M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 1, Winter 2013, pp. 33–44
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.1120.0399

Inventory Pooling to Deliver Differentiated Service

Aydın Alptekinoğlu
Cox School of Business, Southern Methodist University, Dallas, Texas 75275,
aalp@cox.smu.edu

Arunava Banerjee
Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida 32611,
arunava@cise.ufl.edu

Anand Paul
Warrington College of Business Administration, University of Florida, Gainesville, Florida 32611,
anand.paul@warrington.ufl.edu

Nikhil Jain
Servigistics India, Gurgaon, Haryana 122001, India, nikhil.jain@servigistics.com

Inventory pooling is at the root of many celebrated ideas in operations management. Postponement, component commonality, and resource flexibility are some examples. Motivated by our experience in the aftermarket services industry, we propose a model of inventory pooling to meet differentiated service levels for multiple customers. Our central research question is the following: What are the minimum inventory level and optimal allocation policy when a pool of inventory is used in a single period to satisfy individual service levels for multiple customers? We measure service by the probability of fulfilling a customer's entire demand immediately from stock. We characterize the optimal solution in several allocation policy classes; provide some structural results, formulas, and bounds; and also make detailed interpolicy comparisons. We show that the pooling benefit is always *strictly* positive, even when there are an arbitrary number of customers with perfectly positively correlated demands.

Key words: inventory pooling; type 1 service level; inventory allocation policy; aftermarket services; spare parts; pooling benefit; demand correlation
History: Received September 4, 2010; accepted March 16, 2012. Published online in *Articles in Advance* October 24, 2012.

1. Introduction

Inventory pooling, the practice of using a common pool of inventory to satisfy two or more sources of random demand, has been studied in the context of many operationally challenging situations. For example, large streams of literature explore how pooling acts as an essential ingredient in containing the operational costs of high product variety, in mitigating supply chain disruptions, and in striking the right trade-off between operational benefits and fixed costs of product-process flexibility in supply chains (Lee 2004).

In this paper, we analyze a single-period model that captures inventory pooling in an environment where customers' service expectations differ; hence, the policy by which inventory is allocated becomes critical, should one wish to reap the benefits of inventory pooling. We pose a fundamental question: *When a pool of inventory is used to serve customers with varying service-level requirements, what are the minimum inventory level and optimal allocation policy*? We measure service by the probability of meeting a customer's entire demand immediately from stock (type 1 service measure; Silver et al. 1998, p. 245).

This is the essence of a problem that frequently occurs in aftermarket service operations, an industry sector estimated to make up 8% of the gross domestic product in the United States (Cohen et al. 2006), when certain levels of service need to be maintained for a collection of current and relatively long-term contracts at minimum cost. In such settings, the total revenue from trade is fixed because demands are eventually satisfied and prices are contractually set, even though they may vary from one customer to another. Often, service-level requirements of customers differ. We frame the problem as follows: find the combination of inventory level and allocation policy that maintains a set of current contracts most efficiently.

Our model is highly stylized; it assumes a single period in which the firm uses a type 1 service measure and batches demands from multiple customers before attempting to fulfill them. We observed how HOLT CAT, Caterpillar's Texas dealership, manages spare parts inventories. The most important measure

of service it monitors for each store is called *on time in full* (OTIF), the percentage of spare parts orders fully satisfied on time, because its customers often see no value in having only a portion of the parts required to perform a repair (Barry 2006). Moreover, HOLT CAT discourages urgent orders because nonurgent orders for such parts as air filters are typically batch-processed overnight rather than immediately upon order receipt, which is more costly. The totality of HOLT CAT's operation is, of course, much more complicated than what is suggested by our stylized model. For example, the time component of OTIF is only crudely captured, and inventory replenishment and demand batching may not always be synchronized even for nonurgent items. Nevertheless, we hope that our model might serve as a building block for more complex and realistic models in this area.

What piqued our interest the most in industry practice is the decoupling of ordering and allocation decisions. In our model, we treat the ordering decision, which sets the spare parts inventory level, and the allocation decision, which rations the available inventory among customers via a prioritization scheme of some sort, simultaneously. By optimizing jointly over allocation policies as well as inventory levels, we demonstrate the benefits of integrating these decisions.

Our main goal is to find analytical characterizations of the optimal inventory level and allocation policy for customers with different service-level requirements. We define three classes of allocation policies, and we obtain structural results and formulas that optimize jointly over inventory level as well as allocation policy. We find the optimal solution for two customers with arbitrary demand distributions but require independent and identically distributed (iid) demands for three or more customers. We demonstrate the advantages of interlinking inventory and allocation decisions and give insights into when less sophisticated allocation policies are almost as good as the optimal policy. Finally, we have a result that is in contrast with backorder-cost models, in which the pooling benefit is zero when demands are perfectly positively correlated (Eppen 1979). In our service-level-constrained model, we show analytically that the pooling benefit is strictly positive even if demands are perfectly positively correlated. We relegate all proofs to Appendix A (see the online companion, available at http://dx.doi.org/10.1287/msom.1120.0399).

2. Literature Review

In positioning our paper, we find the broad framework presented by Özer and Xiong (2008) useful. They identify four quadrants into which many inventory models can be slotted. The dimensions underlying their matrix are (1) backorder-cost or service-level models and (2) single or multiple demand points. Our paper fits into the fourth quadrant; for completeness we review representative papers in the two quadrants most relevant to our paper: backorder-cost and service-level models with multiple demand points. Even within a given quadrant, researchers make different modeling choices: continuous review versus periodic review, optimizing the parameters of an assumed allocation policy versus finding the form of the optimal allocation policy, and finite horizon versus infinite horizon.

We remark that the setting in the present paper and the settings in the bulk of the literature reviewed below are not directly comparable because they apply to different distribution environments. For instance, in many models, customers carry inventory; the warehouse may (or may not) carry inventory at a central location and allocates inventory to satisfy the replenishment requests from downstream retailers (who sell to end consumers). Further, the literature generally does not assume allocation can be made after demand is realized; that is, demands are not batched during a period but are satisfied in real time. These models capture a context in which it would be too late to wait for demand realizations before making allocation decisions given the positive shipment lead times between the warehouse and the retailers.

Our paper is most closely related to Swaminathan and Srinivasan (1999) and Zhang (2003). Swaminathan and Srinivasan (1999) develop an algorithm to compute the optimal ordering and allocation policies for the same problem that we study. The combinatorial complexity of the problem, and hence the difficulty of obtaining a practical solution efficiently, is evident from their paper. Switching iteratively between binary search and Monte Carlo simulation, their approach is necessarily computational and exponential time, because they pose the problem in its most general form without structuring either the space of policies or the demand distributions. In contrast, we emphasize policies that are intuitive and easy to implement, and to that end we provide some structural results, formulas, and bounds. We also make detailed interpolicy comparisons. Zhang (2003) studies a specific class of allocation policies again in a single period, considering the special case when demand distributions and service levels are such that at most one customer's demand can go unsatisfied.

Eppen and Schrage (1981) study a supplier-depot-multiple-warehouse system in which the warehouses face mutually independent normally distributed demands. At the end of every period, an aggregate replenishment order y is placed with the supplier. Replenishment stock is routed through the depot, where an allocation rule has to be framed for

distributing stock to the warehouses. The following allocation rule is assumed: stock is distributed so as to equalize type 1 service levels at the warehouses. This allocation rule is feasible when demands are stable but may otherwise be infeasible. Assuming that the rule is feasible, the authors develop an expression for the value of y that minimizes the sum of expected holding and backorder costs.

Schwarz et al. (1985) study a one-warehouse multiple-retailer system in which all the entities hold stock and follow continuous-review (Q, R) policies. Each retailer faces independent Poisson demand and receives replenishments from the warehouse, which is replenished by an uncapacitated source. If the retailers run out of stock, they place backorders with the warehouse. The backorders are filled on a first-come, first-serve (FCFS) basis from the warehouse. The problem is to determine lot sizes and reorder points so as to maximize the fill rate at the warehouse subject to an upper bound on the system inventory.

Hopp et al. (1999) model a spare parts distribution system wherein a distribution center (DC) supports a number of customer facilities that generate Poisson demands. The DC as well as the facilities hold stock; the facilities follow a one-for-one replenishment strategy, and the DC follows a continuous-review replenishment strategy. The problem is to determine the parameters of the ordering policies at the DC and the facilities so as to minimize expected inventory-related costs across the system subject to service-level constraints that place upper bounds on the order frequency at the DC and the average delay experienced by each facility.

Caglar et al. (2004) study a distribution system with a similar topology to that in Hopp et al. (1999), but for repairable parts. A fixed number of depots serves customers, each of whom owns a machine with multiple parts subject to failure. Each depot sees a Poisson arrival process of failed parts. Each failed part is replaced by a spare part from stock or backordered. All failed parts are transported to a central warehouse, where they are repaired. Repair times at the warehouse and transportation times between the central warehouse and depots are modeled. The problem is to determine basestock levels at the central warehouse and depots so as to minimize the total systemwide inventory holding cost subject to service-level constraints in the form of bounds on average response time. A computationally efficient heuristic is presented to solve the problem.

Deshpande et al. (2003) study service-level differentiation for two demand classes, each following a Poisson process with different rates. They assume a continuous-review (Q, R) policy for inventory replenishment and a *threshold policy* for inventory allocation, which stipulates that lower-priority customers (those with lower shortage cost) are not served at all if inventory on hand falls below a threshold level. They study optimal policy parameters and backlog clearing mechanisms.

Arslan et al. (2007) study a problem that is quite close to ours, but they model it differently. They aim to find the optimal parameters of a continuous-review (Q, R) inventory policy for a single stocking point and an allocation policy for a number of customers with differentiated service-level requirements. The allocation policy is a natural adaptation of threshold policy to multiple customers. Customer demands are Poisson with different rates. The problem is to find threshold levels and an optimal value of R (for a given Q) such that the probability of a strictly positive inventory level exceeds a certain minimum acceptable level, which varies from customer to customer. The authors present an efficient heuristic to solve the problem.

Özer and Xiong (2008) study a distribution system comprising a warehouse replenishing multiple retailers, each of which operates a continuous-review basestock (one-for-one replenishment) policy. All locations carry inventory. The demand process at each retailer is Poisson; unsatisfied demand is backordered. The warehouse fills requests from the retailers on an FCFS basis. The problem is to determine basestock levels that minimize the system inventory holding cost subject to the following service-level constraint: the probability that a demand at each retailer can be filled from existing stock must exceed a threshold level. Bounds and heuristics are developed to determine optimal basestock levels at each location and the ensuing average cost.

Gallego et al. (2007) study allocation mechanisms whereby a central control point (a manager who has access to systemwide inventory levels and costs) makes stock placement decisions for a set of downstream demand points facing Poisson demand with the objective of minimizing expected cost. The same theme of central-versus-local control is explored in Chen (1998), which studies optimal inventory placement in a serial N-stage system and compares echelon stock (central) and installation stock (local) policies.

In closing, we review a few inventory-pooling models. Eppen (1979) shows that there is benefit to inventory pooling in the face of iid normal demands and studies how this benefit varies as a function of demand correlation and the number of demand points. Erkip et al. (1990) extend the Eppen-Schrage model to the case of correlated demands, both across locations and across time at a given location. Özer (2003) explores the interplay between advance demand information and inventory pooling. Alptekinoğlu and Tang (2005) consider arbitrary numbers of depots and demand locations facing multivariate normal demand.

More broadly, two prominent methods of containing operational costs due to high variety are based on pooling: postponement, also known as delayed product differentiation (Lee and Tang 1997, Aviv and Federgruen 2001), and component commonality (Mirchandani and Mishra 2002, Van Mieghem 2004). Many models of assemble-to-order systems (Akçay and Xu 2004) and resource flexibility (Van Mieghem 1998) have some form of pooling at the core.

3. Problem Formulation

A firm supplies a single product to N customers from a centralized pool of inventory over the duration of a single period. Customer i has a random demand X_i and requires a minimum service level of $\beta_i \in (0,1)$; the probability that X_i is fully satisfied must be β_i or more. The X_i's are continuous positive-valued random variables with distribution functions $F_i(\cdot)$, and their sum has a distribution function $G(\cdot)$.

Events unfold as follows: (1) the firm orders S units of the product in advance so as to receive them at the beginning of the period; (2) actual customer demands, denoted by x_i, realize throughout the period; (3) at the end of the period, the firm allocates the available pool of inventory (S units) among N customers according to an allocation policy and makes the appropriate shipments. Any leftover inventory is discarded.

An *allocation policy* in general is a mapping $\mathbf{A}$: $\mathbb{R}_+^{N+1} \to \mathbb{R}_+^N$ from inventory level and demand realizations $(S, x_1, x_2, \ldots, x_N)$ to inventory allocations $(y_1, y_2, \ldots, y_N)$ resulting in $y_i = \mathbf{A}_i(S, x_1, x_2, \ldots, x_N)$ such that $y_i \le x_i$ (no customer receives more inventory than needed) and $\sum_{i=1}^N y_i = \min\{S, \sum_{i=1}^N x_i\}$ (the firm either depletes its inventory or satisfies all customers), where $\mathbb{R}_+$ denotes the set of nonnegative real numbers. Let Ω be the set of all such mappings.

The firm wants to find the minimum inventory S coupled with an allocation policy $\mathbf{A}$ that together meet the service-level requirements. Both of these decisions are made at the beginning of the period, at which point the outcome of $\mathbf{A}$ in terms of allocating actual quantities to the customers is a priori uncertain. That is, at the time of selecting S and $\mathbf{A}$, demands $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ as well as inventory allocations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ that result from applying $\mathbf{A}$ are uncertain; $Y_i = \mathbf{A}_i(S, \mathbf{X})$, the amount of inventory to be allocated to customer i is a random variable, and customer i's demand is fully satisfied if and only if (iff) the event $Y_i = X_i$ occurs. Service-level requirements are therefore in the form of chance constraints.

The firm's problem can be formally stated as follows (let $P\{\cdot\}$ denote probability):

$$\underset{S\in\mathbb{R}_+,\,\mathbf{A}\in\Omega}{\text{minimize}} \quad S$$

$$\text{subject to} \quad P\{\mathbf{A}_i(S,\mathbf{X}) = X_i\} \ge \beta_i \quad \text{for all } i = 1,2,\ldots,N,$$

where $\Omega \equiv \{\mathbf{A}\colon \mathbb{R}_+^{N+1} \to \mathbb{R}_+^N \mid y_i = \mathbf{A}_i(S, x_1, x_2, \ldots, x_N)$ and $y_i \le x_i$ for $i = 1, \ldots, N$, and $\sum_{i=1}^N y_i = \min\{S, \sum_{i=1}^N x_i\}\}$ is the set of mappings that each specify an allocation of available inventory to customers up to their demands. Note that the mapping $\mathbf{A}$ has to be derived at the beginning, before observing demands, because the firm cannot evaluate the feasibility of S without specifying $\mathbf{A}$.

4. Allocation Policies

We first define a class of allocation policies and show that an optimal policy belongs to this class. A *priority policy* is an allocation policy that leaves, at most, one customer partially satisfied; i.e., the set $\{i \in \{1, \ldots, N\}\colon 0 < y_i < x_i\}$ is either empty or a singleton for all demand realizations.

Theorem 1. *An optimal allocation policy is a priority policy.*

We now offer an alternative definition of a priority policy that is more convenient to work with than the definition based on inventory allocations (y variables). An allocation policy belongs to the class of priority policies if it operates as follows. First, customers are ordered in a *priority list*—the sequence by which inventory is "doled out"—before or after demand realizations are observed. Let $\pi\colon \{1, \ldots, N\} \to \{1, \ldots, N\}$ be a one-to-one correspondence between priority list positions and customers. Each priority list $\Pi = (\pi(1), \ldots, \pi(N))$ is defined by one such correspondence π, with $\pi(j)$ representing the customer in the jth position of the priority list. Second, customer demands are filled from the available inventory in decreasing order of priority; demand from customer $\pi(1)$ is filled first, customer $\pi(2)$ second, and so on. This sequential allocation process stops when all demands are filled or when the available inventory is exhausted, whichever occurs first.

In this paper, we define and analyze two main classes of allocation policies that are differentiated by whether or not they make use of actual demand information when forming the priority list.

4.1. Responsive Priority Policies

The priority list Π is constructed using the demand realization information; e.g., smaller demand is filled first (say customer A's), and then larger demand (customer B's) is filled if there is any stock left over. Because actual demand information is used to determine the priority list, such allocation policies are said to be *responsive*. The set of rules involved in mapping the demand information to a priority list can be quite general. Intuitively, it seems more efficient to fill smaller demands first. At the same time, one needs to recognize that customer demand distributions and

service-level requirements may differ, so a simple rank ordering based on magnitude of demand is in general unlikely to work. This basic tension between efficient use of inventory and ability to differentiate service levels is a recurring theme of our paper.

4.2. Anticipative Priority Policies

The priority list Π is constructed without using the demand realization information. We study two particular variations of *anticipative* policies. The first has a deterministic priority list, fixed a priori independently of demand realizations; e.g., customer A always has higher priority than customer B. Because the priorities are assigned on the basis of a fixed list, we call such a policy a *fixed list policy*. The second uses a randomized priority list, again independently of demands. One of the $N!$ possible permutations, which corresponds to a unique one-to-one correspondence π, is chosen according to a discrete probability distribution over the set of all possible priority lists; e.g., a coin is tossed before the demands are realized, and if it falls heads (tails), customer A's (B's) demand is filled first. In contrast to a fixed list policy, the priority list is decided randomly, so we call such a policy a *randomized list policy*.

The optimal inventory levels within each policy class (indicated with subscripts) are ordered as follows.

THEOREM 2. $S^*_{\text{responsive}} \le S^*_{\text{r-list}} \le S^*_{\text{f-list}}$

In practice it is surely simpler to implement a fixed list policy rather than a randomized list or a responsive policy, and it is cleaner and less information-intensive to operate a randomized list policy rather than a responsive policy. Further, there may be exogenous reasons (such as building long-term relationships with customers) that dictate the adoption of a fixed list policy. For these reasons, we conduct a detailed study of each of these classes of allocation policies, beginning with responsive policies.

5. Responsive Priority Policies

Using a responsive priority policy amounts to allowing the priority list to freely depend on demand realizations. One responsive policy that is intuitively appealing, and straightforward to compute and implement, is to serve the customers in ascending order of demand realizations. We call this allocation policy the *greedy policy* (GP). Given a fixed inventory level and any set of demand realizations, there is no allocation rule that completely satisfies more customers than GP does.

Based on GP, we first develop a lower bound on the optimal inventory level for the general problem with an arbitrary set of customer demand distributions and service-level requirements. Let the order statistics corresponding to demands (ordered from smallest to largest) be $X_{[1]}, \ldots, X_{[N]}$. We define partial convolutions of the order statistics as follows: $Z_n = \sum_{i=1}^{n} X_{[i]}$ for $n \in \{1, \ldots, N\}$. Thus, Z_n represents the sum of the n smallest demands; we denote its distribution function by $H_n(\cdot)$.

THEOREM 3. *The unique solution $\underline{S}$ of the equation $\sum_{n=1}^{N} H_n(S) = \sum_{i=1}^{N} \beta_i$ is a lower bound on the optimal inventory level; i.e., $S^* \ge \underline{S}$.*

If it is not already optimal, this bound represents a good starting point for solving the general problem. The proof uses GP, which is the most efficient policy in using the limited inventory, but GP ignores how service-level requirements of customers are dispersed. Therefore, GP may overserve some customers (the ones who tend to have low demand values) and underserve others (high-demand customers). Intuitively speaking, and based on our experience with numerical examples, $\underline{S}$ is either optimal or near optimal for problem instances where demand distributions and service-level requirements do not differ drastically among customers.

In principle, $H_n(\cdot)$ can be computed analytically by using known facts about the distribution functions of order statistics and their convolutions. But in practice, it is easier to use Monte Carlo simulation to compute it, which is what we did to obtain the lower bound $\underline{S}$.

Without imposing some structure on the demand distributions or a limit on the number of customers, the general problem is difficult because of combinatorics of inventory allocation. In order to glean some structural insights into the problem and obtain analytical characterizations of the optimal solution, we first assume that the demands are iid but otherwise arbitrary. We then analyze the two-customer problem with arbitrary (possibly non-iid) demands.

5.1. Independent and Identically Distributed Demands

We first take the simpler case of undifferentiated service levels.

THEOREM 4. *If the demands $X_1, \ldots, X_N$ are iid random variables and service levels are undifferentiated so that $\beta_1 = \cdots = \beta_N = \beta$, then GP is an optimal allocation policy, and the optimal inventory level is the unique solution $S_{\text{GP}}(\beta)$ to the equation $\sum_{n=1}^{N} H_n(S) = N\beta$.*

Next we analyze the case of customers with iid demands requiring differentiated service levels. We shall see that GP plays an important role in this case also. Suppose $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_N$ with at least one strict inequality, and let $S_{\text{GP}}(\bar{\beta}_n)$ be the stock level required by GP to deliver a service level of exactly $\bar{\beta}_n \equiv (\beta_1 + \cdots + \beta_n)/n$ to customers $1, \ldots, n$ for all $n = 1, \ldots, N$. Theorem 3 implies that, with iid demands,

$S_{GP}(\bar{\beta}_N)$ is a lower bound for the stock level required by the optimal responsive policy. We shall show that, barring a theoretical degenerate case to be spelled out in the next paragraph, the lower bound is in fact attained; the optimal stock level is $S_{GP}(\bar{\beta}_N)$. Further, the optimal allocation policy involves applying GP to demand realizations after first scaling each demand realization x_i by a fixed scale factor K_i.

We assume that the stock level needed to serve a set of customers is a strictly increasing function of the number of customers. A degeneracy arises when this assumption fails to hold—some customers have service levels so low that they can free ride on the remaining stock after all the other customers are served and still have their service-level requirements fulfilled. These free riders are of no practical interest in our model, because we are concerned with customers with contractually committed service levels.

We call an allocation policy a *cardinal greedy policy* (CGP) if, with a given stock level, it satisfies the demands of exactly as many customers as GP would satisfy for every set of demand realizations.

Theorem 5. *Suppose customer demands are iid, service levels need to be differentiated, and there exist no free riders; i.e.,* $S_{GP}(\bar{\beta}_{n+1}) > S_{GP}(\bar{\beta}_n)$ *for* $n = 1, \ldots, N-1$. *Then,* (a) *a CGP is an optimal allocation policy and the optimal inventory level is* $S_{GP}(\bar{\beta}_N)$, *and* (b) *there exists an* N*-vector* $(K_1, \ldots, K_N)$ *such that an optimal allocation policy for each demand realization* $(x_1, \ldots, x_N)$ *is to prioritize customers either in increasing order of* $K_i x_i$ *or in increasing order of* x_i.

This result implies that service-level differentiation does not impose an additional inventory burden when demands are iid; servicing a set of customers with distinct service levels $\beta_1, \ldots, \beta_N$ and servicing the same set of customers with a service level of $\bar{\beta}$ for every customer requires an identical inventory level.

In more detail, the following allocation policy is optimal: (i) observe the demand realizations $(x_1, \ldots, x_N)$; (ii) allocate stock to customer i in increasing order of $K_i x_i$ $(i = 1, \ldots, N)$ while passing on to the next customer in the list if the current customer has a demand realization that exceeds the remaining stock; (iii) count the number of customers N_K whose demands are completely satisfied with this allocation policy and compare it with the number of customers N_G whose demands would be completely satisfied by GP; and (iv) if $N_K = N_G$, use the allocation policy in (ii) above; otherwise, allocate according to GP.

This allocation policy is in the class of CGP policies, is feasible for inventory level $S_{GP}(\bar{\beta}_N)$ and a given set of service levels, and is therefore optimal. We note, however, that Theorem 5(b) is an existence result; it asserts that there is an optimal scaling but does not give us a recipe for finding the optimal scale factors.

5.2. The Two-Customer Case

In this subsection, we show that a particular subclass of responsive policies contains the optimal solution in the two-customer case for any set of service-level requirements and demand distributions (possibly non-iid). We treat the special case of bivariate normal demands in Appendix B (see the online companion), focusing on how the optimal inventory level (S^*) and magnitude of the pooling benefit ($S_1 + S_2 - S^*$) behave as a function of demand correlation and demand variability.

When $N = 2$, the firm's allocation policy just needs to pick for each demand realization the customer that has the first priority (recall from Theorem 1 and the following discussion that it is sufficient to work with priority lists). Let $\hat{\mathbf{A}}$: $\mathbb{R}_+^3 \to \{1, 2\}$ be a mapping from inventory level and demand realizations (S, x_1, x_2) to a customer identity, with $\hat{\mathbf{A}}(S, x_1, x_2)$ specifying the customer who gets the first priority. As in the general formulation (§3), $\hat{\mathbf{A}}$ has to be decided before demand realizations are known; hence the customer with the first priority $\hat{\mathbf{A}}(S, X_1, X_2)$ is a priori uncertain.

There are five possibilities for demand realizations: (i) if $x_1 + x_2 \le S$, who gets priority makes no difference because both customers can be fully satisfied; (ii) if $x_1 \le S$ and $x_2 > S$, only customer 1 can be fully satisfied; (iii) if $x_1 > S$ and $x_2 \le S$, only customer 2 can be fully satisfied; (iv) if $x_1 \le S$, $x_2 \le S$ and $x_1 + x_2 > S$, inventory level is high enough to satisfy either customer individually but not both; and (v) if $x_1 > S$ and $x_2 > S$, neither customer can be fully satisfied. We assume without loss of optimality that $\hat{\mathbf{A}}(S, x_1, x_2) = 1$ when (ii) happens and $\hat{\mathbf{A}}(S, x_1, x_2) = 2$ when (iii) happens (any allocation policy that fails to satisfy these properties for some S can be improved, in the sense of increasing the service level that it delivers to either customer or both using the same inventory). Let $\hat{\Omega}$ be the set of all such mappings. Considering who gets fully satisfied in each of these possibilities, the firm's problem with two customers can be formally stated as

$$\begin{aligned}
\underset{S\in\mathbb{R}_+,\,\hat{\mathbf{A}}\in\hat{\Omega}}{\text{minimize}}\quad & S \\
\text{subject to}\quad & P\{X_1 + X_2 \le S\} + P\{X_1 \le S, X_2 > S\} \\
& \qquad + P\{\omega(S), \hat{\mathbf{A}}(S, X_1, X_2) = 1\} \ge \beta_1, \quad (\mathrm{SL}_1) \\
& P\{X_1 + X_2 \le S\} + P\{X_1 > S, X_2 \le S\} \\
& \qquad + P\{\omega(S), \hat{\mathbf{A}}(S, X_1, X_2) = 2\} \ge \beta_2, \quad (\mathrm{SL}_2)
\end{aligned}$$

where $\omega(S)$ represents the event that the firm can fully satisfy one of the customers but not both; i.e., $X_1 \le S$, $X_2 \le S$, and $X_1 + X_2 > S$. It is only when $\omega(S)$ occurs that the firm's choice of allocation policy matters.

We define a *linear knapsack policy* with parameters (k_1, k_2) and (t_1, t_2), where $k_i \ge 0$ and t_i are scalars,

to be the following procedure for allocating inventory between two customers: (1) apply the linear transformation $\tilde{x}_i = k_i x_i + t_i$ to each of the demand realizations and (2) prioritize customers in increasing order of $\tilde{x}_i$ and allocate S accordingly. The x's can be interpreted as the volume and $\tilde{x}$'s as the cost (linear in volume) of a set of items that could potentially be packed in a knapsack with a total volume S; hence the name linear knapsack. Note that the capacity of the knapsack is also a decision variable here.

To assign the first priority to customer 1 (2) if $x_1 \le S$ and $x_2 > S$ ($x_2 \le S$ and $x_1 > S$), a linear knapsack policy must have a tie for $x_1 = x_2 = S$; i.e., $k_1 S + t_1 = k_2 S + t_2$. This requires the intercepts be linked in a certain fashion: $t_2 - t_1 = S(k_1 - k_2)$. Without loss of generality, we set $k_1 = 1$, $t_1 = 0$, and $t_2 = S(1 - k_2)$. A linear knapsack policy can thus be specified more parsimoniously by one scalar, k_2, and the linear transformations $\tilde{x}_1 = x_1$ and $\tilde{x}_2 = k_2 x_2 + S(1 - k_2)$. In particular, it gives priority to customer 1 over customer 2 iff $x_1 < k_2 x_2 + S(1 - k_2)$; i.e., $\hat{\mathbf{A}}(S, x_1, x_2) = 1$ iff $x_1 < k_2 x_2 + S(1 - k_2)$.

We are now ready to state our main result concerning the two-customer problem.

Theorem 6. *With two customers, the optimal inventory level is S^*, and the linear knapsack policy with $k_1 = 1$ and $k_2 = k^*$ is an optimal allocation policy. The optimal policy parameters are the following*:

	S^*	k^*
Case 1: $\beta_1 > \alpha_1$ *and* $\beta_2 > \alpha_2$	S_0	k_0
Case 2: $\beta_1 > \alpha_1$ *and* $\beta_2 \le \alpha_2$	$F_1^{-1}(\beta_1)$	0
Case 3: $\beta_1 \le \alpha_1$ *and* $\beta_2 > \alpha_2$	$F_2^{-1}(\beta_2)$	∞

with S_0 and k_0 uniquely determined by two implicit expressions:

$$\begin{aligned} &P\{X_1 + X_2 \le S_0\} = \beta_1 + \beta_2 - 1 + P\{X_1 > S_0, X_2 > S_0\}, \\ &P\{\omega(S_0), X_1 < k_0 X_2 + S_0(1 - k_0)\} \\ &\quad = \beta_1 - P\{X_1 + X_2 \le S_0\} - P\{X_1 \le S_0, X_2 > S_0\}, \end{aligned} \tag{1}$$

and the threshold service levels, α_1 and α_2, defined as

$$\alpha_1 \equiv P\{X_1 + X_2 \le S_0\} + P\{X_1 \le S_0, X_2 > S_0\},$$
$$\alpha_2 \equiv P\{X_1 + X_2 \le S_0\} + P\{X_1 > S_0, X_2 \le S_0\}$$

Case 1 represents the mainstream situation without free riders, whereas in Cases 2 and 3, one of the customers (customers 2 and 1, respectively) is able to free ride in the sense that he is satisfied even if he never gets priority in the event of $\omega(S_0)$. These two are extreme cases, where setting the inventory level as if there were only one customer is optimal. The threshold service levels α_i, the probability that customer i faces no contest from the other customer at inventory level S_0, let us precisely specify when a customer's required service level is low enough to qualify him as a free rider. Note that the customers cannot both be free riders.

Cases 1–3 are mutually exclusive and also exhaustive for all practical purposes. There remain two other possibilities: {$\beta_1 \le \alpha_1$ and $\beta_2 \le \alpha_2$ with at least one inequality strict} cannot happen because $\beta_1 + \beta_2 = \alpha_1 + \alpha_2 + P\{\omega(S_0)\}$ by definition; {$\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2$} is a pathological case with $P\{\omega(S_0)\} = 0$, which makes all three solutions equivalent and optimal. (We ignore the latter for ease of exposition.)

Building on Theorem 6, we now establish for the general problem with any number of customers that there is always some benefit to pooling. Let $S_i = F_i^{-1}(\beta_i)$ be dedicated inventory levels in the absence of pooling.

Theorem 7. *The pooling benefit is always strictly positive; i.e., $S^* < S_1 + S_2 + \cdots + S_N$.*

The proof first uses Theorem 6 to show that the pooling benefit is always strictly positive in the two-customer case. It then rests on the following observation: Theorem 6 can be used to develop upper bounds for the general problem. Suppose that the firm pairs customers and solves the ordering and allocation problems for each pair in isolation. The sum of inventory levels obtained for pairs, plus dedicated inventories for nonpaired customers (if any), would be an upper bound on the globally optimal inventory level.

Although it is commonly known that the pooling benefit vanishes as correlation approaches +1 in newsvendor models (Eppen 1979), Theorem 7 holds for perfectly positive correlation also. The intuitive reason is that the responsive policy is able to *respond* to variations in demands efficiently. To see this in a concrete example, take two customers, let demands be perfectly positively correlated $P\{X_1 = X_2\} = 1$, and assume symmetric service-level requirements $\beta_1 = \beta_2 = \beta$. When demand is moderately high ($S/2 < X_1 \le S$), the responsive policy can satisfy one of the customers fully, whereas the comparable no-pooling policy with the total inventory S divided into two dedicated piles of size $S/2$ would not be able to fully satisfy any of the customers. So with the same amount of inventory, no-pooling always achieves less in terms of service. This symmetric two-customer example makes the argument especially transparent, but a similar dynamic drives the result in the asymmetric case also. In fact, the pooling benefit is generally larger in problem instances with asymmetry in demand distributions and/or service-level requirements. In closing, we note that there are resource flexibility and component commonality models that show

pooling benefit under perfectly positive correlation (Van Mieghem 1998, 2004). Their rationale is distinct from ours because it rests on some form of asymmetry, e.g., differences in profitability between products.

6. Anticipative Priority Policies

Anticipative priority policies ignore demand realizations when making up the priority list. This operational simplicity may come at the expense of carrying higher inventory. Anticipative policies are still worthwhile to analyze because they are often observed in industry, especially the fixed list policies.

6.1. Fixed List Policies

A fixed list policy is the simplest allocation policy to design and operate. Customers are put in a fixed priority list, and their demands are filled from the pool of inventory one after the other in the order dictated by the list until there is no more stock left or until all the demands are completely filled.

For a given set of demand distributions and service levels, each of the $N!$ distinct priority lists is associated with a distinct inventory level. The following result identifies those inventory levels and finds the optimal inventory level with its corresponding optimal fixed list policy.

THEOREM 8. *The optimal fixed list policy ranks the customers in decreasing order of their required service levels. Relabel customers such that* $\beta_1 \geq \cdots \geq \beta_N$. *Set* $\pi(k) = k$ *for all* $k \in \{1, \ldots, N\}$. *The optimal priority list is* $\Pi^*_{\text{f-list}} = (1, \ldots, N)$. *The optimal inventory level is* $S^*_{\text{f-list}} = \max\{G_1^{-1}(\beta_1), \ldots, G_N^{-1}(\beta_N)\}$, *where* G_k *is the distribution function of* $X_1 + \cdots + X_k$ *for* $k \in \{1, \ldots, N\}$.

It is surprising that the highest-service-level-first rule is optimal without any conditions on demand distributions. For instance, whether the highest-service-level customer has a low or high demand on average compared with the other customers, it is optimal to give that customer top priority in allocation. This is true even when customer demands are correlated. Hence, a fixed list policy may be the policy of choice in practice, especially when distributional information about customer demands is lacking.

Despite their popularity in practice, fixed list policies do not necessarily guarantee a positive pooling benefit. We show this by counterexample in §7; the pooling benefit can be strictly negative for optimal fixed list policies. For customer demands with multivariate normal distribution, however, the optimal fixed list policy does ensure a nonnegative pooling benefit. Recall that $S_i = F_i^{-1}(\beta_i)$.

THEOREM 9. *The optimal fixed list policy yields a positive pooling benefit, i.e.,* $S^*_{\text{f-list}} \leq S_1 + \cdots + S_N$, *if the demands* $(X_1, \ldots, X_N)$ *follow an arbitrary multivariate normal distribution with means* $(\mu_1, \ldots, \mu_N)$, *standard deviations* $(\sigma_1, \ldots, \sigma_N)$, *and correlation coefficients* $\rho_{ij} \in [0, 1]$ *between the demands of customers* i *and* j.

6.2. Randomized List Policies

A randomized list policy involves a randomization step to generate the priority list, which can be specified by a set of $N!$ positive fractional weights placed on all possible priority lists ($N!$ permutations of N customers) that sum to unity. In this section, we show how to compute the optimal randomized list policy. The case of iid demands is easier to solve, so we analyze it first and then move on to arbitrary demand distributions.

Consider iid demand random variables $X_1, \ldots, X_N$. Let the distribution function of the sum of any n of these random variables be $G_n(\cdot)$. Let the column vectors $(G_1(S), \ldots, G_N(S))^T$ and $(\beta_1, \ldots, \beta_N)^T$ be denoted by $\mathbf{C}(S)$ and $\mathbf{B}$, respectively. Let w_{ij} be the probability that customer i is assigned priority position j, and let $\mathbf{W}$ be the $N \times N$ matrix with w_{ij} in row i and column j. Note that $\mathbf{W}$ is a doubly stochastic matrix, and by Birkhoff's theorem it can be written as a convex combination of $N \times N$ permutation matrices (Marshall and Olkin 1979, p. 19). Hence, $\mathbf{W}$ constitutes a randomized list policy; the permutation matrices and the positive fractional weights summing to 1, which make up the convex combination, determine the priority list.

THEOREM 10. *Suppose the demands are iid random variables.* (a) *The optimal randomized list policy can be found by solving the following problem: minimize S subject to* $\mathbf{W} \cdot \mathbf{C}(S) \geq \mathbf{B}$, *where S and the elements of the matrix* $\mathbf{W}$ *are the decision variables.* (b) *The unique solution* S_c *of the equation* $\sum_{n=1}^{N} G_n(S) = \sum_{i=1}^{N} \beta_i$ *is a lower bound for the optimal stock.* (c) *All the service levels are exactly satisfied if and only if* $\mathbf{C}(S_c)$ *majorizes* $\mathbf{B}$. *The optimal inventory in this case is precisely* S_c.

Solving the optimization problem in (a) and the equation in (b) are both easy because $G_n(S)$ are monotone increasing in S. Further, for a fixed value of S, the mathematical program in (a) is a linear program. Also note that the solution to an equation like $\sum_{n=1}^{N} G_n(S) = \sum_{i=1}^{N} \beta_i$ can be estimated using Monte Carlo simulation software. We have found that problem instances with service levels upward of 70% almost invariably have exact solutions. Once the optimal stock has been found, finding the optimal doubly stochastic matrix, and hence the optimal allocation policy parameters, is a matter of solving linear equations.

When demands are not iid, the problem is significantly more complex. We outline a solution procedure to handle this case. Let $P(\pi_k)$ denote a discrete probability distribution over all $N!$ priority lists $\Pi_k = (\pi_k(1), \ldots, \pi_k(N))$ for $k = 1, \ldots, N!$, such that $\sum_{k=1}^{N!} P(\pi_k) = 1$.

Step 1. Compute the optimal fixed-list inventory level (from Theorem 8), which serves as an upper bound on the optimal randomized-list inventory level (by Theorem 2): $S^*_{\text{f-list}} = \max\{G_1^{-1}(\beta_1), \ldots, G_N^{-1}(\beta_N)\}$. A lower bound is $\max\{F_1^{-1}(\beta_1), \ldots, F_N^{-1}(\beta_N)\}$.

Step 2. Set $S = S^*_{\text{f-list}}$ and attempt to find a probability distribution $P(\cdot)$ over all possible priority lists so that the following inequality is satisfied for all customers $i = 1, \ldots, N$ (let $I\{\cdot\}$ denote the indicator function):

$$\sum_{j=1}^{N}\sum_{k=1}^{N!} I\{\pi_k(j) = i\}P(\pi_k)G_{\{\pi_k(1),\ldots,\pi_k(j)\}}(S) \geq \beta_i,$$

where G_M is the distribution of the sum of demands for customers who belong to set $M \subseteq \{1, \ldots, N\}$.

Step 3. Perform a binary search for the smallest feasible S between the upper and lower bounds computed in Step 1, repeating Step 2 as many times as needed and stopping when we reach an S for which the system of linear inequalities has no solution. The last feasible inventory level is optimal.

The procedure converges because we employ binary search, or interval bisection, between finite upper and lower bounds to find the optimal stock level. The bisection is guaranteed to converge because $\sum_{j=1}^{N}\sum_{k=1}^{N!} I\{\pi_k(j) = i\}P(\pi_k)G_{\{\pi_k(1),\ldots,\pi_k(j)\}}(S)$ is a continuous and monotone function of S. Although the binary search itself is logarithmic, the algorithm is exponential time—$O(c^N)$—because we need to solve for $N!$ variables in Step 2 of the algorithm.

Table 1 Optimal Inventory Level and Pooling Benefit (% Reduction in Inventory Due to Pooling) in Problem Instances with $N = 3$, iid Normal Demands, and Differentiated Service Levels

						Optimal inventory			Pooling benefit (%)		
Demand distribution	β_1 (%)	β_2 (%)	β_3 (%)	Sum (%)	No pooling	Fixed list	Randomized list	Responsive	Fixed list	Randomized list	Responsive
N (10, 2)	75.0	75.0	75.0	225	34.05	32.35	27.69	27.66	4.98	18.67	18.75
	72.5	75.0	77.5	225	34.06	32.10	27.69	27.66	5.74	18.69	18.77
	70.0	75.0	80.0	225	34.08	31.82	27.69	27.66	6.63	18.75	18.83
	67.5	75.0	82.5	225	34.13	31.60	27.69	27.66	7.40	18.86	18.93
	65.0	75.0	85.0	225	34.19	31.35	27.69	27.66	8.31	19.02	19.09
	80.0	80.0	80.0	240	35.05	32.92	29.13	29.13	6.08	16.89	16.89
	77.5	80.0	82.5	240	35.06	32.60	29.13	29.13	7.03	16.92	16.92
	75.0	80.0	85.0	240	35.11	32.35	29.13	29.13	7.85	17.02	17.02
	72.5	80.0	87.5	240	35.18	32.10	29.13	29.13	8.75	17.20	17.20
	70.0	80.0	90.0	240	35.30	31.82	29.13	29.13	9.85	17.47	17.47
	85.0	85.0	85.0	255	36.22	33.59	30.43	30.43	7.26	15.98	15.98
	82.5	85.0	87.5	255	36.24	33.25	30.43	30.43	8.26	16.04	16.04
	80.0	85.0	90.0	255	36.32	32.92	30.43	30.43	9.36	16.22	16.22
	90.0	90.0	90.0	270	37.69	34.43	31.82	31.82	8.65	15.57	15.57
	87.5	90.0	92.5	270	37.74	34.00	31.82	31.82	9.92	15.69	15.69
	85.0	90.0	95.0	270	37.93	33.59	31.82	31.82	11.43	16.10	16.10
	95.0	95.0	95.0	285	39.87	35.70	33.59	33.59	10.46	15.74	15.74
	92.5	95.0	97.5	285	40.09	35.00	33.59	33.59	12.69	16.20	16.20
N (10, 3)	75.0	75.0	75.0	225	36.07	33.50	27.21	26.62	7.13	24.56	26.21
	72.5	75.0	77.5	225	36.08	33.10	27.21	26.62	8.27	24.59	26.23
	70.0	75.0	80.0	225	36.12	32.71	27.21	26.62	9.44	24.67	26.31
	67.5	75.0	82.5	225	36.19	32.36	27.21	26.62	10.58	24.81	26.45
	65.0	75.0	85.0	225	36.29	32.01	27.21	26.62	11.79	25.02	26.65
	80.0	80.0	80.0	240	37.57	34.40	28.93	28.71	8.45	23.01	23.59
	77.5	80.0	82.5	240	37.59	33.95	28.93	28.71	9.70	23.05	23.63
	75.0	80.0	85.0	240	37.66	33.50	28.93	28.71	11.04	23.18	23.76
	72.5	80.0	87.5	240	37.77	33.10	28.93	28.71	12.36	23.40	23.98
	70.0	80.0	90.0	240	37.94	32.71	28.93	28.71	13.79	23.75	24.33
	85.0	85.0	85.0	255	39.33	35.40	30.73	30.66	9.99	21.86	22.03
	82.5	85.0	87.5	255	39.36	34.85	30.73	30.66	11.47	21.93	22.10
	80.0	85.0	90.0	255	39.48	34.40	30.73	30.66	12.86	22.16	22.33
	90.0	90.0	90.0	270	41.53	36.65	32.75	32.73	11.76	21.15	21.19
	87.5	90.0	92.5	270	41.61	36.00	32.75	32.73	13.49	21.30	21.34
	85.0	90.0	95.0	270	41.89	35.40	32.75	32.73	15.49	21.82	21.85
	95.0	95.0	95.0	285	44.80	38.55	35.39	35.39	13.96	21.01	21.01
	92.5	95.0	97.5	285	45.13	37.50	35.39	35.39	16.91	21.59	21.59

Table 2 Optimal Inventory Level and Pooling Benefit (% Reduction in Inventory Due to Pooling) in Problem Instances with $N = 3$, iid Lognormal Demands, and Differentiated Service Levels

						Optimal inventory			Pooling benefit (%)		
Demand distribution	β_1 (%)	β_2 (%)	β_3 (%)	Sum (%)	No pooling	Fixed list	Randomized list	Responsive	Fixed list	Randomized list	Responsive
Log N (10, 5)	75.0	75.0	75.0	225	36.90	34.85	26.96	24.75	5.56	26.94	32.93
	72.5	75.0	77.5	225	36.94	34.10	26.96	24.75	7.69	27.02	33.00
	70.0	75.0	80.0	225	37.07	33.40	26.96	24.75	9.90	27.27	33.23
	67.5	75.0	82.5	225	37.29	32.70	26.96	24.75	12.31	27.70	33.63
	65.0	75.0	85.0	225	37.62	32.10	26.96	24.75	14.68	28.34	34.22
	80.0	80.0	80.0	240	39.93	36.50	29.00	27.23	8.60	27.38	31.82
	77.5	80.0	82.5	240	40.00	35.65	29.00	27.23	10.87	27.50	31.93
	75.0	80.0	85.0	240	40.21	34.85	29.00	27.23	13.32	27.87	32.28
	72.5	80.0	87.5	240	40.57	34.10	29.00	27.23	15.96	28.53	32.90
	70.0	80.0	90.0	240	41.15	33.40	29.00	27.23	18.84	29.53	33.84
	85.0	85.0	85.0	255	43.78	38.60	31.42	30.02	11.84	28.23	31.44
	82.5	85.0	87.5	255	43.90	37.50	31.42	30.02	14.58	28.43	31.63
	80.0	85.0	90.0	255	44.29	36.50	31.42	30.02	17.59	29.06	32.23
	90.0	90.0	90.0	270	49.16	41.40	34.55	33.45	15.78	29.71	31.95
	87.5	90.0	92.5	270	49.44	39.90	34.55	33.45	19.30	30.12	32.34
	85.0	90.0	95.0	270	50.43	38.60	34.55	33.45	23.46	31.49	33.67
	95.0	95.0	95.0	285	58.36	45.90	39.42	38.63	21.35	32.45	33.82
	92.5	95.0	97.5	285	59.68	43.30	39.42	38.63	27.45	33.95	35.28
Log N (10, 10)	75.0	75.0	75.0	225	37.19	36.86	26.68	21.28	0.90	28.27	42.78
	72.5	75.0	77.5	225	37.29	35.40	26.68	21.28	5.07	28.46	42.93
	70.0	75.0	80.0	225	37.59	34.05	26.68	21.28	9.42	29.02	43.39
	67.5	75.0	82.5	225	38.11	32.85	26.68	21.28	13.81	29.99	44.16
	65.0	75.0	85.0	225	38.90	31.70	26.68	21.28	18.51	31.42	45.30
	80.0	80.0	80.0	240	42.75	40.25	29.91	24.59	5.84	30.03	42.47
	77.5	80.0	82.5	240	42.91	38.50	29.91	24.59	10.27	30.29	42.68
	75.0	80.0	85.0	240	43.41	36.86	29.91	24.59	15.08	31.09	43.34
	72.5	80.0	87.5	240	44.31	35.40	29.91	24.59	20.10	32.49	44.49
	70.0	80.0	90.0	240	45.74	34.05	29.91	24.59	25.56	34.61	46.24
	85.0	85.0	85.0	255	50.28	44.65	34.06	28.81	11.19	32.25	42.69
	82.5	85.0	87.5	255	50.58	42.30	34.06	28.81	16.37	32.66	43.04
	80.0	85.0	90.0	255	51.56	40.25	34.06	28.81	21.94	33.94	44.12
	90.0	90.0	90.0	270	61.66	50.95	39.96	34.72	17.37	35.19	43.69
	87.5	90.0	92.5	270	62.42	47.50	39.96	34.72	23.90	35.98	44.38
	85.0	90.0	95.0	270	65.12	44.65	39.96	34.72	31.44	38.64	46.69
	95.0	95.0	95.0	285	83.43	62.10	50.38	44.94	25.57	39.62	46.14
	92.5	95.0	97.5	285	87.41	55.50	50.38	44.94	36.50	42.36	48.59

7. Numerical Comparisons of Allocation Policies

In this section, we present numerical examples with the express purpose of making comparisons between the allocation policy classes analyzed in §§5 and 6. Tables 1 and 2 report the optimal inventory levels that correspond to the no pooling, fixed list, randomized list, and responsive policies when there are three customers with iid demands and differentiated service levels. The underlying demand distributions are normal with a CV (coefficient of variation = standard deviation/mean) of 0.2 and 0.3, and lognormal with a CV of 0.5 and 1. We vary $\bar{\beta}$ from 75% to 95% in increments of 5% and introduce higher service-level differentiation by starting with a uniform set of β's and simultaneously reducing β_1 and increasing β_3 by 2.5% at a time.

Both tables reveal an interesting insight. When demands are relatively stable with CV 0.3 or lower, there is virtually no difference between the inventory levels prescribed by the optimal responsive policy and the optimal randomized list policy. When demands are highly variable, however, the differences widen significantly, and the gulf is particularly marked for lower service levels. These observations suggest that when demands are modeled with stable distributions such as normal, it is sufficient to restrict the search to fixed and randomized list policies. However, in the case of highly unstable or long-tailed distributions, responsive policies may reap

significant inventory savings over fixed and randomized list policies.

We also observe that higher service differentiation consistently leads to a higher pooling benefit for all three policy classes. This is driven by the fact that the no-pooling inventory level is more sensitive to the maximum of the required service levels than the optimal pooling solution in any policy class. The optimal fixed list policy solution is driven by the minimum, whereas the optimal randomized list and responsive policy solutions are driven by the average of the required service levels. Especially within the latter two policy classes, the use of a single pool of inventory better absorbs the stochastic highs and lows in customer demands and allows for a response to typical rather than extreme events. Note that the no-pooling solution requires more inventory with higher differentiation; when inventory is pooled, the optimal fixed list policy requires less, whereas the optimal randomized list and responsive policies require the same amount. A caveat is in order: non-iid demands can alter these observations because in that case, how demands differ (e.g., which customer's mean demand is higher) may interact with the nature of service differentiation in unpredictable ways. Contrary to our basic finding, Deshpande et al. (2003, Table 2, p. 696) observe that higher service-level differentiation leads to a smaller inventory-pooling benefit.

Finally, note that an optimal fixed list policy can suffer from a *strictly negative* pooling benefit; it can require more inventory than no pooling. Our numerical experiments (not reported here) suggest that this happens for highly variable demands and relatively low service levels. For example, if three customers had iid lognormal demands with mean 10 and standard deviation 15, and $\beta_1 = \beta_2 = \beta_3 = 75\%$, the pooling benefit from the optimal fixed list policy would be -6.33%.

8. Concluding Remarks

This paper was motivated by our observations of industry practice in the field of supply chain planning for aftermarket service operations. Firms often have agreements with their clients containing explicit service-level clauses for delivering parts to support products. In this paper, we develop solutions that simultaneously determine the replenishment quantity and the priority rule for optimally allocating inventory to customers demanding different service levels.

We earmark three fundamental classes of allocation policies: fixed list, randomized list, and responsive. We obtain complete solutions for fixed and randomized list policies and partial solutions for responsive policies in the form of bounds and solutions for special cases. We uncover a subclass of responsive policies, called linear knapsack, that is optimal generally for the two-customer case. We show for any number of customers that the pooling benefit is always strictly positive even when demands are perfectly positively correlated. We find that when demands are independent random variables with low to moderate coefficients of variation, there is virtually no difference between the inventory levels prescribed by the optimal responsive policy and the optimal randomized list policy. However, if demands are highly variable, the differences between the optimal prescriptions of the three policy classes become significant, which is when responsive policies are most helpful.

It is immediate that our single-period solutions extend to a periodic-review infinite-horizon model with service defined as the long-run fraction of periods in which a customer's demand is fully satisfied from stock, provided lead time is zero. In that simple multiperiod scenario, each period is effectively decoupled from the next, and the optimal one-period solution can be implemented in every period without any loss of optimality. An extension to a multiperiod model appears to be much more challenging if any of the following features are incorporated into the model: significant supply or demand lead time, different frequencies for inventory replenishment and demand batching, finite planning horizon, and nonstationary allocation policies. We mark out such extensions as worthy problems for future research.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0399.

Acknowledgments

The first and third authors were partially supported by a summer research grant provided by Warrington College of Business Administration, University of Florida, Gainesville.

References

Akçay Y, Xu SH (2004) Joint inventory replenishment and component allocation optimization in an assemble-to-order system. *Management Sci.* 50(1):99–116.

Alptekinoğlu A, Tang CS (2005) A model for analyzing multichannel distribution systems. *Eur. J. Oper. Res.* 163(3): 802–824.

Arslan H, Graves SC, Roemer TA (2007) A single-product inventory model for multiple demand classes. *Management Sci.* 53(9):1486–1500.

Aviv Y, Federgruen A (2001) Design for postponement: A comprehensive characterization of its benefits under unknown demand distributions. *Oper. Res.* 49(4):578–598.

Barry C (2006) How well are you serving your customers? *Multichannel Merchant* (April 26), http://multichannelmerchant.com/opsandfulfillment/advisor/customer_ICOFR/.

Caglar D, Li C-L, Simchi-Levi D (2004) Two-echelon spare parts inventory system subject to a service constraint. *IIE Trans.* 36(7):655–666.

Chen F (1998) Echelon reorder points, installation reorder points, and the value of centralized demand information. *Management Sci.* 44(12):S221–S234.

Cohen MA, Agrawal N, Agrawal V (2006) Winning in the aftermarket. *Harvard Bus. Rev.* 85(5):129–138.

Deshpande V, Cohen MA, Donohue K (2003) A threshold inventory rationing policy for service-differentiated demand classes. *Management Sci.* 49(6):683–703.

Eppen GD (1979) Effects of centralization on expected costs in a multi-location newsboy problem. *Management Sci.* 25(5): 498–501.

Eppen GD, Schrage L (1981) Centralized ordering policies in a multi-warehouse system with leadtimes and random demand. Schwarz LB, ed. *Multi-Level Production/Inventory Systems: Theory and Practice* (North-Holland, New York), 51–67.

Erkip N, Hausman WH, Nahmias S (1990) Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Sci.* 36(3):381–392.

Gallego G, Özer Ö, Zipkin P (2007) Bounds, heuristics, and approximations for distribution systems. *Oper. Res.* 55(3):503–517.

Hopp WJ, Zhang RQ, Spearman ML (1999) An easily implementable hierarchical heuristic for a two-echelon spare parts distribution system. *IIE Trans.* 31(10):977–988.

Lee HL (2004) The triple-a supply chain. *Harvard Bus. Rev.* 82(10): 102–112.

Lee HL, Tang CS (1997) Modelling the costs and benefits of delayed product differentiation. *Management Sci.* 43(1):40–53.

Marshall AW, Olkin I (1979) *Inequalities: Theory of Majorization and Its Applications* (Academic Press, New York).

Mirchandani P, Mishra AK (2002) Component commonality: Models with product-specific service constraints. *Production Oper. Management* 11(2):199–215.

Özer Ö (2003) Replenishment strategies for distribution systems under advance demand information. *Management Sci.* 49(3):255–272.

Özer Ö, Xiong H (2008) Stock positioning and performance estimation for distribution systems with service constraints. *IIE Trans.* 40(12):1141–1157.

Schwarz LB, Deuermeyer BL, Badinelli RD (1985) Fill-rate optimization in a one-warehouse N-identical retailer distribution system. *Management Sci.* 31(4):488–498.

Silver EA, Pyke DF, Peterson R (1998) *Inventory Management and Production Planning and Scheduling* (John Wiley & Sons, Hoboken, NJ).

Swaminathan JM, Srinivasan R (1999) Managing individual customer service constraints under stochastic demand. *Oper. Res. Lett.* 24(3):115–125.

Van Mieghem JA (1998) Investment strategies for flexible resources. *Management Sci.* 44(8):1071–1078.

Van Mieghem JA (2004) Commonality strategies: Value drivers and equivalence with flexible capacity. *Management Sci.* 50(3): 419–424.

Zhang J (2003) Managing multi-customer service level requirements with a simple rationing policy. *Oper. Res. Lett.* 31(6):477–482.