Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Order Ahead for Pickup: Promise or Peril?

Ke Sun, Yunan Liu, Luyi Yang

To cite this article:
Ke Sun, Yunan Liu, Luyi Yang (2026) Order Ahead for Pickup: Promise or Peril?. Manufacturing & Service Operations Management

Published online in Articles in Advance 18 Feb 2026

. https://doi.org/10.1287/msom.2024.0865

https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Articles in Advance, pp. 1–20
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Order Ahead for Pickup: Promise or Peril?

Ke Sun,[a,b] **Yunan Liu,**[c,d] **Luyi Yang**[e,*]

[a] Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada; [b] College of Administrative Sciences and Economics, Koç University, Istanbul 34450, Türkiye; [c] Supply Chain Optimization Technology, Amazon, New York, New York 10001; [d] Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695; [e] Haas School of Business, University of California, Berkeley, Berkeley, California 94720

*Corresponding author

Contact: kelsey@bjtu.edu.cn, https://orcid.org/0000-0002-2095-929X (KS); yunanliu@amazon.com, https://orcid.org/0000-0001-9961-2610 (YL); luyiyang@haas.berkeley.edu, https://orcid.org/0000-0002-5370-6926 (LY)

Received: February 26, 2024
Revised: November 6, 2024; May 19, 2025; October 25, 2025
Accepted: October 29, 2025
Published Online in Articles in Advance: February 18, 2026

https://doi.org/10.1287/msom.2024.0865

Abstract. ***Problem definition*:** Mobile technologies have increasingly enabled remote customers to order ahead at quick-service restaurants. As customers travel to the service facility to pick up their orders, their orders also advance in the food preparation queue. It is widely believed that the ability to order ahead reduces customers' total delay and, therefore, allows restaurants to attract more orders and achieve higher throughput than if customers must order onsite. ***Methodology/results*:** We build a queueing game–theoretic model to study a mixed ordering scheme in which some customers order ahead and some order on-site. Our analysis shows that the common practice of accepting all orders as they come in and requiring all orders to be irrevocable can cause a mixed ordering scheme to surprisingly achieve lower throughput than an on-site–only ordering scheme. The throughput shortfall can persist even when the service provider freely chooses whether to share queue-length information with remote customers. However, if remote customers who order ahead are allowed to cancel unprepared orders when they arrive at the service facility, then such a mixed ordering scheme with cancellation achieves higher throughput than the on-site–only ordering scheme even though it does not uniformly dominate the original mixed ordering scheme (without cancellations). We then study a capped ordering scheme in which the service provider stops accepting new remote orders if the number of outstanding orders reaches a certain threshold. When the service provider optimally sets the cap, the mixed–capped ordering scheme outperforms both the mixed ordering scheme without capping and the on-site–only ordering scheme in throughput but not necessarily the cancellation scheme. Finally, we propose an optimal mechanism in which the service provider determines both the capping and cancellation thresholds subject to customers' individual rationality constraints. ***Managerial implications*:** Our paper highlights the unintended consequences of ordering ahead and provides prescriptive guidance for managing such a service system.

Funding: The work of K. Sun was supported by the Beijing Natural Science Foundation, China [Grant 9244031].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2024.0865.

Keywords: on-demand service • omnichannel service • order ahead • travel • reneging

1. Introduction

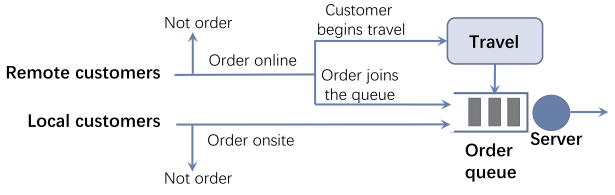
In today's on-demand economy, customers value instant gratification more than ever before and want to have their demand fulfilled as promptly as possible. In response, quick-service restaurants are increasingly enabling customers to order ahead on demand and pick up their orders at the restaurant. Online food ordering is projected to be a $106 billion industry by 2031 (Business Research Insights 2024), and the key ingenuity of ordering ahead is that it allows customer orders to virtually advance in the order processing queue while customers themselves physically travel to the restaurant. By the time they arrive, their orders are near completion or even ready for pickup. This parallel effect contrasts with the tandem nature of a traditional scheme in which customers must travel to the service facility and place orders only after they arrive on-site. Hence, ordering ahead is believed to reduce customers'

total delay, thereby attracting more orders and generating higher "transaction volume" (Pucci 2017) than if customers can only order on-site.

Ordering ahead of time usually also means commitment ahead of time. It is not uncommon for restaurants to require that all orders be final once they are placed. For example, Starbucks (2017) says in the FAQ about its Mobile Order & Pay that "once your order has been placed it cannot be delayed or canceled." Other restaurant chains (Peet's Coffee 2019, Subway 2020) have similar terms of use. Whereas ordering ahead features the lock-in effect, ordering on-site does not require pretravel commitment. Customers who order on-site can postpone their ordering decision until they arrive at the restaurant and see the status of the queue (i.e., how many people are waiting for their orders). Given that allowing customers to order ahead both attracts orders (with less delay) and retains orders (with a no-cancellation policy), one would naturally think that it would increase the restaurant's throughput. Our paper challenges this view.

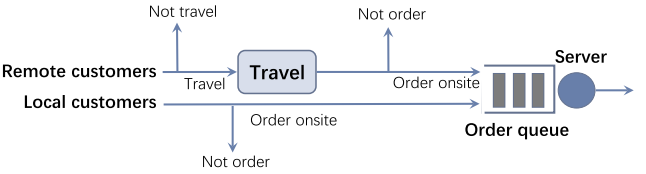
We develop a queueing game–theoretic model in which a service provider faces a mix of remote and local customers, both of whom are delay-sensitive. Remote customers are distanced from the service facility, and it takes time for them to travel to the facility, whereas local customers are nearby and their travel time is negligible. Upon experiencing a need, a remote (local) customer decides whether to place an irrevocable order ahead (on-site). The service provider operates in a mixed mode, taking both remote and on-site orders, and processing them following a first come, first served rule. Remote customers who order ahead travel to the service facility to pick up their orders. For simplicity, price is common across channels, and under the constant pricing assumption, maximizing profit is equivalent to maximizing throughput. We compare the above mixed ordering scheme (illustrated in Figure 1) with an on-site–only ordering scheme in which all customers may only order on-site (illustrated in Figure 2). In both schemes, customers see the number of outstanding orders on-site but not remotely. This is consistent with the practice that restaurants often put up digital screens in-store to show the list of outstanding orders but nevertheless only share with remote customers a wait-time estimate at best rather than the length of the order processing queue.

Figure 1. (Color online) A Mixed Ordering Scheme



Note. Ordering ahead has a parallel structure: as remote customers travel to the service facility, their orders are also advancing in the order-processing queue.

Figure 2. (Color online) An On-Site–Only Ordering Scheme



Note. Remote customers' decision process has a tandem structure: they need to travel to the service facility before placing an order.

Despite the additional convenience of online ordering offered to remote customers, the mixed ordering scheme may yield lower throughput than the on-site–only ordering scheme. This occurs when the market size is intermediate and travel time is short (Theorem 1). The key culprit for this throughput shortfall is the lock-in effect of ordering ahead, which is a double-edged sword. On the one hand, customer demand is secured early on, which increases the workload. On the other hand, remote customers who order ahead may unknowingly place and commit to orders when the queue is already long. This mismatch can increase utilization and exacerbate congestion, and this deters future demand and then reduces the future workload. In fact, when the market size is intermediate (which implies abundant orders and nontrivial congestion) and travel time is short (which implies a limited benefit from parallelization), the mixed ordering scheme winds up with lower throughput than the on-site–only ordering scheme.

To address the throughput shortfall, we study three mitigation strategies:

- Optimally sharing queue information with remote customers.
- Allowing cancellation of remote orders.
- Capping remote orders.

We find that the throughput shortfall can persist even after the first mitigation strategy is put in place (Theorem 2). The service provider's dilemma is that sharing such queue-length information with remote customers causes them to stop ordering when the queue is long but encourages them to order when it is small. When the market size is intermediate, the orders lost when the queue is long do not compensate for the orders gained when the queue is short, prompting the service provider to prefer not sharing information. Hence, the throughput shortfall cannot be eliminated by merely adjusting the information policy.

By contrast, we find that the second mitigation strategy—allowing remote customers to cancel their orders when they see their orders' queue positions upon arrival at the service facility—can eliminate the

throughput shortfall, enabling the mixed ordering scheme to outperform the on-site–only ordering scheme (analytically shown in Theorem 3 for a small buffer system and numerically confirmed more broadly). Allowing for cancellations is a self-regulating mechanism that does not deter remote customers with a long queue at the moment of ordering; rather, customers abandon on their own only if the queue is actually long when they have to wait on-site. Whereas allowing for cancellations in a mixed ordering scheme attains higher throughput than the on-site–only ordering scheme, it leads to lower throughput than the mixed ordering scheme without cancellation when the market size is small (Theorem 4). Hence, allowing for cancellations is not always a good idea.

Next, we turn to the third mitigation strategy, capping remote orders. In such a strategy, the service provider proactively stops accepting new remote orders when the number of existing outstanding orders reaches a threshold that is optimally determined by the service provider to maximize throughput. Remote order capping is practiced by some Starbucks stores that turn off point-of-sale systems used for mobile ordering when the stores are too busy (Dean 2021). The mixed–capped ordering scheme (with an optimal capping threshold) achieves higher throughput than both the mixed ordering scheme (without capping or cancellation) by construction and the on-site–only ordering scheme when queue-length information is not shared with remote customers (Theorem 5). As with information sharing, the mixed–capped ordering scheme regulates congestion by forgoing orders at the outset, but unlike information sharing, the threshold in the mixed–capped ordering scheme can be fine-tuned to ensure that inducing more customers to place orders does not come at the expense of letting go too many orders that have already been placed. We show that the mixed–capped ordering scheme generates a higher throughput than allowing for cancellations when the market size is small or large (Theorem 6) but numerically find that the opposite can be true when the market size is intermediate. This suggests that the deployment of operational strategies (capping orders or allowing for their cancellation) is contingent on the market conditions.

We further propose an optimal scheme, taking the perspective of a dictatorial service provider solving an admission control problem with three sets of levers: (i) whether to accept or cap online orders when remote customers place orders, (ii) whether to keep or cancel online orders when remote customers arrive at the store, (iii) whether to accept or cap on-site orders subject to the individual rationality constraints that both remote customers and local customers have nonnegative expected utility. We observe from our numerical study that, overall, the mixed–capped ordering scheme has the smallest throughput gap from the optimal mechanism among all the simple schemes considered. Figure 3 presents different schemes we study in a tree structure, and Table 1 summarizes the key results.

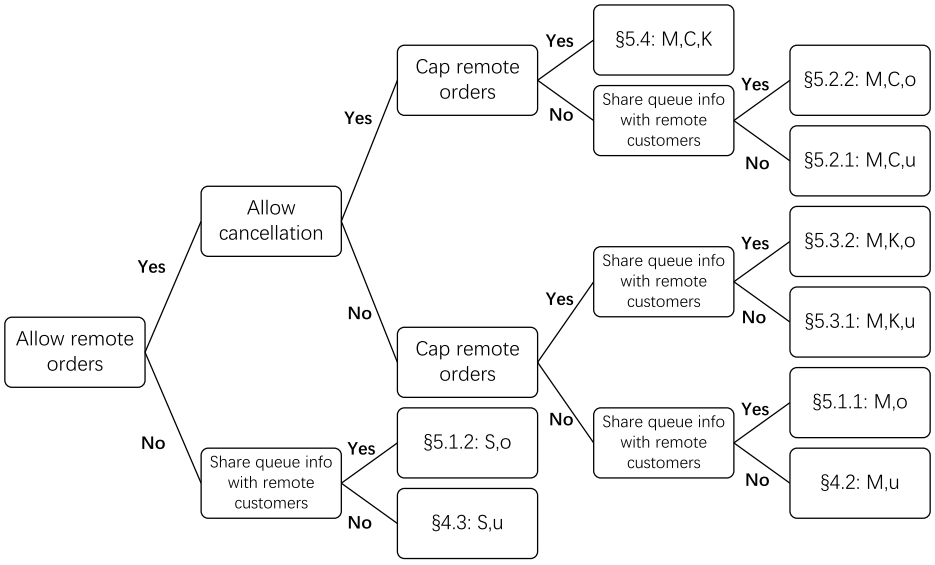
We study three model extensions. The first extension captures food quality degradation that may arise when remote orders are complete before customers arrive at the store. The second extension captures remote customers' channel choice on whether to order ahead, order on-site, or not order at all. The third extension allows the travel speed of remote customers to be heterogeneous.

2. Related Literature

Our paper contributes to the growing literature on omnichannel retail in general and omnichannel service operations in particular. Most relatedly, Gao and Su (2018) show that the adoption of online self-order technologies (such as mobile apps) increases restaurants' throughput. They argue that one driver of this result is the "advance order effect." However, instead of actually modeling the act of ordering in advance, they make the simplifying assumption that ordering online entails a lower waiting cost per unit time than ordering off-line. This same assumption is used by a series of papers that model customer behavior in settings broadly related to omnichannel services, including Baron et al. (2023), Cui et al. (2020), Feldman et al. (2023), and Chen et al. (2022).

Our paper differs from this line of work in both modeling and insights. On the modeling front, our paper explicitly models customers' travel time and, therefore, the parallelism of travel time and waiting time (i.e., the system state evolution during travel). Doing so allows us to capture the advance order effect with higher operational fidelity. Modeling travel time also naturally gives rise to two potential decision epochs spaced by a time lag (the moment a customer's need arises and the moment the customer arrives at the service facility) and enables us to build a particularly novel model of the mixed-ordering-with-cancellation scheme that captures both strategic balking (not placing orders at the first decision epoch) and strategic reneging (canceling orders at the second), a rare combination in the literature. On the insight front, the extant literature (e.g., Gao and Su 2018, Baron et al. 2023) all points to online ordering as a means to increase throughput. In contrast, by carefully modeling the operational subtleties and customer incentives in such a service system, we find that the mixed ordering scheme can counterintuitively result in lower throughput than the on-site–only ordering scheme. We further propose mitigation strategies to overcome the throughput shortfall. Collectively, these new developments highlight the challenges in managing the mixed ordering scheme

Figure 3. A Tree-Based Illustration of Various Schemes



Note. *M*, mixed (on-site and remote) orders; *S*, on-site–only orders; *C*, allow cancellation; *K*, cap remote orders; *o*, observable (queue information shared with remote customers); *u*, unobservable (queue information not shared with remote customers).

and advance our theoretical understanding of such a system.

In the omnichannel service operations space, our paper complements Farahani et al. (2022), who also model travel time in ordering ahead, yet the focus is diametrically different. They study how to manage queues to best meet a prespecified target of pickup time, balancing the trade-off between earliness and tardiness of order readiness. As such, they abstract away from customers' strategic ordering decisions and focus on supply-side interventions. By contrast, our work carefully models customer incentives and examines the demand-side response to different order-ahead mechanisms.

Beyond the application area, our paper contributes to the queueing economics literature that studies customers' strategic behavior in queueing systems pioneered by Naor (1969). We refer to Hassin and Haviv (2003) and Hassin (2016) for comprehensive reviews. In particular, Hassin and Roet-Green (2021) form the basis of

Table 1. Summary of Key Analytical Results

Result	Schemes involved	Key insight
Theorem 1	(S,u), (M,u)	Allowing online orders can reduce throughput when information is not shared with remote customers
Theorem 2	$(S,u),(S,o)$, $(M,u),(M,o)$	Allowing online orders can reduce throughput when information is optimally shared with remote customers
Theorem 3	$(S,u),(S,o)$, $(M,C,u),(M,C,o)$	Mixed ordering with cancellation beats on-site–only ordering under optimal information sharing
Theorem 4	$(M,u),(M,o)$, $(M,C,u),(M,C,o)$	Allowing cancellation of online orders can reduce throughput under optimal information sharing
Theorem 5	$(S,u),(S,o)$, $(M,K,u),(M,K,o)$	Mixed–capped ordering beats on-site–only ordering under optimal information sharing
Theorem 6	$(M,C,u),(M,C,o)$, $(M,K,u),(M,K,o)$	Order cappng can beat order cancellation under optimal information sharing
Proposition 1	(M,u)	Remote customers' equilibrium order placing strategy when information is not shared with remote customers
Proposition 2	(M,o)	Remote customers' equilibrium order placing strategy when information is shared with remote customers
Proposition 3	$(M,u),(M,o)$	Information sharing is optimal in the mixed ordering scheme if and only if the market size is large
Proposition 4	(M,C,u)	Cancellation equilibrium without information sharing
Proposition 5	$(M,K,u),(M,K,o)$	Order capping does not trigger ordering on-site when the capping threshold is high

our benchmark model of the on-site–only ordering scheme in which traveling and waiting are in tandem. They numerically show that, to maximize throughput, the service provider should withhold queue-length information from remote customers when the market size is small but reveal such information when the market size is large. This insight is consistent with the earlier finding established in a simpler setting (Hassin 1986, Chen and Frank 2004). We not only extend the framework of Hassin and Roet-Green (2021) from homogeneous customers to heterogeneous customers (i.e., a mix of local and remote customers who differ in travel time) but also enrich this literature with various models of ordering ahead in which traveling and waiting are in parallel. Further, the ordering on-site nature of Hassin and Roet-Green (2021) precludes strategic reneging, which nevertheless emerges in our cancellation model. The scant literature on strategic reneging has considered reneging triggered by either time-varying service rewards (Hassin and Haviv 1995), nonlinear waiting costs (Haviv and Ritov 2001), or random utility shocks (Ata and Peng 2018). By contrast, our cancellation model is novel in that reneging is driven by information about customers' updated queue position upon arrival at the service facility.

3. Model Setup

We model a service provider (e.g., a restaurant) as a single server that processes customer orders. Order processing times are independent and identically distributed (IID) random variables following an exponential distribution with mean $1/\mu$, where μ is referred to as the capacity of the service provider. Customer needs arise according to a Poisson process with rate Λ, where Λ is referred to as the market size. The market consists of two types of customers who differ in their physical location: remote customers and local customers. Let $\gamma \in (0,1]$ and $1-\gamma$ be the fraction of remote customers and local customers, respectively.

Remote customers are away from the service facility when their needs arise, and their travel times to the service facility are IID random variables following an exponential distribution with mean $1/\beta$, where β is referred to as the travel speed. A remote customer is not entirely certain about the travel time before travel because of potentially unanticipated (elevator/road) traffic. Upon experiencing a need, a remote customer decides whether to place an (irrevocable) order online (from where the customer is located); if the customer places an order, the customer travels to the service facility to pick up the order. Local customers are near the service facility when their needs arise, and their travel times to the service facility are negligible. Upon experiencing a need, a local customer decides whether to place an order on-site. Local customers can be thought of as those who stroll down the street and happen to pass by the store. The service provider processes orders on a first come, first served basis. See Figure 1 for an illustration of the process flow for this mixed ordering scheme.

Customers receive a reward V for having their needs fulfilled on demand. Each customer incurs a delay cost c per unit time between the point when the customer experiences a need and the point when the customer receives the order. Customers are expected-utility maximizers. Consistent with practice, remote customers are provided with a wait-time estimate based on the historical average (e.g., on a mobile app), whereas local customers see the real-time queue length (i.e., the number of outstanding orders) as quick-service restaurants often set up in-store digital screens to display the status of outstanding orders. We later investigate in Section 5 a case in which the service provider can also share this real-time queue-length information with remote customers if doing so increases the system throughput.

To preclude trivial cases in which the service reward is too low for remote customers to ever place orders, we enforce Assumption 1 for the rest of the paper.

Assumption 1. $V > c[1/\mu + 1/\beta - 1/(\beta+\mu)]$.

4. Equilibrium and Comparison with On-Site–Only Ordering

In this section, we first characterize customers' order-placing strategies in equilibrium (in which nobody can strictly increase the customer's own expected utility through unilateral deviation). Next, we compare the equilibrium throughput of the mixed ordering scheme with that of an on-site–only ordering scheme in which both remote and local customers must order on-site (introduced in Section 4.3).

4.1. Preliminaries

This section derives the expected utility of a remote customer if the customer places an order for a given queue length. We first derive the probability distribution of the queue position for a remote customer's order after a random travel time. Suppose that, when a remote customer's need arises, the initial queue length is $n \geq 0$ (i.e., n outstanding orders yet to be processed). If the customer places an order, the customer joins the back of the queue, and the customer's queue position is $n+1$. Let N_n denote the customer's updated queue position upon arrival at the service facility. Thus, N_n is equal to $n+1$ less the number of service completions X up to the customer's own order during the customer's travel. The support of N_n is $\{0,1,\ldots,n+1\}$, where $N_n = 0$ means the order is complete and ready for pickup. Formally, the random variable $N_n \overset{d}{=} [n+1-X]^+$, where

$y^+ \equiv \max\{y, 0\}$ and X is a geometric random variable with $\mathbb{P}(X = i) = [\beta/(\beta+\mu)][\mu/(\beta+\mu)]^i$ for $i = 0, 1, \ldots$. Denote $\sigma \equiv \mu/(\beta+\mu)$. We characterize the probability distribution of N_n in Lemma 1.

Lemma 1 (Updated Queue Position Distribution). *The probability distribution of N_n is*

$$p_n(0) \equiv \mathbb{P}(N_n = 0) = \sigma^{n+1};$$
$$p_n(i) \equiv \mathbb{P}(N_n = i) = (1-\sigma)\sigma^{n-i+1}, \quad i = 1, \ldots, n+1.$$

For a remote customer who places an order when the initial queue length is n, the customer's expected utility conditioned on n, $v(n) = V - c\sum_{i=0}^{(n+1)}(i/\mu)\cdot p_n(i) - c/\beta$. After simplification,

$$v(n) \equiv V - \frac{c}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu}\right). \tag{1}$$

4.2. Equilibrium

This section characterizes customers' order-placing strategies in equilibrium. Local customers place an order if and only if the queue length they see is less than $n_e \equiv \lfloor \mu V/c \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. We refer to n_e (where the subscript "e" stands for equilibrium) as the Naor threshold (Naor 1969). Remote customers place an order with probability $q \in [0,1]$ (determined in equilibrium) based on the expected utility. Next, we characterize this equilibrium order-placing probability q.

Given remote customers' order-placing probability q and local customers' order-placing threshold n_e, let $\rho_T \equiv [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$, $\rho_R \equiv \gamma\Lambda q/\mu$, with subscripts T and R for "total" and "remote," respectively. Thus, for $\rho_R < 1$, the steady-state probability of the number of outstanding orders being i is

$$\pi_{i,M,u}(q) = \begin{cases} \rho_T^i \left(\frac{1-\rho_T^{n_e}}{1-\rho_T} + \frac{\rho_T^{n_e}}{1-\rho_R}\right)^{-1}, & \text{for } i < n_e, \\ \rho_R^{i-n_e}\rho_T^{n_e}\left(\frac{1-\rho_T^{n_e}}{1-\rho_T} + \frac{\rho_T^{n_e}}{1-\rho_R}\right)^{-1}, & \text{for } i \geq n_e. \end{cases} \tag{2}$$

The expected utility for a remote customer who places an order is $U_{M,u}(q) = \sum_{n=0}^{\infty} v(n)\pi_{n,M,u}(q)$. Thus, $q \in (0,1)$ is an equilibrium only if $U_{M,u}(q) = 0$, $q = 1$ is an equilibrium if $U_{M,u}(1) > 0$, and $q = 0$ is an equilibrium if $U_{M,u}(0) < 0$. Proposition 1 characterizes the equilibrium strategy $q_{M,u}$.

Proposition 1 (Equilibrium). *There exist thresholds on market size Λ, $\underline{\lambda}_{M,u}$, and $\overline{\lambda}_{M,u}$ such that remote customers' equilibrium order-placing probability $q_{M,u}$ is*

$$q_{M,u} = \begin{cases} 1, & \text{if } \Lambda \leq \underline{\lambda}_{M,u}, \\ \in (0,1), & \text{if } \underline{\lambda}_{M,u} < \Lambda < \overline{\lambda}_{M,u}, \\ 0, & \text{if } \Lambda \geq \overline{\lambda}_{M,u}. \end{cases} \tag{3}$$

The resulting throughput $TH_{M,u} = \mu(1 - \pi_{0,M,u}(q_{M,u}))$.

When the market size is sufficiently small ($\Lambda \leq \underline{\lambda}_{M,u}$), the service system is not expected to be congested, and thus, all remote customers place orders. Despite this, the resulting throughput is still lower than the market size Λ because local consumers do not always place orders if they see a long queue. When the market size is intermediate ($\underline{\lambda}_{M,u} < \Lambda \leq \overline{\lambda}_{M,u}$), the system is expected to be somewhat congested, causing some remote customers not to place orders. When the market size is sufficiently large ($\Lambda > \overline{\lambda}_{M,u}$), all remote customers stop placing orders because the system is expected to be heavily congested even with local customers only.

4.3. Comparison with an On-Site–Only Ordering Scheme

In this section, we compare the mixed ordering scheme with an on-site–only ordering scheme in which remote customers may place orders (i.e., join the queue) only after they travel to and arrive at the service facility. In the on-site–only ordering scheme, as before, local customers decide whether to place orders based on the observed queue length; remote customers first decide whether to travel to the service facility, and if they choose to travel, then upon arrival at the service facility, they further decide whether to place an order based on the observed queue length, just as do local customers. See Figure 2 for an illustration of the process flow.

When on-site, customers (local or remote) place an order if and only if the queue length they see is less than the Naor threshold n_e. Given the on-site order-placing threshold n_e and remote customers' travel probability $q \in [0,1]$ (determined in equilibrium), the system operates as an $M/M/1/n_e$ queue, and the steady-state probability of the number of outstanding orders being i is $\pi_{i,S,u}(q) = (\rho_T)^i / \sum_{j=0}^{n_e}(\rho_T)^j$, $i = 0, 1, \ldots, n_e$, where $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$. Then, a remote customer's expected utility of joining is $U_{S,u}(q) = \sum_{i=0}^{n_e-1}(V - c(i+1)/\mu)\pi_{i,S,u}(q) - c/\beta$. Thus, a remote customer's equilibrium travel probability $q_{S,u}$ can be characterized in a similar way to Proposition 1. The resulting throughput $TH_{S,u} = \mu(1 - \pi_{0,S,u}(q_{S,u}))$. Notably, when the market size is large enough, remote customers stop traveling to the service facility, let alone placing orders. Theorem 1 compares the throughput of the on-site–only ordering scheme, $TH_{S,u}$, with that of the mixed ordering scheme, $TH_{M,u}$.

Theorem 1 (Mixed Ordering Can Be Worse Than On-Site Only).

i. *When travel speed β is sufficiently low or market size Λ is sufficiently small, the mixed ordering scheme has higher throughput than the on-site–only ordering scheme ($TH_{M,u} > TH_{S,u}$).*

ii. *When β is sufficiently high, for an intermediate range of the market size, the mixed ordering scheme has lower throughput than the on-site only ordering scheme ($TH_{M,u} < TH_{S,u}$).*

Theorem 1 shows that, whether the mixed ordering scheme can achieve higher throughput than the on-site–only scheme depends on both remote customers'

travel speed and the market size. Strikingly, if remote customers travel quickly, then allowing remote customers to order ahead results in lower throughput than if they must order on-site when the market size is intermediate.

Here is the rationale. On the one hand, for remote customers, the mixed ordering scheme parallelizes waiting (for order processing) and traveling (for order pickup). This parallel effect lures more remote customers and puts upward pressure on throughput. On the other hand, the mixed ordering scheme requires remote customers to precommit to their order before they observe the real-time congestion, whereas remote customers in the on-site–only scheme can defer ordering decisions until they see the queue on-site. Precommitment is a double-edged sword for throughput. On the positive side, it secures customer orders early on. This lock-in effect puts upward pressure on the system throughput. On the flip side, remote customers may unknowingly place and commit to orders when the queue is already long. This exacerbates system congestion, which, in turn, deters both local and remote customers from placing orders. As a consequence, this lock-in effect also deters future demand, which then reduces the future workload. When travel is fast and the market size is intermediate, the parallel effect dwindles, but the increased congestion because of the lock-in effect becomes formidable. The downward pressure from the lock-in effect on throughput overshadows the upward pressure from the parallel and lock-in effects combined, causing the mixed ordering scheme to lag behind the on-site–only scheme in throughput.

We supplement Theorem 1 with a numerical trial illustrated by Figure 4. We observe from Figure 4(a) that, when the travel speed is low, the throughput of the mixed ordering scheme always exceeds the throughput of the on-site–only ordering scheme, consistent with Theorem 1(i). We also observe from Figure 4, (b) and (c), that, when the travel speed is not low, the mixed ordering scheme has lower throughput than that of the on-site–only ordering scheme for an intermediate market size, consistent with Theorem 1(ii). Additional numerical studies reveal that the throughput shortfall does not necessarily hinge on the mean travel time being shorter than the mean processing time. In practice, the order processing time depends on the nature of the food: coffee may be quick to make but hot meals may take longer.

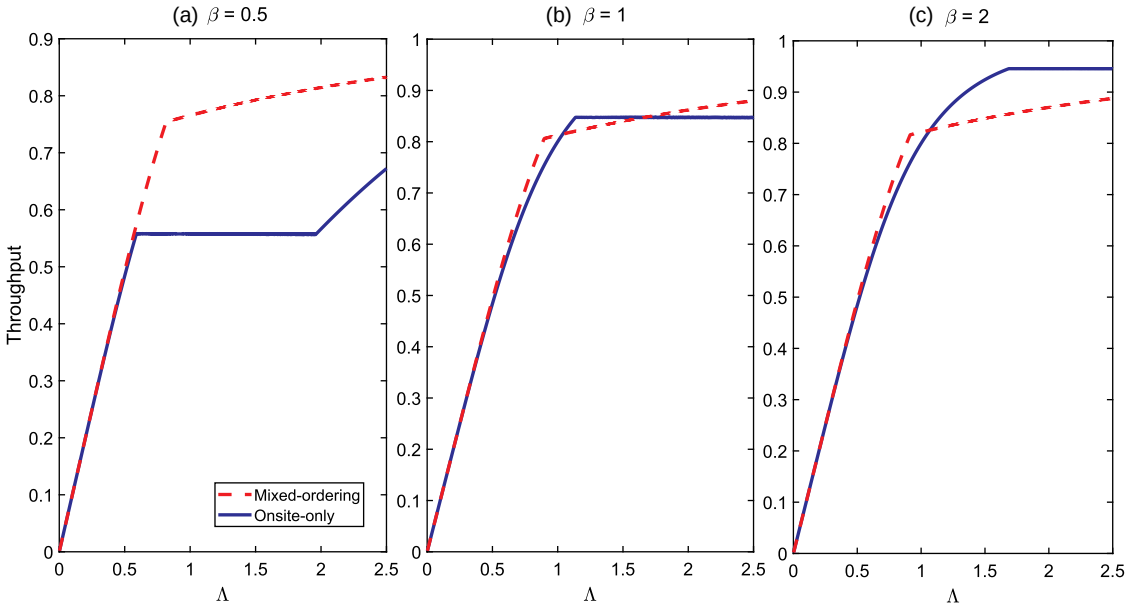
5. Mitigation Strategies

In this section, we consider three strategies that might mitigate the potential throughput shortfall of the mixed ordering scheme: (i) in Section 5.1, we allow the service provider to share queue-length information with remote customers; (ii) in Section 5.2, we allow order cancellation from remote customers; and (iii) in Section 5.3, we allow order capping from the service provider.

5.1. Providing Queue Information to Remote Customers

Recall that a driver for the throughput shortfall of the mixed ordering scheme discussed in Section 4.3 is that remote customers—who have access to wait-time estimates based only on historical averages—may

Figure 4. (Color online) Throughput Comparison of Mixed Ordering vs. On-Site Only



Note. $\mu = 1, V = 2, c = 0.5, \gamma = 0.7$.

unknowingly place orders when the queue is already long. This begs the question of whether the throughput shortfall can be eliminated if the queue information is shared in real time with remote customers. This section explores remote queue-length information as a mitigation strategy.

5.1.1. Mixed Ordering Scheme. We first study the mixed ordering scheme. We start by characterizing remote customers' order-placing strategy when queue-length information is shared with remote customers, that is, when remote customers are informed of the number of outstanding orders when deciding whether to place orders. Note that the expected utility of a remote customer who places an order after observing n outstanding orders, $v(n)$, is derived in (1). Thus, a remote customer places an order if and only this expected utility is nonnegative. Proposition 2 characterizes the order-placing strategy of a remote customer.

Proposition 2 (Remote Customer Strategy with Information). *When queue-length information is shared with remote customers,*

i. *A remote customer places an order if and only if the customer observes a queue length n less than threshold $n_{e,R}$ (i.e., $n < n_{e,R}$), where $n_{e,R}$ is uniquely determined by $n_{e,R} \equiv \min\{n \in \mathbb{N} : v(n) < 0\}$.*

ii. *Threshold $n_{e,R}$ is no greater than the Naor threshold n_e, that is, $n_{e,R} \le n_e$. Specifically, if $\lfloor \frac{\mu V}{c} \rfloor = \frac{\mu V}{c}$, then $n_{e,R} < n_e$ for any $0 < \beta < \infty$; otherwise, $n_{e,R} < n_e$ if and only if travel speed β is low, that is, $\beta < \underline{\beta}$, where $\underline{\beta}$ uniquely solves $V - \frac{cn_e}{\mu} - \frac{c}{\underline{\beta}}\sigma^{n_e} = 0$.*

Whereas the threshold structure of remote customers' order-placing strategy in Proposition 2(i) is intuitive, Proposition 2(ii) may be slightly less straightforward. Here is the rationale. The total delay that remote customers experience is either (i) the travel time (if the order is ready before travel is complete) or (ii) the delay between order generation and order completion (if travel is complete before the order is ready), whichever is longer. Hence, the total delay is expected to be longer than the delay in (ii) alone, which is what local customers endure for the same queue length. This implies that remote customers are generally less receptive to a long queue than their local counterparts (i.e., $n_{e,R} \le n_e$). In particular, when travel time is expected to be long, the total delay is also expected to be much longer than the delay in (ii), inducing remote customers to adopt a strictly lower joining threshold than that of local customers (i.e., $n_{e,R} < n_e$).

Given local customers' order-placing threshold n_e and remote customers' order-placing threshold $n_{e,R}$, let $\rho \equiv \Lambda/\mu$ be the potential traffic intensity. Thus, the steady-state probability of the number of outstanding orders being i is

$$\pi_{i,M,o} = \begin{cases} \rho^i \left(\frac{1-\rho^{n_{e,R}}}{1-\rho} + \frac{\rho^{n_{e,R}}[1-((1-\gamma)\rho)^{n_e-n_{e,R}+1}]}{1-(1-\gamma)\rho} \right)^{-1}, & i = 0,1,\ldots,n_{e,R}, \\ ((1-\gamma)\rho)^{i-n_{e,R}} \rho^{n_{e,R}} \left(\frac{1-\rho^{n_{e,R}}}{1-\rho} + \frac{\rho^{n_{e,R}}[1-((1-\gamma)\rho)^{n_e-n_{e,R}+1}]}{1-(1-\gamma)\rho} \right)^{-1}, & i = n_{e,R}+1,\ldots,n_e. \end{cases}$$

The resulting system throughput is $TH_{M,o} = \mu(1-\pi_{0,M,o})$. Next, we take the perspective of a throughput-maximizing service provider who chooses whether to share queue-length information with remote customers, that is, a service provider that solves the optimization problem $\max\{TH_{M,o}, TH_{M,u}\}$, where throughput $TH_{M,u}$ is defined in Proposition 1. Proposition 3 characterizes the service provider's information-sharing policy in the mixed ordering scheme.

Proposition 3 (Whether to Share Information). *There exists a unique threshold $\tilde{\Lambda}$ such that the throughput-maximizing service provider should not share queue-length information with remote customers ($TH_{M,u} \ge TH_{M,o}$) if $\Lambda \le \tilde{\Lambda}$ and should share information otherwise.*

Proposition 3 generalizes the classic result of Chen and Frank (2004) to a setting of heterogeneous customers. When the market size is small, congestion is light, and withholding queue-length information from remote customers can induce all of them to place orders, whereas if queue-length information is shared, remote customers who happen to see a long queue refrain from ordering. Hence, not sharing information is preferred. By contrast, when the market size is large, congestion is nontrivial, and revealing queue-length information to remote customers induces them to place orders only when the queue is sufficiently short, which keeps the queue length in check and regulates congestion. This, in turn, entices more customers.

5.1.2. On-Site–Only Ordering Scheme. In the on-site–only ordering scheme, customers follow a Naor joining threshold when on-site. However, if queue-length information is shared with remote customers, then remote customers' strategy in deciding the probability of traveling is complex. Hassin and Roet-Green (2021) study a simplified version of this problem with only (homogeneous) remote customers but not local customers. They show that finding remote customers' traveling equilibrium is analytically intractable and instead develop an algorithm to numerically search for the equilibrium. We extend their numerical procedure to our setting of heterogeneous customers and obtain the resulting throughput of the on-site–only ordering scheme $TH_{S,o}$.

5.1.3. Throughput Comparison. Let TH_M^* and TH_S^* denote the maximum throughputs achieved by the optimal remote information-sharing policy in the mixed ordering scheme and the on-site–only scheme, respectively. That is, $TH_i^* \equiv \max\{TH_{i,o}, TH_{i,u}\}$, $i \in \{M, S\}$. Theorem 2 compares TH_M^* with TH_S^*.

Theorem 2 (Mixed Ordering Can Still Backfire). *When travel speed β is sufficiently high, for an intermediate range of the market size, the mixed ordering scheme has lower throughput than the on-site–only ordering scheme even if the service provider optimally chooses whether to share queue-length information with remote customers in each respective scheme ($TH_M^* < TH_S^*$).*

Theorem 2 shows that the potential throughput shortfall of the mixed ordering scheme cannot be eliminated even when the service provider freely decides whether to share queue-length information with remote customers. The conundrum is that withholding queue-length information from remote customers causes a supply–demand mismatch that drives congestion (which is particularly problematic when congestion is already high, i.e., when the market size is large), but sharing the information turns customers away outright (which is particularly problematic when the service provider desperately needs customers, i.e., when the market size is small). Thus, when the market size is neither too small nor too large, the service provider is pushed into a tight corner and cannot salvage the mixed ordering scheme by merely adjusting information.

We supplement Theorem 2 with a numerical study that compares the throughput of the mixed ordering scheme (TH_M^*) with that of the on-site–only ordering scheme (TH_S^*) when the service provider optimally chooses whether to share queue-length information with remote customers in each respective scheme. The result is presented in Figure 5. Consistent with Theorem 2, we observe from Figure 5, (b) and (c), that when the travel speed is not too low and the market size is intermediate, the on-site–only ordering scheme outperforms the mixed ordering scheme. Notably, comparing Figure 5, (b) and (c), with the counterparts of Figure 4 reveals instances in which the service provider switches to sharing queue information with remote customers in the mixed ordering scheme and sticking to no sharing in the on-site–only ordering scheme, yet the throughput of the mixed ordering scheme still falls behind that of the on-site–only ordering scheme.

5.2. Allowing Cancellation of Remote Orders

Recall that another driver for the throughput shortfall of the mixed ordering scheme discussed in Section 4.3 is that remote customers commit to orders when the queue is already long. This begs the question of whether noncommitment (i.e., allowing remote customers to cancel orders when they arrive on-site) helps. This section explores order cancellation as a mitigation strategy.

In the mixed-ordering-with-cancellation scheme, the service provider informs on-site customers of the number of outstanding orders and the queue position of each order (e.g., by displaying a sequence of order IDs on in-store digital screens); further, (remote) customers who order ahead can freely cancel their unfinished orders when they arrive at the service facility. Upon cancellation, an order is withdrawn from the order queue and is no longer processed. Hence, not all orders initially placed are eventually prepared. See Figure 6 for a process flow illustration.

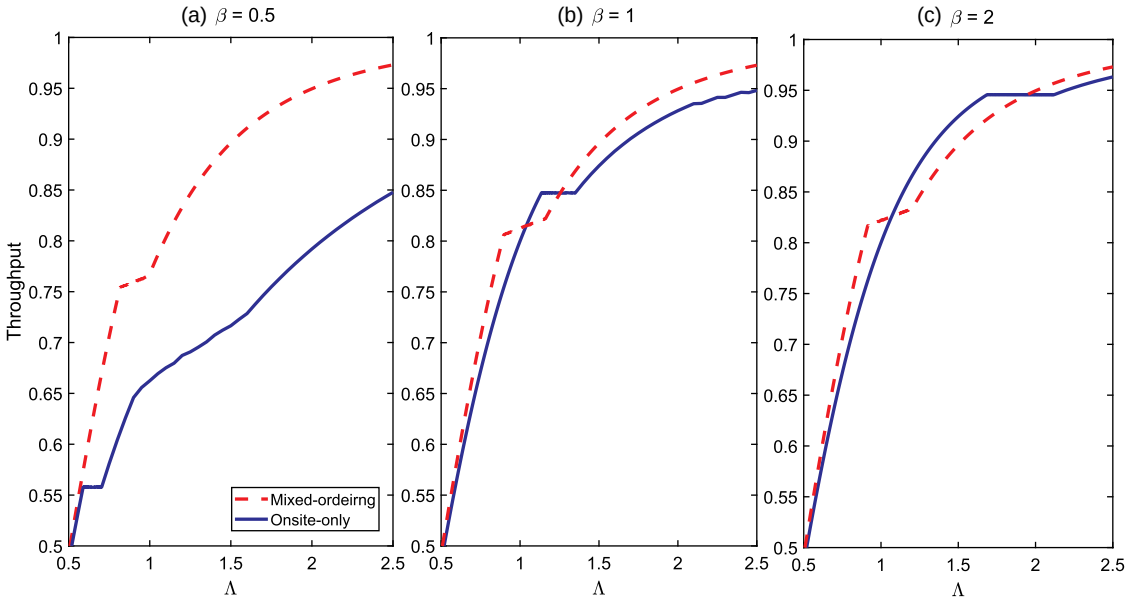
On the surface, this cancellation scheme bears a resemblance to the on-site–only ordering scheme in that remote customers in both may choose not to stay after traveling to the service facility. Nevertheless, the key distinction is that, in the on-site–only ordering scheme, customers may choose to balk from the physical queue before committing to the service, whereas in the mixed ordering scheme with cancellation, remote customers may choose to renege on a previously secured spot in the virtual order queue. As such, our model of the cancellation scheme captures both strategic balking (not placing orders) and strategic reneging (canceling orders).

We first characterize the cancellation strategy of remote customers who arrive at the service facility and observe their queue positions. One complication is that an arriving customer cannot tell whether those with order IDs ahead of the customer's have arrived or not. Customers who have not yet arrived may later cancel their orders upon arrival, thus affecting the focal customer's calculation about the customer's own expected wait time if the customer does not cancel. Thus, each arriving customer needs to think strategically about the cancellation strategies of those who have not arrived. Yet Lemma 2 shows that customers' cancellation strategy has a surprisingly simple threshold structure.

Lemma 2 (Cancellation Strategy). *In the mixed-ordering-with-cancellation scheme, each remote customer cancels the customer's order upon arrival at the service facility if and only if the customer's queue position (the number of outstanding orders ahead of the customer's plus the customer's own order) is greater than n_e.*

Here is the rationale behind Lemma 2. If a customer's queue position does not exceed the Naor threshold n_e, then the customer's dominant strategy is to keep waiting for the order. This result further implies that the first n_e outstanding orders in the queue do not get canceled. Therefore, if a customer's queue position exceeds n_e upon arrival at the service facility, then the customer definitely cancels the order because the customer knows that at least the first n_e orders will not get

Figure 5. (Color online) Throughput Comparison of Mixed Ordering vs. On-Site Only Under Optimal Information



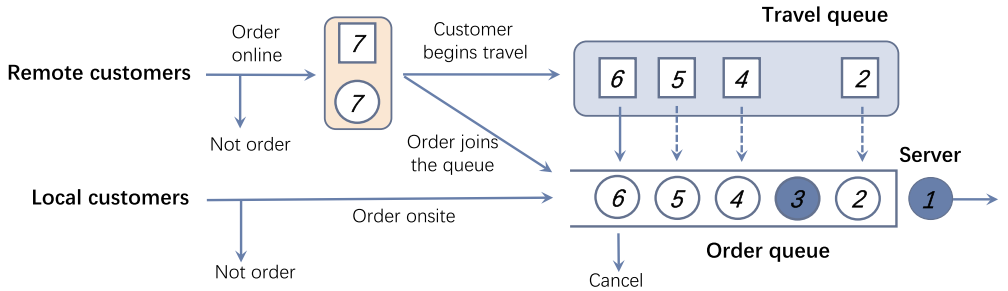
Note. $\mu = 1, V = 2, c = 0.5, \gamma = 0.7$.

canceled and, thus, will be processed before the customer's (regardless of when their order-placing customers arrive), which implies that the reward from getting the customer's own order is not worth the wait. Following the same logic, a local customer places an order if and only if the order queue length is less than n_e.

5.2.1. Queue-Length Information Not Shared with Remote Customers. We next divide our analysis by whether queue-length information is shared with remote customers when they decide whether to order. We start with the case in which such information is not shared.

Given remote customers' cancellation threshold n_e and local customers' joining threshold n_e at the service facility, we now derive remote customers' order-placing probability q when they experience a need. Given q, the system state of the order queue, that is, the number of outstanding orders i, evolves according to a birth–death process with a state-dependent birth rate $\lambda_i(q) = \gamma\Lambda q + (1-\gamma)\Lambda \cdot \mathbf{1}_{\{i \le n_e - 1\}}$ for $i = 0, 1, \ldots$ and a state-dependent death rate μ_i: $\mu_i = \mu + \beta(i - n_e)^+$, $i = 0, 1, \ldots$. Birth rates $\lambda_i(q)$ correspond to order arrivals. When the queue is short ($i \le n_e - 1$), both remote and local customers place orders; hence, the birth rate $\gamma\Lambda q + (1-\gamma)\Lambda$. Otherwise, only remote customers place orders; hence, the birth rate $\gamma\Lambda q$. Next, we explain death rate μ_i. An order cancellation occurs only when a remote customer completes traveling and finds a long order queue ahead of the customer's order at the service facility. A remote customer who sees less than n_e outstanding orders ahead of the customer's does not cancel the order. Such a customer's arrival at the service facility does not trigger a death event in the system. Therefore, when all customers have a short queue

Figure 6. (Color online) Mixed Ordering with Cancellation



Notes. Customers and orders are depicted by squares and circles, respectively. Customers 1 and 3 are on-site, waiting to pick up their orders (represented by the two solid circles); customers 2, 4, and 5 are still traveling; customer 6 is about to arrive at the service facility and cancel the order; customer 7 is about to place an order.

ahead of their orders (i.e., the number of outstanding orders $i \le n_e$), the order queue length can only be decremented by a service completion (which occurs at rate μ). On the other hand, if $i > n_e$, it must imply that $i - n_e$ remote customers have an order queue position above n_e, and furthermore, these customers must all be traveling and have not yet arrived on-site (because, otherwise, they would have already canceled their orders). Hence, in addition to a service-completion event, the order queue length can also be decremented by an order cancellation (which corresponds to one of those $i - n_e$ customers arriving at the service facility) at rate $\beta(i - n_e)$.

Given the birth and death rates, the steady-state probability of the number of outstanding orders being i, $\pi_{i,M,C,u}(q)$, satisfies the flow balance equations $\lambda_i(q)\pi_{i,M,C,u}(q) = \mu_{i+1}\pi_{i+1,M,C,u}(q)$ for $i = 0, 1, \ldots$, from which we obtain the following product form steady-state probabilities:

$$\pi_{i,M,C,u}(q) = \left(\frac{1-\rho_T^{n_e+1}}{1-\rho_T} + \rho_T^{n_e}\sum_{j=1}^{\infty}\prod_{k=1}^{j}\frac{\gamma\Lambda q}{\mu+k\beta}\right)^{-1} \rho_T^{(i\wedge n_e)}\prod_{k=1}^{(i-n_e)^+}\frac{\gamma\Lambda q}{\mu+k\beta}, \quad i = 0,1,\ldots, \tag{4}$$

where $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$. Next, similar to Lemma 1, Lemma 3 characterizes remote customers' updated queue position upon arrival at the service facility (before the customer cancels if at all), denoted by $N_{n,M,C}$, when the customer's queue position is $n+1$ at the moment of ordering.

Lemma 3 (Updated Queue Position Distribution Under Cancellation).

i. *If $n > n_e$, the probability distribution of $N_{n,M,C}$ is $p_{n,M,C}(0) \equiv \mathbb{P}(N_{n,M,C} = 0) = \prod_{k=0}^{n}\frac{\mu_k}{\mu_k+\beta}$; $p_{n,M,C}(i) \equiv \mathbb{P}(N_{n,M,C} = i) = \frac{\beta}{\mu_{i-1}+\beta}\prod_{k=i}^{n}\frac{\mu_k}{\mu_k+\beta}$, $1 \le i \le n+1$, where $\mu_i = \mu + \beta(i-n_e)^+$, $i = 0, 1, \ldots$ and $\prod_{k=i}^{j} x_k = 1$ for $i > j$.*

ii. *If $n \le n_e$, $N_{n,M,C}$ has the same distribution as N_n given by Lemma 1.*

Given q, the expected utility of a remote customer who places an order is

$$U_{M,C,u}(q) \equiv \sum_{i=0}^{n_e-1} v(i)\pi_{i,M,C,u}(q) + \sum_{i=n_e}^{\infty}\left[\sum_{j=0}^{n_e}\left(V - \frac{cj}{\mu}\right)p_{i,M,C}(j) - \frac{c}{\beta}\right]\pi_{i,M,C,u}(q), \tag{5}$$

where $v(i)$, $\pi_{i,M,C,u}(q)$, and $p_{i,M,C}(j)$ are given by Equations (1) and (4) and Lemma 3, respectively. Proposition 4 characterizes remote customers' equilibrium order-placing probability $q_{M,C,u}$.

Proposition 4 (Equilibrium in the Cancellation Model). *In the mixed-ordering-with-cancellation scheme, when queue-length information is not shared with remote customers, there exist thresholds on market size Λ, $\underline{\lambda}_{M,C,u}$, and $\overline{\lambda}_{M,C,u}$ such that remote customers' equilibrium order-placing probability $q_{M,C,u} = \mathbf{1}_{\{\Lambda \le \underline{\lambda}_{M,C,u}\}} + \tilde{q}_{M,C}\cdot\mathbf{1}_{\{\underline{\lambda}_{M,C,u} < \Lambda < \overline{\lambda}_{M,C,u}\}}$, where $\tilde{q}_{M,C} \in (0,1)$ uniquely solves $U_{M,C,u}(\tilde{q}_{M,C}) = 0$, with $U_{M,C,u}(q)$ given in Equation (5). The resulting throughput is $TH_{M,C,u} = \mu[1 - \pi_{0,M,C,u}(q_{M,C,u})]$.*

5.2.2. Queue-Length Information Shared with Remote Customers. We next consider the case in which queue-length information is shared with remote customers when they decide whether to order. Recall from Proposition 2 that, in such a case, remote customers place orders only when the observed queue length is less than $n_{e,R}$ with $n_{e,R} \le n_e$. By the time they arrive on-site, their queue position is only improved (at least no worse than $n_{e,R}$). Because Lemma 2 shows that a customer only cancels if the customer's queue position is worse than n_e, it implies that no customers have the incentive to cancel in the original mixed ordering scheme. Hence, enabling cancellation does not make a difference when queue-length information is shared with remote customers. Hence, the system throughput in this case $TH_{M,C,o}$ equals the throughput in its noncancellation counterpart, $TH_{M,o}$, that is, $TH_{M,C,o} = TH_{M,o}$.

5.2.3. Throughput Comparison. Let $TH^*_{M,C} = \max\{TH_{M,C,o}, TH_{M,C,u}\}$ denote the maximum throughput achieved by optimal information sharing in the mixed-ordering-with-cancellation scheme. We next compare this throughput with that in the on-site–only ordering scheme, TH^*_S.

Theorem 3 (Mixed-Ordering-with-Cancellation vs. On-Site–Only). *When queue-length information is optimally shared with remote customers, for $n_e = 1$, the mixed-ordering-with-cancellation scheme has higher throughput than the on-site–only scheme, that is, $TH^*_{M,C} \ge TH^*_S$.*

Theorem 3 shows that enabling cancellation mitigates the throughput shortfall under optimal information sharing; we prove the result analytically for $n_e = \lfloor V\mu/c \rfloor = 1$, but numerically, we find that it holds for all the problem instances tested (see Section 5.4 for details of our numerical study). Cancellation is a self-regulating mechanism that alleviates the issue of over-congestion created by the lock-in effect without turning remote customers away from the outset just because the queue is long initially (which is the case if the service provider resorts to information sharing alone). Customers abandon only after they arrive at the service facility and actually expect a long wait on-site.

Therefore, cancellation restores the advantage of the mixed ordering scheme over the on-site–only scheme. Next, we investigate the impact of enabling cancellation on the throughput of the mixed ordering scheme.

Theorem 4 (To Cancel or Not to Cancel). *When queue-length information is optimally shared with remote customers, allowing cancellation in the mixed ordering scheme results in lower throughput ($TH^*_{M,C} < TH^*_M$) if market size Λ is small.*

Whereas Theorem 3 suggests that enabling cancellation is a promising solution that allows the mixed ordering scheme to outperform the on-site–only scheme, Theorem 4 shows that the cancellation scheme nevertheless falls short of the noncancellation one when the market size is small. The cancellation scheme forgoes orders that the noncancellation scheme holds on to otherwise. This order loss is critical when there are not too many orders to begin with, that is, when the market size is small. In this case, the cancellation scheme falls short of the noncancellation scheme in retaining orders. Hence, cancellation addresses an existing problem (i.e., the throughput shortfall relative to the on-site–only scheme) by creating a new one (i.e., a potential throughput loss relative to the mixed ordering scheme without cancellation).

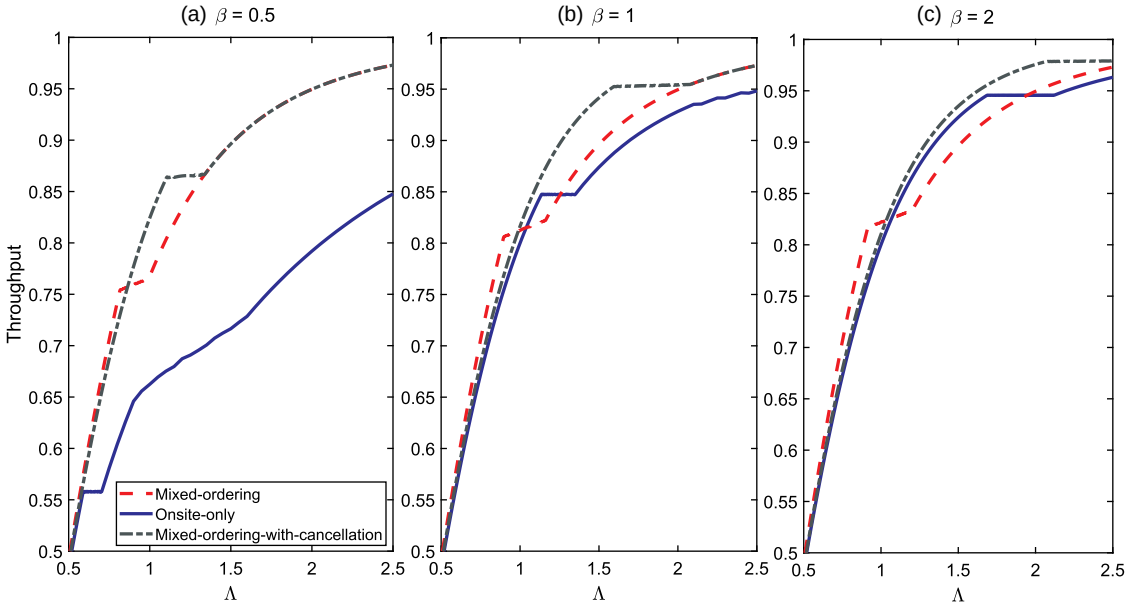
Figure 7 illustrates a three-way throughput comparison of the mixed ordering scheme (TH^*_M), the on-site–only scheme (TH^*_S), and the mixed-ordering-with-cancellation scheme ($TH^*_{M,C}$) when the service provider chooses the optimal information in each respective scheme. We observe that the mixed-ordering-with-cancellation scheme always outperforms the on-site–only ordering scheme, that is, $TH^*_{M,C} > TH^*_S$ (confirming Theorem 3 for general n_e), mitigating the throughput shortfall of the noncancellation scheme. Nevertheless, we also observe that the cancellation scheme falls short of its noncancellation counterpart when the market size is small (confirming Theorem 4). Further, we observe that the cancellation scheme coincides with the noncancellation scheme when the market size is large because both choose to share information and that the former achieves strictly higher throughput than the latter only when the market size is intermediate.

5.3. Capping Remote Orders

The limitations of the previous two mitigation strategies motivate us to explore yet another alternative: a mixed–capped ordering scheme in which the service provider accepts new remote orders (placed by remote customers online) if the total number of outstanding orders is strictly less than a threshold $\nu_K \in \mathbb{N} \cup \{\infty\}$ and stops accepting any new remote orders otherwise. Threshold ν_K is a decision variable of the service provider. We allow remote customers who cannot place an order online (because of order capping) to opt to travel to the store in hopes of ordering on-site based on the information they receive about the queue. As before, once on-site, (remote and local) customers decide whether to order according to the Naor threshold. See Figure 8 for a process flow illustration of the mixed–capped ordering scheme.

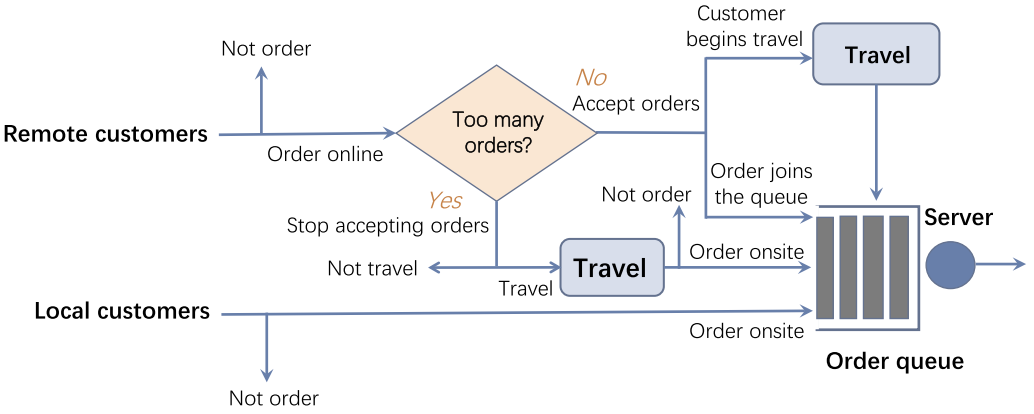
We define the system state by (n_s, n_t), where $n_s \in \{0, 1, \ldots, n_e \vee \nu_K\}$ represents the number of outstanding orders waiting to be served, and $n_t \in \{0, 1, \ldots\}$ denotes

Figure 7. (Color online) Throughput Comparison of Mixed Ordering, On-Site Only, and Mixed Ordering with Cancellation Under Optimal Information



Note. $\mu = 1, V = 2, c = 0.5, \gamma = 0.7$.

Figure 8. (Color online) Mixed–Capped Ordering Scheme



the number of traveling customers who have not placed an order yet (i.e., customers who cannot place an order online because of order capping but opt to travel to the store). Let π_{n_s,n_t} denote the steady-state probability that the system is in state (n_s,n_t).

5.3.1. Queue-Length Information Not Shared with Remote Customers. We start with the case in which queue-length information is not shared with remote customers when they decide whether to order. Remote customers' strategy (when they are still remote) is defined by (q_R,q_W), where $q_R \in [0,1]$ is the probability that remote customers place orders and $q_W \in [0,1]$ is the probability that remote customers travel (or walk) to the store when they cannot order online (because of order capping).

Denote $\lambda_L = (1-\gamma)\Lambda$, $\lambda_R = \gamma\Lambda q_R$, and $\lambda_W = \gamma\Lambda q_R q_W$.

The system transitions from state (n_s,n_t) to (n_s+1,n_t) because of either an order placed by a local customer (which occurs with rate λ_L if $n_s < n_e$) or an order placed by a remote customer (which occurs with rate λ_R if $n_s < \nu_K$). It transitions to state (n_s,n_t+1) when a customer cannot order online because of order capping (i.e., $n_s \ge \nu_K$) and decides to travel (at rate λ_W) provided that $n_s \in \{\nu_K,\dots,n_e \vee \nu_K\}$. It transitions to state (n_s-1,n_t) when an order is completed (at rate μ) or to state (n_s+1,n_t-1) when a traveling customer arrives at the store and orders on-site (at rate $n_t\beta$) for $n_s < n_e$. However, if $n_s \ge n_e$ when a traveling customer arrives, the customer does not order on-site, and the state shifts to (n_s,n_t-1). The following balance equations in (6) describe the above birth-and-death process. For $n_s \in \{0,1,\dots,n_e \vee \nu_K\}$ and $n_t \in \{0,1,\dots\}$, we have

$$\begin{aligned}&(\lambda_L \mathbf{1}_{\{n_s<n_e\}} + \lambda_R \mathbf{1}_{\{n_s<\nu_K\}} + \lambda_W \mathbf{1}_{\{n_s\ge\nu_K\}} + \mu\cdot\mathbf{1}_{\{n_s>0\}} + n_t\beta)\pi_{n_s,n_t}\\ &= \lambda_L\cdot\mathbf{1}_{\{0<n_s\le n_e\}}\pi_{n_s-1,n_t} + \lambda_R\cdot\mathbf{1}_{\{0<n_s\le\nu_K\}}\pi_{n_s-1,n_t}\\ &\quad + \lambda_W\cdot\mathbf{1}_{\{n_s\ge\nu_K,n_t>0\}}\pi_{n_s,n_t-1} + \mu\cdot\mathbf{1}_{\{n_s<n_e\vee\nu_K\}}\pi_{n_s+1,n_t}\\ &\quad + \beta(n_t+1)[\mathbf{1}_{\{0<n_s<n_e\}}\pi_{n_s-1,n_t+1} + \mathbf{1}_{\{n_s\ge n_e\}}\pi_{n_s,n_t+1}]. \qquad (6)\end{aligned}$$

From (6), we compute the steady-state probabilities π_{n_s,n_t}, which depend on remote customers' strategy (q_R,q_W). Define $\pi_{n_s}(q_R,q_W) = \sum_{n_t=0}^{\infty}\pi_{n_s,n_t}(q_R,q_W)$ to be the steady-state probability that there are n_s outstanding orders. Thus, the throughput is $TH_{M,K,u} = \mu(1-\pi_0(q_R,q_W))$.

5.3.1.1. Expected Utility. Given that all remote customers follow strategy (q_R,q_W), a tagged remote customer's expected utility from placing a remote order when the need arises is given by

$$\begin{aligned}U_1(q_R,q_W) &= \sum_{n_s=0}^{\nu_K-1} v(n_s)\pi_{n_s}(q_R,q_W)\\ &\quad + \max\{U_2(q_R,q_W),0\}\sum_{n_s=\nu_K}^{n_e\vee\nu_K}\pi_{n_s}(q_R,q_W),\end{aligned}$$

where $v(n_s)$ is the expected utility conditioned on n_s outstanding orders defined in (1) when the remote order is accepted (if $n_s < \nu_K$), and $U_2(q_R,q_W)$ is the expected utility from traveling to the store to order on-site when remote ordering is capped (if $n_s \ge \nu_K$); $U_2(q_R,q_W)$ is given by

$$U_2(q_R,q_W) = \sum_{n_s=\nu_K}^{\nu_K\vee n_e}\sum_{n_t=0}^{\infty}\frac{\pi_{n_s,n_t}(q_R,q_W)}{\sum_{n_s=\nu_K}^{\nu_K\vee n_e}\pi_{n_s}(q_R,q_W)}\mathcal{U}(n_s,n_t+1\,|\,q_R,q_W),$$

where $\frac{\pi_{n_s,n_t}(q_R,q_W)}{\sum_{n_s=\nu_K}^{n_e\vee\nu_K}\pi_{n_s}(q_R,q_W)}$ is the conditional probability that the system is in state (n_s,n_t) given that $n_s \ge \nu_k$ before the tagged customer travels, and $\mathcal{U}(n_s,n_t+1\,|\,q_R,q_W)$ is the expected utility of traveling given state (n_s,n_t+1) (note that, once the tagged customer travels, there are a total of n_t+1 traveling customers who have not placed an order yet). Denote $\theta = \lambda_R\mathbf{1}_{\{n_s<\nu_K\}} + \lambda_L\mathbf{1}_{\{n_s<n_e\}} + \lambda_W\mathbf{1}_{\{n_s\ge\nu_K\}} + \mu\mathbf{1}_{\{n_s>0\}} + n_t\beta$. Following the logic of the state transitions introduced for setting up balance equations in (6), $\mathcal{U}(n_s,n_t\,|\,q_R,q_W)$ satisfies the following recursive equations

(we suppress its dependency on (q_R, q_W) for brevity): for $n_s \in \{0,1,\dots,n_e \vee \nu_K\}$ and $n_t \in \{1,\dots\}$,

$$\begin{aligned}\mathcal{U}(n_s,n_t) = &-\frac{c}{\theta}+\frac{\lambda_R \mathbf{1}_{\{n_s<\nu_K\}}+\lambda_L \mathbf{1}_{\{n_s<n_e\}}}{\theta}\mathcal{U}(n_s+1,n_t)\\ &+\frac{\lambda_W \mathbf{1}_{\{n_s\ge\nu_K\}}}{\theta}\mathcal{U}(n_s,n_t+1)\\ &+\frac{\mu \mathbf{1}_{\{n_s>0\}}}{\theta}\mathcal{U}(n_s-1,n_t)+\frac{\beta \mathbf{1}_{\{n_s<n_e\}}}{\theta}\left[V-\frac{n_s+1}{\mu}\right]\\ &+\frac{\beta(n_t-1)}{\theta}[\mathcal{U}(n_s+1,n_t-1)\mathbf{1}_{\{0<n_s<n_e\}}\\ &\qquad +\mathcal{U}(n_s,n_t-1)\mathbf{1}_{\{n_s\ge n_e\}}].\end{aligned}$$

In equilibrium, the (q_R, q_W) pair satisfies

$$q_R \in \arg\max_{q\in[0,1]} qU_1(q_R,q_W), \quad q_W \in \arg\max_{q\in[0,1]} qU_2(q_R,q_W).$$

Proposition 5 characterizes conditions under which remote customers do not order on-site.

Proposition 5. *In the mixed–capped ordering scheme, when $\nu_K \ge n_{e,R}$, remote customers who cannot order online because of order capping do not travel to order on-site.*

When customers cannot order ahead because of order capping, they know the queue is at least as long as the capping threshold ν_K. Also, recall from Proposition 2 that customers do not order ahead if there are $n_{e,R}$ outstanding orders. Therefore, if $\nu_K \ge n_{e,R}$, the queue is too long for remote customers to order ahead. We further show that ordering on-site generates even lower utility than ordering ahead, and thus, when $\nu_K \ge n_{e,R}$, remote customers who cannot order online because of order capping have no incentive to travel to the service facility and attempt to order on-site. In this case, n_t always equals zero, and it suffices to track n_s only.

5.3.2. Queue-Length Information Shared with Remote Customers. When queue-length information is shared with remote customers, recall that, without order capping, remote customers use a joining threshold $n_{e,R}$. Hence, order capping is only effective if $\nu_K < n_{e,R}$ (otherwise, the system degenerates into one without order capping). We focus on this case of $\nu_K < n_{e,R}$ in which remote customers who see an order queue less than ν_K order online, those who see an order queue equal to or longer than $n_{e,R}$ cannot order online and do not attempt to order on-site (Proposition 5), and those who see an order queue of length $n_s \in [\nu_K, n_{e,R}]$ cannot order online but may attempt to order on-site. Thus, remote customers' strategy is specified by a probability vector $\mathbf{p} = (p_{\nu_K}, p_{\nu_K+1}, \dots, p_{n_{e,R}-1})$, where $p_{n_s} \in [0,1]$ is the probability that a remote customer who cannot order online because of order capping travels to the store to attempt ordering on-site upon observing n_s orders in the order queue before travel for $\nu_K \le n_s < n_{e,R}$.

Similar to (6), given strategy $\mathbf{p}$, we set up balance equations of steady-state probabilities $\pi_{n_s,n_t,o}$ that describe the evolution of system state (n_s, n_t) in the case in which queue information is shared with remote customers: for $n_s \in \{0,1,\dots,n_e \vee \nu_K\}$ and $n_t \in \{0,1,\dots\}$,

$$\begin{aligned}&(\lambda_L \mathbf{1}_{\{n_s<n_e\}}+\gamma\Lambda \mathbf{1}_{\{n_s<n_{e,R}\wedge\nu_K\}}+\gamma\Lambda p_{n_s}\mathbf{1}_{\{\nu_K\le n_s<n_{e,R}\}}\\ &\quad +\mu\cdot\mathbf{1}_{\{n_s>0\}}+n_t\beta)\pi_{n_s,n_t,o}\\ &=\lambda_L\cdot\mathbf{1}_{\{0<n_s\le n_e\}}\pi_{n_s-1,n_t,o}+\gamma\Lambda\cdot\mathbf{1}_{\{0<n_s\le\nu_K\}}\pi_{n_s-1,n_t,o}\\ &\quad +\gamma\Lambda p_{n_s}\cdot\mathbf{1}_{\{\nu_K\le n_s<n_{e,R},n_t>0\}}\pi_{n_s,n_t-1,o}\\ &\quad +\mu\cdot\mathbf{1}_{\{n_s<n_e\vee\nu_K\}}\pi_{n_s+1,n_t,o}\\ &\quad +\beta(n_t+1)[\mathbf{1}_{\{0<n_s<n_e\}}\pi_{n_s-1,n_t+1,o}+\mathbf{1}_{\{n_s\ge n_e\}}\pi_{n_s,n_t+1,o}],\end{aligned}$$

5.3.2.1. Expected Utility. Given that all other remote customers adopt strategy $\mathbf{p}$, a tagged remote customer's expected utility from traveling to the store to attempt ordering on-site when remote ordering is capped and the customer observes an order queue of length $n_s \in [\nu_K, n_{e,R}]$ is

$$U_o(n_s|\mathbf{p}) = \frac{\sum_{n_t=0}^{\infty}\pi_{n_s,n_t,o}\mathcal{U}_o(n_s,n_t|\mathbf{p})}{\sum_{n_t=0}^{\infty}\pi_{n_s,n_t,o}},$$

where $\pi_{n_s,n_t,o}/\sum_{n_t=0}^{\infty}\pi_{n_s,n_t,o}$ is the conditional probability that the system is in state (n_s, n_t) when the order queue has length n_s; $\mathcal{U}_o(n_s,n_t|\mathbf{p})$ is the remote customer's expected utility in state (n_s, n_t) and satisfies the following recursive equations (we suppress its dependency on $\mathbf{p}$ for brevity and denote $\theta_o = \gamma\Lambda\mathbf{1}_{\{n_s<n_{e,R}\vee\nu_K\}} + \lambda_L\mathbf{1}_{\{n_s<n_e\}} + \gamma\Lambda p_{n_s}\mathbf{1}_{\{\nu_K\le n_s<n_{e,R}\}} + \mu\mathbf{1}_{\{n_s>0\}} + n_t\beta$):

$$\begin{aligned}\mathcal{U}_o(n_s,n_t) = &-\frac{c}{\theta_o}+\frac{\gamma\Lambda\mathbf{1}_{\{n_s<\nu_K\}}+\lambda_L\mathbf{1}_{\{n_s<n_e\}}}{\theta_o}\mathcal{U}_o(n_s+1,n_t)\\ &+\frac{\gamma\Lambda p_{n_s}\mathbf{1}_{\{\nu_K\le n_s<n_{e,R}\}}}{\theta_o}\mathcal{U}_o(n_s,n_t+1)\\ &+\frac{\mu\mathbf{1}_{\{n_s>0\}}}{\theta_o}\mathcal{U}_o(n_s-1,n_t)+\frac{\beta\mathbf{1}_{\{n_s<n_e\}}}{\theta_o}\left[V-\frac{n_s+1}{\mu}\right]\\ &+\frac{\beta(n_t-1)}{\theta_o}[\mathcal{U}_o(n_s+1,n_t-1)\mathbf{1}_{\{0<n_s<n_e\}}\\ &+\mathcal{U}_o(n_s,n_t-1)\mathbf{1}_{\{n_s\ge n_e\}}].\end{aligned}$$

In equilibrium, the traveling probability vector $\mathbf{p} = (p_{\nu_K}, p_{\nu_K+1}, \dots, p_{n_{e,R}-1})$ satisfies

$$p_{n_s} \in \arg\max_{p\in[0,1]} pU_o(n_s|\mathbf{p}), \quad n_s = \nu_K,\dots,n_{e,R}.$$

5.3.3. Throughput Comparison. We next compare the throughput of the capped ordering scheme (with a throughput-maximizing capping threshold), $TH^*_{M,K}$, with those of the three schemes introduced earlier: (i) the mixed ordering scheme (without capping or cancellation), (ii) the on-site–only ordering scheme, and (iii) the cancellation scheme when each scheme has its respective optimal remote information sharing policy. As for (i), the ordering scheme without capping

essentially has a capping threshold of $\nu_K = \infty$ and is, thus, dominated by the mixed–capped ordering scheme with an optimized capping threshold. Hence, it follows that order capping increases throughput, that is, $TH^*_{M,K} \geq TH^*_M$. Theorems 5 and 6 address comparisons (ii) and (iii), respectively.

Theorem 5 (Mixed–Capped Ordering vs. On-Site Only). *When queue-length information is optimally shared with remote customers, for $n_e = 1$, the mixed–capped ordering scheme has higher throughput than the on-site–only ordering scheme, that is, $TH^*_{M,K} \geq TH^*_S$.*

Theorem 5 shows that order capping mitigates the throughput shortfall of the mixed ordering scheme, enabling it to capture more customers than the on-site–only ordering scheme. We prove that this throughput dominance holds for general n_e if queue-length information is not shared with remote customers (i.e., $TH_{M,K,u} \geq TH_{S,u}$ for general n_e), but when information is shared in the on-site–only ordering scheme, the underlying queueing system becomes analytically intractable, and thus, we can only analytically prove this result for $n_e = 1$ even though, numerically, we find that it holds for all the problem instances tested (see Section 5.4 for details of our numerical study). The introduction of order capping keeps the queue length in check. Thus, it regulates congestion and enables customers who successfully order ahead to enjoy the benefit of the parallel effect without worrying about the longer than usual delay they might otherwise encounter in the scheme without order capping. Therefore, customers are more willing to place orders. Moreover, the capping threshold can be fine-tuned to strike the balance between acquisition (getting more customers to place orders) and retention (keeping more orders that have been placed) so that the mixed ordering throughput is indeed higher than the on-site–only throughput. Theorem 6 compares the throughput of the mixed–capped ordering scheme with that of the cancellation scheme.

Theorem 6 (To Cancel or to Cap). *When queue-length information is optimally shared with remote customers, the mixed–capped ordering scheme has higher throughput than the cancellation scheme ($TH^*_{M,K} \geq TH^*_{M,C}$) when the market size is sufficiently small or large.*

When the market size is small, the scheme without capping or cancellation has higher throughput than the cancellation scheme, according to Theorem 4. Therefore, in this case, the mixed–capped ordering scheme (with an optimally determined cap) outperforms the cancellation scheme. When the market size is large, the mixed–capped ordering scheme fends off orders at the outset, and this more sharply regulates congestion that the cancellation scheme that lets customers voluntarily withdraw orders in the process. Hence, in this case, the mixed–capped ordering scheme again outperforms the cancellation scheme. However, when the market size is intermediate, it is unclear whether the mixed–capped ordering scheme still generates higher throughput than the cancellation scheme. We explore this question numerically in Figure 9.

Figure 9 conducts a four-way throughput comparison of the mixed ordering scheme (TH^*_M), the on-site–only ordering scheme (TH^*_S), the mixed-ordering-with-cancellation scheme ($TH^*_{M,C}$), and the mixed–capped ordering scheme ($TH^*_{M,K}$) when the service provider optimally chooses whether to share queue-length information with remote customers in each respective scheme. We observe that the mixed–capped ordering scheme always outperforms both the mixed ordering scheme ($TH^*_{M,K} \geq TH^*_M$), which is by construction, and the on-site–only ordering scheme ($TH^*_{M,K} \geq TH^*_S$), confirming Theorem 5 for general n_e. Hence, as in the cancellation scheme, order capping can be yet another approach to mitigate the throughput shortfall. Yet, unlike the cancellation scheme, order capping does not have the unintended consequence of falling short of the basic scheme without cancellation or capping (because the capping threshold can be optimized). Further, in many instances, the mixed–capped ordering scheme also dominates the cancellation scheme (echoing Theorem 6), but this is not always the case. Figure 9(c) shows that, when travel is fast and the market size is intermediate, the mixed–capped ordering scheme results in lower throughput than the cancellation scheme (although the difference is small). The rationale is that order capping is a more drastic measure of regulating congestion than allowing order cancellation and, therefore, can overshoot.

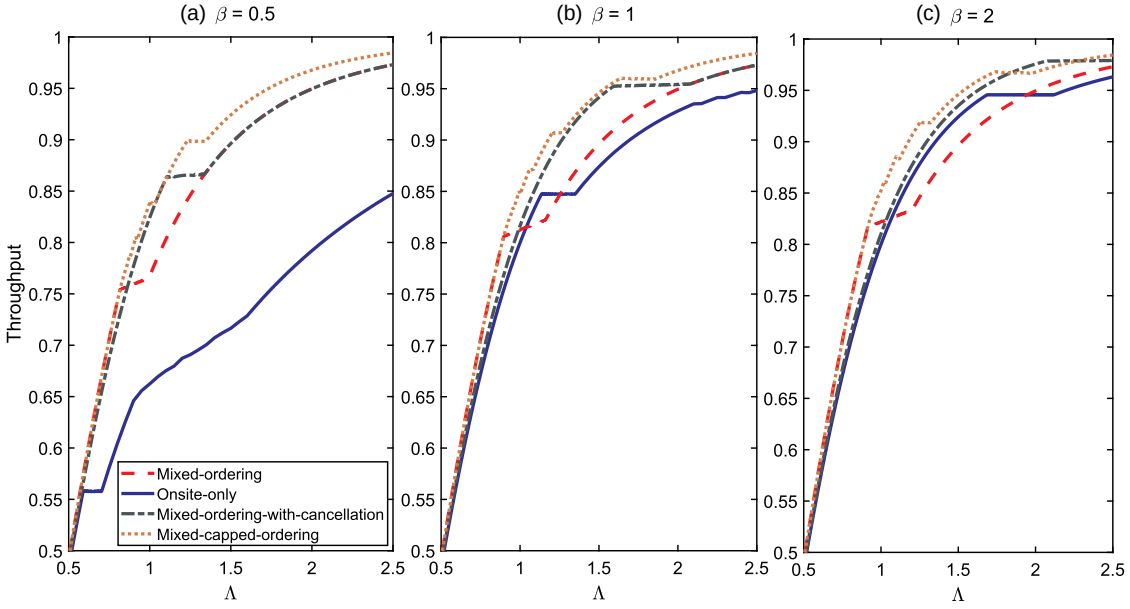
In sum, the mixed–capped ordering scheme holds promise as it attains higher throughput than both the mixed ordering scheme (by construction) and the on-site–only ordering scheme (proved analytically in Theorem 5 for $n_e = 1$ and confirmed numerically for general n_e). Whereas it is not guaranteed to outperform the mixed-ordering-with-cancellation scheme, it tends not to fall far behind (based on numerical observation). However, in order for the mixed–capped ordering scheme to work in practice, the capping threshold must be carefully calibrated to the business characteristics and clearly communicated to remote customers, both of which are not without practical challenges.

5.4. Optimal Scheme

In this section, we propose an optimal scheme in which a dictatorial service provider solves an admission control problem with three sets of levers:

i. Whether to cap online orders: When remote customers place online orders, accept them with probability $q \in [0,1]$ if the total number of outstanding orders is less than $\nu_K \in \mathbb{N} \cup \{\infty\}$ and stop accepting online orders otherwise.

Figure 9. (Color online) Throughput Comparison of Mixed Ordering, On-Site Only, Mixed Ordering with Cancellation, and Mixed–Capped Ordering Under Optimal Information



Note. $\mu = 1, V = 2, c = 0.5, \gamma = 0.7$.

ii. Whether to cancel remote orders: When remote customers arrive at the store to pick up their online orders, if their orders are not prepared, keep the orders with probability $\phi \in [0,1]$ if their queue position is at most $\nu_C \in \mathbb{N} \cup \{\infty\}$ and cancel online orders otherwise.

iii. Whether to cap on-site orders: When local customers place on-site orders, accept them with probability $p \in [0,1]$ if the total number of outstanding orders is less than $\nu_L \in \mathbb{N} \cup \{\infty\}$ and stop accepting on-site orders otherwise.

The service provider chooses $(q, \nu_K, \phi, \nu_C, p, \nu_L)$ to maximize throughput $TH(q, \nu_K, \phi, \nu_C, p, \nu_L)$, subject to the individual rationality constraints that both remote customers' expected utility, $UR_{M,C,K}(q, \nu_K, \phi, \nu_C, p, \nu_L)$, and local customers' expected utility, $UL_{M,C,K}(q, \nu_K, \phi, \nu_C, p, \nu_L)$, are nonnegative. Thus, the service provider's problem is[1]

$$\begin{aligned} \max_{q,\nu_K,\phi,\nu_C,p,\nu_L} \quad & TH_{M,C,K}(q,\nu_K,\phi,\nu_C,p,\nu_L), \\ \text{s.t.} \quad & UR_{M,C,K}(q,\nu_K,\phi,\nu_C,p,\nu_L) \geq 0, \\ & UL_{M,C,K}(q,\nu_K,\phi,\nu_C,p,\nu_L) \geq 0. \end{aligned}$$

The optimal control problem, characterized by six control variables, is inherently difficult to solve. In fact, even for given control variables, the resulting queueing system may not be amenable to analysis as one may need to track the status of every customer (e.g., whether it is a remote customer or a local customer and whether the remote customer is still traveling) to characterize the system evolution. To overcome this challenge, we adopt a simulation-based optimization framework that systematically searches for optimal solutions (Jian and Henderson 2015). For each candidate control profile (i.e., a specified set of control variables), we perform bottom-up Monte Carlo simulations to estimate key performance measures, including steady-state throughput and customer utility. Our simulation is agent-based simulation (Macal and North 2014): besides estimating the system-level performance such as the sizes of the order queue and travel queue, we also track the individual customer dynamics, including customer ID, travel time, queueing position, service time, service outcome (capped, cancelled, or served), and utility.

Using the simulation-optimization approach, we numerically quantify how the previously considered simpler schemes perform relative to the optimal scheme. We generate 625 representative instances from the following parameter combinations: $\mu = 1$, $c = 0.5$, $V \in \{1.5, 2.5, \ldots, 5.5\}$, $\gamma \in \{0.1, 0.3, \ldots, 0.9\}$, $\Lambda \in \{0.5, 1, \ldots, 2.5\}$, $\beta \in \{0.5, 1, \ldots, 2.5\}$ (all satisfying Assumption 1). In each instance, we compute the percentage gap between the throughput of the optimal scheme and the equilibrium throughput of each of the four previously considered schemes under optimal information sharing (the percentage throughput gap is the throughput difference divided by the throughput of the optimal mechanism). We present statistics of these percentage throughput gaps in Table 2, including the mean, median, maximum, minimum, first quartile, and third quartile.

Table 2 indicates that the mixed–capped ordering scheme is overall the closest to the optimal mechanism in throughput with a mean throughput loss of 1.497% (even though it does not always dominate the cancellation scheme). Mixed ordering with cancellation is also effective overall but shows more variability in

Table 2. The Percentage Throughput Gap of Different Schemes Relative to the Optimal Scheme

Scheme	Mean	Median	Maximum	Minimum	First quarter	Third quarter
Mixed–capped ordering	1.4966%	1.4418%	9.5950%	0	0.3421%	2.0037%
Mixed ordering with cancellation	1.9964%	1.6836%	11.3673%	0	0.6550%	2.3557%
Mixed ordering	2.4714%	1.8587%	20.0085%	0	0.6248%	2.7426%
On-site only	3.9178%	2.0557%	90.1412%	0	0.7874%	3.7615%

Note. $\mu = 1, c = 0.5, V \in \{1.5, 2.5, 3.5, 4.5, 5.5\}, \gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}, \beta \in \{0.5, 1, 1.5, 2, 2.5\}, \Lambda \in \{0.5, 1, 1.5, 2, 2.5\}$.

performance. The mixed ordering scheme without capping or cancellation is generally not as effective as the one with capping or the one with cancellation. The on-site–only ordering scheme is overall the least effective in throughput performance, demonstrating the biggest gaps in all statistical measures (even though it outperforms the mixed ordering scheme in some instances). We also observe that the mixed–capped ordering scheme can have a more noticeable gap relative to the optimal scheme when most of the customers are local (e.g., $\gamma = 0.1$) and customers' reward from service is low (i.e., $V = 1.5$). In such instances, capping remote orders is not the focus of the optimal admission control, which instead shifts to the vast majority of local customers who would balk in equilibrium as soon as they see a moderately long queue. The optimal scheme forces these customers to join by essentially not letting them see the queue length; the practicality of such a coercive approach can be up for debate. In more relevant scenarios in which the service reward is not extremely low and remote customers are more significant constituents, the mixed–capped ordering scheme performs well with a consistently small throughput gap. This near-optimality is also clearly illustrated in Figure 10. In sum, our numerical study underscores the value of ordering ahead and points to order capping in the mixed ordering scheme as an effective control lever that balances simplicity and performance.

6. Extensions

6.1. Food Quality Degradation

This extension incorporates the issue of food quality degradation. Specifically, when remote customers order ahead, food can be ready before customers arrive and, thus, may be "soggy" at the time of pickup. Our base model assumes away the disutility caused by soggy food and still finds that the mixed ordering scheme may result in lower throughput than the on-site–only ordering scheme. Incorporating such disutility in ordering ahead implies that remote customers are even less inclined to place orders, leading to even lower throughput, thus only strengthening this key insight. In Online Section EC.1.1, we formally model food deteriorating in quality over time after an order is complete. We find that our most interesting results that occur when travel is fast are particularly robust in that incorporating food quality degradation hardly affects the system throughput of any scheme. This is because, when travel is fast, customers are likely to arrive at the service facility before their order is complete, making food quality degradation a secondary concern.

6.2. Channel Choice

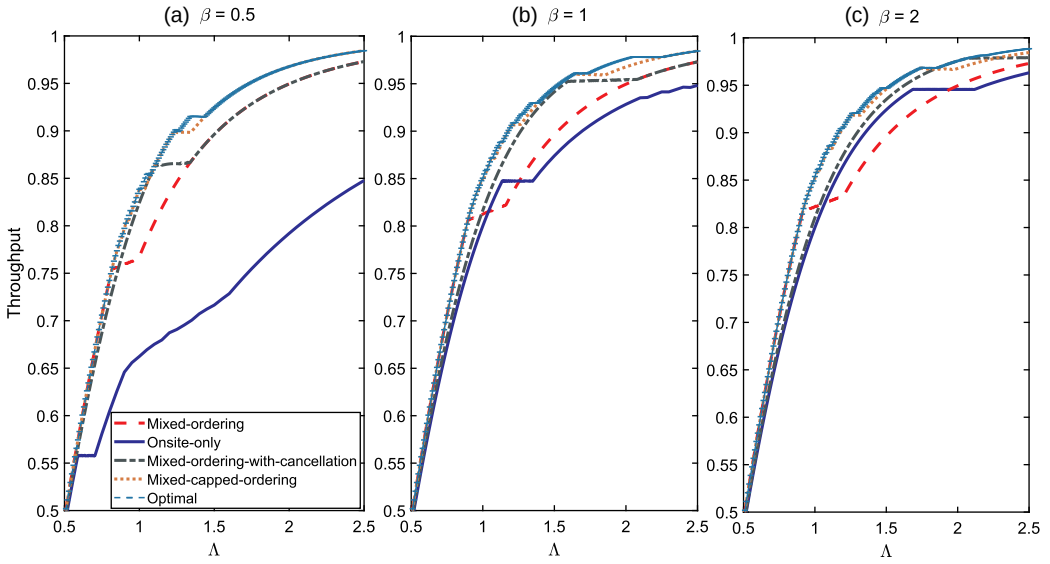
This extension expands remote customers' strategy space and allows for channel choice in mixed ordering schemes. When a need arises, remote customers decide whether to order ahead, order on-site, or not order at all. That is, remote customers not only choose whether to order (as in the base model), but also from which channel to order. When queue information is shared, remote customers do not choose to order on-site, and therefore, our results from the base model carry over. However, when queue-length information is not shared with remote customers, remote customers face a trade-off in the channel choice: ordering ahead allows an order to join the queue earlier, but ordering on-site prevents customers from unknowingly joining a long queue. Thus, a remote customer may choose to order on-site with a certain probability. Nevertheless, if remote customers can cancel their orders upon arrival at the service facility, then they again do not order on-site. Thus, even when queue information is not shared with remote customers, modeling the channel choice may only affect the mixed ordering scheme (with or without capping) but not the cancellation scheme or the on-site–only ordering scheme. In Online Section EC.1.2, we formally characterize the order-placing equilibrium in these affected schemes and demonstrate the robustness of our insights.

6.3. Heterogeneous Travel Speed of Remote Customers

This extension allows remote customers' travel speed to be heterogeneous. Let remote customers' travel speed β be continuously distributed over support $[a, b]$, where $0 \le a < b \le \infty$. For a remote customer with travel speed β, the travel time is drawn from an exponential distribution with rate β. In the mixed ordering scheme, each remote customer chooses whether to order ahead, order on-site, or not order based on the customer's own travel speed β. We set up the model in Online Section EC.1.3 and characterize remote customers' ordering strategy in Proposition 6.

Proposition 6 (Double-Threshold Strategy). *Under heterogeneous travel speed, in the mixed ordering scheme*

Figure 10. (Color online) Throughput Comparison with the Optimal Mechanism

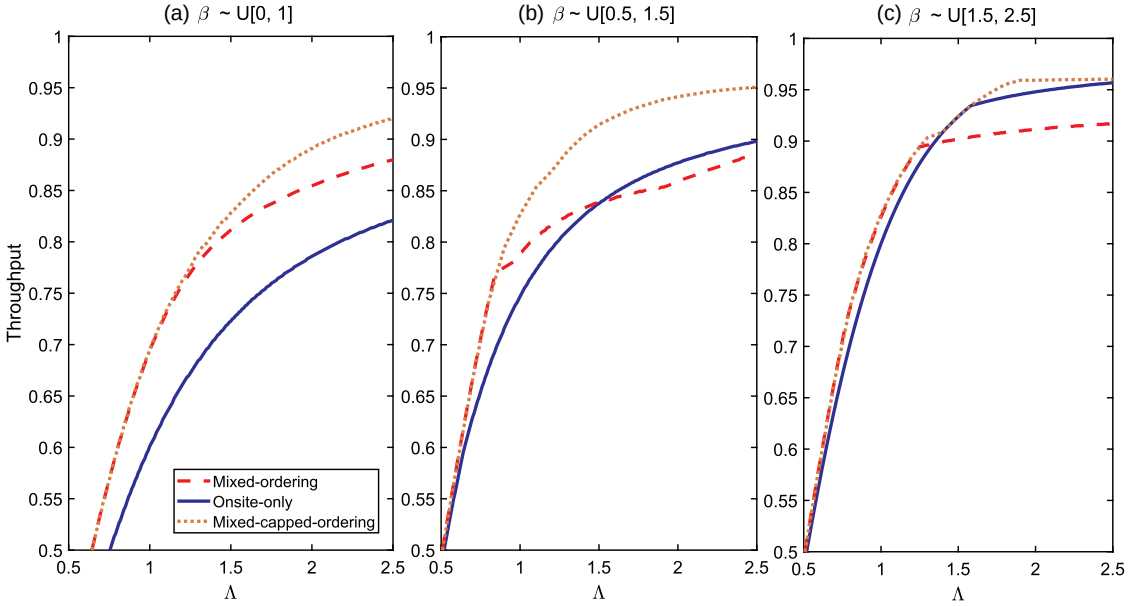


Note. $\mu = 1, V = 2, c = 0.5, \gamma = 0.7$.

without queue-length information shared with remote customers, there exist two thresholds β_1, β_2 *with* $a \le \beta_1 \le \beta_2 \le b$ *such that a remote customer with travel speed* β *does not order if* $\beta < \beta_1$*, orders ahead if* $\beta_1 \le \beta \le \beta_2$*, and orders on-site if* $\beta > \beta_2$.

Proposition 6 shows that customers adopt a double-threshold ordering strategy in the mixed ordering scheme. Those with a high β (those who live near or travel fast, e.g., by car) order on-site because the benefit of ordering ahead (parallelization) for these customers is outweighed by the benefit of ordering on-site (queue-length information); those with an intermediate β order ahead because their time savings from ordering ahead (because of parallelization) is significant enough to prevail over the lack of information; and those with a low β (those who live far or travel slowly, e.g., by foot) do not place orders because they expect too much delay with either ordering mode. We numerically compare the throughputs of three schemes: (i) on-site only, (ii) mixed ordering, and (iii) mixed–capped ordering (with the optimal capping threshold) when queue-length information is not shared with remote customers. We observe from Figure 11 that, consistent with the base model, the mixed ordering scheme can have lower throughput than the on-site–only ordering scheme, yet

Figure 11. (Color online) Throughput Comparison Under Heterogeneous Travel Speed of Remote Customers



Note. $\mu = 1, V = 2, c = 0.5, \gamma = 0.7$.

introducing order capping into the mixed ordering scheme restores its throughput advantage. In fact, we analytically prove this result of throughput dominance in Theorem 7.

Theorem 7. *Under heterogeneous travel speed, when queue-length information is not shared with remote customers, the mixed–capped ordering scheme (with the capping threshold optimized) has higher throughput than the on-site–only ordering scheme.*

We acknowledge that one limitation of this extension is the omission of the cases in which queue-length information is shared with remote customers and those that permit order cancellations because of their intractability. Because of the heterogeneity in remote customers' travel speed, if information is shared or cancellation is allowed, then the computation of throughput requires deriving the steady-state distribution of a high-dimensional Markov chain that tracks the travel speed of every single traveling customer en route to the service facility, which would be best left for future research.

7. Conclusion

A key value proposition of letting customers order ahead is that doing so presumably attracts more orders and achieves higher throughput than if customers must order on-site. Our paper cautions that whether ordering ahead delivers this value hinges on the way it is operationalized. Specifically, a common practice in the field—all orders are final once placed—can generate orders that are placed and locked in when the queue is already long, burdening the service system and prolonging congestion-driven delay, which, in turn, deters customers from placing orders. As a result, a mixed ordering scheme in which remote customers can order ahead may alarmingly achieve lower throughput than if remote customers can only order on-site.

To overcome this throughput shortfall, we consider a variety of mitigation strategies. The first one is to let restaurants optimally choose whether to share queue-length information with remote customers. We find the throughput shortfall can persist after this intervention. The second strategy is to allow remote customers who order ahead to cancel orders after they arrive at the service facility. Whereas this strategy is promising in eliminating the throughput shortfall, it triggers a new problem as allowing cancellation reduces throughput when the market size is small. The third strategy is to stop accepting new remote orders at the outset in the event of too many outstanding orders. Such a mixed–capped ordering scheme outperforms both the one without order capping and the on-site–only ordering scheme but not necessarily the mixed-ordering-with-cancellation scheme. Finally, we formulate an optimal mechanism as an admission control problem subject to individual rationality constraints and numerically find that overall, the mixed–capped ordering scheme has the smallest throughput gap from the optimal mechanism among all the simple schemes considered.

We conclude by discussing the caveats of our model and future research directions. First, a practical concern of order capping and cancellation is the loss of goodwill, which might hurt future business. Second, our paper focuses on on-demand services (Taylor 2018) in which customers value instant gratification and prefer to have their requests fulfilled as soon as possible. That is why our model assumes that customers incur the same unit delay cost regardless of the nature of the delay (on-site or during travel), consistent with Hassin and Roet-Green (2021). Thus, in such an on-demand setting, customers have no incentive to postpone their travel because delay is equally costly regardless of where it occurs. In settings without such a salient on-demand feature, one may argue that waiting at home is less annoying and, thus, less costly than waiting at the service facility and that customers can have an incentive to postpone their travel after placing their order. Such strategic postponement prolongs total delay and can be left for future research.

Acknowledgments

The authors express their gratitude to the review team for their constructive feedback. The authors are grateful to the Associate Editor and the four reviewers, whose detailed suggestions significantly sharpened the focus and improved the quality of this manuscript. A special note of thanks goes to the Department Editor, Guillaume Roels, for his exceptional vision and invaluable guidance that helped us realize the full potential of this work.

Endnote

[1] We thank the review team for suggesting this formulation of the optimal scheme.

References

Ata B, Peng X (2018) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Oper. Res.* 66(1):163–183.

Baron O, Chen X, Li Y (2023) Omnichannel services: The false premise and operational remedies. *Management Sci.* 69(2):865–884.

Business Research Insights (2024) Online food ordering market size, share, growth, and industry analysis regional forecast by 2031. Accessed February 6, 2026, https://www.businessresearchinsights.com/market-reports/online-food-ordering-market-102644.

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

Chen M, Hu M, Wang J (2022) Food delivery service and restaurant: Friend or foe? *Management Sci.* 68(9):6539–6551.

Cui S, Wang Z, Yang L (2020) The economics of line-sitting. *Management Sci.* 66(1):227–242.

Dean G (2021) Former Starbucks workers say the chain's mobile ordering is out of control. *Business Insider Online* (June 26), https://www.businessinsider.com/starbucks-mobile-ordering-app-barista-pandemic-coffee-customers-online-digital-2021-6.

Farahani MH, Dawande M, Janakiraman G (2022) Order now, pickup in 30 minutes: Managing queues with static delivery guarantees. *Oper. Res.* 70(4):2013–2031.

Feldman P, Frazelle AE, Swinney R (2023) Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. *Management Sci.* 69(2):812–823.

Gao F, Su X (2018) Omnichannel service operations with online and offline self-order technologies. *Management Sci.* 64(8):3595–3608.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

Hassin R (2016) *Rational Queueing* (CRC Press, Taylor and Francis Group, Boca Raton, FL).

Hassin R, Haviv M (1995) Equilibrium strategies for queues with impatient customers. *Oper. Res. Lett.* 17(1):41–45.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, vol. 59 (Springer Science & Business Media, New York).

Hassin R, Roet-Green R (2021) On queue-length information when customers travel to a queue. *Manufacturing Service Oper. Management* 23(4):989–1004.

Haviv M, Ritov Y (2001) Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems* 38:495–508.

Jian N, Henderson SG (2015) An introduction to simulation optimization. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds. *Proc. 2015 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1780–1794.

Macal C, North M (2014) Introductory tutorial: Agent-based modeling and simulation. Tolk A, Diallo SY, Ryzhov IO, Yilmaz L, Buckley S, Miller JA, eds. *Proc. 2014 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 6–20.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(6):15–24.

Peet's Coffee (2019) Can I cancel my mobile order after it's been placed? Accessed February 6, 2026, https://faq.peets.com/hc/en-us/articles/360025049112-Can-I-cancel-my-mobile-order-after-it-s-been-placed-.

Pucci R (2017) Mobile order and pay ahead: A new sales channel for restaurants and merchants. Mercator Advisory Group. Accessed February 6, 2026, https://www.mercatoradvisorygroup.com/Press_Releases/Mobile_Order_and_Pay_Ahead__a_New_Sales_Channel__Increases_Volume_for_Restaurants_and_Merchants/.

Starbucks (2017) When will my mobile order be ready? Accessed February 6, 2026, https://customerservice.starbucks.com/app/answers/detail/a_id/3041.

Subway (2020) Subway® website ordering and mobile app terms of use—USA only. Accessed February 6, 2026, https://www.subway.com/en-us/legal/order_apptermsofuse.

Taylor TA (2018) On-demand service platforms. *Manufacturing Service Oper. Management* 20(4):704–720.