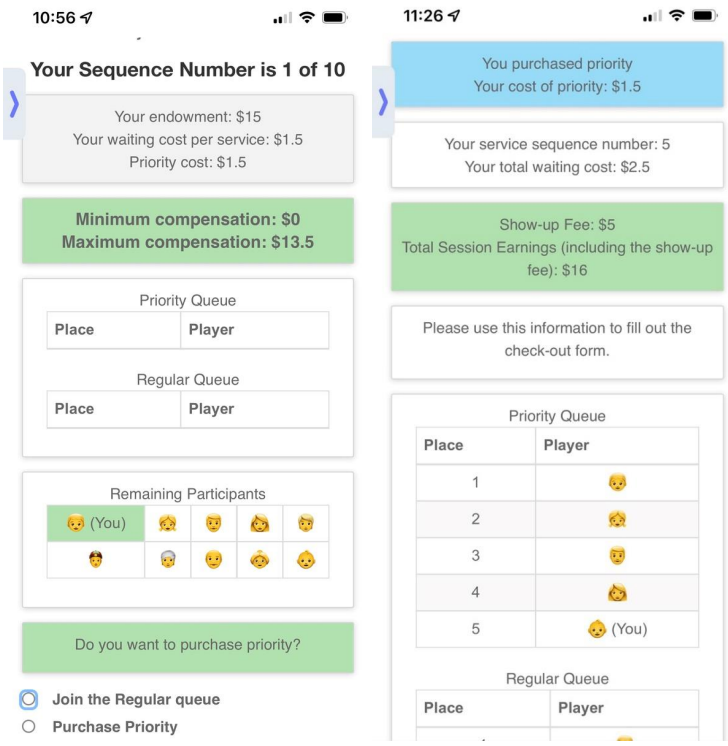


Figure EC.2 Instructions for compensation treatment in the All-Human Study (page 2)