Store Fulfillment Strategy for an Omni-Channel Retailer

Elnaz Jalilipour Alishah[1], Kamran Moinzadeh[2], Yong-Pin Zhou[3]

[1] Microsoft Corporation, Redmond, WA, USA 98052

[2,3] Michael G. Foster School of Business, University of Washington, Seattle, WA, USA

[1] eljalili@microsoft.com, [2] kamran@uw.edu, [3] yongpin@uw.edu

---

In this paper, we study the fulfillment strategies of an omni-channel retailer who wants to leverage her established offline retail channel infrastructure to help online sales. We consider a single product that is sold in both online and offline channels to non-overlapping markets with independent Poisson demand. The offline store can fulfill online demand at an additional handling and fulfillment cost, *k*, but not vice versa. The retailer makes decisions at three different levels: 1) at the strategic level the retailer must establish a fulfillment structure in terms of where to stock inventory in the two channels, 2) at the tactical level, the retailer decides how much inventory to have for each channel before the season starts, 3) at the operational level throughout the season, as demand unfolds and inventory depletes, the retailer makes rationing decision about whether to use offline inventory to fill online order at any moment. We build separate and integrated models to study these decisions, and find that the optimal rationing decision has a threshold-based structure that depends critically on *k* and the mix of demand between the two channels. Two simple rationing heuristics are proposed and shown to be effective. Furthermore, integrating the rationing policy into higher-level decisions, we show that it can have significant impact on the retailer's stocking and fulfillment structure decisions. As a result, we propose an integrated policy, where the retailer builds separate inventory stocks for each channel but can use the offline inventory to back up online sales, subject to a rationing heuristic procedure. The heuristic is simple, effective, and robust. We discuss the various practical implications of our findings. Finally, numerical test results confirm the analytical findings and also guide us to propose expanded heuristics that work well with multiple offline stores.

---

1. Motivation and Introduction

Online shopping appeals to consumers for its convenience, information abundance, and possible lower price. Spurred by rapid development and spread of Internet and mobile technologies, online shopping has expanded exponentially. Whereas total retail sales in the US grew 4.1% from the 4th quarter of 2015 to the 4th quarter of 2016, e-commerce sales grew 14.3% in the same period (U.S. Census Bureau 2017). Online-only retailers, such as Amazon.com, led the charge but traditional retailers, such as Macy's and Walmart, did not stand still, and also invested heavily in their ecommerce expansion. Walmart, for example, bought Jet.com and ModCloth (Kapner 2017) and is close to acquiring online men's fashion retailer Bonobos (Cao 2017).

For traditional retailers, the expansion of online sales presents two crucial problems: First, online orders require different capabilities from the fulfillment center than offline ones. For example, the fulfillment center needs to be able to pick many items quickly in small batches and combine them for shipment, whereas traditional offline warehouses are set up to move products in large quantities to a smaller number of destinations. To manage online demand, many traditional retailers simply built more warehouses configured specifically for online orders. With an entry price of at least $100 million per warehouse (Banjo *et al.* 2014), this infrastructure investment imposes a huge financial burden. Second, the shift to online shopping reduces traffic to the offline stores, making it harder to manage offline store inventory. Reduction in offline inventory would cause more frequent stockouts in the stores.

Offline stores are also costlier for traditional retailers to operate as they require more expensive space, labor, and are less efficient in their use of space, labor, and inventory due to its smaller scale of risk pooling, as compared with online stores. This gives online retailers a sizable competitive advantage. To cope with that, many retailers, such as The Limited, JCPenney, Staples, and Macy's, are forced to close numerous stores to focus on more profitable locations and product categories. Some, including Radio Shack, Bebe, Payless Shoes, and Toys"R"Us, have all recently announced bankruptcy and store closings (for example, see Kapner 2017).

Increasingly, traditional retailers such as Nordstrom, Macy's, Walmart, and Target have turned to other more productive ways to utilize their offline store in the competition with online retailers. They realized that instead of building more and more expensive warehouses, they could use the abundant inventory/storage capabilities they have already built all around the country – in the form of offline stores –to meet the growing online sales (Mattioli 2012). Using this store fulfillment (*a.k.a.* ship-from-store) approach, when stock in the online warehouse(s) runs out, an online order can be routed to an offline store where a clerk will take the order and pick items from shop floor, pack them up in a backroom, and then ship to the customer.

When used properly, these established offline stores present the retailer with a great opportunity to integrate online and offline fulfillments. The shift to – and the subsequent judicious use of – offline inventory could also help them to alleviate the out-of-stock problems plaguing offline stores. Target reported that 30% of its online orders were already fulfilled from the stores, and that its offline store in-stock performance also improved (Chao 2016).

Store fulfillment is part of a broader set of omni-channel strategies that retailers have been pursuing in recent years, in order to strategically position and use inventory resources across both online and offline channels. Another common approach is in-store pickup of online orders. For customers who want instant gratification, this is an attractive option. It also increases traffic to offline store, which may lead to extra sales. In this paper, we focus solely on the store fulfillment, however. Readers interested in the in-store pickup strategy are referred to Gallino (2014) and the references therein.

While the store fulfillment strategy can help the retailer to more effectively use its inventory in both online and offline stores, it also has downsides. At present, many offline stores are not set up/organized for online order picking, packing, and shipping (Baird & Kilcourse 2011). The logistics costs and inefficiency of stores versus warehouses in handling online orders may result in margin erosion (Manhattan Associates 2011, Weedfald 2014). Moreover, it creates more work and inconvenience for store clerks who must now fulfill online orders and help in-store customers (Banjo *et al.* 2014). Store fulfillment can cost three to four times more when compared with that in an online warehouse (Banjo 2012). According to a PwC survey of CEOs (PwC 2014), 67% of all 410 respondents rank fulfillment cost as the highest cost for fulfilling orders. Thus, a successful omni-channel fulfillment strategy must seek balance between satisfying online demand and curbing fulfillment cost. It is our aim in this research to derive efficient and profitable fulfillment strategies and provide insights about managing store fulfillment.

While our research is motivated by the store fulfillment strategy adopted by large retailers, our model is equally applicable to smaller, offline retailers that are moving to become omni-channel. The initial heavy capital requirement for opening an online channel, including building a website with all its associated e-commerce functions (billing, fulfillment, processing returns, etc.), poses a big challenge to small retailers. Seizing this opportunity, a number of e-commerce platforms provide fulfillment services. For example, Fulfillment by Amazon (FBA) offers to manage inventory and fulfillment for independent sellers, who would retain ownership of their inventory, but let FBA handle the physical stocking, handling, and shipping of the products. Facing such choices, sellers must decide whether to use such services and, if so, how to coordinate the management of inventory with their existing offline store. From private communications, we know some Amazon sellers will completely rely on FBA to manage all of their inventory, yet others will divide up inventory between FBA and their own warehouse, and use the latter to satisfy both demand generated by their own website and Amazon demand that exceeds the inventory placed with FBA.

For products with a short sales season, the retailer may also have to make real-time inventory rationing decisions, when both online and offline demand chase after a limited quantity of offline inventory. The Yeti Rambler was such a highly sought after product for the 2015 Christmas season. Some retailers, such as

Illinois-based Ace Hardware Corp. had to cut off using offline inventory to fulfill online orders during the last few weeks of the holiday season, in order to prioritize sales to local, offline customers who they believe are more profitable.

Our paper aims to tackle the fulfillment problem at all three levels described above. At the strategic level, the retailer must decide whether to stock channel-specific inventory or rely on just the offline inventory to fulfill both demand streams. Then, at the tactical level once the fulfillment structure is determined, the retailer must decide the amount of inventory to stock. Finally, at the operational level, the retailer must be able to ration the remaining offline inventory, in real time, between the two demand streams to maximize profit. We refer to these three decisions as the fulfillment structure, stocking, and rationing decisions, respectively.

Our contribution to the academic literature and business practice is four-fold: First, we build an integrated model to tackle all three of the problems described above. Second, our model is set in a realistic continuous-time framework and, we can characterize the structure of the optimal rationing policy through which we develop two simple yet effective heuristics. Third, we are able to provide concrete insights and guidance to the omni-channel retailer regarding its fulfillment structure and stocking decisions; namely, the retailer should shift some of its online inventory to the offline channel and use a judicious rationing policy to achieve profit maximization. Fourth, using an extensive numerical study, we demonstrate the value of integrating all these decisions, and show that our proposed approach is both profit-efficient and robust.

The rest of this paper is organized as follow. In Section 2, we review several related literature streams. In Section 3, we build analytical models to study the retailer's problems. In Section 4, we use an extensive numerical study to further explore the results developed in Section 3, and gain managerial insights into the retailer's decisions. Finally, we conclude in Section 5.

2. Literature Review

Our study of the retailer's fulfillment strategy is closely related to the literature on e-fulfillment and multi-channel distribution (see Agatz *et al.* 2008, de Koster 2002, Ricker and Kalakota 1999). Both Bretthauer *et al.* (2010) and Alptekinoglu and Tang (2005) study static allocation followed by Mahar *et al.* (2009) who consider the dynamic allocation of online sales across supply chain locations. More recently, Mahar *et al.* (2010) and Mahar *et al.* (2012) explore store configuration when in-store pickups and returns are allowed. The paper that comes closest to ours is Bendoly *et al.* (2007) who study whether online orders should be handled in a centralized or decentralized fashion. In our paper, not only do we compare these fulfillment structures, we also integrate this decision with the stocking and rationing decisions. This makes our approach more practical and closer to the omni-channel ideal.

Another stream of literature that's closely related to the fulfillment structure and stocking aspects of our research is that on inventory pooling, which started with Eppen (1997) who showed the benefit of warehouse consolidation in a single-period setting. This seminal work has since been extended to include the examinations of correlated and general demand distributions (Corbett and Rajaram 2006), demand variability (Gerchak and Mossman 1992, Ridder *et al.* 1998, Gerchak and He 2003, Berman *et al.* 2011, Bimpikis and Markakis 2014), and holding and penalty costs (Chen and Lin 1989, Mehrez and Stulman 1984, Jönsson and Silver 1987). Some researchers have identified conditions under which pooling may not be beneficial, such as service levels less than 0.5 (Wee and Dada 2005) and right skewed demand distribution under product substitution (Yang and Scharge 2009). When the demand streams are non-identical, Eynan (1999) shows numerically that if the margins are different, lower margin customers serve as a secondary outlet of leftovers. Ben-Zvi and Gerchak (2012) model demand pooling with different shortage cost, and show that retailers are better off if they pool their inventory and give priority to customers with higher underage cost when allocating inventory *after demand is realized*.

Our model differs in two aspects. First, unlike the above models where demand streams are different in only one dimension, our demand streams are different in several dimensions: not only do they vary in margin and leftover cost, the online orders also incur an extra handling and fulfillment cost if they are filled from the offline store. Second, our inventory rationing is performed as demand unfurls in real time, not after all the demand is realized as is the case in many previous works. Similar to the aforementioned papers, we develop our model in a single-period setting. Readers interested in periodic-review inventory pooling are referred to Scharge (1981), Erkip *et al*. (1990), Benjaafar et al (2005), and Song (1994).

The rationing of inventory between online and offline demand in our model is related to three separate but overlapping streams of research: inventory rationing, transshipment, and substitution.

INVENTORY RATIONING The inventory rationing literature is concerned with how to use pooled inventory to satisfy several classes of demand. Kleijn and Dekker (1998) give a review of early papers in the literature. In the periodic-review setting, Veinott (1965) first proves the optimality of threshold based rationing policy. His work is extended by Topkis (1968), Evans (1968), and Kaplan (1969). In the single-period setting, Nahmias and Demmy (1981) present a model for two demand classes and Moon and Kang (1998) extend it to multiple classes.

Our model differs from the existing literature (*e.g.,* Nahmias & Demmy 1981, Atkins & Katircioglu 1995, Frank *et al.* 2003, Deshpande *et al.* 2003, Melchiors *et al.* 2000) in that demand arrivals and decision epochs are continuous within a single, finite period setting. Chen *et al.* (2011) is the only other paper with a similar setting but they approximate the continuous arrivals by discretizing time. Another distinguishing feature of our model is that demand margins are endogenized by the retailer's rationing decision, because the margin on an online demand is lower if it's satisfied by a unit of offline inventory.

LATERAL TRANSSHIPMENT Under lateral transshipment, if one retail store is out of stock, another store can supply it at a cost. Lee (1987) studies a two-echelon model with one depot and *n* identical stores, and evaluates three rules on choosing which store should be the origin of transshipment. Wee and Dada (2005) find the optimal transshipment origin in a similar two-echelon model with one warehouse and *n* identical stores. Unlike these two papers which assume inventory is monitored in continuous time, the majority of works in the literature study the rationing problem in periodic-review inventory models. Moreover, to simplify analysis, they assume that transshipment occurs either at the end of the period after demand is realized (Krishnan and Rao 1965, Tagaras 1989, Tagaras and Cohen 1992, Robinson 1990; Rudi *et al.* 2001), or at the beginning of each period in anticipation of stockout (Allen 1958, Gross 1963, Karmakar and Patel 1977, Herer and Rashit 1999). In contrast, although we study a single-period inventory model, we allow rationing decisions to be made continuously throughout the period, *as demand arrives*. Only a few other papers in the literature allow transshipment decisions *within* a period in the periodic-inventory setting. Archibald *et al.* (1997) use a finite-horizon continuous-time Markov decision process to study whether to use transshipment or place an emergency order. Axsäter (2003) studies a store that uses a (R,Q) policy to replenish from the supplier, supplemented by lateral transshipment. Due to the complexity of the model, he derives a myopic rationing heuristic, which is still too complicated to be incorporated into the stocking problem. In our paper, not only we are able to characterize the optimal rationing policy, we also develop a simple, effective heuristic that could be used in future modeling work. For a more detailed review on lateral transshipments, please see Paterson *et al.* (2011).

SUBSTITUTION Our paper has similarity to those on firm-driven product substitution, because when online inventory runs out, offline inventory can be used as a perfect substitute, at an extra cost. Pasternack and Drezner (1995) consider two substitutable products with stochastic demand within a single period. They show that total order quantity under substitution may increase or decreases depending on the substitution revenue. Bassok *et al.* (1999) show concavity and submodularity of the expected profit function under various assumptions in a single-period setting with downward product substitution. Deflem and Van Nieuwenhuyse (2013) examine the benefits of downward substitution between two products in a single-period setting. Again, all these papers make the simplifying assumption that substitution occurs at the end of the period after demand is realized. In contrast, in our paper, substitution decisions are made in real time. For a review on the substitution literature, please see Shin *et al.* (2015).

In recent years there have been a few modeling papers that discuss the effect of integrating inventory and pricing decisions across online and offline channels. Aflaki and Swinney (2017) focuses on the impact of such virtual inventory pooling on the pricing in both channels and strategic customers' responses. Harsha *et al.* (2017) uses a mixed integer program heuristic to rebalance inventory of a markdown item within a finite time period across online and offline stores by dynamically adjusting prices in all these stores. When implemented, the heuristic reports 6-12% increase in markdown revenue. Similarly, Lei *et al*. (2016) study the joint pricing and inventory fulfillment decisions for an item in a finite selling season. As the optimal solution is intractable, they propose a number of heuristics. Even though we do not study the pricing decision, our model applies to inventory management of items that can be replenished over time, and we study fulfillment structure and inventory ordering decisions that are absent in these two papers.

3. Model Setup

Consider the management of a single product for an omni-channel retailer with one online and one offline store. The two stores have independent, exogenous, and non-overlapping Poisson customer arrivals with mean rates of λ_0 and λ_1 (throughout the paper, we will use subscript 0 for all the online-related parameters and variables, and 1 for the offline-related ones). We focus on the inventory management of this product during a fixed sales period $[0,T]$. Before the sales period starts, the retailer makes the fulfillment structure and stocking decisions. There is no replenishment during the sales period, so the retailer dynamically rations its fixed inventory between online and offline demands as they unfold during the period, in order to maximize profit.

Let p_n denote the product's fixed unit profit margin at store $n \in \{0,1\}$. When online demand is satisfied by offline inventory, there is an extra unit cost of k (similar to Axsäter 2003), representing the higher handling, overhead, and shipment costs in the offline store compared to warehouses. Many retailers match online and offline sales prices, but some don't. We impose no restriction on the relationship between p_0 and p_1 except that $p_1 > p_0 - k$. This is a reasonable assumption that indicates the offline store prefers satisfying its own customer to an online one (which is clearly the case in the Yeti Rambler example, Nassauer 2015).

While some retailers integrate pricing and inventory decisions (e.g., Aflaki and Swinney 2017, Harsha *et al.* 2017), and it is possible that the retailer's fulfillment, inventory, and rationing decisions can impact customers' choice of purchasing channel (hence the demand rates), these are worthwhile questions beyond the scope of this paper. In our model we have assumed exogenous prices and demand rates in order to focus on the key fulfillment questions.

Let S_n be the initial stocking level at store $n \in \{0,1\}$. Any unsatisfied demand will be lost, and any leftover in store n at the end of the sales period will incur a unit cost of $h_n \le p_n$. This assumption applies to many products such as those in garment or fashion industry, and is commonly used in the literature (see, for example, Yang and Schrage 2009).

As described earlier, using the store fulfillment approach, the retailer needs to make decisions at three levels: strategical, tactical, and operational, which are modeled as follows.

- ***Fulfillment Structure Decision:*** At the strategic level, the retailer must decide whether to stock only S_1 in the offline store and use it to satisfy both online and offline demands (denoted as the Pooling, or P, structure) or stock S_0 and S_1 in the two stores respectively (denoted as the Non-Pooling, or NP, structure).
- ***Stocking Decision:*** At the tactical level, retailer must decide on the appropriate level of S_1 and S_0 (in the P structure $S_0 = 0$).
- ***Rationing Decision:*** At the operational level, the retailer must ration offline inventory, in real time, between the two demand streams with margin of p_1 and $p_0 - k$ to maximize profit.

While these decisions can be analyzed separately, we find the value of integrating these three levels of decision in our model creates more practical insight. In Section 3.1, we focus on the fulfillment structure and stocking decisions assuming no inventory rationing. Then in Section 3.2, we derive the optimal rationing policy and develop two practical, effective heuristics for given inventory level(s). In Section 3.3, we integrate all the decisions.

Although we analyze a stylized model, we also study more general problem settings in later sections:

- In Section 4.4 we numerically investigate the case of non-homogenous Poisson arrivals.
- In Section 4.5 we numerically investigate the case of one online store and multiple offline stores.

We are able to show that the insights generated by the simple stylized model extend to these more general settings.

3.1. Fulfillment Structure and Inventory Stocking Problems with No Rationing

When the retailer makes the fulfillment structure decision, she weighs the pros and cons of the P and NP structures (Figure 1). In the P structure, the offline store reaps the inventory pooling benefit. In the NP structure, online demand can be satisfied using online inventory, thus avoiding the additional cross-channel handling and fulfillment cost k. In this section, we study when each structure should be adopted; furthermore, we derive the associated optimal stocking levels in each structure, assuming first-come first-served among all demand arrivals. As a next step, rationing will be studied and incorporated in Sections 3.2 and 3.3.

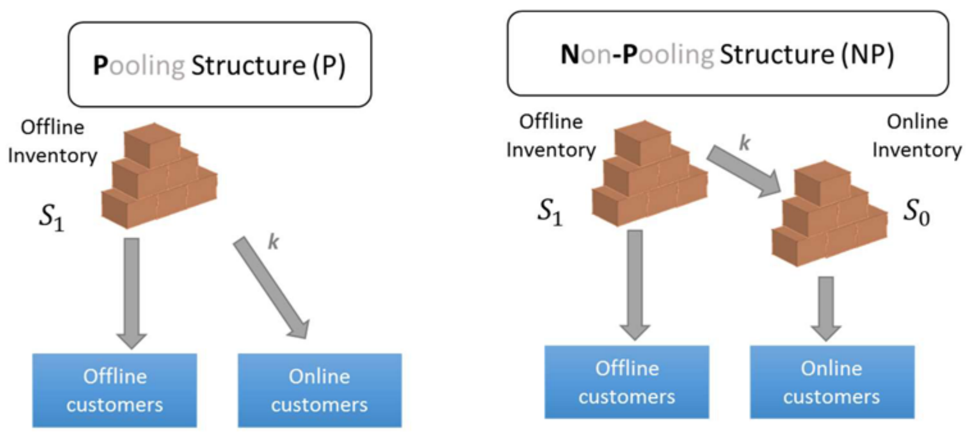

Figure 1 P and NP Fulfillment Structure Designs

We assume that the retailer is risk-neutral and seeks to maximize her expected profit. Let $\Pi^{NP}(S_0,S_1)$ be the retailer's expected profit in the NP structure, given the stocking levels S_0 and S_1, and $\Pi^{P}\left(S_1\right)$ be the retailer's expected profit in the P structure, given the total stocking level S_1.

In the NP structure, the two stores operate as separate newsvendor systems facing independent Poisson demand with average rate λ_n, margin p_n, and leftover cost h_n, $n \in \{0,1\}$. In the P structure, both stores are operated centrally as a single newsvendor system with Poisson demand with average rate $\lambda_{01} = \lambda_0 + \lambda_1$, leftover cost h_1, and a weighted product margin of

$$\omega = \frac{\lambda_1}{\lambda_0+\lambda_1} p_1 + \frac{\lambda_0}{\lambda_0+\lambda_1}\left(p_0 - k\right). \tag{1}$$

Clearly, the cost of using offline inventory to satisfy online demand, k, reduces the overall expected product margin. When k is high enough such that $\omega \le 0$ then the fulfillment structure decision becomes trivial as retailer will lose money under P structure, and thus NP will be the best choice. Therefore, we only consider $\omega \ge 0$.

The newsvendor results from Hadley and Whitin (1963) are summarized in Proposition 1. Following their notation, we denote the PDF and the complementary CDF of a Poisson random variable with a rate of λ by $p\left(j,\lambda\right)$ and $P\left(j,\lambda\right)$, respectively. When necessary we use superscript P and NP to indicate the two fulfillment structures.

Proposition 1 (Hadley and Whitin 1963, Chapter 6.2, pages 297-299)

(1) In the NP structure, the optimal inventory level for store $n \in \{0,1\}$, S_n^{NP}, is the largest S such that:

$$P\left(S,\lambda_n T\right) \ge \frac{h_n}{h_n + p_n}. \tag{2}$$

Furthermore, the retailer's optimal total profit function is:

$$\Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right) = \sum_{n=0}^{1}\left(p_n S_n^{NP} - \left(p_n + h_n\right)\sum_{y=0}^{S_n^{NP}}\left(S_n^{NP} - y\right) p\left(y,\lambda_n T\right)\right). \tag{3}$$

(2) In the P structure, the retailer's optimal inventory level at the offline store, S_1^P, is the largest S such that:

$$P\left(S,\lambda_{0,1} T\right) \ge \frac{h_1}{h_1 + \omega}. \tag{4}$$

Furthermore, the retailer's optimal total profit function is:

$$\Pi^{P}\left(S_1^{P}\right) = \omega S_1^P - \left(\omega + h_1\right)\sum_{y=0}^{S_1^P}\left(S_1^P - y\right) p\left(y,\lambda_{0,1} T\right). \tag{5}$$

Proposition 1 shows the optimal stocking levels in the NP and P structures. The next proposition compares these quantities. All proofs in this paper can be found in the Electronic Companions.

Proposition 2

(1) $S_1^P \geq S_1^{NP}$.

(2) $S_0^{NP} + S_1^{NP} - S_1^P$ is strictly decreasing in k. Moreover, there exists $\dot{k} \geq 0$ such that $S_1^P \leq S_0^{NP} + S_1^{NP}$ for all $k \geq \dot{k}$.

In the P structure, the offline inventory is used to satisfy both online and offline demand. Thus, the offline inventory must increase accordingly, which explains (1) in Proposition 2. This result is expected, but its proof is non-trivial due to the presence of extra handling and fulfillment cost k.

The second part of Proposition 2 requires some additional explanation. Since there is no rationing, all demands are filled on a first-come-first-served basis. Therefore, for larger k, the retailer has lower product margin ω; thus, the base stock level is lower, from (4). When k is sufficiently high (*i.e.*, $k \geq \dot{k}$), the P structure has lower base stock (and thus, lower inventory levels) than that in the NP structure. However, this is driven by the smaller product margin ω due to k, which is different from the traditional pooling benefits in the absence of k.

On the other hand, for smaller values of k, rationing has a smaller impact on the system. Thus, the comparison of P and NP mirrors that in the traditional inventory pooling literature, where pooling can reduce inventory if the products being pooled are similar to each other. Otherwise, pooling can counter-intuitively increase inventory levels (e.g., see Ben-Zvi and Gerchak 2012 for numerical examples). In our setting, we also show that pooling reduces inventory when the two stores are identical ($p_0 = p_1$ and $h_0 = h_1$). This idea is extended in the following corollary, where we also give a simple sufficient condition under which pooling always reduces the total system inventory. Please note that in practice $h_0 \leq h_1$ and it's also very common to have $p_0 = p_1$.

Corollary 1 When $p_0 \leq p_1$, $h_0 \leq h_1$, and $\dfrac{h_0}{p_0} \leq \dfrac{h_1}{p_1}$, $S_1^P \leq S_0^{NP} + S_1^{NP}$. A special case is when $p_0 = p_1$ and $h_0 \leq h_1$.

Next, we compare the retailer's expected profit in these two structures.

Proposition 3

(1) $\Pi^P\left(S_1^P\right) - \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$ is decreasing in k.

(2) There exists a finite $\bar{k}$ such that $\Pi^{P}\left(S_1^{P}\right)-\Pi^{NP}\left(S_0^{NP},S_1^{NP}\right)>0$ if and only if $k<\bar{k}$.

Proposition 3 states that the preference of one structure to the other has a threshold form: smaller values of k favor P and larger ones favor NP, with the threshold being $\bar{k}$. (When the NP structure is always preferred, $\bar{k}$ is set to be zero.) A numerical example is presented in Figure 2. This result is intuitive as large values of k impose heavy penalty for every fulfillment of online demand by offline inventory, pushing the retailer to carry online-specific inventory.

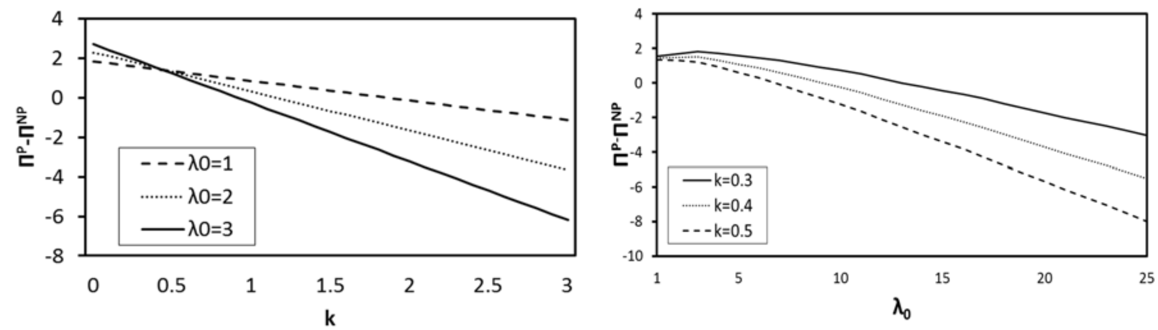

Figure 2 Difference between the optimal P profit and the optimal NP profit as k and λ_0 increase

$(\lambda_1=10,\, p_0=10,\, p_1=10,\, h_0=1,\, h_1=1)$

From Figure 2, we further observe that the threshold $\bar{k}$ is decreasing in λ_0. That is, when online demand is large, it makes more sense to have online-specific inventory in order to avoid the cross-channel handling and fulfillment cost k. The next proposition gives theoretical support to this observation.

Proposition 4 $\Pi^{P}\left(S_1^{P}\right)-\Pi^{NP}\left(S_0^{NP},S_1^{NP}\right)$ is submodular in k and λ_0.

Submodularity means that the threshold on k found in Proposition 3 is decreasing in λ_0. Figure 3 depicts a typical dominance map of the P and NP structures. There is a monotonically decreasing switching curve in the k-λ_0 space. Below the curve, inventory pooling (P structure) is preferred and above it channel-specific inventory (NP structure) is preferred. A similar threshold on λ_0 is numerically observed by Bendoly *et al.* (2007).

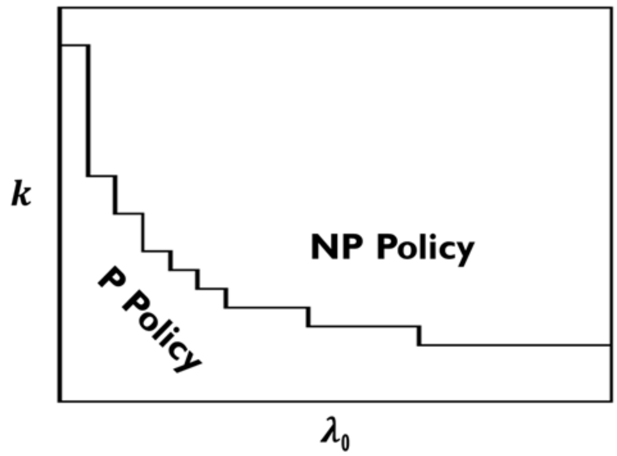

Figure 3 P versus NP Structure Decision Map ($\lambda_1 = 10, p_0 = 10, p_1 = 10, h_0 = 1, h_1 = 1$)

The P structure always leads to lower inventory (Proposition 2), but not necessarily higher profit (Proposition 4), due to the cross-channel fulfillment cost k which is an important distinguishing feature of our model. This result adds to the existing literature on inventory pooling, which has explored the benefits and effect of pooling in regard to demand distribution, demand correlation, and cost parameters asymmetry (Yang and Schrage 2009, Gerchak and Mossman 1992, Pasternack and Drezner 1991).

3.2 Inventory Rationing Problem

Our analysis in Section 3.1 does not incorporate any real-time inventory rationing. However, since a unit of offline inventory gets a higher margin when it is used to satisfy an offline demand ($p_1 > p_0 - k$), it may be more profitable to protect some offline inventory for possible future offline customers, rather than using them to satisfy immediate online customers (Nassauer 2015). Therefore, any offline demand is always satisfied as long as there is offline inventory, but this is not necessarily true with online demand. Let θ $(0 \le \theta < T)$ denote the start time of rationing (*i.e.,* when online demand starts to be routed to the offline store.) In the P structure, $\theta = 0$ which means rationing starts at the beginning of the season. In the NP structure, θ represents the instant online store runs out of inventory. From θ onward, all online demand is routed to the offline store but due to the rationing policy, it is not always filled.

Our inventory rationing model is similar to the multi-class revenue management (RM) capacity allocation problem (Talluri and Van Ryzin 2004). A common solution approach is to approximate the problem using discrete-time setting dynamic program, where each time interval has a demand of only 0 or 1. However in this paper we employ a continuous-time framework and seek an exact solution.

Let $t \in [0,T]$ denote the elapsed time from the beginning of the sales season. We formulate the rationing problem as a dynamic program. Because of the memory-less property of Poisson arrival, the state variable for decision making is (i,t) where *i* is the offline inventory level at time *t*. The Bellman equation can be written as follow:

$$V(i,t) = \max_{u\in\{0,1\}} \int_t^T \left(\alpha_0\left(u\left(p_0-k\right)+V\left(i-u,\tau\right)\right)+\alpha_1\left(p_1+V\left(i-1,\tau\right)\right)\right)\lambda_{0,1}p\left(0,\lambda_{0,1}\left(\tau-t\right)\right)d\tau \qquad (6)$$

where u represents the decision to accept ($u=1$) or reject online demand ($u=0$). If an online demand is routed to the offline store, the retailer can accept it and make a profit p_0-k, or reject it and protect the inventory unit for possible future offline use. The latter is optimal if and only if the expected future value of the (protected) marginal inventory unit exceeds p_0-k.

Intuitively, the marginal value of an extra unit of inventory should be higher when there is more time left in the sales season to sell it. Moreover, for a fixed time, this value should be decreasing in the existing inventory level: more inventory means a higher probability for this unit to be left over. Therefore, we expect $V(i,t)$ to be concave in i, and submodular in (i,t). The following lemma formalizes this observation. It is based on Liang (1999), which studies a revenue management problem in a continuous-time setting.

Lemma 1 (Liang 1999) $V(i,t)$ is concave in i and submodular in (i,t).

As discussed above, when employing a rationing policy, the retailer compares the expected value of a marginal inventory with p_0-k. The concavity in i shows decreasing marginal value in i, therefore if it is optimal to accept an online demand when on hand inventory is *i*, then it will also be optimal to accept online demand at inventory levels larger than *i*. We define a policy with such a property as a threshold-based rationing policy, as follow:

There exists an inventory threshold $\tau(t)$ for all $t \in [0,T]$ where an online demand at time $t>0$ is accepted if and only if the offline inventory at time t is above $\tau(t)$.

The following theorem is a key result in our study of the rationing problem.

Theorem 1 There exist a positive integer *l* and points in time $\{t_j\}_{j=0}^{l}$ where $0=t_l<t_{l-1}\dots<t_1<t_0<T$, such that a threshold-based rationing policy is optimal for the offline inventory, with the thresholds defined as follows: $\tau(t)=0$ on $t \in [t_0,T]$ and $\tau(t)=j$ on $t \in [t_j,t_{j-1})$, $\forall 1\le j\le l$.

The main idea behind Theorem 1 is that the marginal value of an additional unit of inventory is decreasing not only in inventory but also in the remaining time. Thus, the inventory-based threshold should also decrease with time. To characterize these thresholds, we first consider t just before the end of the season (*i.e.,* $t \approx T$). Any offline inventory at that point is almost certain to end up as leftover. Therefore, if an online demand occurs, the offline store should accept it and take the sure profit $p_0 - k$. As we move backward from T, more time is left in the sales season, and the probability of the marginal unit being leftover decreases. There comes a time, denoted by t_0, when the marginal value of protecting an offline inventory equals to $p_0 - k$. Then t_0 is the first point (going backwards in time) when the retailer becomes indifferent between protecting and not protecting the last unit of offline inventory. Moving further back from t_0, the retailer now strictly prefers to protect the last unit, so the threshold jumps to one. This procedure can be applied recursively to find the other *indifference times*, t_j, $(1 \le j \le l)$. At those time points, it is equally optimal for the retailer to protect j units and j+1 units. Hence the retailer is indifferent between protecting or not protecting the (j+1)$^{\text{th}}$ unit. For expositional purpose, we break the tie and define the retailer to protect j units at t_j, but not the (j+1)$^{\text{th}}$.

Theorem 1 simplifies the computation of the optimal policy. Instead of calculating the optimal decision $u(i,t)$ for all the (i,t) values, it now suffices to find all the indifference points. Theorem 1 also allows us now to express the value function in a straightforward way. For $t \in [t_j, t_{j-1}]$, the optimal profit function can be written as:

$$\begin{aligned} V(i,t) = &\sum_{m=0}^{i-j-1} \left(\omega m + V\left(i-m, t_{j-1}\right)\right) p\left(m, \lambda_{0,1}\left(t_{j-1}-t\right)\right) \\ &+ \int_{\xi=t}^{t_{j-1}} \left(\omega(i-j) + \sum_{n=0}^{\infty} \left(p_1 (n \wedge j) + V\left((j-n)^+, t_{j-1} \right) \right) p\left(n, \lambda_1\left(t_{j-1}-\xi\right)\right) \right) \lambda_{0,1} p\left(i-j-1, \lambda_{0,1}(\xi - t)\right) d\xi. \end{aligned} \tag{7}$$

We can now use (7) and the indifference property $V(i+1,t) = V(i,t)$ to recursively find all the t_i, $i = 0,1,2,\ldots$.

Figure 4 illustrates the structure of the optimal rationing policy via a numerical example, for different values of λ_0. As expected, when λ_0 increases, the retailer stocks more inventory at time 0, and the optimal rationing policy employs more threshold levels, making it both computationally and practically difficult to use. This makes the case for developing simple yet effective heuristics for practical use.

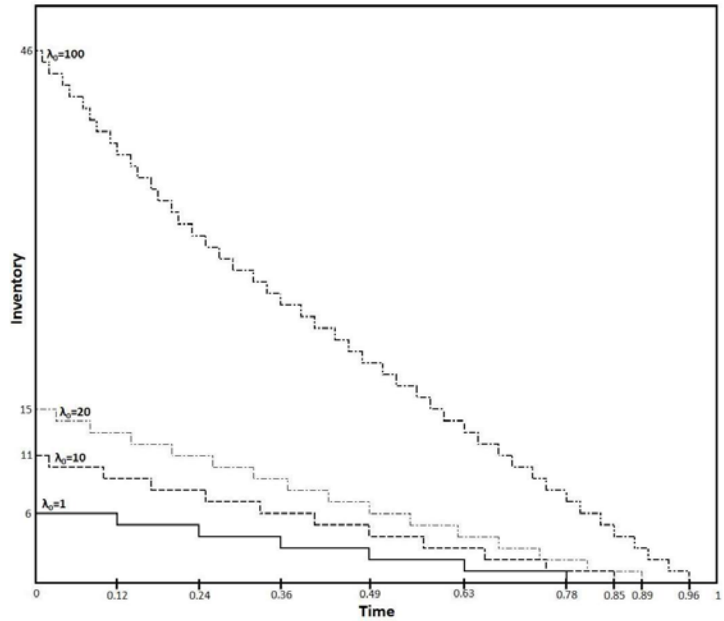

Figure 4 Optimal Rationing Policy change as λ_0 increases ($\lambda_1 = 10, p_0 = 10, p_1 = 10, h_1 = 1, k = 1$)

We develop heuristics using two different approaches. In the first, we limit the number of thresholds to a single one. In the second, we retain the multi-threshold structure of the optimal policy, but use a simple function to approximate $V(i,t)$ in (7). This function does not need to be evaluated recursively, so the indifference points are easier to find. We present these two heuristics in 3.2.1 and 3.2.2 respectively.

3.2.1. Single Threshold (ST) Heuristic

In the optimal rationing policy, the threshold varies over time; therefore, a natural simplification is to use a fixed threshold throughout the season. This replaces the staircase shape of the optimal policy with a fixed horizontal line. Once this single threshold is set at time θ, it is used for the rest of the season. Because of this, the threshold should reflect the level of inventory, i, at time θ: if the inventory level at time θ is high, the retailer should be more concerned about having leftover at the end of the season; therefore, a low value of single threshold should be set, and vice versa. We define this single threshold (ST) heuristic as follows:

Let there be i units of inventory in the offline store at time θ. Under a single threshold heuristic, the offline store accepts an online demand at time $t > \theta$ if and only if the offline inventory at time t is above a calculated threshold level $\tau^{ST}(i,\theta)$.

Under the ST heuristic, the threshold $\tau^{ST}(i,\theta)$ is chosen to maximize the offline store's expected profit for $[\theta,T]$ (note that the online store is already out of stock, this also represents the system profit):

$$\tau^{ST}\left(i,\theta\right)=\arg\max_{\tau} H^{ST}\left(\tau\middle|i,\theta\right)=\sum_{m=0}^{i-\tau-1}\left(\omega m-h_1\left(i-m\right)\right)p\left(m,\lambda_{0,1}(T-\theta)\right)$$
$$+\int_{\xi=\theta}^{T}\left(\omega\left(i-\tau\right)+p_1\tau-\left(p_1+h_1\right)\sum_{n=0}^{\tau}\left(\tau-n\right)p\left(n,\lambda_1\left(T-\xi\right)\right)\right)\lambda_{0,1}p\left(i-\tau-1,\lambda_{0,1}(\xi-\theta)\right)d\xi. \tag{8}$$

Even though $\tau^{ST}\left(i,\theta\right)$ appears more complicated, it is much easier to use than the optimal policy because the threshold needs to be calculated only once and used for the rest of the season. In particular, we note that unlike the profit function of the optimal policy, $V\left(i,t\right)$ in (7), evaluation of $H^{ST}\left(\tau\middle|i,\theta\right)$ in (8) is not recruvie. Thus, the ST heuristic is much simpler to compute than the optimal rationing policy.

Lemma 2

(1) For any fixed i and θ, $H^{ST}\left(\tau\middle|i,\theta\right)$ is concave in τ

(2) For any fixed i, $H^{ST}\left(\tau\middle|i,\theta\right)$ is submodular in $\left(\theta,\tau\right)$.

(3) For any fixed θ, $H^{ST}\left(\tau\middle|i,\theta\right)$ is submodular in $\left(i,\tau\right)$.

The concavity of $H^{ST}\left(\tau\middle|i,\theta\right)$ allow us to find $\tau^{ST}\left(i,\theta\right)$ for any given i and θ using the first order condition. The submodularity properties in parts (2) and (3) imply that $\tau^{ST}\left(i,\theta\right)$ is decreasing in i or θ, when the other is fixed. This makes intuitive sense, as more inventory and less time left to sell should lead the inventory protection level to be lower.

Proposition 5 If the online store runs out of inventory at time θ, $0\le\theta\le T$, and the offline inventory at that time is i, then the offline store should use a single threshold heuristic where $\tau^{ST}\left(i,\theta\right)=\max\left\{\tau:\Delta_{\tau}H^{ST}\left(\tau,\theta\right)\ge 0\right\}$.

3.2.2. Newsvendor Thresholds (NT) Heuristic

We now take a different approach, which simplifies how the thresholds are calculated. To do so, we assume once the threshold is hit, rest of the season all inventory will be protected for offline store customers. As a result, “future” value function $V\left(i,t\right)$ in the right-hand side of dynamic program (7) is replaced by the offline store newsvendor profit function, $G_1\left(i,t\right)=p_1 i-\left(p_1+h_1\right)\sum_{j=0}^{i}\left(i-j\right)p\left(i,\lambda_1\left(T-t\right)\right)$. This way, the calculation of indifference point simplifies to:

$$H^{NT}\left(i,t\right)=\sum_{m=0}^{i-j-1}\left(\omega m+H^{NT}\left(i-m,t_{j-1}\right)\right)p\left(m,\lambda_{0,1}\left(t_{j-1}-t\right)\right)$$
$$+\int_{\xi=t}^{t_{j-1}}\left(\omega\left(i-j\right)+\sum_{n=0}^{\infty}\left(p_1\left(n\wedge j\right)+G_1\left(\left(j-n\right)^+,t_{j-1}\right)\right)p\left(n,\lambda_1\left(t_{j-1}-\xi\right)\right)\right)\lambda_{0,1}p\left(i-j-1,\lambda_{0,1}\left(\xi-t\right)\right)d\xi.$$

We refer to this heuristic, which is based on optimizing $H^{NT}\left(i,t\right)$, the newsvendor thresholds (NT) heuristic.

Lemma 3

(1) $H^{NT}\left(i,t\right)$ is concave in i and submodular in $\left(i,t\right)$.

(2) $\tau^{NT}\left(t\right)=\max\left\{i:\Delta_i H^{NT}\left(i,t\right)\geq p_0-k\right\}$ *for all* $t\in[\theta,T]$.

Again, the concavity of $H^{NT}\left(i,t\right)$ in part (1) of Lemma allows us to use the first order condition to characterize the optimal control in part (2). Therefore, we now know that the NT heuristic works as follows:

Under the NT heuristic, an online demand at time $t>\theta$ is accepted if and only if the offline inventory at time t is above $\tau^{NT}\left(t\right)$.

Note that the threshold under the ST heuristic, $\tau^{ST}\left(i,\theta\right)$, depends on the timing of θ and the offline inventory level at θ, but it is calculated once and used for the rest of the season. In contrast, the NT heuristic employs thresholds that vary over time, but their computations are much simpler due to the use of newsvendor value function $G_1\left(i,t\right)$ which makes the profit function evaluation of the NT heuristic non-recursive. We are, thus, able to compute the indifference points in closed-form solutions, which require a straightforward inversion of a Poisson complementary CDF.

The following proposition shows that these thresholds behave in a similar fashion to the optimal ones and decrease in unit steps over time.

Proposition 6 There exists a positive integer *n* and time points $\{t_j'\}_{j=0}^{n}$, where $0=t_n^{'}<\ldots<t_1^{'}<t_0^{'}<T$, such that the optimal NT threshold is $\tau^{NT}\left(t\right)=j$ for $t\in[t_j^{'},t_{j-1}^{'})$ and $\tau^{NT}\left(t\right)=0$ for $t\in[t_0,T]$. Moreover, the time points $\{t_j'\}_{j=0}^{n}$ are solutions to the following equation:

$$P\left(j,\lambda_1\left(T-t_{j-1}^{'}\right)\right)=\frac{p_0-k+h_1}{p_1+h_1}.$$

Proposition 6 provides the indifference points that are key to the complete characterization of the NT heuristic. The following corollary further establishes that at any time *t*, the NT heuristic thresholds are no greater than the optimal policy thresholds. Thus, in addition to having easier-to-computer thresholds, the NT heuristic also will have fewer thresholds to compute.

Corollary 2 The indifference points under the NT heuristic are no greater than those under the optimal policy. That is, $t_j \geq t_j^{'}$ for all j.

3.3. Integrated Fulfillment Structure, Stocking, and Rationing Policy

In Sections 3.1 and 3.2 we analyzed the fulfillment structure, stocking, and rationing decisions separately. Next, we integrate them into a coherent inventory policy. To that end, we use superscript ^{X,Y} to indicate an integrated policy X,Y where $X \in \left\{P,NP\right\}$ represents the fulfillment structure and $Y \in \left\{\varnothing, OPT\right\}$ represents the no rationing and optimal rationing policies. In the numerical studies in Section 4, we will extend Y to include all the rationing policies, therefore $Y \in \left\{\varnothing, ST,NT,OPT\right\}$ then.

First, we investigate how the use of the optimal rationing policy could impact the retailer's overall profit as well as her inventory stocking levels.

Proposition 7

(1) $\Pi^{P,\varnothing}(S_1) \leq \Pi^{P,OPT}(S_1)$ and $\Pi^{NP,\varnothing}(S_0,S_1) \leq \Pi^{NP,OPT}(S_0,S_1)$ for any fixed S_0 and S_1.

(2) $S_1^{NP,\varnothing} \leq S_1^{NP,OPT}$ and $S_0^{NP,\varnothing} \geq S_0^{NP,OPT}$.

Even though part (1) is stated for any fixed S_0 and S_1, once we maximize over all possible S_0 and S_1, we can see that the profit improvement under rationing holds in optimality as well. This profit increase is expected: Since no rationing is always a feasible action for the optimal rationing policy, profit should increase from the use of optimal rationing. However, we numerically show in Section 4 that the magnitude of such a profit improvement can be substantial, especially in the NP structure.

At the same time, the optimal rationing policy also impacts the retailer's inventory stocking levels. Within the NP structure, the use of optimal rationing helps the retailer to shift inventory from the online store to the offline store where its use is more flexible and, thus, more valuable. We see this inventory shift in practice where large retailers are putting a higher emphasis on using offline stores to satisfy online demand and moving inventory in that direction (Chao 2016). Part b) in the proposition formalizes this intuition.

While the results in Proposition 7 are very intuitive, their proofs are not trivial as they require the establishment of concavity and submodularity of the profit functions in the inventory stocking levels. These are provided in a lemma in the Electronic Companions.

Proposition 7 demonstrates the effects of using the optimal rationing policy on the retailer's inventory and profit, within a fixed fulfillment structure, P or NP. Next, we discuss the impact of the rationing policy on the retailer's fulfillment structure itself.

Broadly speaking, rationing offers benefits at both the operational and the strategic levels. On the operational level, it helps the retailer to better utilize its limited offline inventory, in much the same way the expected marginal seat revenue (EMSR) model helps airlines to sell tickets in the RM literature. By differentiating the two types of customers and rejecting the lower-margin online customers at the appropriate time, the retailer can achieve a higher margin and better inventory utilization. This is how rationing helps in the P structure.

On the strategic level, rationing allows the retailer to combine the benefits of both P and NP as it can now place some inventory in the online store, in order to minimize cross-channel fulfillment costs caused by k. The retailer can choose this amount to be suitably low to minimize leftovers, knowing that it can always use offline inventory and rationing to handle any excess online demand. This is how rationing helps in the NP structure.

For any fixed stocking level(s), either benefit (operational or strategic) could dominate. However, when the retailer sets its inventory level(s) optimally, we expect the operational benefit to be minimized, and the strategic benefit to dominate. Therefore, rationing should offer bigger profit improvement to a retailer that uses the NP fulfillment structure than to one that uses the P fulfillment structure. This will be numerically confirmed in Section 4.

4. Numerical Studies

In this section, we study the retailer's fulfillment structure, stocking, and rationing decisions using a numerical experiment. The three subsections in this section mirror those in Section 3: In 4.1 we compare the performance of P and NP structures when no rationing us employed (*i.e.,* $Y = \varnothing$). In 4.2 we investigate the effectiveness of the rationing policies and study their impact on the retailer's average profit, margin, and inventory stocking level. Finally, 4.3 will be devoted to the integrated policies.

In the numerical study, we normalize $p_0 = p_1 = 10$, and $T = 1$. Then we set λ_0 and k at:

- $\lambda_1 \in \{10, 20\}$ and $\frac{\lambda_0}{\lambda_1} = \lambda_\alpha \in \{0.2,\ 0.5,\ 0.8,\ 1,\ 2\}$,
- $k \in \{0.02, 0.05, 0.1, 0.2, 0.5\} * p_0$.

Since p_0 is fixed in the numerical experiment, setting the online leftover cost h_0 is equivalent to setting service level $SL_0 = \frac{p_0}{p_0 + h_0}$. Thus, set

- $SL_0 \in \{0.65, 0.75, 0.85, 0.95\}$, which implies $h_0 \in \left\{\frac{7}{13}, \frac{1}{3}, \frac{3}{17}, \frac{1}{19}\right\}$.

The offline leftover cost is generally higher than its online counterpart, thus we set:

- $\frac{h_1}{h_0} \in \{1,\ 1.25,\ 1.6\}$, which leads to the following offline store service levels for fixed values of online service level:

SL_0	0.65			0.75			0.85			0.95		
SL_1	0.54	0.60	0.65	0.65	0.71	0.75	0.78	0.82	0.85	0.92	0.94	0.95

In total, 600 cases were considered.

4.1. Fulfillment Structure and Inventory Stocking Decisions

In Section 3.1, we prove that there is a switching curve on the λ_0-k space that demarcates whether the retailer should hold store-specific inventory, or pool all of her inventory at the offline store. Using the 600 instances from the test bed, we studied the switching curve and the performance deviation (%) between the two structures on profit, product margin, and inventory stocking level. To streamline the presentation, we express the results as (NP,ϕ)'s performance percentage deviation from (P,ϕ), defined as:

$$Performance\ Dev = \frac{Performance^{NP,\varnothing} - Performance^{P,\varnothing}}{Performance^{P,\varnothing}},$$

where $Performance \in \{Profit, Margin, Inventory\}$. Note that margin is defined as profit per unit of inventory stocked, and inventory represents the initial inventory stocking level. In the tables to follow, the average performance deviations over the 600 cases are presented.

Table 1 represents the average performance deviation for 25 possible combinations of λ_α and k considered in the study. As can be observed, the results presented in Proposition 3 are confirmed. That is: P is a better fulfillment structure for small values of λ_α and k; furthermore, the switching curve between the two fulfillment structures is monotonically decreasing in both λ_α and k.

Table 1 Policy (NP,ϕ)'s Average Profit Deviation from Policy (P,ϕ) by k and λ_α

Online : offline demand ratio (λ_α)	Handling Cost (k): 0.2	0.5	1	2	5
2	-0.61%	1.54%	5.32%	13.79%	49.75%
1	-1.76%	-0.17%	2.59%	8.60%	31.67%
0.8	-2.08%	-0.67%	1.76%	7.01%	26.54%
0.5	-2.60%	-1.55%	0.25%	4.05%	17.39%
0.2	-2.86%	-2.34%	-1.45%	0.37%	6.24%

It should be noted that throughout our analysis we have ignored the fixed cost associated with building online warehouse, if the retailer adopts the NP structure. Clearly, this is a sunk cost and should not be involved in the fulfillment structure decision when the retailer has already built the online warehouse for other purposes or other items. It is also a reasonable assumption when the NP structure corresponds to using fulfillment services provided by an online platform, since the vast majority of the cost charged by such services (*i.e.,* Amazon Fulfillment) is variable based on volume of fulfillment as the fixed monthly fee is nominal. In cases when the retailer needs to build and operate the online warehouse herself, the fixed cost would only affect the threshold of determining preference of the P or NP structure as the switching curve would be shifted upward and the analysis remain intact.

Of the 600 cases studied, in 385 cases (NP,ϕ) and in the remaining 215 cases (P,ϕ) were the preferred structures. Table 2 presents the average performance deviations of profit, margin, and inventory when a fulfillment structure is preferred. Note that the profit figures in Table 2 can be aggregated to obtain the corresponding ones in Table 1. The other two performance measures can be similarly aggregated.

Table 2 (NP,ϕ)'s Average Performance Deviation from (P,ϕ)

	Policy (NP,ϕ) **Is Preferred (385 cases)**	**Policy** (P,ϕ) **Is Preferred (215 cases)**
Profit Dev	11.22%	-2.14%
Margin Dev	2.55%	-7.28%
Inventory Dev	8.48%	5.62%

To summarize, when (NP,ϕ) is preferred, the profit deviation is significant (11.22% on average). This is accompanied by a big jump in inventory (8.48%), which can be explained by the lack of inventory pooling. Moreover, margin is modestly higher (2.55%). In the remaining cases when either λ_0 or k is low and (P,ϕ) is preferred, (P,ϕ) outperforms (NP,ϕ) by a modest amount (2.14%), but as a result of inventory pooling employment, on average there is a reduction of 5.62% in inventory, leading to a 7.28% increase in product margin.

These observations are in line with our expectation: when the NP structure is preferred, it uses a dedicated online stock to reduce the fulfillment cost of online orders, but that comes at the expense of increased inventory. Thus, whereas profit is higher significantly, the product margin enhancement is modest. Conversely, when the P structure is preferred, the profit improvement is limited, but it does so with lower inventory and a much improved product margin.

In closing, the results presented in 4.1 indicate that in the absence of a rationing policy, selecting the proper fulfillment structure could lead to significant improvement in profit, product margin, and/or inventory.

4.2. Inventory Rationing Decision

BASE TESTS

Next, to isolate the impact of the rationing policy, we fixed the fulfillment structure, and studied the performance of various rationing policies compared to the baseline case when no rationing is employed. Define the performance deviation to be:

$$Performance\ Dev = \frac{Performance^{X,Y} - Performance^{X,\varnothing}}{Performance^{X,\varnothing}},$$

where $Performance \in \{Profit, Margin, Inventory\}$, $X \in \{P, NP\}$, and $Y \in \{ST, NT, OPT\}$. Table 3 provides a summary of the results when the retailer sets the stocking level(s) optimally for each Y.

Table 3 Policy (X,Y)'s Average Performance Deviation from Policy (X,ϕ) (Using Optimal Stocking)

	X = P				**X = NP**		
	Y = ST	**Y = NT**	**Y = OPT**		**Y = ST**	**Y = NT**	**Y = OPT**
Profit Dev	0.01%	0.09%	0.12%		2.13%	2.15%	2.16%
Margin Dev	0.08%	0.34%	0.54%		6.31%	6.38%	6.42%
Inventory Dev	-0.07%	-0.24%	-0.40%		-3.85%	-3.90%	-3.93%

Two notable observations can be made from Table 3. First, employing rationing in the NP fulfillment structure makes a difference. This is true irrespective of which rationing policy (the optimal one or the ST, NT heuristics) is employed. On average, inventory is reduced around 4% and profit is improved by more than 2%. In contrast, while rationing is still beneficial under P, its impact is minimal. This is consistent with our conjecture in Section 3.3 that the benefit of rationing is more significant in the NP structure than in the P structure.

Second, we note that both heuristics perform well in general. Under the NP structure, the two heuristics performances are very close to that of the optimal policy. Let us provide an intuitive explanation for this observation: Under the NP structure, when the retailer sets inventory levels optimally, the rationing start time in the season, θ, usually happens towards the end of the sales period. In such cases, only a few of the optimal policy's thresholds will effectively be used. Therefore, the simplification to a single threshold (in the case of ST) and the approximation of the value function (in the case of NT) will not deviate too much from the optimal policy.

In practice, the two heuristics offer different advantages: ST is simpler and has analytical properties that make it more suitable to be used in analytical modeling. In contrast, NT performs better and is more effective when the rationing start time, θ , happens early in the season – either because the P fulfillment structure is used, or due to an early occurrence of online stockout.

ADDITIONAL TESTS – LOW SERVICE LEVELS

Intuitively, rationing is most valuable when there is abundant demand in the presence of limited inventory. In the following, we extend our experimentations when service levels are small in both channels. Keeping all the other parameters as before, we considered the following combinations of service levels:

SL_0	0.25			0.35			0.45			0.55		
SL_1	0.17	0.21	0.25	0.25	0.30	0.35	0.34	0.40	0.45	0.43	0.49	0.55

This gives us another 600 test cases. Table 4, provides a summary of the magnitude of the performance of the rationing policies (optimal and the two proposed heuristics) under the two fulfillment structures:

Table 4 Policy (X,Y)'s Average Performance Deviation from Policy (X,ϕ) (Using Optimal Stocking)

	X = P				**X = NP**		
	Y = ST	**Y = NT**	**Y = OPT**		**Y = ST**	**Y = NT**	**Y = OPT**
Profit Dev	0.01%	0.36%	0.54%		7.70%	7.81%	7.88%

Note that the profit improvements in Table 4 are significantly higher than those in Table 3, especially under the NP fulfillment structure. This confirms our intuition that rationing is more valuable for lower service level targets – characteristic of products that have either higher leftover cost h or lower margin p. This is because at low service levels, the risk of offline inventory going unsold is low. Thus, it is more important to use a proper rationing policy to protect offline inventory for offline customers.

SUMMARY

The results of the numerical tests considered in 4.2 consistently suggest that rationing could have a significant impact on the retailer's inventory levels and profit, especially in the NP fulfillment structure and/or when service levels are low. Moreover, both heuristics perform extremely well in the NP structure. In the P structure, because there is a longer rationing time window, the multi-threshold NT heuristic tends to perform better than the relatively less flexible, single-threshold ST heuristic. However, both heuristics are shown to be effective.

4.3. Impact of Rationing on Fulfillment Structure and Stocking Decisions

In the previous section, we studied the impact of rationing in the P and NP fulfillment structures. Now, we integrate all the decisions and investigate the impact of rationing on the fulfillment structure as well. Parallel to Table 2, Table 5 summarizes the performance of the operating measures over the two fulfillment structures (P and NP) when the optimal rationing is employed ((P,OPT) and (NP,OPT)).

Table 5 Policy (NP,OPT)'s Average Performance Deviation From (P,OPT)

	(NP,OPT) Is Preferred (595 Cases)	**(P,OPT) Is Preferred (5 Cases)**
Profit Dev	8.56%	-0.13%
Margin Dev	4.59%	-0.13%
Inventory Dev	3.64%	0.00%

As conjectured in Section 3.3, employment of rationing will make the NP fulfillment structure attractive. In fact, one can observe from In the previous section, we studied the impact of rationing in the P and NP fulfillment structures. Now, we integrate all the decisions and investigate the impact of rationing on the fulfillment structure as well. Parallel to Table 2, Table 5 summarizes the performance of the operating measures over the two fulfillment structures (P and NP) when the optimal rationing is employed ((P,OPT) and (NP,OPT)).

Table 5, that in 595 out of 600 cases considered, (NP,OPT) dominated (P,OPT). This is quite an improvement when one notes that (NP, ϕ) was dominant in only 385 cases in Table 2. Moreover, in the remaining 5 cases where (P,OPT) was dominant, the performance of the (NP,OPT) policy was nearly as good.

Note that the overwhelming dominance of the (NP,OPT) policy has practical implications. First, in many practical situations, estimation of system parameters are not accurate. More importantly, while our analysis and results are on a single product, in practice, a retailer handles a large number of such single-seasonal products and thus, it is impractical to expect the retailer to tailor its fulfillment structure for each individual product. Therefore, retailers can achieve near optimal results by choosing the NP fulfillment structure and apply rationing policies when managing stocking and fulfillment of their product families.

Similar to Table 1, Table 6 below breaks down the profit deviation by the k and λ_0 parameter values. Comparing the two tables, we can see clearly that the use of rationing moves the switching curve closer to the lower left corner, just as discussed in section 3.3. The dominance of the (NP,OPT) policy is also apparent: it is optimal in almost all the cases. Depending on the k and λ_0 parameters, the profit advantage of the (NP,OPT) policy also could be very substantial. Table 6 shows a heat map of the profit deviations.

Table 6 Policy (NP,OPT)'s Average Profit Deviation from Policy (P,OPT) by k And λ_α

$\lambda_\alpha = 2$	1.41%	3.40%	6.96%	14.99%	48.89%
$\lambda_\alpha = 1$	0.94%	2.35%	4.86%	10.45%	31.96%
$\lambda_\alpha = 0.8$	0.77%	1.95%	4.08%	8.82%	26.53%
$\lambda_\alpha = 0.5$	0.45%	1.23%	2.66%	5.84%	17.25%
$\lambda_\alpha = 0.2$	0.14%	0.56%	1.32%	3.02%	9.14%
	$\boldsymbol{k = 0.2}$	$\boldsymbol{k = 0.5}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 2}$	$\boldsymbol{k = 5}$

We already know from Table 3 that both (NP,ST) and (NP,NT) are effective policies within the NP fulfillment structure. Results of a further analysis, presented in Table 7 below, shows that (NP,ST) and (NP,NT) have nearly as good a performance against (P,OPT). Therefore, we conclude that the retailer should consider using either (NP,ST) or (NP,NT) as its stocking-rationing policy in all situations.

Table 7 Policy (NP,ST)/(NP,OPT)'s Average Performance Deviation From Policy (P,OPT)

	(NP,ST) or (NP,NT) Is Preferred (595 Cases)	**(P,OPT) Is Preferred (5 Cases)**
Profit Dev.	8.53% / 8.55%	-0.13% / -0.13
Margin Dev.	4.47% / 4.55%	-0.13% / -0.13
Inventory Dev.	3.73% / 3.67%	0.00% / 0.00

4.4. Extension to the Cases When Demand Follows Non-Homogenous Poisson Process

In this section we extend our numerical experimentation to cases where demands are non-homogenous Poisson processes (NHPP), and study how time-varying demand patterns would favor P or NP structure.

First, note that in the absence of rationing, the base stock levels only depend on the total demand during the season and not on how it is spread over the season. When rationing decisions are allowed, however, having different demand rate patterns will make a difference. Using a standard time-scaling technique (Law and Kelton 2000), any NHPP can be generated from an appropriately defined HPP, and events in both processes have one-to-one correspondences. Consequently, we can show that the optimal policy for an NHPP arrival process has the same time-varying threshold structure as its HPP counterpart. Moreover, instead of directly solving the optimization problem with the NHPP, we can simply formulate its corresponding HPP, and solve for the indifference time points of optimal rationing policy using Equation (7). Their corresponding time points in the NHPP are the optimal indifferent time points there, and can be found via the same scaling transformation. For details of how this is done in a similar inventory problem, please see Jain *et al.* (2012).

Let $\lambda_n(t)$ represent the demand rate of store n at time t. Figure 5 shows the four demand rate patterns we consider for analysis; we call them patterns 1, 2, 3, and 4 respectively. To fairly compare results across different demand patterns – including the homogenous Poisson Process (HPP) results presented earlier –

we keep the expected total demand throughout the season equal to $\lambda_n T$ (*i.e.*, $\int_{t=0}^{T} \lambda_n(t)dt = \lambda_n T$) in all the four patterns. For the other parameters, we choose the same values used as in the HPP case.

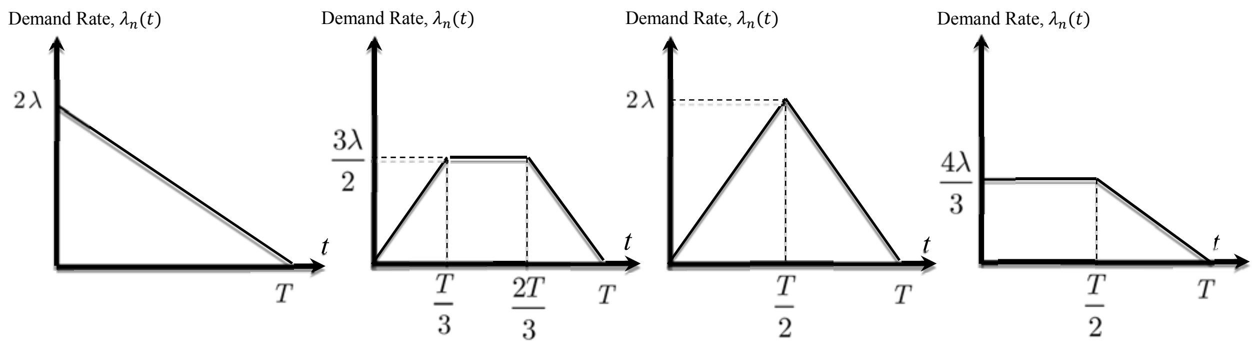

Figure 5 Demand Rate Patterns

In Section 4.3, we discussed that with HPP, rationing has a more pronounced impact in the NP structure than the P structure, making (NP,OPT) an attractive strategy for retailers considering implementing store fulfillment strategy. We also showed that under the NP fulfillment structure, the two heuristic rationing procedures proposed result in near optimal profits. We conducted a numerical test to explore robustness of our findings under the NHPP demand scenarios.

The results verify the robustness of our earlier findings. For brevity, we will not provide the specific results in this paper. It is worth noting that under the NHPP demand scenarios, fulfillment structure of (NP, OPT) outperforms (P, OPT) in all demand patterns, and both heuristics again perform very close to the OPT rationing policy.

4.5. Extension to the Cases with Multiple Offline Stores

Thus far, we have considered the case with one offline and one online store. However, in practice, most retailers will have multiple offline stores. In this section, we extend our model to such cases. The purpose of this section is two-fold:

1) We have shown that the ST and NT heuristics are effective rationing policies with one offline store. We extend the ideas behind these heuristics to the cases with multiple offline stores and one online store and evaluate their effectiveness via a numerical experiment.
2) How would the presence of multiple offline stores affect the fulfillment structure under these rationing policies? In particular, are many offline stores needed and how often will rationing take place at each offline store?

Given the robustness of the NP fulfillment structure, we only focus on extending (NP,ST) and (NP,NT) heuristics and investigate their effectiveness compared to when rationing is not employed (*i.e.,* (NP, $\varnothing$)). Let:

$$Performance\ Dev = \frac{Performance^{NP,ST/NP,NT} - Performance^{NP,\varnothing}}{Performance^{NP,\varnothing}},$$

where: $Performance \in \{profit, fill\ rate, inventory\}$. In addition, we track a measure called *O2O (online-to-offline) fulfillment,* defined as the percentage of online orders handled by each offline store (trivially, *O2O fulfillment* for (NP, $\varnothing$) is zero).

We assume there are N offline stores that are geographically dispersed and operated independently from each other. The stores have *i.i.d.* Poisson demand processes with identical rate λ, profit margin (p) and leftover cost (h), but different handling cost for online orders (k), which reflects the readiness of each offline store for online order fulfillment. Some stores do not have suitable space and facility to implement an efficient pick-and-ship operation; thus, their unit handling cost k should be higher than the others.

In the following numerical tests, we analyze up to 4 offline stores. As before, we normalize $p = 10$, and $T = 1$, and set the rest of the parameters as follows. There are 243 cases in total.

- $\lambda \in \{5, 10, 20\}$ and $\frac{\lambda_0}{\lambda} \in \{0.6,\ 1,\ 1.4\}$,
- $k_1 \in \{0.02, 0.1, 0.2\} * p_0$ and $\frac{k_{i+1}}{k_i} = 1.05$,
- $SL_0 \in \{0.8, 0.9, 0.99\}$, which implies $h_0 \in \left\{\frac{5}{2}, \frac{10}{9}, \frac{10}{99}\right\}$,
- $h \in \{1,\ 1.1, 1.2\} * h_0$, which leads to the following offline store service levels for values of online service level:

SL_0	0.8			0.9			0.99		
SL	0.8	0.78	0.77	0.9	0.89	0.88	0.99	0.99	0.99

4.5.1. Store Fulfillment Policy

With multiple stores, each offline store will develop its own rationing policy to determine whether to satisfy an online order with its inventory. Before that, however, the retailer must first determine which offline store should handle an online order; we call this the routing policy.

A routing policy can be static or dynamic. Under a static routing policy, stores are ranked before the sales period starts. When an online order is to be sent to an offline store, it is sent to the offline stores by their predetermined ranking. For example, suppose each offline store uses an inventory-based threshold rationing policy. Once the online store runs out-of-stock at time θ, the retailer will forward an incoming online order to store 1 if it has inventory above its rationing threshold. If not, then the order will be forwarded to store 2, and so on, until the order is accepted by one offline store or rejected by all.

A static routing policy, by its nature, does not change once the season starts; so it's easy to deploy. This also constitutes its major drawback, however. It could happen that store 1's inventory is only slightly above its threshold, but store 2's inventory is excessively above its threshold. While store 1 has a lower handling cost, store 2's inventory is more likely to incur leftover cost at the end of the season. Therefore, a good policy would also account for each store's remaining inventory as well as time left in the period when making routing decisions. We call such policies dynamic routing policies.

In the next two sections, we will study a static and a dynamic routing policy respectively. Each offline store uses either the ST or the NT rationing policy.

4.5.1.1. Static Routing Policy

Since our offline stores differ only in the handling cost, a static routing policy would rank them by k such that $n=1$ is the store with smallest k and $n=N$ is the store with largest k. We call this a *k-ranked* static routing heuristic.

With the routing policy in place, each offline store must establish its own rationing decision for when online orders are routed its way. Given their effectiveness in earlier numerical studies, we will assume that the offline stores use ST and NT rationing policies, adjusted for the fact that there are multiple offline stores as follows:

- For ST heuristic: when online orders are routed to store n ($n=1,\ldots,N$) for the first time, the offline store will calculate a single ST threshold as in described 3.2.1. If its inventory is below the threshold, it'll stop rationing for the rest of the season, and online orders will be routed to store $n+1$ instead.

If its inventory is above the threshold, however, it will accept online orders until its inventory drops below the threshold – after which point store n will stop rationing and all online orders will be routed to $n+1$, and so on. When $n=N$ and the threshold is reached, then all online orders will be rejected from that point on.

- For NT heuristic: each time an online order needs to be routed to an offline store, each offline store's inventory is checked against its threshold (as in Section 3.2.2). The store with the lowest rank whose inventory is above its threshold will be routed the online demand and satisfy it.

Note that under ST, each offline store has a single threshold, once it's reached, that offline store is closed to all future online demand. So, the routing will be sequential. Under NT, however, because each store's threshold changes over time, a closed store n may be open later, so the online orders can be routed to different offline stores in a back-and-forth order.

For each of the 243 cases, we jointly optimize the retailer's inventory levels at the online and all the offline stores using a numerical search. Their performance comparison with the (NP, ∅) policy is given in Table 8 below.

Table 8 Retailer's Average Profit Deviation % / Optimal Inventory Deviation % from the (NP, ∅) Policy

	N=1	N=2	N=3	N=4
(NP,ST) Policy	2.29% / 1.40%	2.09% / 1.22%	1.86% / 0.99%	1.57% / 0.81%
(NP,NT) Policy	2.30% / 1.41%	2.25% / 1.22%	1.91% / 1.00%	1.60% / 0.81%

Table 8 shows that while rationing by itself increases profit between 0.81% and 1.41%, optimizing stock levels contributes an additional 0.8% to profit increase. It is notable that as the number of offline stores used to back up online sales increases, profit deviation % decreases. Since figures are reported as percentages, this decrease is driven by an increased denominator (profit of all N stores) and not by a decreased numerator (the profit deviation is higher in absolute values as N increases). Nonetheless, this decreased % average profit improvement has important managerial implications. It illustrates that adding more offline stores to the store fulfillment program has a diminishing return. As it is costly to make a traditional offline store capable of fulfilling online demand (*e.g.*, the facility and training for pick, pack, and ship), the diminishing return in Table 8 suggests that it's not always optimal to convert as many offline stores as possible to online backups.

This is expected as there is limited additional online demand that can be captured by using offline stores as backup. Therefore, as number of offline stores increase at some point there will be more than enough backup inventory available.

Also since offline stores are ranked by k for online orders fulfillment, stores on the bottom of the list with high k contribute marginally to online order fulfillment. Therefore, the percentage of online orders fulfilled by offline stores is also decreeing by store rank (Table 9). Furthermore, Table 10 shows that when 4 offline stores are ready to serve as the online store's backup, it reduces the optimal amount of inventory that should be carried in the online store. Thus, more inventory will be carried in the offline stores. This certainly is consistent with the trend in practice regarding store fulfillment. Note that while the first two store's inventory is adjusted up meaningfully, stores 3 and 4 do not see much increase in inventory. Together with observations from Tables 8 and 9, we conclude that, in this example, adding the 4^{th} offline store as an online backup does not provide meaningful inventory backup or profit improvement. It may not be needed at all!

Table 9 Expected O2O Fulfillment for *N*=4

	Store 1	Store 2	Store 3	Store 4
(NP,ST) Policy	13.87%	2.00%	0.53%	0.18%
(NP,NT) Policy	13.62%	2.08%	0.56%	0.19%

Table 10 Optimal Store Inventory Average Deviation (%) from (NP, $\varnothing$) Stocking Level for *N*=4

	Store 0	Store 1	Store 2	Store 3	Store 4
(NP, ST) Policy	-34.24%	9.50%	1.56%	0.62%	0.46%
(NP, NT) Policy	-34.28%	9.18%	1.56%	0.62%	0.46%

Table 10 shows, in the case of *N*=4, how inventory gets redistributed for a retailer using store fulfillment strategy. In Table 11 we can see further that the total inventory across all the stores is reduced by about 4.25% as a result of store fulfillment and rationing. It also shows how this total inventory reduction varies with the number of offline stores. It is not a monotone function, however. As *N* increases, marginal inventory reduction diminishes, but the total inventory increases (as there are more stores). The big-denominator effect leads to smaller inventory reduction percentage.

Table 11 Average Total Inventory Deviation (%) from (NP, ∅)

	N=1	N=2	N=3	N=4
(NP, ST) Policy	-4.87%	-5.35%	-5.06%	-4.26%
(NP, NT) Policy	-5.32%	-5.48%	-5.03%	-4.25%

4.5.1.2. Dynamic Routing Policy

With a dynamic routing policy, many factors must be considered when deciding which offline store's inventory should be used to satisfy an online demand. These should include dynamic information such as how much time is left in the sales period and which offline store has the most amount of "excess" inventory (and thus most likely to have leftover).

Next, we develop a simple dynamic policy that routes an online order to the offline store with the most excess inventory at that time, where excess is defined as the number of inventory units above the rationing threshold level at that time. Specifically, an online order at time *t* will be routed to store *n*, which maximizes the following quantity:

$$\max_n \left(i_n(t) - \tau^{NT/ST}(t) \right)$$

where $i_n(t)$ represents inventory level in store *n* at time *t*, and $\tau^{NT/ST}(t)$ is the optimal protection threshold under NT or ST policy at time *t*. We numerically calculated each store's threshold levels, optimal rationing (within the given ST or NT heuristics) and stocking decisions for each of the 243 cases. The overall performance measures are given in Tables 14 and 15 (for the ST and NT heuristics respectively). We provide static and dynamic routing policies side-by-side for easier comparison. Since the routing policy doesn't come into play when there is only one offline store, we report results for $N \geq 2$.

Table 12 Average Performance Measures for (NP,ST) Policy: Deviation % from (NP, ∅)

	N=2		N=3		N=4	
	Dynamic	*Static*	*Dynamic*	*Static*	*Dynamic*	*Static*
Profit Dev.	2.24%	2.09%	2.16%	1.86%	1.85%	1.57%
Inventory Dev.	-5.53%	-5.35%	-5.48%	-5.06%	-4.93%	-4.26%
O2O Fulfillment	24.74%	14.01%	36.21%	16.01%	43.15%	16.04%

Table 13 Average Performance Measures for (NP,NT) Policy: Deviation % from (NP, $\varnothing$)

	N=2		*N*=3		*N*=4	
	Dynamic	*Static*	*Dynamic*	*Static*	*Dynamic*	*Static*
Profit Dev.	2.84%	2.25%	2.88%	1.91%	2.73%	1.60%
Inventory Dev.	-6.92%	-5.48%	-7.11%	-5.03%	-6.72%	-4.25%
O2O Fulfillment	24.77%	14.04%	36.31%	16.02%	43.63%	16.59%

As shown in Table 12 and Table 13, the proposed dynamic routing policy can significantly improve the retailer's performance. Between the two rationing policies, we observe that NT does much better than ST. This is not surprising, however, as routing decision can be applied to every online demand if the stores use NT rationing. Under ST rationing, once a store is chosen, all online demand will be routed only to that store until its threshold level is reached. Then the routing moves on to the next-ranked store. There is thus less routing flexibility under ST.

In conclusion, we show in this section that the ST and NT rationing policies are easily extendable to an N-Store case setup. Using a wide range of parameters we show that both extended policies (NP,ST) and (NP, NT) perform very well when compared with no rationing. Managerially we find that it is not always necessary to have as many offline stores to back up the online store. Using a static routing policy, retailers can choose a few very efficient offline stores to add to its store fulfillment infrastructure and achieve most of the store fulfillment benefits. Using a dynamic routing policy, the retailer should add more offline stores to the store fulfillment infrastructure, and its performance should also improve considerably.

As an indirect validation of our parameter values, we note that the range of O2O fulfillment is in line with that being reported in practice. For example, Target reported in February 2016 that around 30% of its online orders are filled by offline stores (Chao 2016).

5. Summary and Future Research Directions

In this paper we study the inventory management problem of an omni-channel retailer who already has an established offline store and is looking to leverage it to help with online sales. Our model incorporates relevant decision factors at three different levels: fulfillment structure (strategic), inventory (tactical), and rationing (operational). We derive the optimal rationing policy structure and develop two simple heuristics that we demonstrate, through an extensive numerical test, to be very effective. Integrating the rationing policy into higher-level decisions, we showed that it can have significant impact on the retailer's stocking and fulfillment structure decisions. The integrated (NP, ST) and (NP, NT) policies – where the retailer has an inventory stock dedicated to online sales, but can also use offline inventory as backup when needed, subject to the ST or NT rationing heuristic – is proved to be simple, effective, and robust.

Being a first model in our attempt to analyze the omni-channel strategies, this paper also points to several directions for future research. The current paper focuses on the store fulfillment approach, where inventory backup is uni-directional; it would be interesting and important to extend the study to bi-directional backup and rationing. This requires the retailer to invest in store staffing to actively capture potential lost sale in the store. The retailer could also invest in technology (*i.e.,* QR codes, apps, online portal in the store) so that customers can order items online directly when shopping in the offline store (*e.g.,* Athletha). This trend is slowly taking hold in practice, so a rigorous analytical study would offer guidance.

Another possibility is to incorporate the option of in-store pickup into the analytical model. We can extend our model by having three classes of demand to the offline store. A nested threshold policy may be optimal, but a rigorous study is needed to find out its impact on the stocking decisions as well as the overall cross-channel fulfillment structure for the retailer.

Finally, we consider a retailer that owns both channels, so the decisions are centralized to maximize total profit. Practically, however, each channel may have its own profit target and consideration, even within the same retailer. In such a case, one must consider incentive issues: for example, when an offline store fills an online order who gets the credit and how much credit? Getting a good handle on this is essential to the success of an omni-channel approach.

6. References

Aflaki, A., Swinny, R. 2017. Inventory Integration with Rational Consumers. *Working paper*, Duke University. Retrieved from *SSRN*: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2842182 September 25.

Agatz, Niels A.H., Moritz Fleischmancn, Jo A.E.E. van Nunen. 2008. E-fulfillment and multi-channel distribution – A review. *European Journal of Operational Research* **187**(2) 339–356.

Allen, S. G. 1958. Redistribution of total stock over several user locations. *Naval Research Logistics Quarterly* **5** 51–59.

Alptekinoglu, Aydın, Christopher S. Tang. 2005. A model for analyzing multi-channel distribution systems. *European Journal of Operational Research* **163**(3) 802-824

Amazon. 2014. *Annual Reports and Proxies.* Retrieved from Amazon.com: http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-reportsannual

Archibald , T. W., Sassen , S. A. E. and Thomas , L. C. 1997. An Optimal Policy for a Two Depot Inventory Problem with Stock Transfer, *Management Science* **43(2)** Feb. 1997, 173-183.

Atkins, Derek, Kaan Katircioglu. 1995. Managing inventory for multiple customers requiring different levels of service. Working Paper 94-MSC-015, University of British Columbia, Vancouver, BC, Canada.

Axsäter, Sven. 2003. A New Decision Rule for Lateral Transshipments in Inventory Systems. *Management Science* **49**(9) 1168–1179.

Baird, N., & Kilcourse, B. 2011. Omni-channel fulfillment and the future of retail supply chain. *Retail Systems Research*.

Banjo, S. 2012. Wal-Mart Delivery Service Says to Amazon: Bring It. Retrieved from *The Wall Street Journal*: http://www.wsj.com/articles/SB10000872396390444897304578046461338658772

Banjo, S., Kapner, S., & Ziobro, P. 2014. Can Wal-Mart Clercks Ship as Fast as Amazon Robots? Retrieved from *The Wall Street Journal*: http://www.wsj.com/article_email/canwalmartclerksshipasfastasamazonrobots1418930087lMyQjAxMTE1ODEwNjcxNDY1Wj

Bassok, Yehuda, Ravi Anupindi, Ram Akella. 1999. Single-Period Multiproduct Inventory Models with Substitution. *Operations Research* **47(4)** 632-642.

Bendoly, Elliot, Doug Blocher, Kurt M. Bretthauer, M.A. Venkataramanan. 2007. Service and Cost Benefits Through Clicks-and-mortar Integration: Implications for The Centralization/Decentralization Debate. *European Journal of Operational Research* **180**(1) 426–442.

Benjaafar, Saif, William L Cooper, Joon-Seok Kim. 2005. On the Benefits of Pooling in Production-Inventory Systems. *Management Science* **51(4)** 548-565.

Ben-Zvi, N., Y. Gerchak. 2012. Inventory Centralization in a Newsvendor Setting When Shortage Costs Differ: Priorities and Costs Allocation. T-M. Choi, ed. *Handbook of Newsvendor Problems*. Springer, New York, Ch.11, 263–276.

Berman, Oded, Krass, Dmitry and Tajbakhsh, M. Mahdi. 2011. On the Benefits of Risk Pooling in Inventory Management. *Production and Operations Management* **20(1)** 57–71.

Bimpikis, Kostas, Markakis, Mihalis G. 2014. Inventory Pooling under Heavy-Tailed Demand. Forthcoming in *Management Science*.

Bretthauer, Kurt M., Stephen Mahar, M.A. Venakataramanan. 2010. Inventory and distribution strategies for retail/e-tail organizations. *Computers & Industrial Engineering* **58**(1) 119–132.

Cao, Jing. 2017. Wal-Mart Close to Acquiring Bonobos for About $300 Million. *Bloomberg*. April 17. https://www.bloomberg.com/news/articles/2017-04-17/wal-mart-said-close-to-acquiring-bonobos-for-about-300-million

Cattani, Kyle D., Gilvan Souza. 2002. Inventory Rationing and Shipment Flexibility Alternatives for Direct Market Firms. *Production and Operations Management* **11**(4) 441–457.

Chan, Jose P. 2013. The Promise of Digital Technology in Brick and Mortar Retail. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Chao, Loretta. 2016. Target Says Online Sales Surge Tied to Store Inventories. *Wall Street Journal*. Feb 24.

Chen, Jian, Youhua Chen, Mahmut Parlar, Yongbo Xiao. 2011. Optimal Inventory and Admission Policies for Drop-shipping Retailers Serving In-store and Online Customers. *IIE Transactions* **43**(5) 332-347.

Chen, M.S., C.T. Lin, 1990. An Example of Disbenefits of Centralized Stocking. *Journal of the Operational Research Society* **41(3)**: 259-262.

Chen, M.-S., C.-T. Lin. 1989. Effects of Centralization on Expected Costs in a Multi-location Newsboy Problem. *Journal of the Operational Research Society* **40(6)** 597–602.

Cherikh, M. 2000. On the Effect of Centralization on Expected Profits in a Multi-location Newsboy Problem. *Journal of the Operational Research Society* **51(6)**755–761.

Corbett, Charles J., Kumar Rajaram. 2006. A Generalization of the Inventory Pooling Effect to Nonnormal Dependent Demand. *Manufacturing & Service Operations Management* **8(4)** 351-358.

De Koster, M.B.M., 2002. The Logistics Behind the Enter Click. In Klose, A., Van Wassenhove, L.N. (Eds.), *Quantitative Approaches to Distribution Logistics and Supply Chain Management*. Springer, Berlin, pp. 131–148.

Deflem, Yannick, Van Nieuwenhuyse, Inneke. 2013. Managing Inventories with One-Way Substitution: A Newsvendor Analysis. *European Journal of Operational Research* **228(3)** 484–493.

Deshpande, Vinayak, Morris A. Cohen, Karen Donohue. 2003. A Threshold Inventory Rationing Policy for Service-Differentiated Demand Classes. *Management Science* **49(6)** 683–703.

Eppen, Gary D. 1979. Note: Effects of Centralization on Expected Costs in a Multi-location Newsboy Problem. *Management Science* **25(5)** 498-501.

Erkip, Nesim, Hausman, Warren H. and Nahmias, Steven. 1990. Optimal Centralized Ordering Policies in Multi-Echelon Inventory Systems with Correlated Demands. *Management Science* **36(3)** 381-392.

Evans, Richard V. 1968. Sales and Restocking Policies in a Single Item Inventory System. *Management Science* **14**(7) 463-472.

Eynan, Amit. 1999. The Multi-location Inventory Centralization Problem with First-Come, First-Served Allocation, *European Journal of Operational Research* **114** 38-49.

Fisher, Marshall, Jan Hammond, Walter R. Obermeyer, Ananth Raman. 1994. Making Supply Meet Demand in An Uncertain World. *Harvard Business Review* May-June.

Frank, Katia C., Rachel Q. Zhang, Izak Duenyas. 2003. Optimal Policies for Inventory Systems with Priority Demand Classes. *Operations Research* **51**(6) 1993–2002.

Gallego, Guillermo, Garrett van Ryzin. 1994. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons. *Management Science* **40**(8) 999-1020.

Gallino, S., A. Moreno. 2014. Integration of Online and Offline Channels in Retail: The Impact of Sharing Reliable Inventory Availability Information. *Management Science* **60** 1434-1451.

Gerchak, Yigal, David Mossman. 1992. On the Effect of Demand Randomness on Inventories and Costs. *Operations Research* **40(4)** 804-807.

Gerchak, Yigal, Qi-Ming He. 2003. On the Relation Between the Benefits of Risk Pooling and the Variability of Demand. *IIE Transactions* **35(11)** 1027-1031.

Gross, D. 1963. Centralized Inventory Control in Multilocation Supply Systems. H. E. Scarf, D. M. Gilford, M. W. Shelly, eds. *Multistage Inventory Models and Techniques*. Stanford University Press, Stanford, CA, 47–84.

Hadley, G., T.M. Whitin. 1963. *Analysis of Inventory Systems*. Prentice Hall.

Harsha, P., Subramanian, S., and Uichanco, J. 2017. Dynamic pricing of omnichannel inventories. *Working Paper*, IBM. Retrieved from https://www.researchgate.net/publication/301805713_Dynamic_pricing_of_omnichannel_inventories September 25.

Herer, Y. T., A. Rashit. 1999. Lateral Stock Transshipments in a Two Location Inventory System with Fixed and Joint Replenishment Costs. *Naval Research Logistics* **46** 525–547.

Jain, Apurva, Kamran Moinzadeh, and Yong-Pin Zhou. 2012. A Single-Supplier, Multiple-Retailer Model with Single-Season, Multiple-Ordering Opportunities, and Fixed Ordering Cost. *Operations Research* 60(5): 1098–1110.

Johnson, L.A., Montgomery D.C. 1974. Operations research in production planning, scheduling and inventory control. New York: John Wiley and Sons, Inc.

Jönsson, H., E. A. Silver. 1987. Analysis of a Two-Echelon Inventory Control System with Complete Redistribution. *Management Sci*ence **33(2)** 215–227.

Kaplan, Alan. 1969. Stock Rationing. *Management Science* **15**(5) 260–267.

Kapner, Suzanne. 2017. Store Closings Accelerate. *Wall Street Journal*, April 22. https://www.wsj.com/articles/brick-and-mortar-stores-are-shuttering-at-a-record-pace-1492818818

Karmarkar, U. S., N. Patel. 1977. The One-Period, N-location Distribution Problem. *Naval Research Logistics Quarterly* **24** 559–575.

Kleijn, M.J., Dekker, R., 1998. An Overview of Inventory Systems with Several Demand Classes. In Speranza, M.G., Stahly, P. (Eds.), *New Trends in Distribution Logistics*. Springer, Berlin, pp. 253–265.

Krishnan, K. S. and Rao, V. R. K. 1965. Inventory Control in N Warehouses. *Journal of Industrial Engineering* **16** 212-215.

Law A.M., and W.D. Kelton. 2000. Simulation Modeling and Analysis. *McGraw-Hill*, New York.

Lee, H. 1987. A Multi-Echelon Inventory Model for Repairable Items with Emergency Lateral Transshipments. *Management Science* **33**(10) 1302-1316.

Lee, H.L., Nahmias S. 1993. Single-product, single-location models. In: Graves SC, Kan AHGR, Zipkin PH, editors. Handbook in operations research and management science, Volume on logistics of production and inventory. Amsterdam: North Holland; p. 3–58.

Lei, M., Jasin, S., and Sinha, M. 2016. Dynamic Joint Pricing and Order Fulfillment for E-Commerce Retailers. Working Paper, University of Michigan. Retrieved from https://www.researchgate.net/publication/314377333_Dynamic_Joint_Pricing_and_Order_Fulfillment_for_E-Commerce_Retailers. September 25.

Lentini, D., & Zhou, Y.-P. 2015. Toys“R”Us Asia: Ecommerce Strategy. Teaching Case Study, Univeristy of Hong Kong, Asian Case Research Center.

Liang, Y. 1999. Solution to the Continuous Time Dynamic Yield Management Model. *Transportation Science* **33(1)** 117-123.

Mahar S, Wright D, Bretthauer K. 2010. An Analysis of In-Store Pickups and Returns in Retail/E-TailOrganizations. Working Paper, UNC Wilmington.

Mahar, Stephen, Kurt M. Bretthauer, M. A. Venkataramanan. 2009. The Value of Virtual Pooling in Dual Sales Channel Supply Chains. *European Journal of Operational Research* **192** 561-575.

Mahar, Stephen, Peter A. Salzarulo, P. Daniel Wright. 2012. Using online pickup site inclusion policies to manage demand in retail/E-tail organizations. *Computers & Operations Research* **39**(5) 991-999.

Mak, Ho-Yin, Zuo-Jun Max Shen. 2014. Pooling and Dependence of Demand and Yield in Multiple-location Inventory Systems. *Manufacturing & Service Operations Management* **16(2)** 263-269.

Manhattan Associates. 2011. Maximizing Margins in an Omni-Channel World. *RIS News* Whitepaper.

Mattioli, D. 2012. Macy's Regroups in Warehouse Wars. Retrieved from *The Wall Street Journal*: http://www.wsj.com/articles/SB10001424052702303505504577404123150295102

Mehrez, A., Stulman, A., 1984. Priority Centralization of the One Period Multi-location Inventory System. *Zeitschrift für Operations Research* **28** B297-B304.

Melchiors, P., Rommert Dekker, M.J. Kleijn. 2000. Inventory Rationing in an (s, Q) Inventory Model with Lost Sales and Two Demand Classes. *Journal of the Operational Research Society* **51** (1) 111–122.

Nahmias, Steven, W. Steven Demmy. 1981. Operating Characteristics of an Inventory System with Rationing. *Management Science* **27**(11) 1236–1245.

Nassauer, Sarah. 2015. Search for Elusive 'Yeti' Daunting for Holiday Shoppers. *Wall Street Journal* Dec 22.

Netessine, Serguei, Nils Rudi. 2006. Supply Chain Choice on the Internet. *Management Science* **52**(6) 844–864.

Olsen, T.L., Nahmias, S. 2015. Production and Operations Analysis, Waveband Press (pages 306-307).

Pasternack, B., Drezner, Z. 1991. Optimal Inventory Policies for Substitutable Commodities with Stochastic Demand. *Naval Research Logistics* 38, 221–240.

Paterson, Colin, Gudrun Kiesmüller, Ruud Teunter, Kevin Glazebrook. 2011. Inventory Models with Lateral transshipments: A review. *European Journal of Operational Research* **210**(3) 125-136.

PCMag. 2011. This Is How China Does Best Buy. Retrieved from *PCMag*: http://www.pcmag.com/slideshow/story/266121/this-is-how-china-does-best-buy June 24

Pentico, D. 2008. The Assortment Problem: A Survey. *European Journal of Operational Research*, **190** 295–309.

PwC. 2014. Global Retail and Consumer Goods CEO Survey: The Omni-Channel Fulfillment Imperative. Retrieved from *JDA*: http://now.jda.com/ceo2015.html, December.

Ricker , F.R., Kalakota , R. 1999. Order Fulfillment: The Hidden Key to E-Commerce Success. *Supply Chain Management Review* 60–70.

Ridder, Ad, Erwin Van Der Laan, Marc Salomon. 1998. How Larger Demand Variability May Lead to Lower Costs in the Newsvendor Problem. *Operations Research* **46(6)** 934-936.

Robinson, L. W. 1990. Optimal and Approximate Policies in Multi-period Multi-location Inventory Models with Transshipments. *Operations Research* **38** 278-295.

Rudi, Nils, Sandeep Kapur, David F. Pyke. 2001. A Two-Location Inventory Model with Transshipment and Local Decision Making. *Management Science* **47**(12) 1668–1680.

Scottsdale, A. 2015. Retailers Investing Heavily in Omni-Channel Selling – But Where Are the Profits? Retrieved from *JDA*: http://www.jda.com/view/pressrelease/ April 14.

Shin, Hojung, Park, Soohoon, Lee, Euncheol, and Benton, W. C. 2015. A Classification of The Literature on the Planning of Substitutable Products. *European Journal of Operational Research* **246(3)** 686–699.

Song, Jing-Sheng. 1994. The Effect of Leadtime Uncertainty in A Simple Stochastic Inventory Model. *Management Science* **40(5)** 603-613.

Tagaras, G. and M. A. Cohen. 1992. Pooling in Two-location Inventory Systems with Non-negligible Replenishment Lead Times. *Management Science* **38** 1067-1083.

Tagaras, George. 1989. Effects of Pooling on the Optimization and Service Levels of wo-Location Inventory Systems. *IIE Transactions*, **21(3)**.

Talluri, Kalyan T., Garrett J. van Ryzin. 2004. The Theory and Practice of Revenue Management. Springer.

The Boston Company. 2014. From Bricks to Clicks, A Retail Evolution. The Boston Company, Asset Management, LLC.

Topkis, Donald M. 1968. Optimal Ordering and Rationing Policies in a Nonstationary Dynamic Inventory Model with *n* Demand Classes. *Management Science* **15**(3) 160–176.

U.S. Census Bureu. 2017. Quarterly Retail E-Commerce Sales 4th Quarter 2016. https://www2.census.gov/retail/releases/historical/ecomm/16q4.pdf

Veinott, Arthur F. 1965. Optimal Policy in a Dynamic, Single Product, Nonstationary Inventory Model with Several Demand Classes. *Operations Research* **13**(5) 761–778.

Wee, Kwan Eng, and Maqbool Dada (2005) Optimal Policies for Transshipping Inventory in a Retail Network. Management Science 51(10):1519-1533

Weedfald, P. 2014. Omni-Channels Demand Omni-Margins. Retrieved from http://www.dealerscope.com/post/omni-channels-demands-omni-margins/ Feb 9.

Yang, Hongsuk and Linus Schrage. 2009. Conditions That Cause Risk Pooling to Increase Inventory. *European Journal of Operational Research* **192** 837–851.

Electronic Companions

Proof of Proposition 1

Because there is no rationing, online and offline demands are satisfied on a first-come-first-served basis. Therefore, retailer in the P structure can view demand as coming from only one source with a weighted margin of $\omega = \frac{\lambda_1}{\lambda_0 + \lambda_1} p_1 + \frac{\lambda_0}{\lambda_0 + \lambda_1}\left(p_0 - k\right)$, similar to Ben-Zvi and Gerchak (2012). Moreover, the retailer behaves as a newsvendor with margin of ω, leftover cost of h_1 and Poisson demand with rate of $\lambda_{0,1}$. Therefore optimal stocking level of retailer can be calculated using critical ratio. Please refer to Hadley and Within (1963) on page 298 for proof of optimal stocking level for newsvendor model.

Proof of Proposition 2

For this proof, we need some technical results on the Poisson distribution function, and will prove them in Lemma A1 and Lemma A2 below.

Lemma A1: $p(n,\lambda) < \frac{1}{2}$ for all n and λ.

Proof: For any given $n > 0$, $\frac{\partial p(n,\lambda)}{\partial \lambda} = p(n-1,\lambda) - p(n,\lambda) = p(n-1,\lambda)\left(1 - \frac{\lambda}{n}\right)$. So $p(n,\lambda)$ is non-decreasing in λ for $\lambda \le n$ and non-increasing in λ for $\lambda \ge n$. It achieves its maximum at $\lambda = n$. Further, because $p(n-1,n) = p(n,n)$ and $p(n-1,n) + p(n,n) < 1$, we must have $p(n,n) < 1/2$. Therefore, $p(n,\lambda) \le p(n,n) < 1/2$ for all n and λ. □

Lemma A2: $P(n,\lambda)[1 - P(n,\lambda)] \le \lambda p(n-1,\lambda)$ for all n and λ.

Proof: Define $L(\mathrm{n},\lambda) \triangleq -2P(n,\lambda) - (n-1) + \lambda$. Then $\frac{\partial L(n,\lambda)}{\partial \lambda} = -2p(n-1,\lambda) + 1 > 0$ due to Lemma A1. Moreover, $L(n,0) \le 0$ and $\lim_{\lambda \to \infty} L(n,\lambda) = \infty$. This means there exists a finite A such that $L(n,\lambda) < 0$ for $\lambda < A$, $L(n,A) = 0$, and $L(n,\lambda) > 0$ for $\lambda > A$.

Next, define $G(n,\lambda) \triangleq P(n,\lambda)[1-P(n,\lambda)] - np(n,\lambda)$. Since $\dfrac{\partial G(n,\lambda)}{\partial \lambda} = L(n,\lambda)p(n-1,\lambda)$, $G(n,\lambda)$ is decreasing for $\lambda < A$, and increasing for $\lambda > A$. Noting that $G(n,0) = \lim_{\lambda\to\infty} G(n,\lambda) = 0$, we conclude that $G(n,\lambda) \le 0$ for n and λ. Since $\lambda p(n-1,\lambda) = np(n,\lambda)$, we complete the proof. □

Now, we return to the proof of Proposition 2. As a reminder, k is assumed to be bounded such that $\omega \ge 0$.

(1) Define $J(\lambda_0) \triangleq \dfrac{\omega P\left(S_1^{NP},\lambda_{0,1}\right)}{1-P\left(S_1^{NP},\lambda_{0,1}\right)}$ where $\lambda_{0,1} = \lambda_0 + \lambda_1$. We know

$$\lim_{\lambda_0\to 0} J(\lambda_0) = \frac{p_1 P\left(S_1^{NP},\lambda_1 T\right)}{\sum_{y=0}^{S_1^P-1} p\left(y,\lambda_{0,1}T\right)} \ge \frac{p_1\dfrac{h_1}{p_1+h_1}}{\sum_{y=0}^{S_1^P-1} p\left(y,\lambda_{0,1}T\right)} \ge \frac{p_1\dfrac{h_1}{p_1+h_1}}{\dfrac{p_1}{p_1+h_1}} = h_1 \quad \text{and} \quad \lim_{\lambda_0\to\infty} J(\lambda_0) = \infty.$$

Moreover,

$$\begin{aligned}\frac{\partial J(\lambda_0)}{\partial \lambda_0} &= \frac{-\left(\omega - p_0 + k\right)P\left(S_1^{NP},\lambda_{0,1}\right)\left[1-P\left(S_1^{NP},\lambda_{0,1}\right)\right] + \omega\lambda_{0,1}p\left(S_1^{NP}-1,\lambda_{0,1}\right)}{\lambda_{0,1}\left[1-P\left(S_1^{NP},\lambda_{0,1}\right)\right]^2} \\ &= \frac{\omega\left\{\lambda_{0,1}p\left(S_1^{NP}-1,\lambda_{0,1}\right) - P\left(S_1^{NP},\lambda_{0,1}\right)\left[1-P\left(S_1^{NP},\lambda_{0,1}\right)\right]\right\}}{\lambda_{0,1}\left[1-P\left(S_1^{NP},\lambda_{0,1}\right)\right]^2} \ge 0 \quad \text{due to Lemma A2 and } p_0 > k.\end{aligned}$$

Therefore, $J(\lambda_0) > h_1$ for all S_1^{NP} and λ_0, which means $P\left(S_1^{NP},\lambda_{0,1}\right) < \dfrac{h_1}{h_1+\omega}$. Furthermore, $\omega \le p_1$ implies $\dfrac{h_1}{h_1+\omega} \ge \dfrac{h_1}{h_1+p_1}$. These two facts together lead to $S_1^P \ge S_1^{NP}$.

(2) Define $B\left(k\right) = S_1^P - \left(S_0^{NP} + S_1^{NP}\right)$. As $S_0^{NP} + S_1^{NP}$ is independent of k, $\dfrac{\partial B\left(k\right)}{\partial k} < 0$ holds because $\dfrac{\partial S_1^P}{\partial k} < 0$. In the limit of $k \to \infty$, $\lim_{k\to\infty} S_1^P = \lim_{k\to\infty} P^{-1}\left(\dfrac{h_1}{h_1+\omega},\lambda_{0,1}T\right) \overset{\omega=0}{=} P^{-1}\left(1,\lambda_{0,1}T\right) = 0$.

Thus, $\lim_{k\to\infty} B(k) < 0$. There are two possible cases, depending on the value of S_1^P at $k=0$:

a) $S_1^P > S_0^{NP} + S_1^{NP}$. In this case, we define $\dot{k} > 0$ to be the unique solution to $S_1^P = S_0^{NP} + S_1^{NP}$. Due to strict monotonicity in k, we must have $S_1^P > S_0^{NP} + S_1^{NP}$ for all $k < \dot{k}$ and $S_1^P \le S_0^{NP} + S_1^{NP}$ for all $k > \dot{k}$.

b) $S_1^P \leq S_0^{NP} + S_1^{NP}$. In this case we simply define $\dot{k} = 0$ and automatically get, due to strict monotonicity in k, that $S_1^P \leq S_0^{NP} + S_1^{NP}$ for all $k \geq 0$.

Therefore, we have shown that $S_1^P \leq S_0^{NP} + S_1^{NP}$ for all $k \geq \dot{k}$.

Proof of Corollary 1

We will first show that $S_1^P \leq S_0^{NP} + S_1^{NP}$ when

$$\frac{h_1}{h_1 + \omega(k)} \leq \alpha_1 \frac{h_1}{h_1 + p_1} + \alpha_0 \frac{h_0}{h_0 + p_0}. \tag{A1}$$

To see that, we apply the mean value theorem and find that there exists $(u,v) = c\left(S_0^{NP}, \lambda_0\right) + (1-c)\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right)$ for some $0 \leq c \leq 1$ such that

$$P\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right) = P\left(S_0^{NP}, \lambda_0\right) + \left.\frac{\partial P}{\partial S}\right|_{(u,v)} \left(S_1^{NP}\right) + \left.\frac{\partial P}{\partial \lambda}\right|_{(u,v)} \left(\lambda_1\right). \tag{A2}$$

Using a continuous $P(S,\lambda)$, we know $\frac{\partial P(S,\lambda)}{\partial S}|_{(u,v)} = -p(u,v)$ and $\frac{\partial P(S,\lambda)}{\partial \lambda}|_{(u,v)} = p(u,v)$. So (A2) becomes

$$P\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right) = P\left(S_0^{NP}, \lambda_0\right) - p(u,v)\left(S_1^{NP} - \lambda_1\right). \tag{A3}$$

Similarly, there exists $(x,y) = d\left(S_1^{NP}, \lambda_1\right) + (1-d)\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right)$ for some $0 \leq d \leq 1$ such that

$$P\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right) = P\left(S_1^{NP}, \lambda_1\right) - p(x,y)\left(S_0^{NP} - \lambda_0\right). \tag{A4}$$

A weighted sum of (A3) and (A4) gives us

$$P\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right) = \alpha_0 P\left(S_0^{NP}, \lambda_0\right) + \alpha_1 P\left(S_1^{NP}, \lambda_1\right) - \alpha_0 p\left(u,v\right)\left(S_1^{NP} - \lambda_1\right) - \alpha_1 p\left(x,y\right)\left(S_0^{NP} - \lambda_0\right).$$

Because $S_n^{NP} \geq \lambda_n$ for $n = 0,1$,

$$P\left(S_0^{NP} + S_1^{NP}, \lambda_{0,1}\right) \leq \alpha_0 P\left(S_0^{NP}, \lambda_0\right) + \alpha_1 P\left(S_1^{NP}, \lambda_1\right)_0 = \alpha \frac{h_0}{h_0 + p_0} + \alpha_1 \frac{h_1}{h_1 + p_1}. \tag{A5}$$

Combine (A1) and (A5) we get $P\left(S_0^{NP}+S_1^{NP},\lambda_{0,1}\right)\le\frac{h_1}{h_1+\omega}$, and $S_1^P\le S_0^{NP}+S_1^{NP}$ follows. □

Now return to the proof of Corollary 1. We see that, after some algebraic manipulation, (A1) is equivalent to

$$k\ge\frac{[p_0-p_1]h_1\alpha_1(h_0+p_0)+p_0[h_0-h_1]\alpha_0(p_1+h_1)+(p_1+h_1)\alpha_1[p_1h_0-h_1p_0]}{\alpha_1h_1(h_0+p_0)+\alpha_0h_0(p_1+h_1)}. \tag{A6}$$

When the right hand side of (A6) is non-positive, then for any $k\ge 0$ we have $S_1^P\le S_0^{NP}+S_1^{NP}$; that is, pooling reduces inventory. This condition is satisfied, in particular, when $p_0\le p_1$, $h_0\le h_1$, and $\frac{h_0}{p_0}\le\frac{h_1}{p_1}$. Since $h_0\le h_1$ is typically true in practice a special case is when the retailer matches online and offline prices (i.e., $p_0=p_1$). □

Proof of Proposition 3

Part (1) Please note that for any fixed k, $\Pi^P(S)$ can be also written as $\Pi^P(S)=\omega(k)\left[D(\lambda_{0,1}T)\wedge S\right]-h_1\left(S-D(\lambda_{0,1}T)\right)^+$. Since k is a variable now, $\Pi(k)=\max_S\Pi^P(S,k)$.

$$\frac{\partial\Pi(k)}{\partial k}=\frac{\partial\Pi^P(S,k)}{\partial S}\frac{\partial S_k^P}{\partial k}+\frac{\partial\Pi^P(S,k)}{\partial k}. \tag{A7}$$

Please note $\Pi^P(S,k)$ is strictly concave in S. Considering that $\left.\frac{\partial\Pi^P(S,k)}{\partial S}\right|_{S=S_k^P}=0$, we can simplify (A7) as follow:

$$\frac{\partial\Pi(k)}{\partial k}=\frac{\partial\Pi^P\left(S_k^P,k\right)}{\partial k}=\frac{\partial\omega}{\partial k}\left[D(\lambda_{0,1}T)\wedge S_k^P\right]<0. \tag{A8}$$

Because under the NP structure there is no handling cost *k*, we have $\frac{\partial\Pi^{NP}\left(S_0^{NP},S_1^{NP}\right)}{\partial k}=0$. Therefore, the profit difference $\Pi^P\left(S_1^P\right)-\Pi^{NP}\left(S_0^{NP},S_1^{NP}\right)$ is decreasing in *k*.

Part (2) Moreover, it is clear that as $k \to \infty$, we have $\omega \to -\infty$, hence $\Pi^P\left(S_1^P\right) \to 0$. So $\lim_{k\to\infty} \Pi^P\left(S_1^P\right) - \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right) = -\Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right) < 0$. At $k = 0$, there are two possible cases:

a) $\Pi^P\left(S_1^P\right) > \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$. In this case, we define $\bar{k}$ to be the unique solution to $\Pi^P\left(S_1^P\right) = \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$. Due to strict monotonicity in k, we must have $\Pi^P\left(S_1^P\right) > \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$ for all $k < \bar{k}$ and $\Pi^P\left(S_1^P\right) < \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$ for all $k > \bar{k}$.

b) $\Pi^P\left(S_1^P\right) \le \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$. In this case we simply define $\bar{k} = 0$ and automatically get, due to strict monotonicity in k , that $\Pi^P\left(S_1^P\right) < \Pi^{NP}\left(S_0^{NP}, S_1^{NP}\right)$ for all $k > 0$. □

Proof of Proposition 4

Both k and λ_0 are variables in this proposition, so we include them as additional function arguments whenever necessary. For example: we denote $\Pi\left(\lambda_0, k\right) = \max_S \Pi^P\left(S, \lambda_0, k\right)$ and let its optimizer be denoted as S_{k,λ_0}^P ; thus, $\Pi\left(\lambda_0, k\right) = \Pi^P\left(S_{k,\lambda_0}^P, \lambda_0, k\right)$.

$$\frac{\partial \Pi\left(\lambda_0, k\right)}{\partial k} = \frac{\partial \Pi^P\left(S_{k,\lambda_0}^P, \lambda_0, k\right)}{\partial k} = \frac{\partial \omega}{\partial k}\left(S_{k,\lambda_0}^P - \int_{x=0}^{S_{k,\lambda_0}^P} \left(S_{k,\lambda_0}^P - x\right) p\left(x, \lambda_{0,1} T\right)\right) \tag{A9}$$

Then:
$$\frac{\partial^2 \Pi\left(\lambda_0, k\right)}{\partial \lambda_0 \partial k} = \left.\frac{\partial^2 \Pi^P\left(S, k, \lambda_0\right)}{\partial k \partial S}\right|_{S=S_{k,\lambda_0}^P} \frac{\partial S_{k,\lambda_0}^P}{\partial k} + \left.\frac{\partial^2 \Pi^P\left(S, k, \lambda_0\right)}{\partial \lambda_0 \partial k}\right|_{S=S_{k,\lambda_0}^P} \tag{A10}$$

which using (A9) can be written as follow:

$$\frac{\partial^2 \Pi\left(\lambda_0, k\right)}{\partial \lambda_0 \partial k} = \frac{\partial^2 \omega}{\partial \lambda_0 \partial k}\left(S_{k,\lambda_0}^P - \int_{x=0}^{S_{k,\lambda_0}^P} \left(S_{k,\lambda_0}^P - x\right) p\left(x, \lambda_{0,1} T\right)\right) + \frac{\partial \omega}{\partial k}\left(\frac{\partial S_{k,\lambda_0}^P}{\partial \lambda_0} \int_{x=S_{k,\lambda_0}^P}^{\infty} p\left(x, \lambda_{0,1} T\right)\right) \tag{A11}$$

Since $\frac{\partial S_{k,\lambda_0}^P}{\partial \lambda_0} \ge 0$, $\frac{\partial^2 \omega}{\partial \lambda_0 \partial k} \le 0$, and $\frac{\partial \omega}{\partial k} \le 0$, the proof is complete. □

Proof of Theorem 1

Based on Lemma 1, protection thresholds are non-increasing unit step function throughout the season. At the indifference points we have

$$\Delta_i V\left(i, t_{i-1}\right) = p_0 - k. \tag{A12}$$

It is intuitive as it says a unit of inventory should be open to share with online when $\Delta_i V\left(i, t\right) < p_0 - k$ and protected when $\Delta_i V\left(i, t\right) \geq p_0 - k$. Since Lemma 1 shows that $\Delta_i V\left(i, t\right)$ is decreasing in t and i, therefore $t_l < t_{l-1} ... < t_1 < t_0$. So it remains to show that there exist a t_0 close enough to end of season such that retailer would prefer to protect no unit after that. Because $\Delta_i V\left(i, t\right)$ is decreasing in i, if it is optimal to not protect last unit of inventory close enough to the end of season, it will also be optimal to not protect if we have more inventory available ($i > 1$). Therefore, it suffices to show t_0 exist when only one unit of inventory left, i.e. $\Delta_i V\left(i, t\right)$ is at its maximum for a given t.

Now consider an online demand that arrives at time $t > t_0$ when the offline store has one unit of inventory. If the retailer accepts the online demand, then he receives a sure $p_0 - k$ from that unit of inventory. If the retailer rejects the online demand, however, he may expect to sell it later at a higher profit of p_1. This happens with a probability of at most $P\left(1, \lambda_1\left(T - t\right)\right)$ ("at most" because policy may reject other online demands that arrive after t). On the other hand, the unit may be left over at the end of the season and incur a cost of $-h_1$. This happens with a probability of *at least* $p\left(0, \lambda_1\left(T - t\right)\right)$.

Therefore, if we define $\bar{t} = T + \frac{1}{\lambda_1} \ln\left(\frac{p_1 - \left(p_0 - k\right)}{p_1 + h_1}\right)$, then for any $t > \bar{t}$, we have $p_0 - k \geq p_1 P\left(1, \lambda_1\left(T - t\right)\right) - h_1 p\left(0, \lambda_1\left(T - t\right)\right)$.

Thus, the retailer would make less expected profit if he rejects the online demand at $t > \bar{t}$. Defining $t_0 = \min\{\bar{t} \mid \text{it is optimal to accept an online demand for all } t \in [\bar{t}, T]\}$, then we know it's optimal not to protect any offline inventory after t_0. □

Proof of Lemma 2

The first order difference of $H^{ST}\left(\tau \middle| i,\theta\right)$ with respect to τ is as follow:

$$\begin{aligned}\Delta_{\tau} H^{ST}\left(\tau \middle| i,\theta\right) &= -h_1 + \left(p_1 + h_1\right) P\left(i-\tau, \lambda_{0,1}\left(T-\theta\right)\right) \\ &\quad -\left(p_1 + h_1\right)\left(\alpha_1\right)^{-(i-\tau)} \sum_{n=0}^{\tau-1} p\left(n+i-\tau, \lambda_1\left(T-\theta\right)\right) \ {}_1F_1\left(i-\tau, n+i-\tau+1, -\lambda_0\left(T-\theta\right)\right),\end{aligned}$$

where by convention ${}_1F_1\left(a;b;z\right)$ indicates a confluent hypergeometric function. Then $\Delta_{\tau}^2 H^{ST}\left(\tau \middle| i,\theta\right) = -\left(p_1 + h_1\right)\left(\alpha_1\right)^{-(i-\tau)} p\left(i-1, \lambda_1\left(T-\theta\right)\right) \ {}_1F_1\left(i-\tau, i-1+1, -\lambda_0\left(T-\theta\right)\right) \leq 0.$ Moreover,

$$\Delta_i \Delta_{\tau} H^{ST}\left(\tau \middle| i,\theta\right) = -\left(p_1 + h_1\right)\left(\alpha_1\right)^{-(i-\tau-1)} p\left(i-1, \lambda_1\left(T-\theta\right)\right) \ {}_1F_1\left(i-\tau, i, -\lambda_0\left(T-\theta\right)\right) \leq 0.$$

$$\partial \frac{\Delta_{\tau} H^{ST}\left(\tau \middle| i,\theta\right)}{\partial\theta} = -\lambda_{0,1}\left(p_1 + h_1\right)\left(\alpha_1\right)^{-(i-r-1)} p\left(i-1, \lambda_1\left(T-\theta\right)\right) \ {}_1F_1\left(i-r, i, -\lambda_0\left(T-\theta\right)\right) \leq 0. \qquad \square$$

Proof of Proposition 5

The second order difference of $H^{ST}\left(\tau \middle| i,\theta\right)$ with respect to τ is as follow:

$$\begin{aligned}\Delta_r^2 H^{ST}\left(\tau \middle| i,\theta\right) &= -\alpha_0\left(p_0 - k_1 + h_1\right) p\left(i-\tau+1, \lambda_{0,1}\left(T-\theta\right)\right) \\ &\quad -\left(\alpha_0\right)^2\left(\alpha_1\right)^{\tau-1}\left(p_1 + h_1\right) p\left(i+1, \lambda_{0,1}\left(T-\theta\right)\right) \ {}_1F_1\left(\tau, I+2, \lambda_0\left(T-\theta\right)\right)\end{aligned}$$

Both terms are negative, as a result $\Delta_r^2 H^{ST}\left(\tau \middle| i,\theta\right) \leq 0$, meaning $\Delta_{\tau} H^{ST}\left(\tau \middle| i,\theta\right)$ is decreasing in τ.

Therefore, there are two cases:

Case 1) If $\Delta_{\tau} H^{ST}\left(\tau = 1 \middle| i,\theta\right) < 0$ then $\tau^{ST}\left(i,\theta\right) = 0$.

Case 2) If $\Delta_{\tau} H^{ST}\left(\tau = 1 \middle| i,\theta\right) \geq 0$ then $\exists \tau' : \Delta_{\tau} H^{ST}\left(\tau \middle| i,\theta\right) < 0$ for $\forall \tau > \tau'$. Therefore $\tau^{ST}\left(i,\theta\right) = \max\left\{\tau : \Delta_{\tau} H^{ST}\left(\tau,\theta\right) \geq 0\right\}$.

Proof of Lemma 3

To prove $H^{NT}(i,t)$ is concave in i and submodular in (i,t), we use induction. We start from the end of the season (T) and work our way backward with Δt steps. Let's first look at $t \in [T-\Delta t, T]$. Assume Δt is small enough such that there is only a single protection level active in this period ($\tau^{NT}(t)$). When $0 \le i \le \tau^{NT}(t)$, we can write $\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t}$ and $\Delta_i^2 H^{NT}(i,t)$ as follow:

$$\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t} = -\lambda_1 (p_1 + h_1) p(i-1, \lambda_1 (T-t)), \tag{A13}$$

$$\Delta_i^2 H^{NT}(i,t) = -(p_1 + h_1) p(i-1, \lambda_1 (T-t)). \tag{A14}$$

When $i > \tau^{NT}(t)$, we can write $\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t}$ and $\Delta_i^2 H^{NT}(i,t)$ as follow:

$$\begin{aligned}\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t} &= -\lambda_0 (p_0 - k + h_1) p(i - \tau^{NT}(t) - 1, \lambda_{0,1}(T-t)) \\ &-\lambda_{0,1}(p_1 + h_1)(\alpha_1)^{-(i-\tau^{NT}(t)-2)} p(i-1, \lambda_{0,1}(T-t)) \; {}_1F_1(i-\tau^{NT}(t), i, \lambda_0 (T-t)),\end{aligned} \tag{A15}$$

$$\begin{aligned}\Delta_i^2 H^{NT}(i,t) &= -\alpha_0 (p_0 - k + h_1) p(i - \tau^{NT}(t) - 1, \lambda_{0,1}(T-t)) \\ &-(p_1 + h_1)(\alpha_1)^{-(i-\tau^{NT}(t)-2)} p(i-1, \lambda_1 (T-t)) \; {}_1F_1(i-\tau^{NT}(t), i, -\lambda_0 (T-t)).\end{aligned} \tag{A16}$$

(A13), (A14), (A15), (A16) tell us that $\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t} \le 0$ and $\Delta_i^2 H^{NT}(i,t) \le 0$ for all i, therefore $\Delta_i H^{NT}(i,t)$ is decreasing in t and i on $[t,T]$ where $T - t \le \Delta t$.

Next we extend the analysis to show that $\Delta_i H^{NT}(i,t)$ is decreasing in t and i on $[t,T]$ where $T-t \le 2\Delta t$. When $0 \le i \le \tau^{NT}(t)$, we can write $\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t}$ and $\Delta_i^2 H^{NT}(i,t)$ as follow:

$$\begin{aligned}\partial \frac{\Delta_i H^{NT}(i,t)}{\partial t} &= \lambda_1 (\Delta_i G(1, T-\Delta t) - p_1) p(i-1, \lambda_1 (T - \Delta t - t)) \\ &+\lambda_1 \sum_{n=0}^{i-2} \Delta_i^2 G(i-n, T-\Delta t) p(n, \lambda_1 (T-\Delta t - t)),\end{aligned} \tag{A17}$$

$$\Delta_i^2 H^{NT}\left(i,t\right) = \left(\Delta_i G\left(1, T-\Delta t\right) - p_1\right) p\left(i-1, \lambda_1\left(T-\Delta t - t\right)\right) + \sum_{n=0}^{i-2} \Delta_i^2 G\left(i-n, T-\Delta t\right) p\left(n, \lambda_1\left(T-\Delta t-t\right)\right). \tag{A18}$$

When $i > \tau^{NT}\left(t\right)$, we can write $\partial \dfrac{\Delta_i H^{NT}\left(i,t\right)}{\partial t}$ and $\Delta_i^2 H^{NT}\left(i,t\right)$ as follow:

$$\begin{aligned}\frac{\partial \Delta_i H^{NT}\left(i,t\right)}{\partial t} &= -\lambda_{0,1}\alpha_0\left(p_0 - k - \Delta_i G\left(\tau^{NT}\left(t\right)+1, T-\Delta t\right)\right) p\left(i-\tau^{NT}\left(t\right)-1, \lambda_{0,1}\left(T-\Delta t-t\right)\right) \\ &+\lambda_{0,1}\sum_{m=0}^{i-\tau^{NT}(t)-1}\left(\Delta_i^2 H^{NT}\left(i-m, T-\Delta t\right)\right) p\left(m, \lambda_{0,1}\left(T-\Delta t-t\right)\right) \\ &+\lambda_1\left(\alpha_1\right)^{-\left(i-\tau^{NT}(t)-1\right)}\sum_{n=i-\tau^{NT}(t)}^{i-1}\Delta_i^2 G\left(i-n, T-\Delta t\right) p\left(n, T-\Delta t-t\right)\ {}_1F_1\left(i-\tau^{NT}\left(t\right), n+1, -\lambda_0\left(T-\Delta t-t\right)\right),\end{aligned} \tag{A19}$$

$$\begin{aligned}\Delta_i^2 H^{NT}\left(i,t\right) &= -\alpha_0\left(p_0 - k - \Delta_i G\left(\tau^{NT}\left(t\right)+1, T-\Delta t\right)\right) p\left(i-\tau^{NT}\left(t\right)-1, \lambda_{0,1}\left(T-\Delta t-t\right)\right) \\ &+\sum_{m=0}^{i-j-1}\left(\Delta_i^2 H^{NT}\left(i-m, T-\Delta t\right)\right) p\left(m, \lambda_{0,1}\left(T-\Delta t-t\right)\right) \\ &+\left(\alpha_1\right)^{-\left(i-\tau^{NT}(t)-2\right)}\sum_{n=i-\tau^{NT}(t)}^{i-1}\Delta_i^2 G_1\left(i-n\right) p\left(n, \lambda_1\left(T-\Delta t-t\right)\right)\ {}_1F_1\left(i-\tau^{NT}\left(t\right), n+1, -\lambda_0\left(T-\Delta t-t\right)\right).\end{aligned} \tag{A20}$$

By definition, $\Delta_i G\left(\tau^{NT}\left(t\right)+1, T-\Delta t\right) < p_0 - k$. Previously, we showed that $\Delta_i^2 G_1\left(i,t\right) \le 0$, and $\Delta_i^2 H^{NT}\left(i,t\right) \le 0$ on $t \in \left[T-\Delta t, T\right]$. Therefore, $\Delta_i H^{NT}\left(i,t\right)$ is decreasing in i and t on $\left[t,T\right]$ where $T-t \le 2\Delta t$. We can repeat this process for $3\Delta t$, $4\Delta t$, … and eventually cover the whole season, $\left[0,T\right]$. Therefore $H^{NT}\left(i,t\right)$ is submodular in i and t, also concave in i on $t \in \left[0,T\right]$. □

Part (2): We know $\lim_{t\to 0}\Delta_i H^{NT}\left(1,t\right) \sim p_1$ and $\lim_{t\to T}\Delta_i H^{NT}\left(1,t\right) \sim -h_1$. In part (1), we showed $\dfrac{\partial \Delta_i H^{NT}\left(i,t\right)}{\partial t} \le 0$, therefore $\Delta_i H^{NT}\left(i,t\right)$ is decreasing in $t \in \left[0,T\right]$. As a result, $\Delta_i H^{NT}\left(1,t\right)$ for $t \in \left(0,T\right)$ is either below or above $p_0 - k$. When $\Delta_i H^{NT}\left(1,t\right) < p_0 - k$, $\tau^{NT}\left(t\right) = 0$ for $\forall i \ge 1$ since $\Delta_i^2 H^{NT}\left(i,t\right) \le 0$. When $\Delta_i H^{NT}\left(1,t\right) \ge p_0 - k$, then $\exists j : \Delta_j H^{NT}\left(j+1,t\right) \le p_0 - k$. Therefore $\tau^{NT}\left(t\right) = \max\left\{i : \Delta_i H^{NT}\left(i,t\right) \ge p_0 - k\right\}$ for all $t \in [0,T]$. □

Proof of Proposition 6

Please recall that solution to indifference points of optimal rationing policy is identified using $\Delta_i V\left(i,t_{i-1}\right) = p_0 - k$. $\Delta_i H^{NT}\left(i,t'_{i-1}\right) = p_0 - k$ will be used to calculate indifference points for NT policy. At indifference points, we know $H^{NT}\left(i,t_{i-1}\right) = G_1\left(i,t_{i-1}\right)$ since all units will be protected for remaining of season. Therefore $\Delta_i H^{NT}\left(i,t\right)$ on $t \in \left[0,T\right]$ is simplified as follow, just as in the classic newsvendor model:

$$\Delta_i H^{NT}\left(i,t\right) = p_1 - \left(p_1 + h_1\right)\sum_{n=0}^{i-1} p\left(n,\lambda_1\left(T-t\right)\right). \tag{A21}$$

By Lemma 3, $\Delta_i H^{NT}\left(i,t\right)$ is decreasing in t. Therefore at a given i either $\Delta_i H^{NT}\left(i,t\right)$ is always below $p_0 - k$ or there exist a time (t'_{i-1}) after which it goes below $p_0 - k$. When the former happens $t'_{i-1} = 0$, while in the latter case $t'_{i-1} > 0$. We know at the limits (A21) becomes:

$$\lim_{t\to T} \Delta_i H^{NT}\left(i,t\right) = -h_1,$$

$$\lim_{T\to\infty}\lim_{t\to 0} \Delta_i H^{NT}\left(i,t\right) = \lim_{T\to\infty}\left(p_1 - \left(p_1 + h_1\right)\sum_{n=0}^{i-1} p\left(n,\lambda_1\left(T-t\right)\right)\right) = p_1,$$

since $\Delta_i H^{NT}\left(i,t\right)$ is monotonically decreasing in i as proved earlier. Therefore, starting from $i = 1$ there exit a t'_0 such that $\Delta_i H^{NT}\left(1,t'_0\right) = p_0 - k$. We know $\Delta_i H^{NT}\left(i,t\right)$ is also decreasing in i (i.e. concavity) therefore following holds:

$$\Delta_i H^{NT}\left(2,t'_0\right) \le \Delta_i H^{NT}\left(1,t'_0\right) = p_0 - k.$$

This means that there exists a $t'_1 \le t'_0$ such that $\Delta_i H^{NT}\left(2,t'_1\right) = p_0 - k$. This procedure continues until we reach a level of inventory ($i = n + 2$) where $t'_{n+1} \le 0$. Then all the indifference points have been identified as follows: $0 = t'_n < \ldots < t'_1 < t'_0 < T$.

Moreover, Equation (A21) using $\Delta_i H^{NT}\left(i,t'_{i-1}\right) = p_0 - k$ can be further simplified as $\left(p_1 + h_1\right)\sum_{n=0}^{i-1} p\left(n,\lambda_1\left(T-t\right)\right) = p_1 - p_0 + k$. In the last step, it can be written as $P\left(i,\lambda_1\left(T-\overline{t}_i\right)\right) = \frac{p_0 - k + h_1}{p_1 + h_1}$ to complete the proof. □

Proof of Corollary 2

We first prove the following lemma which is used in the proof of Corollary 2.

Lemma A3: $\Delta_i V\left(i,\theta\right) \geq \Delta_i G_1\left(i,\theta\right)$ for any θ and $i \geq 0$.

Proof: Let $\delta\left(i,\theta\right) = V\left(i,\theta\right) - G_1\left(i,\theta\right)$. Since $V\left(i,\theta\right)$ is the optimal value function on $\left[\theta,T\right]$, we know that $\delta\left(i,\theta\right) \geq 0$ for all θ and i. It remains to be show that $\Delta_i\delta\left(i,\theta\right) \geq 0$ for all θ and $i \geq 0$.

- When $i = 1$, $\Delta_i V\left(i,\theta\right) \geq \Delta_i G_1\left(i,\theta\right)$ holds easily because $V\left(0,\theta\right) = G_1\left(0,\theta\right) = 0$.
- When $i > 1$, the proof is more complicated. We follow a backward induction and start from end of the season where $\theta \in \left[t_0,T\right]$. In this period, first order and second order differences with respect to i can be written as follow:

$$\Delta_i\delta\left(i,\theta\right) = \left(\omega + h_1\right)P\left(i,\lambda_{0,1}\left(T-\theta\right)\right) - \left(p_1 + h_1\right)P\left(i,\lambda_1\left(T-\theta\right)\right) \quad \forall i \geq 0,$$

$$\Delta_i^2\delta\left(i,\theta\right) = -\left(\omega + h_1\right)p\left(i-1,\lambda_{0,1}\left(T-\theta\right)\right) + \left(p_1 + h_1\right)p\left(i-1,\lambda_1\left(T-\theta\right)\right) \quad \forall i \geq 1.$$

Now define a ratio $R_i = \frac{\left(\omega + h_1\right)p\left(i-1,\lambda_{0,1}\left(T-\theta\right)\right)}{\left(p_1 + h_1\right)p\left(i-1,\lambda_1\left(T-\theta\right)\right)} = \frac{\omega + h_1}{p_1 + h_1}e^{-\lambda_0\left(T-\theta\right)}\left(1 + \frac{\lambda_0}{\lambda_1}\right)^{i-1}$. It is easy to see that $\Delta_i^2\delta\left(i,\theta\right) \leq 0 \Leftrightarrow R_i \geq 1$. It is also easy to see that R_i is positive and strictly increasing in i. Depending on the initial value of R_i, there are two possible cases:

Case I- $R_1 \geq 1$. In this case $R_i \geq 1, \forall i$, because it's increasing in *i*. Therefore,

$$\left(\omega_1 + h_1\right)p\left(\mathrm{n},\lambda_{0,1}\left(T-\theta\right)\right) \geq \left(p_1 + h_1\right)p\left(n,\lambda_1\left(T-\theta\right)\right) \quad \text{for } \forall n \geq 0.$$

Summing over *n* we get:

$$\left(\omega_1+h_1\right)\sum_{n=i}^{\infty}p\left(n,\lambda_{0,1}\left(T-\theta\right)\right)\geq\left(p_1+h_1\right)\sum_{n=i}^{\infty}p\left(n,\lambda_1\left(T-\theta\right)\right)\quad\text{for }\forall\text{i}\geq 0.$$

That is, $\left(\omega_1+h_1\right)P\left(i,\lambda_{0,1}\left(T-\theta\right)\right)\geq\left(p_1+h_1\right)P\left(i,\lambda_1\left(T-\theta\right)\right)$. Thus, $\Delta_i\delta\left(i,\theta\right)\geq 0$.

Case II- $R_1<1$. Because R_i is increasing and $\lim_{i\to\infty}R_i=\infty$, there must exist an i' such that $R_i<1$ for $0\leq i<i'$ and $R_i\geq 1$ for $i\geq i'$.

Case II(1): For $i\geq i'$, $R_i\geq 1$. As in Case I, we immediately have $\Delta_i\delta\left(i,\theta\right)\geq 0$.

Case II(2): For $0\leq i<i'$, $R_i<1$ and $\delta\left(i,\theta\right)$ is convex in *i*. Because we know $\delta\left(0,\theta\right)=0$, and $\delta\left(i,\theta\right)\geq 0$, it must be the case that $\delta\left(i,\theta\right)$ is increasing in i for all $0\leq i<i'$. (That is, the decreasing part of a convex function is not possible here.)

Therefore, combining both cases we get $\Delta_i\delta\left(i,\theta\right)\geq 0$ for all ***i*** and $\theta\in\left[t_0,T\right]$.

We now extend the analysis to $\theta\in\left[t_1,t_0\right]$. Since $\tau\left(t\right)=1$ on $\theta\in\left[t_1,t_0\right]$, we know $\Delta_iV\left(1,\theta\right)=\Delta_iG_1\left(1,\theta\right)$. Similar to what we did earlier, we can show that $\delta\left(i,\theta\right)=V\left(i,\theta\right)-G_1\left(i,\theta\right)$ is always increasing in ***i*** for $i\geq 2$. Knowing $\Delta_i\vartheta\left(1,\theta\right)=0$, we conclude that $\Delta_i\vartheta\left(i,\theta\right)\geq 0$ for all i on $\theta\in\left[t_1,t_0\right]$.

This proof process can be repeated for $\theta\in\left[t_2,t_1\right]$ … to show that $\Delta_i\vartheta\left(i,\theta\right)\geq 0$ for all i and θ. □

Now we return to the proof of Corollary 2. As a recall, we describe how the indifference points are calculated under the optimal rationing policy. Then we compare that with NT heuristic.

For the optimal rationing policy, indifference points are calculated using $\Delta_iV\left(i,t_i\right)=p_0-k$. To find the first indifference point, t_0, we will use following:

$$\Delta_iV\left(i=1,t_0\right)=p_1-\left(p_1+h_1\right)p\left(0,\lambda_1\left(T-t_0\right)\right)=p_0-k. \tag{A22}$$

Then we calculate $V\left(i,t_0\right)$ for $\forall i\geq 1$. The second indifference point, t_1, is calculated using following:

$$\Delta_iV\left(i=2,t_1\right)=p_1+\sum_{n=0}^{1}\left[\Delta_iV\left(2-n,t_0\right)-p_1\right]p\left(n,\lambda_1\left(t_0-t_1\right)\right)=p_0-k. \tag{A23}$$

Then we calculate $V\left(i,t_1\right)$ for $\forall i \geq 1$. The procedure continuous till $V\left(i,0\right)$ for $\forall i \geq 1$ is calculated.

Now, we compare the procedure with that of NT heuristic. While equation (A22) stays the same, equation (A23) changes to:

$$\Delta_i H^{NT}\left(i=2,t_1'\right) = p_1 + \sum_{n=0}^{1}\left[\Delta_i G_1\left(2-n,t_0'\right) - p_1\right] p\left(n,\lambda_1\left(t_1' - t_0'\right)\right) = p_0 - k\,. \qquad \text{(A24)}$$

Therefore, both policy match up until first indifference point ($t_0 = t_0'$), which means:

$$\tau^{NT}\left(t\right) = \tau^{OPT}\left(t\right) = 0 \;\; \forall t \in \left[t_0', T\right].$$

In the next step to find t_1' using (A24), we need marginal value at $t_0 = t_0'$ which is calculated using $\Delta_i V\left(i,t_0\right)$ and $\Delta_i G_1\left(i,t_0'\right)$ respectively in optimal and NT heuristic. Comparing (A23) and (A24), we find $\sum_{n=0}^{1}\left[\Delta_i G_1\left(2-n,t_0'\right) - p_1\right] p\left(n,\lambda_1\left(t_1' - t_0'\right)\right) = \sum_{n=0}^{1}\left[\Delta_i V\left(2-n,t_0\right) - p_1\right] p\left(n,\lambda_1\left(t_0 - t_1\right)\right)$.

Because of Lemma A3 and $t_0 = t_0'$, it is clear that $t_1' \geq t_1$. This means under NT heuristic, protection level of one is optimal action for a longer period compared to that of optimal policy. Similarly, this analysis can be extended to show that $\bar{t}_n \geq t_n$. Therefore, there are fewer protection levels under the NT heuristic than under the optimal policy. □

Proof of Proposition 7

We first prove the following lemma which is used in the proof of Proposition 7:

Lemma A4: $\Pi^{NP,OPT}(S_0,S_1)$ is concave in S_1 and submodular in S_0 and S_1.

Recall that we define G_1 to be the newsvendor profit function for the offline store:

$$G_1\left(i,t\right) = p_1 i - \left(p_1 + h_1\right)\sum_{j=0}^{i}\left(i-j\right)p\left(i,\lambda_1\left(T-t\right)\right).$$

We now similarly define G_0 to be the newsvendor profit function for the online store:

$$G_0\left(i,t\right) = p_0 i - \left(p_0 + h_0\right)\sum_{j=0}^{i}\left(i-j\right)p\left(i,\lambda_0\left(T-t\right)\right).$$

We can write the (NP,OPT) profit function as follows:

$$\Pi^{NP,OPT}(S_0,S_1) = G_0(S_0,0) + G_1(S_1,0)\big(1 - P\big(S_0,\lambda_0 T\big)\big) + \int_{\theta=0}^{T} p_1 E\big[D_1(\theta)\wedge S\big]\lambda_0 p\big(S_0 - 1,\lambda_0\theta\big)d\theta$$
$$+\int_{\theta=0}^{T}\left\{\sum_{u=0}^{S_1} V\big(S_1 - u,\theta\big)p(u,\lambda_1\theta)\right\}\lambda_0 p(S_0 - 1,\lambda_0\theta)d\theta$$

Its first order difference with respect to S_1 is as follow:

$$\Delta_{S_1}\Pi^{NT,OPT}(S_0,S_1) = \Delta_{S_1} G_1(S_1,0)\big(1 - P\big(S_0,\lambda_0 T\big)\big)$$
$$+\int_{\theta=0}^{T}\left\{p_1 P(S_1,\lambda_1\theta) + \sum_{u=0}^{S_1-1}\Delta_{S_1} V\big(S_1 - u,\theta\big)p(u,\lambda_1\theta)\right\}\lambda_0 p(S_0 - 1,\lambda_0\theta)d\theta$$

And the second order difference with respect to S_1 is as follow:

$$\Delta_{S_1}^2\Pi^{NT,OPT}(S_0,S_1) = \Delta_{S_1}^2 G_1(S_1,0)\big(1 - P\big(S_0,\lambda_0 T\big)\big)$$
$$+\int_{\theta=0}^{T}\left\{-p_1 p(S_1 - 1,\lambda_1\theta) + \sum_{u=0}^{S_1-1}\Delta_{S_1}^2 V\big(S_1 - u,\theta\big)p(u,\lambda_1\theta)\right\}\lambda_0 p(S_0 - 1,\lambda_0\theta)d\theta.$$

Since both $G_1(S_1,0)$ and $V\big(S_1 - u,\theta\big)$ are concave (Propositions 1 and 5), we conclude that $\Pi^{NT,\mathrm{OPT}}(S_0,S_1)$ is concave in S_1.

Next, the (NP,OPT) profit function can be also written as follow:

$$\Pi^{NP,OPT}(S_0,S_1) = \Pi^{NP,\varnothing}(S_0,S_1) + \int_{\theta=0}^{T}\left\{\sum_{u=0}^{S_1}\Big(V\big(S_1 - u,\theta\big) - G_1\big(S_1 - u,\theta\big)\Big)p(u,\lambda_1\theta)\right\}\lambda_0 p(S_0 - 1,\lambda_0\theta)d\theta.$$

We first show that the following two inequalities hold:

$$p(S_0 - 1,\lambda_0\theta) - p(S_0 - 2,\lambda_0\theta) \le 0\,, \tag{A25}$$

$$\Delta_{S_1} V\big(S_1 - u,\theta\big) - \Delta_{S_1} G_1\big(S_1 - u,\theta\big) \ge 0\,. \tag{A26}$$

Considering our assumption that the service levels are more than 0.5, we have $S_n^{NP,\varnothing} \ge \lceil \lambda_n T\rceil \ge \lceil \lambda_0\theta\rceil$, and (A25) holds. (A26) follows from Lemma A3.

Next, the profit function's cross difference with respect S_0 and S_1 can be written as follow:

$$\Delta_{S_0}\Delta_{S_1}\Pi^{NP,OPT}(S_0,S_1) = \int_{\theta=0}^{T}\left\{\sum_{u=0}^{S_1-1}\left(\Delta_{S_1}V\left(S_1-u,\theta\right)-\Delta_{S_1}G_1\left(S_1-u,\theta\right)\right)p(u,\lambda_1\theta)\right\}\lambda_0\left(p(S_0-1,\lambda_0\theta)-p(S_0-2,\lambda_0\theta)\right)d\theta.$$

Thus $\Delta_{S_0}\Delta_{S_1}\Pi^{NP,OPT}(S_0,S_1)\leq 0$ because of (A25) and (A26). That is, the profit function is submodular. □

Now we return to the proof of Proposition 7.

Part (1) Because ϕ is a feasible rationing policy, its performance by definition is no better than the optimal rationing policy. Therefore, we have $\Pi^{P,\phi}(S_1)\leq\Pi^{P,OPT}(S_1)$ and $\Pi^{NP,\varnothing}(S_0,S_1)\leq\Pi^{NP,OPT}(S_0,S_1)$.

Part (2) Please note the following:

$$\Delta_{S_1}\Pi^{NP,OPT}(S_0,S_1) = \Delta_{S_1}G_1\left(S_1,0\right)+\int_{\theta=0}^{T}\left\{\sum_{u=0}^{S_1-1}\left(\Delta_{S_1}V\left(S_1-u,\theta\right)-\Delta_{S_1}G_1\left(S_1-u,\theta\right)\right)p(u,\lambda_1\theta)\right\}\lambda_0 p(S_0-1,\lambda_0\theta)d\theta.$$

Plugging in $S_1=S_1^{NP,\varnothing}$, we note $\Delta_{S_1}G_1(S_1^{NP,\varnothing},0)=0$ and Lemma A3 imply that the LHS is also positive: $\Delta_{S_1}\Pi^{NP,OPT}(S_0,S_1^{NP,\varnothing})\geq 0$. Then, because $\Delta_{S_1}\Pi^{NP,OPT}(S_0,S_1^{NP,OPT})\geq 0$, the concavity of $\Pi^{NP,OPT}(S_0,S_1)$ in S_1 (Lemma A4) means $S_1^{NP,OPT}\geq S_1^{NP,\varnothing}$.

Please also note the following:

$$\Delta_{S_0}\Pi^{NP,OPT}(S_0,S_1) = \Delta_{S_0}G_0(S_0,0)+\int_{\theta=0}^{T}\left\{\sum_{u=0}^{S_1}\left[V\left(S_1-u,\theta\right)-G_1\left(S_1-u,\theta\right)\right]p(u,\lambda_1\theta)\right\}\lambda_0\left[p(S_0-1,\lambda_0\theta)-p(S_0-2,\lambda_0\theta)\right]d\theta.$$

For any $S_0\geq S_0^{NP,\varnothing}$, the assumption of $S_L\geq 0.5$ means $S_0\geq S_0^{NP,\varnothing}\geq\lceil\lambda_n T\rceil\geq\lceil\lambda_0\theta\rceil$. So $p(S_0-1,\lambda_0\theta)\leq p(S_0-2,\lambda_0\theta)$. Plugging in $S_0=S_0^{NP,\varnothing}$, we get $\Delta_{S_0}\Pi^{OPT}(S_0^{NP,\varnothing},S_1)\leq 0$ for all $S_0\geq S_0^{NP,\varnothing}$, because $\Delta_{S_0}G_0(S_0^{NP,\varnothing},0)=0$ and $V\left(S_1-u,\theta\right)\geq G_1\left(S_1-u,\theta\right)$. Since $\Delta_{S_0}\Pi^{NP,OPT}(S_0,S_1)$ is decreasing in S_0 (Lemma A4) on $S_0\geq S_0^{NP,\varnothing}$, we must have $S_0^{NP,OPT}\leq S_0^{NP,\varnothing}$. □