Managing a Retailer's Shelf Space, Inventory, and Transportation

Gerard Cachon
1300 SH/DH, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104
cachon@wharton.upenn.edu
http://opim.wharton.upenn.edu/~cachon/

Retailers must constantly strive for excellence in operations; extremely narrow profit margins leave little room for waste and inefficiency. This article reports a retailer's challenge to balance transportation, shelf space, and inventory costs. A retailer sells multiple products with stochastic demand. Trucks are dispatched from a warehouse and arrive at a store with a constant lead time. Each truck has a finite capacity and incurs a fixed shipping cost, no matter the number of units shipped. There is a per unit shelf-space cost as well as holding and backorder penalty costs. Three policies are considered for dispatching trucks: a minimum quantity continuous review policy, a full service periodic review policy, and a minimum quantity periodic review policy. The first policy ships a truck when demand since the previous shipment equals a fixed fraction of a truck's capacity, i.e., a minimum truck utilization. The exact analysis of that policy is the same as the analysis of reorder point policies for the multiechelon problem with one-warehouse, multiple retailers, and stochastic demand. That analysis is not computationally prohibitive, but the minimum quantity level can be chosen with a simple economic order quantity (EOQ) heuristic. An extensive numerical study finds the following: Either of the two periodic review policies may have substantially higher costs than the continuous review policy, in particular when the warehouse to store lead time is short; the EOQ heuristic performs quite well; the minimum quantity policy's total cost is relatively insensitive to the chosen transportation utilization, and its total cost is close to a lower bound developed for this problem.
(*Inventory Management; Stochastic Demand; Joint Setup Cost*)

Do not envy retailers, for they have a very tough time earning profits (Guar et al. 1999). Hence, excellence in operations is critical for them. An important task for retailers is the balancing of transportation, shelf space, and inventory costs. For instance, a retailer could choose to increase transportation utilization, thereby lowering its transportation cost, but that also increases the time interval between deliveries to a store. To account for less frequent deliveries, the store will either need to expand shelf space and inventory or sacrifice customer service. This research studies the challenge of managing these interactions.

The setting considered here is a single retail store that sells multiple products with stochastic demand. There are linear inventory holding and backorder costs as well as a linear shelf-space cost. The latter reflects the expense of acquiring and maintaining a larger store as total shelf space is expanded. The store is replenished from a warehouse via trucks. Each truck has a finite capacity, C, and incurs a fixed delivery charge independent of the number of units it actually delivers. The warehouse always has enough trucks and inventory to fill the store's replenishment requests, but any shipment requires a constant transportation time from the warehouse to the store. The retailer must assign a shelf-space quantity for each of

1523-4614/01/0303/0211$05.00
1526-5498 electronic ISSN

its N products, choose a replenishment policy, and schedule truck dispatches to minimize total expected costs per unit time. Let S_i be the amount of shelf space assigned to product i, and let $S = \{S_1, \ldots, S_N\}$.

Three policies for dispatching trucks are considered. The *full service periodic review* policy, or (S, T) policy for short, reviews the inventory status of the store every T units of time and dispatches enough trucks to completely replenish the store's shelves; a base-stock policy is the replenishment policy for each product, where the shelf space is the base-stock level.[1] This is a full service policy because each product's order (i.e., its demand since the last review epoch) is always shipped. The parameter T controls the transportation cost, but it is possible that some deliveries will have a low transportation utilization: Because of the full service guarantee, a truck might be dispatched with only one unit. One desirable feature of this policy is that shelf space is minimized (for the given review interval T): Because there is no uncertainty in the supply process, each product's shelf space need only buffer the demand uncertainty during the transportation lead time.

If the retailer does not have the ability to choose a review interval, T, the retailer could use a *minimum quantity periodic review* policy, or a $(Q, S|T)$ policy for short: Every T units of time (an exogenous parameter) the retailer reviews its inventory and dispatches trucks so long as one truck has at least Q units and the other trucks are full. (One might be tempted to require that at each review epoch the average shipment is no less than Q units, but that constraint results in a more complex analysis.) Due to the Q constraint, some portion of the products' orders might not be filled, where each product's order equals the difference between its shelf space and its inventory position (on-hand inventory minus backorders plus on-route inventory). Hence, this policy requires an allocation rule to determine what portion of each order is actually shipped. This allocation rule creates a supply uncertainty that must be accounted for in the shelf-space decision, in addition to demand uncertainty. Due to the complexity of that problem, a heuristic is developed to choose policy parameters.

With a periodic review policy the store manager can plan the stock replenishment process (e.g., getting extra labor to help with unloading and shelf replenishment) because the store receives shipments on a regular schedule. The key deficiency of a periodic review policy is that it might delay some truck shipments: Once there are enough orders to fill a truck, a truck should be shipped immediately, i.e., waiting to ship a full truck only raises costs. A *minimum quantity* policy, or (Q, S) policy for short, eliminates that problem: Inventory is reviewed continuously and a truck is dispatched when Q units have been ordered, where each product's order equals its demand since the last shipment. (Note that all three truck dispatching policies are coupled with base-stock policies to govern the products' replenishment decisions, where each product's base-stock level equals its shelf space.) With this policy every shipment contains exactly Q units, and so the transportation utilization is constant across all shipments, Q/C. As with the $(Q, S|T)$ policy, this policy creates supply uncertainty, but it does so in a way that is analogous to the supply uncertainty generated by reorder point policies in a single product two-echelon supply chain with one warehouse and multiple retailers. Axsäter (1993) provides an exact analysis of that model, and so those results are incorporated into this model to provide a method to determine the optimal (Q, S) policy.

The analysis of the (Q, S) policy is computationally tractable, but a simple heuristic for choosing Q (which, recall, determines the transportation utilization) is desirable. It is shown that the cost function is the sum of a decreasing hyperbolic function (transportation cost) and an approximately increasing linear function (nontransportation costs). That is the same form as the cost function in the well-known economic order quantity (EOQ) problem. Thus, the heuristic developed is analogous to the EOQ with an adjusted holding-cost rate. Because the EOQ cost function is relatively ''flat'' about its optimum, it is reasonable to conjecture that the retailer's costs are insensitive to the chosen transportation utilization. The numerical study validates that conjecture.

[1]This policy resembles Albert Heijn's policy for dispatching trucks from its central warehouse to its stores. Albert Heijn is a large grocery retailer in the Netherlands.

In addition to the sensitivity of the cost function about the optimum, a manager would also want to know how the optimal control variables, Q and S, are impacted by changes in the model's parameters: lead time, product-line breadth, total-demand volume. For example, one could argue that the optimal transportation utilization and the lead time from the warehouse to the store are substitutes: As the lead time is decreased the retailer can take advantage of the faster deliveries by waiting to fill its trucks with additional units. Alternatively, they are complements if the retailer should take advantage of the shorter lead time by not waiting to fill its trucks with additional units so as to reduce supply uncertainty. In fact, the latter is correct. Overall, it is shown that transportation utilization is a complement to lead time and total demand volume but a substitute to the product line breadth.

Whereas it can be conjectured that the (Q, S) policy does better at managing the costs explicitly included in this model than the two periodic review policies, the numerical study provides an indication of the magnitude of the advantage. It is shown that the periodic review policies perform almost as well as the (Q, S) policy in some scenarios, but they may perform significantly worse in other scenarios. In particular, the $(Q, S|T)$ policy performs poorly if T is too large, where a reasonable benchmark for ''too large'' is T greater than the average time for total demand to equal the truck capacity. However, even if T is not too large the average performance of the $(Q, S|T)$ policy deteriorates as the warehouse-to-store lead time is reduced. The (S, T) policy also performs poorly when the warehouse-to-store lead time is short but performs reasonably well when that lead time is long.

It is possible that a policy exists that is even better than the (Q, S) policy. (The optimal policy is not known.) A lower bound is developed for this model to get a sense of how much better an optimal policy might be. The numerical study finds that the (Q, S) policy provides a cost that is not much greater than the lower bound if there is a long lead time or if the ratio of backorder penalty cost to the shell-space cost is small, i.e., if shelf space is relatively expensive. The gap between the feasible cost and the lower bound is significant in particular with a low lead time and a high backorder penalty cost, again, relative to the shelf-space cost. Nevertheless, the overall performance of the (Q, S) policy is quite good.

The remainder of the article is organized as follows: the next section provides a review of the related literature, §3 details the exact evaluation of reorder point policies and also provides the EOQ heuristic procedure, §4 evaluates the fixed interval policies, §5 describes the lower bound, §6 presents the numerical study, and §7 summarizes the conclusions.

1. Literature Review

The retailer's problem is related to the joint replenishment problem (JRP). While there are many versions of the JRP, the key features are that each item/product incurs its own fixed charge (a minor setup cost) whenever it is ordered and the system incurs a fixed charge (a major setup cost) whenever there is an order, no matter the number of items in that order or which items are in the order. The retailer does not incur an item-specific fixed charge, but the fee per truck delivery is similar to the system fixed charge. The only difference, albeit a significant one, is that the truck delivery fee is a fixed charge for a limited number of units, i.e., the capacity of the truck, whereas in the JRP the fixed charge is truly a fixed charge, i.e., there is no capacity limit. Furthermore, the JRP literature does not consider a shelf-space cost.

Several authors study the JRP with stochastic demand. Balintfy (1964) proposed can-order policies: A can-order, a must-order, and an order-up-to level are specified for each product; inventory is reviewed continuously, and an order is placed whenever there is an item with an inventory position at or below its must-order level; included in the order is any item with an inventory position at or below its can-order level, and for each of those items the order raises its inventory position to its order-up-to level. Silver (1981) and Federgruen et al. (1984) propose algorithms to choose can-order policy parameters. A possible deficiency of the can-order policy is the poor coordination across items: An item might trigger an order when there are very few other items that need a replenishment.

To improve coordination across items, Atkins and Iyogun (1988) study two periodic review replenishment policies. In the first policy, items are ordered up to a base-stock level at every review epoch. The decision parameter is the length of the time between review epochs. With the second policy, there is a set of items that are reviewed at every epoch, whereas the other items are reviewed less frequently (but still only at review epochs). The second policy is designed to account for differences in item-specific fixed charges. When there are no item-specific fixed charges, as in this model, the two policies are the same. The full service fixed interval policy in §4 is the same as Atkins and Iyogun's first policy. However, their cost evaluation is approximate, whereas this article provides an exact analysis.

Pantumsinchai (1992) develops a heuristic to choose parameters for the *QS* policy introduced by Renberg and Planche (1967): Whenever the combined inventory position of all items reaches a reorder point, an order is placed to raise the inventory position of all items to their base-stock levels. That policy is the same as the (Q, S) policy considered in this article. However, here the optimal reorder point policy is found. Viswanathan (1997) considers $P(s, S)$ policies: every T units of time each product is ordered based on an (s, S) policy, where T, s and S are chosen parameters. He shows in a numerical study that the $P(s, S)$ policies generally perform better than the other policies mentioned.

Atkins and Iyogun (1988) develop a lower bound for the JRP, which decomposes the problem into N independent problems by allocating the system fixed charge (the major setup cost) among the products. In this article, a new idea is used to develop a lower bound: Instead of allocating the fixed cost across products, demands are allocated across products. Specifically, each system demand is divided among the products proportional to their average demand rates.

Pryor et al. (1999) study the single-item inventory problem with transportation setup costs. That problem is closely related to the one considered here with the key distinction being that they concentrate on the single-item problem. In fact, under certain conditions they find an optimal policy. They also propose a heuristic policy for the two-item problem.

There is an extensive literature on the JRP with deterministic demand: e.g., Jackson et al. (1985), Anily and Federgruen (l991), Federgruen and Zheng (1992), Viswanathan and Mathur (1993), and Bramel and Simchi-Levi (1995). With deterministic demand the timing and quantity of future orders can be anticipated, so it is not clear how to compare those policies with those designed for stochastic demand.

Speranza and Ukovich (1994) consider the deterministic version of the retailer's problem: A firm manages transportation and inventory along a single link (e.g., between a warehouse and a retail store), there are multiple products, trucks have finite capacity, there are inventory holding costs, and there is a fixed cost per delivery. For that problem Blumenfeld et al. (1985) show that the EOQ model can be used to choose the delivery frequency. As Speranza and Ukovich (1994) discover, that method does not provide a good solution if the firm has a limited set of feasible delivery frequencies (e.g., it cannot ship every $\sqrt{2}$ units of time). This article demonstrates that a EOQ heuristic does provide good solutions in a stochastic demand setting, assuming no constraint is imposed on when the firm can dispatch trucks.

There is a significant literature on managing vehicle routing along with inventory costs. Much of that literature assumes deterministic demand. Federgruen and Zipkin (1984a), McGavin et al. (1993), Adelman and Kleywegt (1999), and Reiman et al. (1999), are exceptions. This article does not consider vehicle routing.

There has been some recent literature on periodic review policies in the multiechelon inventory problem with multiple retailers and stochastic demand: Cachon (1999), Chen and Samroengraja (1996, 1999), and Graves (1996). A retailer in that work is analogous to a product in this model. But those papers do not include a joint ordering/transportation cost.

Several authors study the allocation of shelf space across multiple products when customers may switch their demand among products if their preferred product is unavailable (see Mahajan and van Ryzin 1999 for a review). This article does not consider that behavior, i.e., the demand rate for each product is independent of the shelf-space allocation. Gerchak and

Wang (1994) consider a model in which the mean demand rate for a product is increasing in the product's shelf space, but in this model each product's demand rate is exogenous.

2. Model

A single retailer manages a warehouse and one retail store. Trucks are used to transport inventory from the warehouse to the retail store. Each truck has a capacity of C units and costs K per delivery, independent of the amount transported. Once a truck is dispatched from the warehouse, it arrives at the retail store in exactly L units of time. The time to load and unload a truck is ignored. Trucks may be dispatched at any time, and there is no limit on the number of trucks available.

The retailer sells N products. Demands are observed continuously. Let D_i^t be stochastic demand (in units) for product i over any interval of time of length t, and let D^t be total demand over the same interval (also in units). Let λ_i be the mean demand rate, $\lambda_i t = E[D_i^t]$, and let Λ be the total demand rate, $\Lambda = \sum_{i=1}^{N} \lambda_i$. Let $f_i(x, t)$ and $F_i(x, t)$ be the density and distribution functions of D_i^t. D_i^t has a Poisson distribution. In some retail settings the Poisson distribution is not the best representation of the demand process (see Agrawal and Smith 1994), but it does provide analytical tractability.

Product i is charged h_i per unit of inventory at the retail store per unit time. A warehouse inventory holding cost is not charged. (The warehouse probably serves multiple retail stores, but this model focuses on the cost to operate a single store.) Neither is there a holding cost for pipeline inventory, because the retailer cannot influence that cost. Product i is also charged p_i per unit backordered per unit time. It is assumed that all demands are backordered, which is doubtful for most retailers. However, introducing lost sales would render the problem computationally intractable. Further, for large p_i values the retailer will choose policies that lead to high fill rates, and thus, the behavior of this model is an approximation of the behavior in a model with lost sales.

Let $S_i \geq 0$ be the amount of shelf space the retailer allocates to product i; the retailer cannot hold any more than S_i units of product i at its store, nor can product i be stored in another product's shelf space. (For example, a grocery retailer does not want to stock cans of soup in the shelf space designated for diapers. Further, it is too costly to continuously change the products' shelf-space allocations.) One consequence of the shelf-space constraint is that the retailer cannot load onto a truck more units than can fit on the store's shelves if the truck were to arrive at the store immediately.[2]

The retailer incurs a charge of a per unit of shelf space allocated to any product. There is no constraint imposed on the total shelf-space allocation; the model is best applied before the retailer has constructed its store. (A shelf-space constraint can be accommodated, as described in §3.) Note that the shelf-space cost cannot be incorporated into the product's holding cost because, whereas the average holding cost for a product is based on that product's average inventory, the shelf-space cost is based on the product's maximum inventory position.

The retailer's objective is to choose a truck dispatching policy and a shelf-space allocation and an inventory policy to minimize total expected cost per unit time.

Some math notation: $\lfloor x \rfloor$ is the greatest integer less than or equal to x; $\lceil x \rceil$ is the smallest integer greater than or equal to x; $[x]^+ = \max\{0, x\}$; and $[x]^- = \max\{0, -x\}$.

3. The (Q, S) Policy

An intuitive policy for dispatching trucks is to ship a truck whenever the cumulative orders across the products equals a constant threshold, where an order for one unit of product i is generated with each demand for product i; i.e., products are ordered using a base-stock policy and each product's base-stock level equals its shelf space, S_i. Let Q be the truck thresh-

[2]Some stores have backrooms where any product may be stored, thereby allowing the retailer to load more units onto a truck than can be placed on the shelves. However, units in the backroom are not immediately available to customers. Hence, a backroom acts like a warehouse.

old level, where $Q \in [1, C]$: It is never optimal to ship empty trucks, which rules out $Q < 1$; nor is it optimal to delay shipping a full truck, which rules out $Q > C$. When $Q = C$, the (Q, S) policy is a ''full truck'' policy, i.e., only full trucks are dispatched.

The (Q, S) policy is simple to describe, but in practice it may be difficult to implement. The retailer would require an information system that continuously and accurately monitors inventory at the store and communicates that information to the warehouse. In addition, the warehouse must have the capability to respond to orders without the benefit of a periodic shipment schedule.

Although it may not be initially obvious, the (Q, S) policy is a subset of the reorder point policies Axsäter (1993) considers in a two-echelon inventory system. Axsäter (1993) provides a recursive algorithm to exactly evaluate expected inventory and backorder costs and to choose optimal reorder point policies. This section next links this model to Axsäter's model. For clarity, only the necessary results to evaluate costs and to find the optimal policy are explained.[3] The section concludes with a simple heuristic for choosing each product's shelf space and a heuristic for choosing Q.

In Axsäter (1993) there is one warehouse and N retailers. There is a constant lead time from the inventory source to the warehouse, L_w, and a constant lead time from the warehouse to each retailer, L_r. Demand at each retailer is Poisson. Axsäter assumes an identical demand rate across the retailers, but that assumption is not needed when the retailers implement one-for-one ordering. (Axsäter 1990 derives an exact analysis with non-identical retailers and one-for-one ordering at both echelons.) Retailers use (R_i, Q_i) reorder point policies to manage their inventory, and the warehouse inventory is managed with an (R_w, Q_w) reorder point policy.

[3]Axsäter's results do require several straightforward modifications. He assumes identical retailers (in this setting each retailer corresponds to a product), because he considers the possibility of batch ordering by the retailers. The identical retailer assumption is not necessary when the retailers use one-for-one ordering, even if the warehouse uses batch ordering. Also, Axsäter does not consider an ordering/transportation cost nor a shelf-space cost, so those costs must be included in the analysis.

To connect the models, let retailer i correspond to product i. Because a base-stock model manages each product's inventory, let $Q_r = 1$ and $R_i + 1 = S_i$. In Axsäter, each warehouse order contains Q_w units. In this model, each warehouse order is a truck that contains Q units. So, a truck in this model corresponds to a warehouse order in Axsäter, i.e., set $Q_w = Q$. In Axsäter, the warehouse orders a batch from its source, which ships all orders immediately, whenever its inventory position is R_w. In this model, inventory is immediately available to the warehouse, so the warehouse has no on-order inventory, $L_w = 0$. Further, the warehouse inventory does not incur holding costs, so the warehouse in this model has no on-hand inventory either. Hence, in this model, the warehouse's inventory position equals the absolute value of the number of units backordered. Thus, a warehouse order (a truck) is placed (dispatched) when there are Q backorders at the warehouse, which corresponds to $R_w = -Q$ in Axsäter. Trucks require time to travel to the retail store, so $L = L_r$. Note that in Axsäter, R_w is a choice parameter and Q_w is exogenous, whereas in this model both are choice parameters, but $Q = -R_w = Q_w$.

Now consider the evaluation of expected costs for an (Q, S) policy. (The following notation is different from that in Axsäter to streamline the presentation and to provide consistency with the rest of the article.) This is done by relating a unit's arrival time at the retailer with the time the unit is demanded: If the unit arrives before its demand, holding cost is incurred; whereas if the unit arrives after its demand, backorder cost is incurred. Averaging over all units yields the average cost per unit time.

A unit of product i ordered and shipped at time τ arrives at time $\tau + L$. That unit satisfies the S_ith demand to occur after time τ. Let $\hat{g}_i(S_i, L)$ be the expected holding and backorder costs incurred by a unit of product i if the unit is shipped once it is ordered,

$$\hat{g}_i(y, t) = p_i \int_0^t \gamma_i(y, x)(t - x)\, dx + h_i \int_t^\infty \gamma_i(y, x)(x - t)\, dx, \qquad (1)$$

where $\gamma_i(y, t)$ is the density function of the Erlang (λ_i, y) distribution. Note that

$$\int_0^t \gamma_i(y, x)x\,dx = \frac{y}{\lambda_i}\int_0^t \gamma_i(y+1, x)\,dx = \frac{y}{\lambda_i}(1 - F_i(y, t)),$$

since the probability that $(y + 1)$th demand after time τ occurs before time $\tau + t$ is the same as the probability that $(y + 1)$ or more demands occur over the interval $[\tau, \tau + t]$. So (1) can be written as:

$$\hat{g}_i(y, t) = \frac{1}{\lambda_i}[y(h_i + p_i)F_i(y, t) - \lambda_i t(h_i + p_i)F_i(y - 1, t) + p_i(\lambda_i t - y)].$$

Now suppose a unit of product i is ordered at time τ, but it is not shipped until time $\tau + t$, $t \geq 0$. If m demands occur at product i over the interval $[\tau, \tau + t]$, then expected costs for that unit are $\hat{g}_i(S_i - m, L)$: The unit shipped at time $\tau + t$ will arrive at the retailer at time $\tau + t + L$ and satisfy the $(S_i - m)$th subsequent demand after time $\tau + t$. Let n be the number of products ordered in the interval $(\tau, \tau + t]$ (including possibly product i), where $n \in [0, Q - 1]$. ($n = 0$ means that the unit triggers its own truck shipment, i.e., $t = 0$.) Given the set of n demands, the number of product i demands in that set is binomially distributed, where n is the number of draws and $\delta_i = \lambda_i/\Lambda$ is the probability of success. Let $Z_i(n)$ be that random variable,

$$\Pr(Z_i(n) = y) = \binom{n}{y}(\delta_i)^y(1 - \delta_i)^{n-y}.$$

Finally, n is the realization of a uniformly distributed random variable on the interval $[0, Q - 1]$: $-Q + 1 + n$ is the warehouse's inventory position, and Axsäter (1993) demonstrates that the warehouse inventory position is uniformly distributed on the interval $[-Q_w + 1, 0]$. Thus, the expected holding and backorder cost per unit of product i is

$$\frac{1}{Q}\sum_{n=0}^{Q-1}\sum_{m=0}^{n}\Pr(Z_i(n) = m)\hat{g}_i(S_i - m, L).$$

The shelf-space cost for product i occurs at rate aS_i. Transportation cost per unit is K/Q, so transportation cost is incurred at an average rate $\Lambda(K/Q)$. Overall, let $\Pi(S)$ be expected cost per unit time,

$$\Pi(Q, S) = \Lambda\frac{K}{Q} + \sum_{i=1}^{N}\left(aS_i + \frac{1}{Q}\sum_{n=0}^{Q-1}\sum_{m=0}^{n}\Pr(Z_i(n) = m) \times \lambda_i\hat{g}_i(S_i - m, L)\right). \quad (2)$$

Axsäter (1993) demonstrates that the latter term in $\Pi(Q, S)$ is convex in S_i. Because the shelf-space term is linear in S_i, $\Pi(Q, S)$ is convex in S_i. Hence, for a fixed Q it is easy to evaluate the optimal shelf space for each product. (If there is a shelf-space constraint, then a greedy algorithm finds, for a fixed Q, the optimal shelf space for each product: Start each product with zero shelf space, allocate one unit of shelf space at a time to the product that generates the greatest marginal cost reduction, stop when the shelf-space constraint is binding). Let $S_i^*(Q)$ be product i's optimal shelf space given Q. It is intuitive that $S_i^*(Q)$ is nondecreasing in Q: As truck utilization is increased, Q/C, the retailer never reduces a product's shelf space. Finding the optimal shipment quantity, Q^*, requires a search over the feasible interval, $[1, C]$: $\Pi(Q, S_1^*(Q), \ldots, S_N^*(Q))$ may not be convex in Q.

Because $\Pi(Q, S)$ is not well behaved in Q, it is not possible to definitively characterize the behavior of the optimal policy with respect to the parameters (e.g., L, Λ, N). Nevertheless, clues are available to suggest trends. Consider the relationship between L and the optimal Q. L has no impact on transportation costs, so its interaction with Q occurs with the nontransportation costs. Focus on the behavior of the $\hat{g}_i$ function. Note that

$$\hat{g}_i(y + 1, t) - \hat{g}_i(y, t) = \frac{1}{\lambda_i}[(h_i + p_i)F_i(y + 1, t) - p_i].$$

Hence, $y^*(t)$ minimizes $\hat{g}_i(y, t)$ if $y^*(t)$ is the largest integer such that $F_i(y^*(t) + 1, t) \leq p_i(h_i + p_i)$. Because $F_i(y, t)$ is stochastically increasing in t (i.e., $F_i(y, t) \leq F_i(y, t')$ for all $t < t'$), the following tends to hold for $L < L'$ and integer values of x:

$$\hat{g}_i(y^*(L) + x, L) - \hat{g}_i(y^*(L), L) > \hat{g}_i(y^*(L') + x, L') - \hat{g}_i(y^*(L'), L').$$

In words, the $\hat{g}_i$ function becomes ''flatter'' around its

minimum as L increases.[4] If the $\hat{g}_i$ function becomes flatter, then the *marginal* change in nontransportation costs with respect to Q decreases, i.e., as L increases, increasing Q generates a smaller marginal increase in the nontransportation costs. That suggests the optimal Q is increasing in L: Because an increase in L has no impact on the marginal benefit of an increase in Q (a lower transportation cost) and decreases the marginal cost of an increase in Q, the optimal Q tends to increase. The lesson for a manager is that the optimal transportation utilization (Q/C) should decrease if faster deliveries between the warehouse and the store become available.

The same argument can be applied to the total-demand parameter, Λ, and the number of products parameter, N. An increase in Λ, holding all else constant, tends to increase each product's demand rate, which makes the $\hat{g}_i$ functions flatter, i.e., it has the same qualitative impact as an increase in the lead time. Hence, an increase in Λ should lead to a higher optimal Q. On the other hand, an increase in the breadth of the product line, again holding all else constant, lowers the average product's demand rate, which makes the $\hat{g}_i$ functions steeper. Hence, an increase in N should lead to a lower optimal Q. To summarize, high volume retailers with narrow product lines and long lead times should have high transportation utilization. That hypothesis is consistent with the recommendation in Fisher (1997) that companies with innovative products (i.e., ones with high demand variation) should implement market responsive supply chains, one consequence of which includes low transportation utilization.

3.1. Two Heuristics

Although it is not computationally difficult to evaluate the optimal (Q, S) policy, it would be useful to construct simple heuristics to choose Q and S. Those heuristics could provide a retailer with a quick check on the quality of current performance, and they also provide some qualitative insights. Both heuristics are derived by replacing the stochastic variables in the cost function (2) with their mean.

If product i's demand occurred at a deterministic rate λ_i, then at time τ the expected holding and backorder costs at time $\tau + L$ is $g_i^d(x)$, where x is the inventory position at time τ,

$$g_i^d(x) = h_i[x - \lambda_i L]^+ + p_i[x - \lambda_i L]^-.$$

Note that $g_i^d(x) \leq \lambda_i \hat{g}_i(x, L)$ for all x and $g_i^d(x) \cong \lambda_i \hat{g}_i(x, L)$ for large $|x|$. Replacing $\lambda_i \hat{g}_i(x, L)$ with $g_i^d(x)$ in (2) gives

$$\Lambda\frac{K}{Q} + \sum_{i=1}^{N}\left(aS_i + \frac{1}{Q}\sum_{n=0}^{Q-1}\sum_{m=0}^{n}\Pr(Z_i(n) = m)g_i^d(S_i)\right).$$

To approximate the above, replace the binomial random variable with its mean value,

$$\Lambda\frac{K}{Q} + \sum_{i=1}^{N}\left(aS_i + \frac{1}{Q}\sum_{n=0}^{Q-1} g_i^d(S_i - \delta_i n)\right). \tag{3}$$

The above is a convex function in S_i, so the optimal S_i is not difficult to find. But working with discrete S_i values is cumbersome, so construct the continuous approximation of (3), $\tilde{\Pi}(Q, S)$,

$$\tilde{\Pi}(Q, S) = \Lambda\frac{K}{Q} + \sum_{i=1}^{N}\left(aS_i + \frac{1}{\delta_i Q}\int_{S_i - \delta_i Q}^{S_i} g_i^d(y)\right).$$

$\tilde{\Pi}(Q, S)$ is convex in S_i. Let $\tilde{S}_i(Q)$ be an optimal shelf space for product i given the cost function $\tilde{\Pi}(Q, S)$,

$$\tilde{S}_i(Q) = \begin{cases} 0 & a \geq p_i \\ \lambda_i L + \delta_i Q\left(\dfrac{p_i - a}{p_i + h_i}\right) & a < p_i. \end{cases}$$

Note that $\tilde{S}_i(Q)$ is linear in Q and greater than the mean lead time demand. ($S_i^*(Q^*)$ can be less than mean lead time demand, especially if a is large relative to p_i.)

From the envelope theorem, $\tilde{\Pi}(Q, \tilde{S}_i(Q))$ is the sum of a decreasing hyperbolic function and N increasing linear functions. That is the same structure as the well-known EOQ. Hence, $\tilde{\Pi}(Q, \tilde{S}_i(Q))$ is convex in Q. Let $\tilde{Q}$ minimize $\tilde{\Pi}(Q, \tilde{S}_i(Q))$,

$$\tilde{Q} = \min\left\{\sqrt{\frac{\Lambda K}{\sum_{i=1}^{N}\left(\frac{\delta_i}{2}\right)\left(p_i - \frac{(p_i - \min\{a, p_i\})^2}{p_i + h_i}\right)}}, C\right\}.$$

[4]That relationship does not always hold. For certain critical fractiles, $p_i/(h_i + p_i)$, it is possible that the $\hat{g}_i$ function is flatter near its minimum for smaller lead times.

The Q-heuristic sets $Q = \lfloor \tilde{Q} + 0.5 \rfloor$ and the S-heuristic sets $S_i = \lfloor \tilde{S}_i(\tilde{Q}) + 0.5 \rfloor$, i.e., both are rounded off to the nearest integer. Better rounding procedures could be developed, but that level of precision is unnecessary for this exercise, i.e., in practice the optimal (Q, S) policy should be implemented because the exact cost function is not computationally demanding.

It is well known that the EOQ cost function is quite flat about its minimum. Hence, to the extent that $\tilde{\Pi}(Q, S)$ provides a good approximation for the true cost function, it is reasonable to conjecture that $\Pi(Q, S_i^*(Q))$ is also relatively flat about its minimum, i.e., costs are relatively insensitive to the chosen transportation utilization, Q/C. The numerical study evaluates that conjecture along with the quality of the two heuristics.

4. Periodic Review Policies

With a periodic review policy the retailer reviews its inventory every T units of time. Two versions are considered. With a full service policy, an (S, T) policy, the retailer dispatches a sufficient number of trucks at a review epoch to replenish all demand since the previous review epoch. With a minimum quantity policy, a $(Q, S|T)$ policy, the retailer requires that one of the dispatched trucks has at least Q units and the remaining trucks are full. To control its transportation cost, the retailer chooses T in the (S, T) policy. In the minimum quantity policy T is exogenous, so the retailer controls the transportation cost with the parameter Q. (If T and Q were both choice parameters, then the retailer would surely choose $T = 0$, i.e., it would choose a (Q, S) policy.) With both periodic review policies, orders for product i are generated with a base-stock policy, where S_i is the base-stock level.

Expected cost with an $(Q, S|T)$ policy is evaluated in two main steps. The first evaluates the expected transportation cost, and the second evaluates the nontransportation costs. The expected cost of an (S, T) policy is the same as the expected cost of a $(Q, S|T)$ policy with $Q = 1$. As with the (Q, S) policy, a $(Q, S|T)$ policy is a "full truck" policy when $Q = C$.

Begin with more notation. Let $\overline{IP}_i$ be product i's inventory position (on-hand plus on-route inventory minus backorders) immediately before a review epoch. Let IP_i be product i's inventory position immediately after a review epoch. Define $\bar{B}_i$ to be product i's outstanding orders immediately before a review epoch, $\bar{B}_i = S_i - \overline{IP}_i$, and let $\bar{B}$ be the total number of outstanding orders at that time,

$$\bar{B} = \sum_{i=1}^{N} \bar{B}_i.$$

Note that $\bar{B}$ is the total amount of available shelf space immediately before a review epoch. Let B_i be the available shelf space for product i immediately after a review epoch, and let B be the total available shelf space; $\bar{B}_i = B_i + D_i^T$ and

$$\bar{B} = B + D^T, \tag{4}$$

where $B \in [0, Q - 1]$.

At the review epoch in consideration, the retailer will dispatch $m(\bar{B})$ trucks, where

$$m(x) = \left\lceil \frac{x - Q}{C} \right\rceil + 1.$$

The probability that at least one truck will be dispatched at the review epoch is $\Pr(\bar{B} \geq Q)$. The expected truck utilization, $\rho(Q)$, is therefore

$$\rho(Q) = \frac{1}{\Pr(\bar{B} \geq Q)} \sum_{x=Q}^{\infty} \Pr(\bar{B} = x) \times \left(m(x) - 1 + \frac{C - [m(x)C - x]^+}{C} \right) \frac{1}{m(x)}. \tag{5}$$

The expected transportation cost per unit is $K/\rho(Q)C$, and the expected transportation cost per unit time is $\Lambda K/\rho(Q)C$.

The distribution function of $\bar{B}$ is required to evaluate (5). From (4), $\bar{B}$ is a simple convolution of B and D^T, because they are independent. So, it remains to evaluate the distribution function of B.

There are three cases to consider: $Q = 1$, $1 < Q < C$, and $Q = C$. When $Q = 1$, a full service policy is implemented: All outstanding orders at a review epoch are shipped. In that case

$$\Pr(B = x) = \begin{cases} 1 & x = 0 \\ 0 & x > 0. \end{cases}$$

With a full truck policy, $Q = C$, the warehouse op-

erates as if it is using a periodic review, reorder point policy with the reorder point equal to $-C$: Every C demands triggers a truck dispatch at the subsequent review epoch. In that case, B is uniformly distributed on the interval $[0, Q - 1]$.

When $1 < Q < C$, the warehouse's outstanding orders immediately after a review epoch is a Markov process. Simulation is one technique to evaluate the distribution function of B. The analytical approximation that follows is an alternative.

Let B^j be the number of outstanding orders immediately after a review epoch when the jth previous review epoch had zero outstanding orders and all review epochs after that one had at least one outstanding order. For clarification, the 0th previous epoch is the epoch in consideration; hence, $\Pr(B^0 = 0) = 1$.

Define the state of the system to be the number of successive review epochs to have occurred in which all of the epochs had a positive number of outstanding orders; hence the system is in state 0 at time τ when the last review epoch to occur before time τ had zero outstanding orders. Let θ_j be the proportion of time in which the system is in state j. The system either transitions from state j to state $j + 1$ or the system transitions from state j to state 0. Let β_j be the probability the system transitions from state $j - 1$ to state j, $j \geq 1$. Therefore, $1 - \beta_{j+1}$ is the probability the system transitions from state j to state 0.

The β_j probabilities and the distribution functions B^j are evaluated with a system of recursive equations. Define $\bar{B}^j = B^{j-1} + D^T$: $\bar{B}^j$ is the number of outstanding orders immediately before a review epoch in which there were B^{j-1} outstanding orders immediately following the previous review epoch. It then follows that

$$\beta_j = \sum_{m=0}^{\infty} [\Pr(\bar{B}^j \leq mC + Q - 1) - \Pr(\bar{B}^j \leq mC)] \quad \text{and} \tag{6}$$

$$\Pr(B^j \leq x) = \frac{1}{\beta_j} \sum_{m=0}^{\infty} [\Pr(\bar{B}^j \leq mC + x) - \Pr(\bar{B}^j \leq mC)]. \tag{7}$$

Because $\Pr(B^0 = 0) = 1$, the recursion begins with $\bar{B}^1 = D^T$. Next, from (6), β_1 is evaluated and then, from (7), B^1 is evaluated. The remaining recursion is then apparent: β_2, B^2, β_3, B^3, etc.

There are an infinite number of states in this Markov chain, but the probability of reaching state j decreases in j. Therefore, as an approximation, suppose that the system always transitions from state M to state 0 for some large M, i.e., $\beta_{M+1} = 0$. The accuracy of the approximation increases in M, but so does the computational effort. It follows that

$$\theta_i = \begin{cases} \sum_{i=0}^{M-1} \theta_i(1 - \beta_{i+1}) + \theta_M & i = 0 \\ \beta_i \theta_{i-1} & i > 0 \end{cases}$$

and

$$\sum_{i=1}^{M} \theta_i = 1.$$

Solving that system of equations yields

$$\theta_i = \begin{cases} \dfrac{1}{1 + \sum_{j=1}^{M-1} \prod_{k=1}^{j} \beta_k} & i = 0 \\ \dfrac{\prod_{j=1}^{i} \beta_j}{1 + \sum_{j=1}^{M-1} \prod_{k=1}^{j} \beta_k} & i > 0. \end{cases}$$

Finally,

$$\Pr(B \leq x) = \sum_{i=0}^{M} \theta_i \Pr(B^i \leq x).$$

Attention is now turned to the nontransportation costs. Let $g_i(y, t)$ be expected holding and backorder costs for product i at time $\tau + t$, $t \geq L$, when the product's inventory position is y at time τ and no additional shipments will be made before time $\tau + t$:

$$g_i(y, t) = E\{h_i[y - D_i^t]^+ + p_i[y - D_i^t]^-\} = h_i(y - \lambda_i t) + (h_i + p_i) \sum_{x=y+1}^{\infty} (x - y) f_i(x, t).$$

Note that

$$\sum_{x=y+1}^{\infty} xf_i(x, t) = \sum_{x=y+1}^{\infty} \lambda_i t f_i(x - 1, t)$$
$$= \lambda_i t[1 - F_i(y - 1, t)],$$

so the above can be simplified further,

$$g_i(y, t) = (h_i + p_i)[yF_i(y, t) - \lambda_i t F_i(y - 1, t)] - p_i(y - \lambda_i t). \quad (8)$$

Let $G_i(y, T)$ be expected costs over the interval of time $[\tau + L, \tau + L + T]$ when product i's inventory position is y at time τ and that is a review epoch,

$$G_i(y, T) = \int_L^{L+T} g_i(y, t)\, dt.$$

The above is easy to evaluate for $y < 0$, since then

$$G_i(y < 0, T) = -p_i \int_L^{L+T} (y - \lambda_i t)\, dt$$
$$= p_i T\left(\lambda_i\left(L + \frac{T}{2}\right) - y\right). \quad (9)$$

To evaluate $G_i(y, T)$ for $y > 0$, first differentiate (8) with respect to t,

$$\frac{dg_i(y, t)}{dt} = -\lambda_i[g_i(y, t) - g_i(y - 1, t)],$$

where note that $dF_i(y, t)/dt = -\lambda_i f_i(y, t)$. Therefore,

$$G_i(y, T) - G_i(y - 1, T)$$
$$= \int_L^{L+T} [g_i(y, t) - g_i(y - 1, t)]\, dt$$
$$= -\frac{1}{\lambda_i} \int_L^{L+T} \frac{dg_i(y, t)}{dt}\, dt$$
$$= -\frac{1}{\lambda_i}[g_i(y, L + T) - g_i(y, L)].$$

The above immediately provides a recursive equation to evaluate $G_i(y, T)$ for $y > 0$, where (9) provides $G_i(0, T)$,

$$G_i(y) = G_i(y - 1) - \frac{g_i(y, L + T) - g_i(y, L)}{\lambda_i}.$$

Thus, expected holding and backorder costs occur at an average rate $G_i(y)/T$ over the interval.

Expected holding and backorder costs per unit time for product i occur at rate

$$\frac{1}{T}E[G_i(S_i - B_i, T)] = \frac{1}{T}\sum_{j=0}^{Q-1} \Pr(B_i = j)G_i(S_i - j, T).$$

The distribution function of B_i is required to evaluate the above. Those distributions depend on the allocation policy, which is the policy for deciding which products will be shipped and which will not be shipped. A simple allocation policy is first-come-first-serve: products are loaded into trucks in the sequence in which they are ordered. In that case B_i is binomially distributed with success probability δ_i and B draws

$$\Pr(B_i = x) = \sum_{j=x}^{Q-1} \Pr(B = j)\Pr(Z_i(j) = x). \quad (10)$$

There are probably better allocation policies than first-come-first-serve. Those policies would prioritize the products based on the demand and cost characteristics so that the ''neediest'' products would be assured priority in any shipment. Unfortunately, with those policies the analytical evaluation of B_i is very cumbersome. Thus, if a more complex allocation algorithm were used, either (10) can be taken as an approximation or B_i could be evaluated via simulation.

Now it is possible to express the expected average cost of a $(Q, S|T)$ policy, $\Pi(Q, S, T)$,

$$\Pi(Q, S, T) = \Lambda\frac{K}{\rho(Q)C} + \sum_{i=1}^{N}\left(aS_i + \frac{1}{T}E[G_i(S_i - B_i, T)]\right).$$

The first term is the expected transportation cost, the second is the shelf-space cost, and the third term is the expected holding and backorder costs. For fixed Q and T, it is straight-forward to find the optimal S because $\Pi(Q, S, T)$ is convex in each S_i. (Note that B_i is independent of S_i.) The optimal $(Q, S|T)$ policy is found via a search over the interval $Q \in [1, C]$. The optimal (S, T) policy is found via a search over the parameter T. Although an upper bound on the search interval has not been developed, it is intuitive that the search can be terminated when T is substantially

greater than C/Λ: Average total demand over an interval of length C/Λ equals C, so there will be several full trucks waiting to be shipped at each review epoch when $T \gg C/\Lambda$.

5. Lower Bound

The optimal policy is not known for this problem. Nevertheless, the policies described in the previous sections are intuitively reasonable and analytically tractable. The objective of this section is to determine how much better an optimal policy could be relative to those feasible policies. This is done by evaluating a lower bound over all feasible policies.

The retailer's problem is complex because the cost of ordering one product depends on the ordering decision of the other products: Because a truck costs K per delivery no matter the number of products delivered (up to the capacity of C units), the cost of ordering a product may be high if the order triggers another delivery, or it may be low if the order merely fills space in an already committed delivery. Of course, this complication disappears if there is only one product; a single-product retailer would be unusual. It also disappears if trucks carry at most one product; in that case, product i's ordering costs only depend on its order quantity and not on the order quantity of the other products. The latter insight is the foundation for the lower bound proposed by Atkins and Iyogun (1988): Each product i is delivered with its own truck of capacity C and incurs a delivery charge $\alpha_i K$, where $\Sigma_{i=1}^N \alpha_i = 1$. That is, their bound decomposes the problem into N independent problems, and in each of those problems the optimal policy is known. (It is an order point, order-up-to policy).

The bound developed in this section allocates demand among the products instead of the delivery cost. The basic idea is simple. Under actual operations each customer demands precisely one unit from one product. Under demand allocation each customer demands $\delta_i = \lambda_i/\Lambda$ from product i, so each customer's total demand is still one unit, $\Sigma_{i=1}^N \delta_i = 1$. Hence, under actual operations there are two components to demand uncertainty—the timing of customer arrivals and each customer's product choice—whereas with demand allocation there is only component to demand uncertainty—the timing of customer arrivals.

The next step is to show that the minimum cost under demand allocation is indeed a lower bound for cost under actual operations. Let $IP_i(\tau)$ be product i's inventory position at time τ. Assume the (unknown) optimal policy is implemented. Under that policy

$$\sum_{i=1}^{N} E[g_i(IP_i(\tau) - D_i^t, L)]$$

is the expected sum of holding and backorder costs at time $\tau + L$. Define the following cost function

$$g_i^b(y) = g_i(\lfloor y \rfloor, L) + (y - \lfloor y \rfloor)(g_i(\lceil y \rceil, L) - g_i(\lfloor y \rfloor, L)),$$

where note that $g_i(y, L) = g_i^b(y)$ for integer values of y, otherwise $g_i^b(y)$ is a weighted average of $g_i(\lfloor y \rfloor, L)$ and $g_i(\lceil y \rceil, L)$. Because $g_i^b(y)$ is convex, it follows that

$$\sum_{i=1}^{N} E[g_i(IP_i(\tau) - D_i^t, L)] \geq \sum_{i=1}^{N} E[g_i^b(IP_i(\tau) - \delta_i D^t)]. \quad (11)$$

The latter term is the expected sum of holding and backorder costs under demand allocation. So costs under demand allocation are never greater than optimal costs under actual operations; the optimal policy under demand allocation is a lower bound.

It remains to evaluate the optimal policy under demand allocation. Begin with the problem of minimizing product i's inventory and shelf-space costs under demand allocation, i.e., when each customer demands δ_i units of product i. Furthermore, impose the constraint that the average shipment quantity should be no less than q_i units, where q_i is some integer multiple of δ_i units. When $a = 0$, a reorder point policy minimizes the inventory cost subject to the shipment constraint because g_i^b is convex. Let $r_i(q_i)$ be that optimal reorder point,

$$r_i(q_i) = \min_r \frac{1}{q_i/\delta_i} \sum_{j=1}^{q_i/\delta_i} g_i^b(r + j\delta_i).$$

The problem is more complex when $a > 0$. Shelf space is charged in unit increments, but because the product's inventory position changes only in multiples of $\delta_i < 1$ units, the product's maximum inventory position under the optimal policy may be less than

its shelf space. Let $\bar{s}_i$ be the required shelf space if a $(r_i(q_i), q_i)$ reorder point policy is implemented, $\bar{s}_i = \lceil r_i(q_i) + q_i \rceil$, where unused shelf space is possible, $\bar{s}_i > r_i(q_i) + q_i$. It is never optimal to have more than $\bar{s}_i$ units of shelf space: the $(r_i(q_i), q_i)$ reorder point policy would minimize the inventory costs, so the extra shelf space would be wasted. Further, if $\bar{s}_i$ units of shelf space are assigned, then the $(r_i(q_i), q_i)$ reorder point policy is optimal: Increasing or decreasing the reorder point, while requiring $\bar{s}_i$ units of shelf space, only increases the inventory cost. If fewer than $\bar{s}_i$ units of shelf space are assigned, then the optimal policy is still a reorder point policy, but in this case the maximum inventory position must equal the shelf space.[5] Hence, let $r_i^*(q_i)$ be the optimal reorder point given q_i,

$$r_i^*(q_i) = \min_r a\lceil r + q_i \rceil + \frac{1}{q_i/\delta_i} \sum_{j=1}^{q_i/\delta_i} g_i^b(r + j\delta_i)$$

s.t.

$$r \in \{-q_i, -q_i + 1, \ldots, \bar{s_i} - 1 - q_i, \bar{s_i} - q_i\}$$

Note that $r < -q_i$ is never optimal because $g_i^b(y)$ decreases linearly for $y \leq 0$. The optimal shelf space for product i is $s_i^*(q_i) = \lceil r_i^*(q_i) + q_i \rceil$.

Now consider the problem of minimizing the sum of all costs under demand allocation. For any policy the average transportation cost per unit is K/Q^b, where Q^b is the average shipment quantity, and the average transportation cost per unit time is $(K/Q^b)\Lambda$. Given Q^b, the optimal policy must minimize the sum of the product's inventory and shelf-space costs. That is achieved with the reorder point $r_i^*(q_i)$, the shelf space $s_i^*(q_i)$, and the order quantity $q_i = \delta_i Q^b$ for each product i. With that policy a truck is dispatched every Q^b customer arrivals with a total of Q^b units on board, q_i units for product i are included on each shipment, and each truck dispatch raises each product's inventory position to $r_i^*(q_i) + q_i$, i.e., the inventory positions of the products are synchronized so that they all reach their reorder point on the same customer arrivals. Because it is never optimal to delay the shipment of a full truck, the best Q^b is found via search,

$$Q^b = \operatorname*{argmin}_{Q \in [1,C]} \Lambda\frac{K}{Q} + \sum_{i=1}^{N}\left(as_i^*(\delta_i Q) + \frac{1}{Q}\sum_{j=1}^{Q} g_i^b[r_i^*(\delta_i Q) + j\delta_i]\right).$$

The above policy is optimal under demand allocation; its expected cost is the sought after lower bound.

In the multiechelon inventory literature, several lower bounds have been developed around the idea of free-inventory rebalancing: At any moment in time, inventory can be moved instantly from one location/product to another without cost (see Federgruen and Zipkin 1984b, 1984c, 1984d; Chen and Zheng 1994). In effect, for each customer arrival the retailer is able to choose which of the N products the customer demands. Given that ability, to minimize costs the retailer selects the product that has the least cost impact. That bound could be evaluated in this system, but in some settings that bound is not as effective as the demand allocation bound. To explain, consider an extreme setting in which $L = 0$ and $p_i = p_j = p$, where p is very large. If K were sufficiently low, the retailer would assign no shelf space across all products and ship trucks with only one unit. Each product's inventory position would always equal zero (due to $L = 0$ and the single unit shipments), and so average cost per unit time would just be the average transportation cost, ΛK. Now suppose K is sufficiently large so that it may be necessary to ship more than one product per truck. Because backorders are costly, the retailer must allocate some shelf space to store product. It is expensive to allocate one unit of shelf space across all N products. Furthermore, it is unnecessary. Given the ability to choose which product each customer demands, the retailer need assign only one unit of shelf space to the product with the lowest per unit holding cost. Then, as customer arrivals occur the retailer always selects that product for its customers: Only one unit of shelf space is required to accommodate the two unit shipments. But, that strategy does not work if each customer demands a little bit

[5] A reorder point policy is optimal because (1) if the maximum inventory level were less than the shelf space, then the average inventory cost could be reduced by shifting the inventory level so that the shelf-space constraint binds, and (2) because g_i^b is convex, a reorder point policy minimizes the inventory cost subject to the constraint that the average shipment quantity is no less than q_i.

Table 1 Average Truck Utilization with Optimal (*Q, S*) Policy

	$C/\Lambda = 1$			$C/\Lambda = 4$						
	L			*L*				*N*		
K/C	0	1	4	0	1	4	Λ	4	16	32
0.25	27.4%	37.0%	43.0%	14.0%	18.2%	20.3%	4	55.7%	48.0%	47.2%
1	56.0%	71.1%	78.6%	31.3%	34.8%	39.2%	16	57.6%	54.6%	52.8%
4	91.0%	98.1%	99.5%	64.8%	69.7%	74.6%	32	57.6%	56.2%	54.6%

Table 2 Distribution of the Ratio of the Heuristic Cost to the Optimal (*Q, S*) Policy Cost

Policy	Minimum	Median	90th Percentile	95th Percentile	Maximum	Average
Q-heuristic/*S*-optimal	1.000	1.001	1.044	1.071	4.500	1.037
Q-optimal/*S*-heuristic	1.000	1.010	1.489	1.810	2.796	1.157
Q-heuristic/*S*-heuristic	1.000	1.054	1.520	1.984	4.500	1.196

(δ_i) of every product, because then each customer demand generates backorders for the $N - 1$ products that have not been allocated shelf space. Hence, with the demand allocation bound it is necessary to allocate shelf space across all products; it yields a better bound.

6. Numerical Study

This section details a numerical study that evaluates the policies developed in §§3 and 4 as well as the lower bound developed in the previous section.

From all combinations of the following sets, 972 scenarios were constructed:

$$h_i = \{1\} \qquad N = \{4, 16, 32\} \qquad C = \{\Lambda, 4\Lambda\}$$

$$p_i = \{4, 16, 32\} \qquad \Lambda = \{4, 16, 32\} \qquad K = \{C/4, C, 4C\}$$

$$a = \{1, 4\} \qquad \lambda_i = \{\Lambda/N\} \qquad L = \{0, 1, 4\}.$$

In all scenarios the products are identical (same mean demand, holding and backorder cost rates). Truck capacity is chosen relative to total system demand; when $C = \Lambda$, average total demand fills a truck each unit of time, whereas average demand takes four times longer to fill a truck when $C = 4\Lambda$. The transportation cost is defined as the minimum possible transportation cost per unit: Each unit incurs a $K/C = \{¼, 1, 4\}$ transportation cost if 100% utilization is maintained.

For each scenario, the optimal (Q, S) policy was evaluated. Table 1 presents data on truck utilization, Q/C, with the optimal (Q, S) policy. As expected, truck utilization increases sharply with the minimum transportation cost per unit K/C. As conjectured, truck utilization increases with L and Λ, and decreases as the product line becomes more fragmented (N increases). However, the impact of Λ or N is less significant than the impact of either K/C or L.

Table 2 presents data on the performance of three heuristic (Q, S) policies, where the policies differ on which parameters are chosen by heuristic. The first policy uses the Q-heuristic but chooses the optimal shelf space, i.e., $Q = \tilde{Q}$ and $S_i = S_i^*(Q)$. That policy provides excellent performance relative to the optimal (Q, S) policy: Average cost across the scenarios is only 3.7% higher than the optimal (Q, S) policy cost, and for 95% of the scenarios that policy's cost is within 7.1% of the optimal cost. The second policy uses the optimal Q and uses the S-heuristic to choose shelf space. Although median performance of this policy is reasonable (within 1% of the optimal), there are a number of scenarios in which performance is poor: Ten percent of scenarios have costs that are at least 48.9% higher than optimal. The third policy uses both heuristics. That policy yields the worst performance: The median cost increase over the optimal (Q, S) policy is 5.4%, but for 10% of the scenarios that policy's cost is more than 52% higher than optimal. To sum-

Figure 1 **Cost Function with a (*Q*, *S*) Policy**

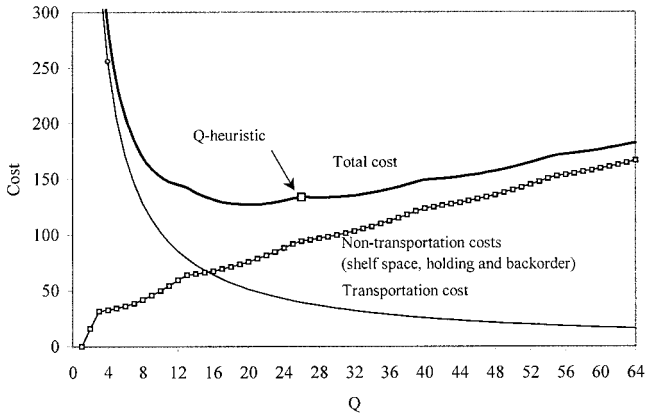
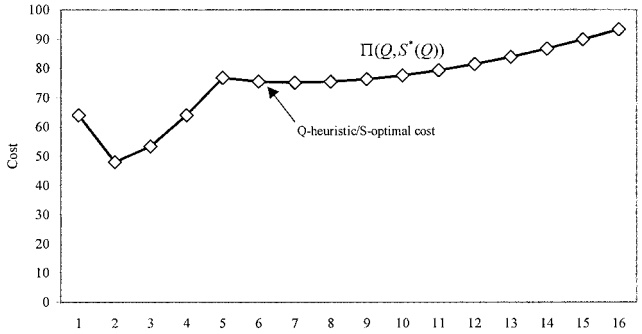


Figure 2 **A Scenario in Which the *Q*-Heuristic/*S*-Optimal Policy Makes a Poor Choice ($N = 32$, $L = 4$, $a = 1$, $p = 32$, $L = 0$, $C = 16$, $K = 16$)**



marize, the *Q*-heuristic provides a good choice for *Q*, hence, it provides retailer with an easy way to check its transportation utilization, Q/C. However, the *S*-heuristic does not provide sufficiently robust performance.

Figure 1 displays the cost function for one scenario in which the first heuristic policy (*Q*-heuristic/*S*-optimal) performs well ($N = 16$, $\Lambda = 16$, $a = 1$, $p = 32$, $L = 0$, $C = 64$, $K = 64$). It is clear from the figure why the *Q*-heuristic is effective: The nontransportation cost is an approximately linear increasing function.

Even with the first heuristic policy there are a few scenarios in which performance is significantly worse than optimal: There are six scenarios (out of 972) in which the policy's cost is more than 50% higher than the optimal cost. Table 3 indicates when the first policy performs well: long lead time and narrow product line (low N). Figure 2 graphs expected cost for a scenario in which the first heuristic policy performs poorly. It is clear that expected cost is not convex in *Q*. The heuristic makes a poor choice because it fails to recognize the benefit of operating with very low transportation utilization.

Figures 3 and 4 reveal the sensitivity of costs to the chosen *Q*. In Figure 3 two cases are considered: *Q* is set 25% above the optimal, $Q = \min\{1.25Q^*, C\}$, or *Q* is set 50% above the optimal, $Q = \min\{1.5Q^*, C\}$. Two scenarios are also considered in Figure 4: *Q* is set to 75% of the optimal, $Q = 0.75Q^*$, or *Q* is set to half of the optimal, $Q = 0.5Q^*$. In all cases *Q* is rounded to the nearest integer, and the optimal shelf space is chosen given *Q*. Scenarios are placed into 10 groups, based on their optimal policy transportation utilization. The figures display averages and maximums for each group of scenarios. Each modification of *Q* generally increases costs by less than 10%. However, there are some scenarios in which a significant penalty can occur by increasing *Q* if the optimal transportation utilization is quite low (say 25%). There may also be significant penalties for choosing *Q* too low if the optimal transportation utilization is quite high (say 95%). Nevertheless, the data indicate that costs are relatively insensitive around the optimal, Q^*, assuming the optimal shelf space is chosen with the implemented *Q*.

Table 4 provides data on the performance of $(Q, S|T)$ policies for the following values of $T/(C/\Lambda) \in$

Table 3 **The Ratio of the *Q*-Heuristic Policy Cost (with Optimal Shelf-Space Choice) to the Optimal (*Q*, *S*) Policy Cost**

	L	0	0	0	1	1	1	4	4	4
	N	4	16	32	4	16	32	4	16	32
Average		1.040	1.121	1.153	1.002	1.004	1.007	1.002	1.002	1.001
Maximum		2.375	4.500	4.500	1.022	1.036	1.064	1.011	1.009	1.013

Figure 3 Sensitivity of Costs to Increases in Q, Assuming Optimal Shelf-Space Assignment

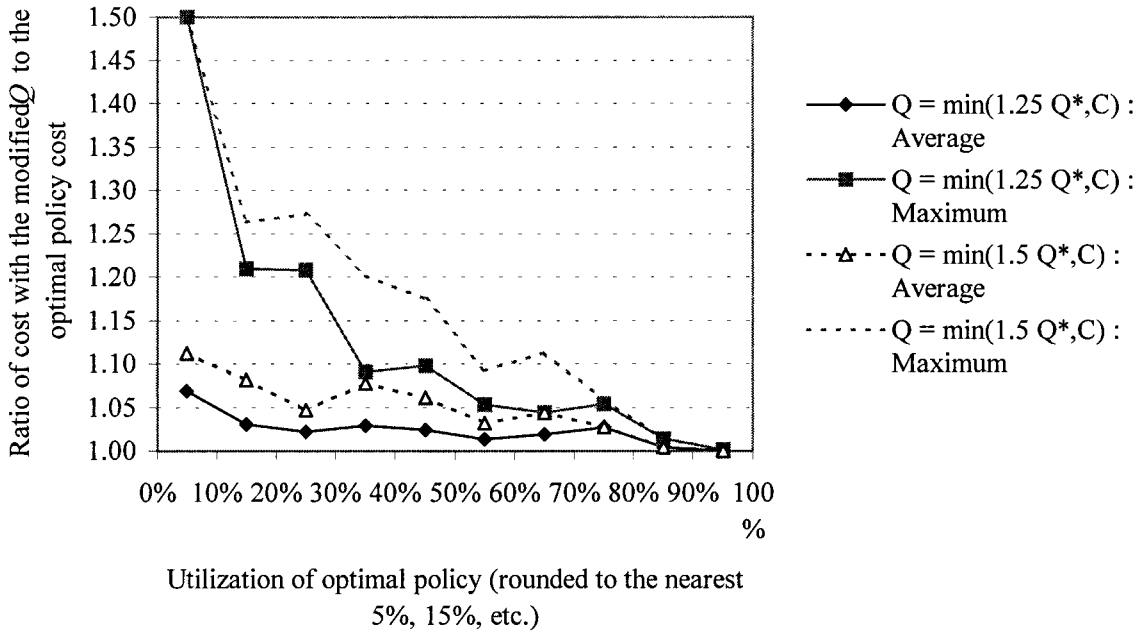


Figure 4 Sensitivity of Costs to Decreases in Q, Assuming Optimal Shelf-Space Assignment

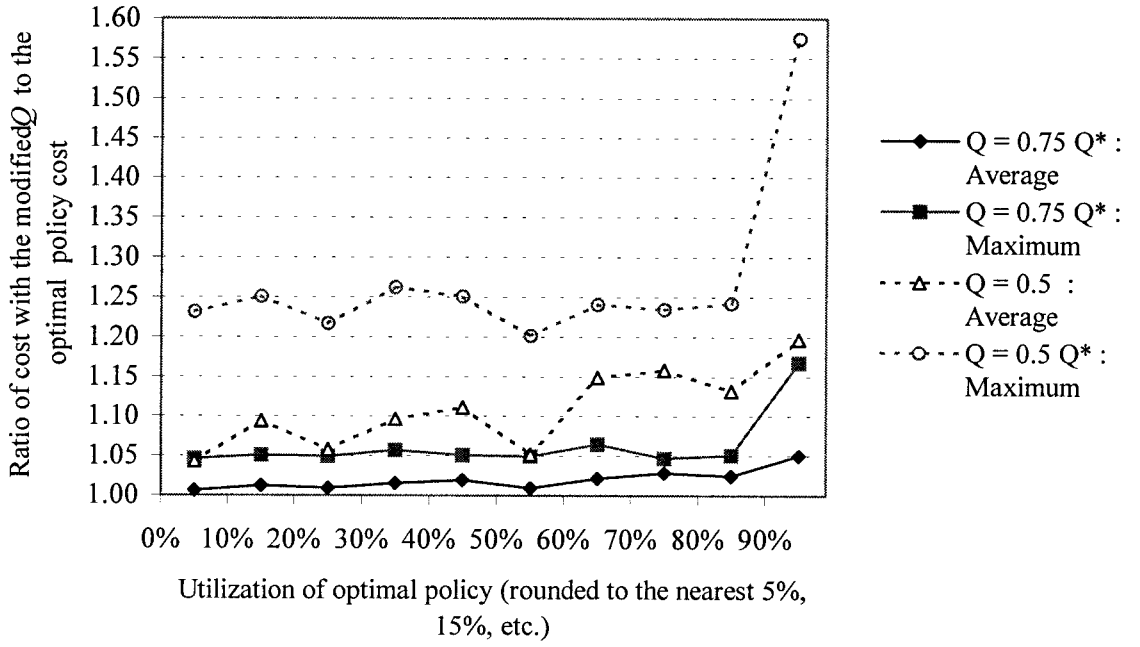


Table 4 **Ratio of $(Q, S \mid T)$ Policy Cost to Optimal (Q, S) Policy Cost**

$T/(C/\Lambda)$	Minimum	5%	Average	95%	Maximum
0.1	1.000	1.001	1.056	1.104	4.101
0.2	1.001	1.003	1.102	1.281	5.010
0.5	1.003	1.007	1.224	1.929	8.539
1	1.017	1.031	1.540	3.067	16.353
2	1.055	1.114	2.219	5.620	32.297

Table 5 **Average Ratio of $(Q, S \mid T)$ Policy Cost to Optimal (Q, S) Policy Cost**

	L		
$T/(C/\Lambda)$	0	1	4
0.1	1.152	1.011	1.004
0.2	1.267	1.029	1.010
0.5	1.553	1.088	1.031
1	2.229	1.283	1.109
2	3.562	1.768	1.327

{0.1, 0.2, 0.5, 1, 2}.[6] The T values are chosen relative to the ratio C/Λ because C/Λ is the time required for average total demand to fill a truck. The table indicates that periodic review policies may perform reasonably well for low values of T, but generally perform quite badly for large values of T. Thus, if a retailer chooses to operate with a periodic shipping interval during which mean demand approximately equals one truck load, then that retailer's cost probably could be reduced substantially if it were to switch to a continuous review shipping policy. According to Table 5, this is particularly true if the retailer has a small lead time between its warehouse and its store; when the warehouse to store lead time is short, each product's inventory cost is sensitive to deviations about the ideal inventory position, so an increase in T is particularly costly in that case because increasing T reduces the retailer's ability to keep each product's inventory position close to its ideal.

Table 6 displays data on the performance of full service periodic review policies ($Q = 1$) when the retailer can choose T. Table 7 gives the optimal period length relative to the average time to fill a truck. These policies do quite well (even in the worse case) when there are long lead times and low transportation costs. In contrast, their performance deteriorates sharply as L decreases. Hence, if the retailer has the capability to make quick deliveries between its warehouse and its store, it should be wary of operating a full service periodic review policy. However, if the retailer believes that the benefit of these policies (e.g., the operational simplicity of knowing that every order will always be filled) outweighs the cost, the retailer should generally choose a period length that is significantly smaller than the average time to fill a truck (Table 7).

Among the set of feasible policies considered, the continuous review (Q, S) policy clearly performs the best. Table 8 indicates how that policy performs relative to the best lower bound. The gap between the best feasible policy and the lower bound is quite small when the ratio p/a is small. Indeed, in about 15% of the scenarios (141 of them) the bound is tight, which means that the (Q, S) policy is in fact optimal. Nevertheless, the gap increases significantly as the p/a ratio increases. Backorder costs dominate when p/a is large, and so, the optimal (Q, S) policy will tend to assign a significant amount of shelf space to each product. It is possible that there exists a better feasible policy for managing that shelf space; however, it is also possible that the bound is poor in those scenarios. Additional research is needed to resolve that issue. (Incidentally, for all of the scenarios tested, the demand allocation bound provided a better bound than either the setup cost allocation bound or the inventory rebalancing bound.)

7. Conclusion

This research studied the management of transportation, shelf space, and inventory costs for a retailer that sells multiple products with stochastic demand. Three operating polices were compared. The continuous review, minimum quantity policy, or (Q, S) policy, performed better than the two periodic review policies. It even compared well against a lower bound developed for this model. However, the advantage of the (Q, S) policy over the periodic review policies should be tempered by the additional implementation

[6]First-come, first-serve allocation is assumed. To evaluate the B distribution function, M is chosen such that the probability of reaching state M is less than 0.000001.

Table 6 **Ratio of Full Service Policy Minimum Cost to Optimal (Q, S) Policy Minimum Cost**

	$L = 0$			$L = 1$			$L = 4$		
K/C	Min.	Average	Max.	Min.	Average	Max.	Min.	Average	Max.
0.25	1.021	1.431	4.028	1.006	1.035	1.111	1.002	1.011	1.029
1	1.010	1.152	2.012	1.007	1.040	1.132	1.003	1.016	1.065
4	1.006	1.132	1.407	1.004	1.080	1.226	1.002	1.045	1.137

Table 7 **Ratio of Full Service Policy T to C/Λ**

	$L = 0$			$L = 1$			$L = 4$		
K/C	Min.	Average	Max.	Min.	Average	Max.	Min.	Average	Max.
0.25	0.06	0.20	0.46	0.06	0.27	0.54	0.10	0.31	0.62
1	0.14	0.41	0.78	0.18	0.48	0.94	0.18	0.53	0.98
4	0.26	0.70	1.34	0.34	0.75	1.38	0.38	0.79	1.54

Table 8 **Ratio of Lower Bound Cost to Optimal (Q, S) Policy Cost**

		L		
	p/a	0	1	4
Minimum	1	1.000	1.000	1.000
	4	0.753	0.870	0.944
	8	0.668	0.878	0.956
	16	0.588	0.837	0.918
	32	0.513	0.804	0.918
Average	1	1.000	1.000	1.000
	4	0.930	0.967	0.988
	8	0.856	0.967	0.990
	16	0.789	0.940	0.977
	32	0.728	0.935	0.975
Maximum	1	1.000	1.000	1.000
	4	1.000	1.000	1.000
	8	1.000	1.000	1.000
	16	1.000	1.000	1.000
	32	1.000	1.000	1.000

challenges of operating with continuous (i.e., real time) inventory review and truck dispatching. Furthermore, the retailer should be aware that the advantage of the (Q, S) policy depends strongly on the warehouse to store lead time: The advantage is small when the lead time is long, but the advantage grows significantly as the lead time is reduced. Hence, if a retailer is able to reengineer its supply chain so that its warehouse to store lead time is decreased, the retailer will gain additional operational benefits by switching from periodic truck replenishment to continuous review truck replenishment.

Even though the model studied has stochastic demand, the behavior of the model is remarkably like the well-known deterministic demand economic order quantity (EOQ). In particular, the retailer's cost is relatively insensitive to the optimal transportation utilization: For example, if a retailer operates with a transportation utilization that is one and a half times greater than the optimal transportation utilization, then (for the scenarios tested) the retailer's total expected cost is generally no more than 10% higher than optimal. In addition, the EOQ structure leads to a simple, but very effective, heuristic for choosing the retailer's transportation utilization, i.e., the Q in the (Q, S) policy.

In addition to the connection to the EOQ model, there is a strong relationship between this model and the multiechelon inventory models with multiple retailers and stochastic demand. Indeed, the analysis of the (Q, S) policy is exactly the same as the analysis of reorder point policies in Axsäter (1993). However, although the lower bounds developed for the multiechelon inventory models could be applied to this setting, a better bound for this model was developed. That bound relaxes the constraint that each demand occurs only for one product, i.e., in the demand allocation lower bound each system demand is propor-

tionally divided among all of the products. Future research will determine if the demand allocation lower bound can improve upon the current bounds for multiechelon inventory models.

References

Adelman, D., A. Kleywegt. 1999. Price directed inventory routing. Working paper, University of Chicago, Chicago, IL.

Agrawal, N., S. A. Smith. 1994. Estimating negative binomial demand for retail inventory management with lost sales. *Naval Res. Logistics.* **43** 839–861.

Anily, S., A. Federgruen. 1991. Capacitated two-stage multi-item production/inventory model with joint setup costs. *Oper. Res.* **39** (3) 443–455.

Atkins, D., P. Iyogun. 1988. Periodic versus ''can-order'' policies for coordinated multi-item inventory systems. *Management Sci.* **34** (6) 791–796.

Axsäter, S. 1990. Simple solution procedures for a class of two-echelon inventory problems. *Oper. Res.* **38** (1) 64–69.

——. 1993. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Oper. Res.* **41** (4) 777–785.

Balintfy, J. 1964. On a basic class of multi-item inventory problems. *Management Sci.* **10** 287–297.

Blumenfeld, D., L. Burns, J. Diltz, C. Dagano. 1985. Analyzing tradeoffs between transportation, inventory and production costs on freight networks. *Transportation Res.* **19B** 361–380.

Bramel, J., D. Simchi-Levi. 1995. A location based heuristic for general routing problems. *Oper. Res.* **43** 649–660.

Cachon, G. 1999. Managing supply chain demand variability with scheduled ordering policies. *Management Sci.* **45** (6) 843–856.

Chen, F., R. Samroengraja. 1996. A staggered ordering policy for one-warehouse multi-retailer systems. forthcoming *Oper. Res.*

——, ——. 1999. Order volatility and supply chain costs. Working paper, Columbia University, New York.

——, Y-S. Zheng. 1994. Lower bounds for multi-echelon stochastic inventory systems. *Management Sci.* **40** (11) 1426–1443.

Federgruen, A., H. Groenevelt, H. Tijms. 1984. Coordinated replenishments in a multi-item inventory system with compound Poisson demands and constant lead times. *Management Sci.* **30** 344–357.

———, Y. S. Zheng. 1992. The joint replenishment problem with general joint cost structures. *Oper. Res.* **40** 384–403.

———, P. Zipkin. 1984a. A combined vehicle routing and inventory allocation problem. *Oper. Res.* **32** (5) 1019–1037.

———, ———. 1984b. Approximations of dynamic, multilocation production and inventory problems. *Management Sci.* **30** 60–84.

———, ———. 1984c. Computational issues in an infinite-horizon, multiechelon inventory model. *Oper. Res.* **32** (4) 818–836.

———, ———. 1984d. Allocation policies and cost approximation for multi-location inventory systems. *Naval Res. Logistics.* **31** 97–131.

Fisher, M. 1997. What is the right supply chain for your product? *Harvard Business Rev* (Mar–Apr) 105–116.

Gerchak, Y., Y. Wang. 1994. Periodic review inventory models with inventory-level-dependent demand. *Naval Res. Logistics.* **41** 99–116.

Graves, S. 1996. A multiechelon inventory model with fixed replenishment intervals. *Management Sci.* **42** (1) 1–18.

Guar, V., M. Fisher, A. Raman. 1999. What explains superior retail performance. Presented at the 1999 INFORMS Spring meeting in Cincinnati, OH.

Jackson, P., W. Maxwell, J. Muckstadt. 1985. The joint replenishment problem with power-of-two intervals. *IIE Transactions.* **17** 25–32.

Mahajan, S., G. van Ryzin. 1999. Retail inventories and consumer choice. *Quantitative Models for Supply Chain Management.* S. Tayur, R. Ganeshan and M. Magazine, eds. Boston, Kluwer.

McGavin, E., L. Schwarz, J. Ward. 1993. Two-interval inventory allocation policies in a one-warehouse N-identical retailer distribution system. *Management Sci.* **39** (9) 1092–1107.

Pantumsinchai, P. 1992. A comparison of three joint ordering inventory policies. *Decision Sci.* **23** 111–127.

Pryor, K., R. Kapuscinski, C. White. 1999. A single item inventory problem with multiple setup costs assigned to delivery vehicles. Working paper, University of Michigan, Ann Arbor.

Reinman, M., R. Rubio, L. Wein. 1999. Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transportation Sci.* **33** (4) 361–380.

Renberg, B., R. Planche. 1967. Un modèle pour la gestion simultanèe des n articles d'un stock. *Revue Francaise d'Informatique et de Recherche Opérationnelle.* **6** 47–59.

Silver, E. 1981. Establishing reorder points in the (S, c, s) coordinated control system under compound Poisson demand. *International J. Production Res.* **19** 743–750.

Speranza, M. G., W. Ukovich. 1994. Minimizing transportation and inventory costs for several products on a single link. *Oper. Res.* **42** (5) 879–896.

Viswanathan, S. 1997. Periodic review (s,S) policies for joint replenishment inventory systems. *Management Sci.* **43** (10) 1447–1453.

———, K. Mathur. 1997. Integrating routing and inventory decisions in one-warehouse multi-retailer multiproduct distribution systems. *Management Sci.* **43** (3) 294–312.

The consulting Senior Editor for this manuscript was Leroy B. Schwarz. This manuscript was received on July 7, 1999, and was with the author 136 days for 2 revisions. The average review cycle time was 97 days.