OPERATIONS RESEARCH
Vol. 52, No. 5, September–October 2004, pp. 707–722
ISSN 0030-364X | EISSN 1526-5463 | 04 | 5205 | 0707

informs

DOI 10.1287/opre.1040.0132

Order Volatility and Supply Chain Costs

Fangruo Chen
Graduate School of Business, Columbia University, Uris Hall, 3022 Broadway,
New York, New York 10027-6902, fc26@columbia.edu

Rungson Samroengraja
Pitney Bowes, Walter Wheeler Drive, MSC 1103, Stamford, Connecticut 06926, rungson.samroengraja@pb.com

The bullwhip effect (amplification of order variance from a downstream stage in a supply chain to an upstream stage) is widely observed in practice, and is generally considered a major cause of supply chain inefficiencies. But are supply chains always better off with strategies that are designed to dampen the bullwhip effect? This paper considers a model where a single product is sold through multiple retail outlets. The retailers replenish their inventories from a factory, which in turn replenishes its own finished-goods inventory through production. The factory's production capacity is finite, and there are transportation economies of scale in replenishing the retailer inventories. We study two types of replenishment strategies that are widely used in practice, and show that a replenishment strategy that reduces the volatility of orders received by the factory does not necessarily reduce the total costs in the supply chain.

Subject classifications: inventory/production: multiechelon, operating characteristics, stochastic, approximations/heuristics.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received February 1999; revisions received June 2002, March 2003; accepted July 2003.

1. Introduction

Interest in supply chain management has grown tremendously over the past decade. This trend has been driven by competitive pressures to improve efficiencies and enabled by advances in information technology. Several industry studies have indicated that the potential for improvement is great. For example, a study by Kurt Salmon Associates (1993) indicates that the grocery industry can reap huge benefits by employing best supply chain management practices. Another study by EHCR (1996), a consortium of North American suppliers, distributors, and providers in the healthcare industry, concludes that the healthcare supply chain can save $11 billion annually by efficient materials management and information sharing. A collective research effort is underway to identify major supply chain diseases and propose effective remedies.

The so-called bullwhip effect refers to the phenomenon that the replenishment orders placed by a supply chain member are often more volatile than the demand it faces, and this amplification of order variance propagates throughout the entire supply chain. Many explanations have been offered as to why the bullwhip effect occurs; see, e.g., Forrester (1961), Blanchard (1983), Caplin (1985), Blinder (1986), Kahn (1987), and Lee et al. (1997). It is generally believed that the bullwhip effect is a sign of bad inventory management.

This paper shows that a replenishment strategy that dampens the bullwhip effect does not necessarily reduce the supply chain costs. We consider a supply chain where a single product is sold through multiple retail outlets. The retailers replenish their inventories from a factory, which in turn replenishes its own finished-goods inventory through production. The factory's production capacity is finite. There are economies of scale in transporting finished goods from the factory to the retailers. We study two types of replenishment strategies at the retail level, the staggered policy and the (R, Q) policy. Under a staggered policy or (T, Y) policy, each retailer places an order to increase its inventory position to a base-stock level Y every T periods, a shipment can only be sent to a retailer on its order occasions, and the reorder intervals of different retailers are staggered so as to smooth the aggregate order process the factory faces. Under an (R, Q) policy, in each period, each retailer orders a minimum integer multiple of Q units from the factory to increase its inventory position to above R. These two types of policies are often used in practice when there are significant economies of scale in replenishment. While the (R, Q) policy is known to be optimal for many single-location inventory settings, the (T, Y) policy has the advantage of giving the factory a smoother demand process. On the other hand, the factory always follows a modified base-stock policy, attempting to increase its inventory position to a constant level S every period subject to a capacity constraint. The replenishment policy parameters, (T, Y, S) or (R, Q, S), are chosen to minimize the total supply chain costs. That is, the supply chain is under centralized control. (It is useful to note that the optimal policy for the two-echelon system considered here is unknown.)

For both types of policies, we develop procedures for computing the long-run average systemwide costs in the supply chain. We then use these procedures to determine the optimal control parameters within each policy type. In a

numerical study, we compared the (T, Y) policy with the (R, Q) policy. We found that the aggregate retailer order process under the optimal (T, Y) policy is often much less volatile than that under the optimal (R, Q) policy. However, in many examples, the (R, Q) policy provides lower supply chain costs. The numerical study was also used to illustrate how the relative performance of the (T, Y) and (R, Q) policies depends on several key system parameters. Finally, we study, again via numerical examples, the impact of the factory allocation policy and the value of centralized demand information in the (T, Y) system.

That (T, Y) policies dampen the bullwhip effect is not new. Lee et al. (1997) identify several causes for the increased demand volatility faced by an upstream supply chain member. One of these causes is the so-called order batching: Due to economies of scale, retailers tend to order in batches, and as a result, the factory sees a demand process that is lumpier than the demand process seen by the retailers. They consider several possible order patterns for the retailers and find that the staggered structure minimizes the upstream demand volatility.

The fact that minimizing the demand variance at the supplier does not necessarily minimize the supply chain's costs has been noted in Cachon (1999). In Cachon's model, the retailers order once every T periods, and the order size is some integer multiple of Q. The retailers' order intervals are staggered. When $T = 1$, his policy resembles the above (R, Q) policy; when $Q = 1$, his policy is essentially the above (T, Y) policy. (There are some differences at the factory/supplier level.) Cachon experiments with different values of T and Q to reach his conclusions. The key differences between the current setup and Cachon's are that we explicitly model (a) the capacity constraint at the factory (the supplier), thus giving more benefits to variance reduction at the factory; and (b) the fixed costs incurred in shipping inventories to the retailers. Moreover, our conclusion is based on the minimum supply chain costs achievable by either the (T, Y) class or the (R, Q) class of policies.

This paper has brought together two streams of research in the multiechelon inventory literature. For many multiechelon stochastic inventory systems, the optimal policy remains unknown. Consequently, research effort has been directed at various heuristic policies. Generally speaking, there are two categories of heuristic policies. In one, the policy has a fixed reorder interval but allows flexible order quantities; in the other, the policy allows flexible reorder intervals but restricts the order quantity in one way or another. For the former, see, e.g., Eppen and Schrage (1981), Federgruen and Zipkin (1984), Jackson (1988), Graves (1996), Aviv and Federgruen (1997), and Chen and Samroengraja (2000). For the latter, see, e.g., Deuermeyer and Schwarz (1981), De Bodt and Graves (1985), Svoronos and Zipkin (1988), Axsäter (1993a), Chen and Zheng (1994, 1997), and Cachon (2001). This paper provides a plausible setting so that one can compare the policies in the above two categories. The multiechelon inventory literature is reviewed in Axsäter (1993b) and Federgruen (1993). For more recent developments of supply chain models, see Tayur et al. (1998).

The rest of this paper is organized as follows. Section 2 describes the model and the two types of replenishment policies. Section 3 provides an exact (respectively approximate) procedure for evaluating the long-run average systemwide costs under the (T, Y) (respectively (R, Q)) policy. Section 4 reports the numerical comparisons between the two policies. Section 5 considers a different allocation policy and the use of centralized demand information in the (T, Y) system and reports related numerical results. Section 6 concludes.

2. Preliminaries

Consider a two-echelon production/distribution system. A factory produces a single product and distributes it to N retailers. Customer demand occurs at each retailer according to a simple Poisson process. Unsatisfied demands are completely backlogged. Although demand can occur at any time, production and distribution decisions are made periodically. The quantity that the factory can produce in a period is limited by a capacity constraint. Production lead time is constant. Each order by a retailer incurs a fixed cost and the transportation lead time from the factory to the retailer is constant. (Lateral transshipment between retailers is not allowed.) Holding costs are assessed for inventories in the system, which include on-hand inventory at the factory, inventories in transit to the retailers, and on-hand inventories at the retailers. Penalty costs are assessed for customer backorders at the retail level. The retailers are assumed to be identical: The demand processes at the retailers are independent with the same average arrival rate, the factory-to-retailer lead times are identical across retailers, and the holding and penalty cost parameters at the retailers are not retailer specific. The objective is to minimize the long-run average systemwide costs. We consider two classes of replenishment policies.

The first class of policies uses a (T, Y) policy at the retail level in a staggered manner. Each retailer orders every T periods. T is restricted to be either a multiple or a divisor of N. If $T \geqslant N$, then the time between consecutive retailer orders is given by T/N. Consider $N = 4$ and $T = 8$; each retailer orders every eight periods, and the retailers are staggered so that the factory receives a retailer order once every two periods. If $T < N$, then the retailers are divided into T groups, and each group has N/T retailers. Each group orders every T periods, and the factory receives a group order every period. For example, with $N = 6$ and $T = 2$, there are two groups of three retailers each, and each group orders every two periods, with the factory receiving a group order every period. Each retailer follows a base-stock policy with order-up-to level Y, based on its nominal inventory position (orders placed but not yet received plus on-hand inventory minus customer backorders). The retailer orders

are filled at the factory according to an allocation policy to be defined later. A shipment can only be sent to a retailer on its order occasions. Each period the factory attempts to increase its inventory position (work-in-process inventory plus on-hand inventory minus unsatisfied retailer orders) to a base-stock level S ($\geqslant 0$), subject to a constraint that the production quantity does not exceed C units. (This modified base-stock policy is optimal in some single-location settings; see Federgruen and Zipkin 1986a, b.) Therefore, the above policy has three control parameters (T, Y, S); for brevity, we will refer to it as a (T, Y) policy.

The second class of policies uses an (R, Q) policy at the retail level. Each retailer follows a periodic-review (R, Q) policy based on its nominal inventory position (orders placed but not yet received plus on-hand inventory minus customer backorders). In each period, if a retailer's nominal inventory position is equal to or less than R, it orders a minimum integer multiple of Q from the factory to increase the nominal inventory position to above R. (This type of policy is often called an (R, nQ) policy, with n standing for the number of batches, each of size Q, that are included in an order.) An allocation policy, to be defined later, specifies how the factory inventory is used to satisfy retailer orders. Shipments from the factory to the retailers are made periodically. As in the (T, Y) system, the production decisions at the factory are made periodically according to a base-stock policy with order-up-to level S ($\geqslant 0$) subject to capacity C. Therefore, the above policy has three control parameters (R, Q, S) and, for brevity, will be referred to as an (R, Q) policy.

To help describe the material flow in the supply chain, imagine that there are $N + 1$ bins numbered $0, \ldots, N$ in the factory. Bin 0 holds the available on-hand inventory of the factory, while the remaining bins are used to hold inventories committed to retailers $1, \ldots, N$, i.e., inventories that have been allocated but not yet shipped. Generally, whenever the factory receives an order from retailer n, it attempts to fill this order by moving inventory from bin 0 to bin n. Inventory, once moved to bins $1, \ldots, N$, may not be reallocated.

To describe the material flow under the (T, Y) system, first consider the special case with $T = N$. In this case, the factory receives an order from one retailer every period. Upon receipt of an order, say from retailer n, the factory attempts to fill this order as much as possible by transferring inventory from bin 0 to bin n. In case bin 0 has insufficient inventory, the factory creates an outstanding order for retailer n for the unfilled amount. When a replenishment batch becomes available at the factory, it is transferred to bin 0 and then used to fill any outstanding orders for the retailers on a first-come, first-served (FCFS) basis. The inventory used to fill the outstanding orders for retailer j is placed in bin j, $j = 1, \ldots, N$. The amount shipped to retailer n on its order occasion is the inventory in bin n after the factory completes the inventory allocation for the period. Now consider the case where $T < N$. Recall that in this case the factory receives orders from N/T retailers every period. If the inventory in bin 0 is enough to satisfy these orders, transfers are made to the bins for the retailers. Otherwise, inventory is allocated on an FCFS basis by sequencing the units being ordered according to the arrival times of the corresponding demands. (Because each retailer follows a base-stock policy, every unit in an order is triggered by a demand, the corresponding demand for the unit. Therefore, the allocation policy requires the factory to have access to the point-of-sale data at the retail level.) An outstanding order is created for the remaining units, with their sequence retained for later allocation. When a replenishment batch becomes available at the factory, it is used to satisfy the outstanding orders on an FCFS basis (by the sequence in which these outstanding orders were created). In case an outstanding order cannot be filled completely, we allocate according to the retained sequence of the individual units. Finally, if $T > N$, the material flow is essentially the same as when $T = N$. The only difference is that the factory receives an order from one retailer every T/N periods. (An alternative allocation policy will be considered in §5.)

The material flow in the (R, Q) system is simpler. Whenever a replenishment batch becomes available at the factory, it goes directly to bin 0. Because the retailers order at the same time, an allocation policy is required in case the factory has insufficient inventory to satisfy all retailer orders. To specify the allocation policy, imagine that the retailers follow a continuous-review (R, Q) policy; i.e., each retailer places an order for Q units as soon as its nominal inventory position falls to R. (This is possible because each customer demands a single unit, i.e., no batch demands.) It is clear that the total number of Qs ordered by a retailer in a period under the continuous-review scenario is the same as the integer multiple of Qs ordered by the same retailer under the period-review case. The advantage of the continuous-review scenario is that now no two retailers will order at the same time, and the factory will be able to satisfy the retailer orders on an FCFS basis. This is the allocation policy for the (R, Q) system. (To implement this allocation policy in the periodic-review scenario, simply associate each sub-batch Q in a retailer's order with a time index that is the arrival time of the demand unit that "triggers" the order for the sub-batch. Allocation is then based on this time index on an FCFS basis.) As before, the factory fills the retailer orders by transferring inventory from bin 0 to the retailer bins, but the inventory transfers must now be in integer multiples of Q. (This integer-multiples assumption is primarily for technical convenience.) If the factory is unable to fill a retailer order completely, the remainder becomes an outstanding order. At the beginning of each period, the contents in bins $1, \ldots, N$ are shipped to the retailers.

Under the (T, Y) system, the following events occur sequentially at the beginning of each period unless otherwise stated.

1. The designated retailers place an order. When $T \geqslant N$, there is at most one designated retailer; otherwise, there are N/T designated retailers.

2. At the factory, the replenishment batch due this period becomes available and goes to bin 0. The outstanding orders for the retailers, if any, are filled according to the allocation scheme described above. An outstanding order, or any portion thereof, once filled, is no longer considered outstanding even if it is not yet shipped.

3. The factory fills the current order from the designated retailers, if any, as much as possible according to the allocation scheme described above. If bin 0 has insufficient inventory, the factory creates an outstanding order for the unfilled portion.

4. The factory reviews its inventory position and places a replenishment order, if necessary.

5. The contents in the bins for the designated retailers are shipped.

6. The shipments due to arrive at the retailers this period are received. Outstanding customer backorders, if any, are satisfied. (Demand arrives throughout the period at the retailers and is satisfied from their on-hand inventories, or otherwise backlogged.)

Under the (R, Q) system, the following events occur sequentially at the beginning of each period unless otherwise stated.

1. Each retailer reviews its nominal inventory position and places an order (for an integer multiple of Q).

2. At the factory, the replenishment batch due this period becomes available and goes to bin 0. The inventory in bin 0 is then used to fill outstanding orders according to the allocation policy described above. This is accomplished by transferring inventory to the retailer bins.

3. The factory reviews its inventory position and places an order, if necessary.

4. The contents in the retailer bins are shipped.

5. The shipments due to arrive at the retailers this period are received. Outstanding customer backorders, if any, are satisfied. (Demand arrives at the retailers throughout the period and is satisfied from their on-hand inventories, or otherwise backlogged.)

The system parameters are:

$\lambda =$ average demand arrival rate at a retailer.
$C =$ production capacity per period at the factory.
$L =$ production lead time at the factory, a nonnegative integer representing a number of periods.
$H =$ holding cost per unit per period at the factory.
$K =$ cost per retailer order, regardless of order size.
$l =$ transportation lead time from the factory to a retailer, a nonnegative integer representing a number of periods.
$h =$ echelon holding cost per unit per period at the retailers, $h \geqslant 0$ (thus, the installation holding cost at the retailers is $h + H$).
$p =$ backorder penalty cost per unit per period at the retailers.

We assume that $C > N\lambda$ for the system to be stable.

To describe the inventory state of the system, we define for any time t:

$IP_0(t) =$ factory inventory position
$=$ inventory on hand at the factory (i.e., in bin 0) plus work in progress minus outstanding orders for the retailers.
$X(t) =$ factory shortfall
$= S - IP_0(t)$.
$I_0(t) =$ inventory on hand at the factory (i.e., in bin 0).
$B_0(t) =$ total outstanding orders for the retailers at the factory.
$IL_0(t) =$ factory inventory level
$= I_0(t) - B_0(t)$.
$B_0^n(t) =$ total outstanding orders for retailer n at the factory.
$IC_n(t) =$ inventory committed to retailer n at the factory (i.e., inventory in bin n).
$IT_n(t) =$ inventory in transit to retailer n.
$IP_n(t) =$ inventory position at retailer n
$=$ inventory on hand at retailer n plus inventory in transit to retailer n minus customer backorders at retailer n.
$NIP_n(t) =$ nominal inventory position at retailer n
$=$ outstanding orders for retailer n at the factory plus inventory committed to retailer n at the factory (i.e., in bin n) plus the inventory position at retailer n.
$I_n(t) =$ inventory on hand at retailer n.
$B_n(t) =$ customer backorders at retailer n.
$IL_n(t) =$ inventory level at retailer n
$= I_n(t) - B_n(t)$.

Note that $IP_0(t)$ and $IL_0(t)$ do not include the inventories in the retailer bins. In other words, for the purpose of determining the factory inventory position/level, one can imagine that the retailer bins are "outside" of the factory.

For convenience, the time index t, if it is an integer, marks the beginning of period t after all the replenishment events for that period have occurred, but before demand. We write t^-, with an integer-valued t, for the beginning of period t before all the events.

3. Cost Evaluation

The objective of this section is to compute the long-run average systemwide costs for both (T, Y) and (R, Q) policies. We begin with an accounting scheme for assessing the holding and backorder costs in the system.

For any time t, determine the system on-hand inventory and charge H for each unit. The system on-hand inventory consists of the on-hand and committed inventories at the factory (i.e., the contents in bins $0, \ldots, N$), inventories in transit to the retailers, and inventories on hand at the retailers. (It is easy to see that the long-run average holding cost associated with the inventories in transit to the retailers

is constant and is independent of the control parameters. The inclusion of this cost component, while not essential for determining the optimal control parameters, simplifies presentation.) Second, charge h for each unit of on-hand inventory at the retailers. Finally, charge p for each unit of customer backorder at the retailers. Therefore, the rate at which the total holding and backorder costs accrue at time t is given by

$$H\left[I_0(t)+\sum_{n=1}^{N}(IC_n(t)+IT_n(t)+I_n(t))\right] + h\sum_{n=1}^{N}I_n(t)+p\sum_{n=1}^{N}B_n(t).$$

Subtracting and adding $H\sum_{n=1}^{N}B_0^n(t)$, which is equal to $HB_0(t)$ by definition, to the above expression, we have

$$H[I_0(t)-B_0(t)]+H\sum_{n=1}^{N}[B_0^n(t)+IC_n(t)+IT_n(t)+I_n(t)] +h\sum_{n=1}^{N}I_n(t)+p\sum_{n=1}^{N}B_n(t),$$

which, after subtracting and adding $H\sum_{n=1}^{N}B_n(t)$, becomes

$$HIL_0(t)+H\sum_{n=1}^{N}NIP_n(t)+\sum_{n=1}^{N}[hI_n(t)+(p+H)B_n(t)]. \quad (1)$$

Below, we determine the long-run average values of these three components for each policy type.

3.1. Exact Evaluation for the (T, Y) Policy

We begin by determining the long-run average value of $HIL_0(t)$. For any $t_1 \leqslant t_2$, let $Z(t_1, t_2]$ be the total orders received by the factory in the time interval $(t_1, t_2]$. Because $IL_0(t+L)=IP_0(t)-Z(t, t+L]$, and by definition $IP_0(t)=S-X(t)$ for any t, it suffices to determine the long-run average values of $Z(t, t+L]$ and $X(t)$.

First, consider the long-run average value of $Z(t, t+L]$. Because the expected customer demand in the system in a period is $N\lambda$, the long-run average value of $Z(t, t+L]$ is equal to the expected system demand over L periods, which is $NL\lambda$. This is due to flow conservation. To determine the long-run average value of $X(t)$, consider two cases.

Case 1: $T \leqslant N$. Let X be $X(t)$ in steady state. The distribution of X can be determined by solving

$$X=\max\{X+Z-C, 0\}, \quad (2)$$

where Z represents the size of the total retailer orders received in a period. The steady-state distribution of the shortfall has been characterized by Tayur (1992), Glasserman and Tayur (1996), Glasserman (1997), and Roundy and Muckstadt (1997), and it is easy to compute. Therefore, the long-run average value of $HIL_0(t)$ is

$$H(S-E[X]-NL\lambda), \quad T \leqslant N. \quad (3)$$

Case 2: $T > N$. To determine the long-run average value of $X(t)$ in this case, note that the factory receives an order from a retailer every T/N periods, and each order is a Poisson random variable with mean $T\lambda$. Define an order cycle to be the time interval between successive retailer orders. Thus, the length of an order cycle is T/N periods. Let t be the beginning of an order cycle. Let $W(t)$ be the shortfall at t^-; i.e., $X(t^-)$, whose steady-state distribution can be obtained by solving $W=\max\{W+Z-T/N\cdot C, 0\}$, where Z is a Poisson random variable with mean $T\lambda$ and $T/N\cdot C$ is the total production capacity over an order cycle. This equation is essentially the same as (2) and thus can be easily solved. We say that period $t+k$ is of type k, $k=0, \ldots, T/N-1$. Let X_k be the shortfall in a type-k period in steady state. Note that

$$X_k=\max\{W+Z-(k+1)C, 0\}, \quad k=0, \ldots, T/N-2$$

and $X_{T/N-1}=W$. Consequently, the long-run average value of $HIL_0(t)$ is

$$H\left(S-\frac{\sum_{k=0}^{T/N-1}E[X_k]}{T/N}-NL\lambda\right), \quad T > N. \quad (4)$$

We proceed to consider the long-run average value of $NIP_n(t)$ for any retailer n. Suppose retailer n places an order at time 0. Thus, $NIP_n(0)=Y$. Now take any $t < T$. Note that $NIP_n(t)$ is equal to Y minus the customer demands in the interval $(0, t]$. Thus, $E[NIP_n(t)]=Y-\lambda t$. Consequently, the long-run average value of $NIP_n(t)$ is

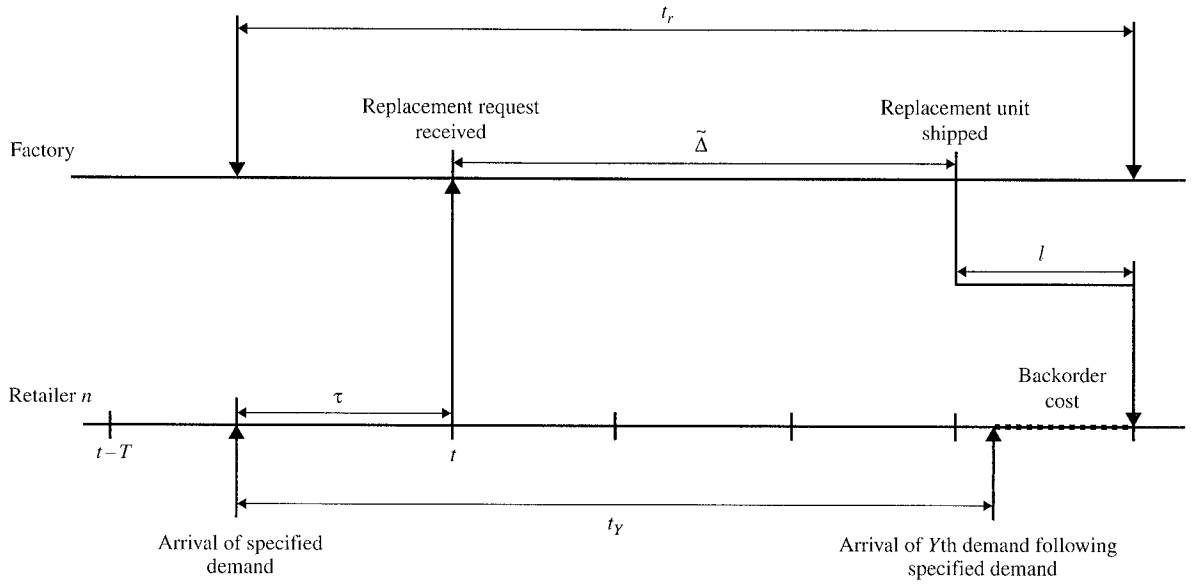
$$\frac{1}{T}\int_0^T(Y-\lambda t)\,dt=Y-\frac{T\lambda}{2}.$$

Summing across all the retailers, we have the long-run average value of $H\sum_{n=1}^{N}NIP_n(t)$ as

$$HN\left(Y-\frac{T\lambda}{2}\right). \quad (5)$$

To determine the long-run average holding and backorder costs at the retailers, we use an approach developed by Axsäter (1990) that determines the long-run average holding and backorder costs associated with an arbitrary unit of customer demand. The idea is to "follow" the sequence of events generated by the arrival of an arbitrary customer demand. This demand triggers a later request (by a retailer) for a replacement unit. This replacement unit will be used to satisfy a future demand. We determine the time that this replacement unit will arrive at the retailer that ordered it as well as the arrival time of the future demand that this unit satisfies. If the replacement unit arrives ahead of the future demand, a holding cost is incurred; otherwise, a backorder cost is incurred. This cost is said to be associated with the demand that triggered the replacement unit. Multiplying the expected holding and backorder cost incurred by an arbitrary demand by the total demand rate at all

Figure 1. Diagram of events.



retailers yields the long-run average holding and backorder costs at the retailers. We now describe this procedure in detail.

Take an arbitrary demand. This demand, hereafter referred to as the *specified demand*, occurs at, say, retailer n. It arrives τ units of time before the next order occasion by the retailer, which happens at time t. Recall that when $T < N$, there are additional retailers ordering at t. Let i be the number of demands that have occurred at *all* the retailers that order at t since their last order occasion (at time $t - T$), but before the arrival of the specified demand. Figure 1 illustrates the timeline of the events generated by the specified demand.

The specified demand triggers a request for a replacement unit for retailer n, hereafter referred to as the *replacement request*. To characterize the arrival time of the replacement unit at the retailer, note that at time t an order containing the request is transmitted to the factory. Because all retailers follow a base-stock policy, the number of units ordered is equal to the number of demands that have occurred since the last order occasion. Based on the allocation policy presented in the previous section, the factory fills retailer orders in the following sequence: (1) all the retailer orders before t, (2) all the requests triggered by the i units of demand that preceded the specified demand, and (3) the replacement request. If the factory is able to fill all these orders/requests by time t, then the replacement unit is shipped immediately. Otherwise, the replacement unit will be shipped at the next order occasion by retailer n or even later. The time from t to the shipment of the replacement unit is the delay, $\tilde{\Delta}$. Once shipped, it takes an additional l periods for the replacement unit to reach the retailer. Thus, the total time from the arrival of the specified demand to the arrival of the replacement unit at the retailer that ordered it is $\tau + \tilde{\Delta} + l \stackrel{\text{def}}{=} t_r$.

The future demand that the replacement unit is used to satisfy is the Yth demand at retailer n following the specified demand. Let t_Y be the time from the specified demand to the Yth demand. If the Yth demand occurs before the replacement unit arrives, then backorder costs are incurred. Otherwise, holding costs are incurred. The holding and backorder costs associated with the specified demand can be expressed as

$$h(t_Y - t_r)^+ + (p+H)(t_Y - t_r)^-.$$

Note that t_Y is an Erlang random variable with parameters Y and λ because the demand process at retailer n is Poisson with rate λ. Let $g^Y(t)$ be the density function of t_Y; i.e.,

$$g^Y(t) = \frac{\lambda^Y t^{Y-1} e^{-\lambda t}}{(Y-1)!}, \quad t \geqslant 0.$$

Let $c^Y(t_r)$ be the expected holding and backorder costs associated with the specified demand conditioned on the replacement unit arrival time t_r. Because t_Y is independent of t_r (which is determined by the demand processes up to time $t - \tau$ at the retailers ordering at time t as well as the demand processes at the other retailers), we have

$$c^Y(t_r) = (p+H)\int_0^{t_r} g^Y(s)(t_r - s)\,ds + h\int_{t_r}^{\infty} g^Y(s)(s - t_r)\,ds, \quad Y \geqslant 1,$$

$$c^0(t_r) = (p+H)(t_r).$$

Following Axsäter (1990), we can simplify the above expression:

$$c^Y(t_r) = e^{-\lambda t_r}\frac{p+H+h}{\lambda} \cdot \sum_{k=0}^{Y-1}\frac{(Y-k)}{k!}t_r^k\lambda^k + (p+H)(t_r - Y/\lambda),$$

where $0! = 1$. It remains to determine the distribution of t_r.

We begin with the distribution of $\tilde{\Delta}$ given τ and i. Recall that $\tilde{\Delta}$ is the time from when the request for the replacement unit is received by the factory until the replacement unit is shipped. Because shipments to a retailer are only made every T periods, it is possible that a replacement unit is filled, but not yet shipped. Define Δ to be the time from when the request for the replacement unit is received at the factory (at time t) until it is filled. The distribution of Δ uniquely determines the distribution of $\tilde{\Delta}$:

$$\Pr(\tilde{\Delta}=0 \mid \tau, i)=\Pr(\Delta=0)=1-\Pr(\Delta>0),$$

$$\Pr(\tilde{\Delta}=jT \mid \tau, i)=\Pr((j-1)T<\Delta\leqslant jT)$$
$$=\Pr(\Delta>(j-1)T)-\Pr(\Delta>jT),$$
$$j=1,2,\ldots.$$

Below, we derive $\Pr(\Delta>j)$, $j=0,1,\ldots$.

Recall that the factory receives orders from the retailers periodically. The allocation policy described in the previous section establishes a sequence in which these orders are to be filled. This sequence is determined by the following rules: (1) retailer orders received by the factory at time t_1 are filled before retailer orders received by the factory at time t_2 for any $t_1<t_2$; (2) the units contained in the orders received by the factory in the same period are sequenced according to the arrival times of the corresponding demands. Take any periods, t_1 and t_2, with $t_1<t\leqslant t_2$. Recall that the replacement unit is contained in an order to the factory in period t. Define $V^-(t_1)$ to be the total retailer orders received by the factory after period t_1 (exclusive) but filled before the replacement unit. Define $V^+(t_2)$ to be the total retailer orders filled by the factory after the replacement unit, but received before period t_2 (inclusive). Thus,

$$Z(t_1,t_2]=V^-(t_1)+1+V^+(t_2),$$

where "1" corresponds to the replacement request.

Consider $\Pr(\Delta>j)$, $j=0,1,\ldots$. Note that $(\Delta>j)$ if and only if the backlog at the factory at time $t+j$ exceeds all the outstanding orders that are filled after the replacement request; i.e., $B_0(t+j)>V^+(t+j)$. Because $B_0(t)>0$ implies $I_0(t)=0$, and thus $IL_0(t)=-B_0(t)$,

$$\Pr(\Delta>j)=\Pr(IL_0(t+j)<-V^+(t+j)). \quad (6)$$

To obtain the value of the right-hand side, we distinguish between two cases.

Case 1: $j<L$. In this case,

$$IL_0(t+j)=IP_0(t+j-L)-Z(t+j-L,t+j]$$
$$=IP_0(t+j-L)-V^-(t+j-L)-1-V^+(t+j).$$

Therefore, from (6),

$$\Pr(\Delta>j)$$
$$=\Pr(IP_0(t+j-L)-V^-(t+j-L)$$
$$-1-V^+(t+j)<-V^+(t+j))$$
$$=\Pr(IP_0(t+j-L)-V^-(t+j-L)-1<0)$$
$$=\Pr(S-X(t+j-L)-V^-(t+j-L)-1<0). \quad (7)$$

Note that $X(t+j-L)$ is independent of $V^-(t+j-L)$. To determine the distributions of these variables, consider the following two cases.

Case A: $T\leqslant N$. In this case, the factory receives retailer orders every period and the order quantities received in different periods are independent and identically distributed. Let X be the shortfall in steady state. Note that $V^-(t+j-L)$ is equal to $Z(t+j-L,t-1]+i$, where $Z(t+j-L,t-1]$ is the total retailer orders received by the factory in $L-j-1$ consecutive periods. Let Z^k be the total retailer orders received by the factory in k consecutive periods. Thus, Z^k is a Poisson random variable with mean $k\cdot(N/T)\cdot T\lambda=kN\lambda$. From (7), we have

$$\Pr(\Delta>j)=\Pr(X+Z^{L-j-1}>S-i-1),$$
$$j=0,1,\ldots,L-1 \quad \text{if } T\leqslant N.$$

Case B: $T>N$. In this case, the factory receives an order from one retailer every T/N periods, while the order quantities received at different order occasions are still independent and identically distributed. Recall that there are T/N period-types and that X_k is the shortfall in a type-k period in steady state, $k=0,\ldots,T/N-1$. Note that period t (an order occasion) is of type 0. Let t' be the last order occasion before or at time $t+j-L$. Let $t+j-L-t'=m$. Then, period $t+j-L$ is of type m. Similarly, $Z(t+j-L,t-1]$ represents a total of

$$\left\lfloor \frac{L-j-1}{(T/N)} \right\rfloor \stackrel{\text{def}}{=} q$$

retailer orders, where $\lfloor x\rfloor$ is the largest integer $\leqslant x$. Redefine Z^k to be the sum of k retailer orders, which is a Poisson random variable with mean $kT\lambda$. From (7), we have

$$\Pr(\Delta>j)=\Pr(X_m+Z^q>S-i-1),$$
$$j=0,1,\ldots,L-1 \quad \text{if } T>N.$$

Case 2: $j\geqslant L$. Again, we have,

$$IL_0(t+j)=IP_0(t+j-L)-Z(t+j-L,t+j].$$

Take any $k=0,1,\ldots,j-L$. The event $(\Delta>j)$ implies $(\Delta>k+L)$, which in turn implies $(IL_0(t+k+L)<-V^+(t+k+L))$; see (6). Because $IL_0(t+k+L)=IP_0(t+k)-Z(t+k,t+k+L]$, we have

$$IP_0(t+k)-Z(t+k,t+k+L]<-V^+(t+k+L).$$

Because $Z(t+k,t+k+L]<V^+(t+k+L)$, we have $IP_0(t+k)<0$. However, the factory base-stock level is $S\geqslant 0$. Therefore, the factory production in period $t+k$ must be at its capacity, C, $k=0,1,\ldots,j-L$. This observation leads to

$$IP_0(t+j-L)=IP_0(t-1)-V^-(t-1)-1$$
$$-V^+(t+j-L)+(j-L+1)\cdot C.$$

Therefore,

$$\begin{aligned}
&\Pr(\Delta > j)\\
&\quad = \Pr\big(IP_0(t+j-L) - Z(t+j-L, t+j] < -V^+(t+j)\big)\\
&\quad = \Pr\big(IP_0(t+j-L) < -V^+(t+j-L)\big)\\
&\quad = \Pr\big(IP_0(t-1) - V^-(t-1) - 1 - V^+(t+j-L)\\
&\qquad\qquad + (j-L+1)\cdot C < -V^+(t+j-L)\big)\\
&\quad = \Pr\big(S - X(t-1) - V^-(t-1) - 1 + (j-L+1)\cdot C < 0\big).
\end{aligned}$$

Note that $V^-(t-1)$ is equal to i because it does not include any orders received by the factory before time t. When $T \leqslant N$, $X(t-1)$ in steady state is distributed as X, the steady-state shortfall. When $T > N$, $X(t-1)$ in steady state is distributed as $X_{T/N-1}$, the steady-state shortfall in a type-$(T/N-1)$ period. Therefore,

$$\begin{aligned}
&\Pr(\Delta > j)\\
&\quad = \begin{cases} \Pr\big(X > S - i - 1 + (j-L+1)\cdot C\big), & T \leqslant N,\\ \Pr\big(X_{T/N-1} > S - i - 1 + (j-L+1)\cdot C\big), & T > N. \end{cases}
\end{aligned}$$

Note that the distribution of Δ depends on the values of τ and i. Due to the Poisson demand processes, τ is uniformly distributed over $(0, T)$. Recall that i is the number of demands that have arrived at the retailers that order at time t in an interval of length $T-\tau$. Thus, the probability mass function of i is

$$p(i,\tau) = \frac{e^{-\mu}\mu^i}{i!}, \quad i = 0, 1, 2, \ldots,$$

where $\mu = \lambda(T-\tau)$ if $T \geqslant N$ and $\mu = \lambda(N/T)(T-\tau)$ otherwise.

The expected holding and backorder costs associated with the specified demand are given by

$$\frac{1}{T}\int_0^T \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} c^Y(\tau + jT + l)\Pr(\tilde{\Delta} = jT \mid \tau, i)p(i,\tau)\,d\tau.$$

(Note that this cost expression also depends on S.) Because the demand rate is λ units per period per retailer, the long-run average holding and backorder costs per period at all the retailers are

$$N\lambda\frac{1}{T}\int_0^T \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} c^Y(\tau + jT + l)\Pr(\tilde{\Delta} = jT \mid \tau, i)p(i,\tau)\,d\tau. \quad (8)$$

Finally, note that the long-run average systemwide ordering costs are $N\cdot K(1-e^{-\lambda T})/T$ per period. Therefore, the long-run average total costs per period in the system are the sum of (3), (5), (8), and $N\cdot K(1-e^{-\lambda T})/T$ when $T \leqslant N$. Substitute (4) for (3) when $T > N$.

3.2. Approximate Evaluation for the (R, Q) Policy

As with the (T, Y) policy, the systemwide holding and backorder costs have three components; see (1). First, we derive the long-run average value of $HIL_0(t)$. To this end, we again use the relation $IL_0(t+L) = IP_0(t) - Z(t, t+L]$. Because for any t, $IP_0(t) = S - X(t)$, we have

$$IL_0(t+L) = S - X(t) - Z(t, t+L].$$

The same flow conservation argument used to determine the long-run average value of $Z(t, t+L]$ in the (T, Y) system applies in the (R, Q) system as well; i.e., the long-run average value of $Z(t, t+L]$ is simply the expected customer demand over all retailers in L periods, $NL\lambda$.

However, the steady-state distribution of $X(t)$ is more difficult to obtain than in the (T, Y) system. The problem lies in the order process. In the (T, Y) system, the orders received by the factory are independent and identically distributed, while the orders received by the factory in different periods in the (R, Q) system are not. To see this, imagine a system where the demand rate is low and Q is large. If a retailer places an order in a period, it is quite unlikely that the same retailer will place an order again in the following period. As a result of this dependency, the factory shortfall at time t, which is affected by the retailer orders in period $t-1$, contains information about the retailer orders in period t. Consequently, the factory shortfall in steady state, X, is difficult to characterize analytically. We rely on simulation for its distribution. The long-run average value of $HIL_0(t)$ is then

$$H(S - E[X] - NL\lambda). \quad (9)$$

The long-run average value of $NIP_n(t)$, for any n, is easy to obtain. Because retailer n uses an (R, Q) policy and its demand follows a Poisson process, it is well known that the steady-state distribution of the nominal inventory position is uniform in the interval $[R+1, R+Q]$. Thus, the long-run average value of $NIP_n(t)$ is simply $R + (Q+1)/2$. The long-run average value of $H\sum_{n=1}^{N} NIP_n(t)$ is given by

$$HN\left(R + \frac{Q+1}{2}\right). \quad (10)$$

The methodology for determining the long-run average value of $\sum_{n=1}^{N}[hI_n(t) + (p+H)B_n(t)]$ is similar to that used in the (T, Y) case. The sequence of events generated by an arbitrary demand is used to determine the costs that will be associated with it. We are interested only in demands that trigger an order (of batch-size Q). (Therefore, we are using the continuous-review scenario in the following derivation; see §2.) Consider an arbitrary demand that arrives at retailer n and triggers an order for a replacement batch. This replacement batch will be used to satisfy Q future demands at retailer n. We compute the time this replacement batch will arrive at retailer n. We also determine the arrival times of the Q future demands that this batch satisfies. For each of the Q demands, we determine whether a holding or backorder cost is incurred. The total cost is associated with the demand that triggered the replacement batch. Multiplying this cost by the rate at which orders

occur at all the retailers yields the long-run average holding and backorder costs at the retailers. We now describe this procedure in detail.

Consider an arbitrary demand. This demand occurs at, say, retailer n. It arrives τ units of time before the next shipping occasion, which happens at time t. (Thus, t is the beginning of the first period after the demand occurrence.) Suppose this demand, hereafter referred to as the *specified demand*, triggers an order.

To determine the costs associated with the specified demand, note that the order it triggers, hereafter referred to as the *replacement request/batch*, is placed just after retailer n's nominal inventory position reached R. Thus, the units in the replacement batch will meet the $(R+y)$th successive demands, $y = 1, \ldots, Q$. Holding and backorder costs are assessed depending on whether the replacement batch arrival time is before or after the demands to be satisfied by the units in this batch. These costs are associated with the specified demand.

Upon receipt of a retailer order (for Q units), the factory attempts to fill this order (i.e, transfer inventory from bin 0 to bin n). However, the factory can ship to a retailer only at the beginning of a period. Thus, the replacement request is delayed at least τ units of time. If the factory is unable to fill this request before the next shipping occasion (at time t), additional delay ensues. Let Δ be the number of periods from the first shipping occasion (at time t) until the replacement batch is shipped from the factory. The total time from the arrival of the specified demand to the arrival of the replacement batch at the retailer that ordered it is given by $\tau + \Delta + l \stackrel{\text{def}}{=} t_b$. Let t_y, $y = 1, \ldots, Q$, be the time from the arrival of the specified demand to the arrival of the $(R+y)$th successive demand. Thus, the total costs associated with the specified demand is given by

$$\sum_{y=1}^{Q} [h(t_y - t_b)^+ + (p+H)(t_y - t_b)^-].$$

It remains to derive the distributions of t_y, $y = 1, \ldots, Q$, and the stochastic components of t_b: τ and Δ.

First, because the demand process at each retailer is Poisson, the distribution of t_y, $y = 1, \ldots, Q$, is Erlang with parameters λ and $R+y$. Let $g_y(t)$ be the density function of t_y; i.e.,

$$g^y(t) = \frac{\lambda^{R+y} t^{(R+y-1)} e^{-\lambda t}}{(R+y-1)!}, \quad t \geqslant 0, \quad y = 1, \ldots, Q.$$

Let $c^y(t_b)$ be the expected holding and backorder costs for the yth unit in the replacement batch, given t_b. Note that $c^y(t_b)$ is equivalent to $c^Y(t_r)$ in the (T, Y) system with $R + y = Y$ and $t_b = t_r$. Thus,

$$c^y(t_b) = e^{-\lambda t_b} \frac{p+H+h}{\lambda} \cdot \sum_{k=0}^{R+y-1} \frac{(R+y-k)}{k!} t_b^k \lambda^k + (p+H)(t_b - (R+y)/\lambda),$$

where $0! = 1$. Given t_b, the total expected holding and backorder costs associated with the specified demand is $\sum_{y=1}^{Q} c^y(t_b)$.

We proceed to derive the distribution of Δ given τ. Take any periods t_1 and t_2 with $t_1 < t \leqslant t_2$. Redefine $V^-(t_1)$ to be the total retailer orders (in terms of individual units) received by the factory after time t_1 (exclusive), but before the replacement request. Similarly, redefine $V^+(t_2)$ to be the total retailer orders after the replacement request that are received by the factory by time t_2 (inclusive). Thus, $Z(t_1, t_2]$ is equal to $V^-(t_1) + Q + V^+(t_2)$, where Q represents the replacement request.

Consider $\Pr(\Delta > j)$, $j = 0, 1, \ldots$. The event $(\Delta > j)$ is equivalent to $B_0(t+j) > V^+(t+j)$ because retailer orders are filled on an FCFS basis. To establish a linkage between $B_0(t+j)$ and $IL_0(t+j)$, recall that retailer orders cannot be partially filled. Therefore, it is possible that $I_0(t)$ and $B_0(t)$ are both positive, which is the case only if $I_0(t) \leqslant Q - 1$. The following observations are true for any t':

1. $B_0(t') = mQ$, $m \geqslant 0$, integer;
2. $I_0(t') \geqslant 0$;
3. If $m \geqslant 1$, then $I_0(t') < Q$;
4. $IL_0(t') \geqslant 0 \Rightarrow B_0(t') = 0$.

Let $V^+(t+j) = mQ$ $(\geqslant 0)$. Suppose that $B_0(t+j) > mQ$. From the above observations, $I_0(t+j) \leqslant Q - 1$. Therefore,

$$\begin{aligned} IL_0(t+j) &= I_0(t+j) - B_0(t+j) \\ &\leqslant (Q-1) - (m+1)Q \\ &\leqslant Q - 1 - mQ - Q \\ &< -mQ. \end{aligned}$$

Conversely, if $IL_0(t+j) < -mQ$, then $B_0(t+j) = I_0(t+j) - IL_0(t+j) \geqslant -IL_0(t+j) > mQ$. Consequently,

$$\Pr(\Delta > j) = \Pr(IL_0(t+j) < -V^+(t+j)). \tag{11}$$

The value of the right-hand side of (11) is obtained by using arguments almost identical to the ones used for the right-hand side of (6). We distinguish between two cases.

Case 1: $j < L$. In this case,

$$\begin{aligned} IL_0(t+j) &= IP_0(t+j-L) - Z(t+j-L, t+j] \\ &= IP_0(t+j-L) - V^-(t+j-L) - Q - V^+(t+j). \end{aligned}$$

Therefore, from (11),

$$\begin{aligned} &\Pr(\Delta > j) \\ &= \Pr\big(IP_0(t+j-L) - V^-(t+j-L) \\ &\qquad -Q - V^+(t+j) < -V^+(t+j)\big) \\ &= \Pr\big(IP_0(t+j-L) - V^-(t+j-L) - Q < 0\big) \\ &= \Pr\big(S - X(t+j-L) - V^-(t+j-L) - Q < 0\big). \end{aligned} \tag{12}$$

To determine the distribution of $V^-(t+j-L)$, recall that it equals the orders placed by all retailers in the interval

$(t+j-L, t-\tau)$. Let Z_n^δ be the total amount ordered by retailer n in the interval $(t-\tau-\delta, t-\tau)$. Let Z_{-n}^δ be the total amount ordered by all retailers except retailer n in the same interval. Thus,

$$V^-(t+j-L) = Z_n^{L-j-\tau} + Z_{-n}^{L-j-\tau}.$$

First, consider Z_n^δ. Because retailer n places an order at $t-\tau$, the nominal inventory position just before the order placement is equal to R. Going backward in time, an order by retailer n is triggered every Q demands. Let D^δ be the number of demands that occur at any retailer in an interval of length δ. Thus,

$$\Pr(Z_n^\delta = 0) = \Pr(D^\delta < Q),$$
$$\Pr(Z_n^\delta = zQ) = \Pr\big(zQ \leqslant D^\delta < (z+1)Q\big), \quad z = 1, 2, \ldots.$$

Now consider Z_{-n}^δ. Take any retailer n', $n' \neq n$. Let A^δ be the total amount ordered by retailer n' in the interval $(t-\tau-\delta, t-\tau)$. Note that the nominal inventory position of retailer n' at $t-\tau-\delta$ is uniformly distributed over the interval $[R+1, R+Q]$ in steady state. By conditioning on this nominal inventory position, we have

$$\Pr\big(A^\delta = 0 \mid NIP_{n'}(t-\tau-\delta) = R+y\big) = \Pr(D^\delta < y),$$
$$\Pr\big(A^\delta = zQ \mid NIP_{n'}(t-\tau-\delta) = R+y\big)$$
$$= \Pr\big(y+(z-1)Q \leqslant D^\delta < y+zQ\big)$$

for $z = 1, \ldots$. Thus,

$$\Pr(A^\delta = zQ) = \frac{1}{Q}\sum_{y=1}^{Q} \Pr\big(A^\delta = zQ \mid NIP_{n'}(t-\tau-\delta) = R+y\big), \quad z = 0, 1, \ldots.$$

Because the order processes at all the retailers except retailer n are independent with identical characteristics, the distribution of Z_{-n}^δ is obtained by taking the $(N-1)$-fold convolution of A^δ.

To evaluate the right-hand side of (12), recall that $X(t+j-L)$ contains information about the retailer orders in period $t+j-L$. Therefore, $X(t+j-L)$ is not independent of $V^-(t+j-L)$. To simplify computation, however, we assume that they are independent. Under this approximation, we have

$$\Pr(\Delta > j) = \Pr\big(X + Z_n^{L-j-\tau} + Z_{-n}^{L-j-\tau} > S-Q\big), \quad j = 0, 1, \ldots, L-1,$$

which is evaluated by convolving the three random variables X, $Z_n^{L-j-\tau}$, and $Z_{-n}^{L-j-\tau}$.

Case 2: $j \geqslant L$. Using the arguments developed under (T, Y), we have

$$IP_0(t+j-L) = IP_0(t-1) - V^-(t-1) - Q - V^+(t+j-L) + (j-L+1)\cdot C.$$

Because $IL_0(t+j) = IP_0(t+j-L) - Z(t+j-L, t+j]$, we have from (11),

$$\begin{aligned}\Pr(\Delta > j) &= \Pr\big(IP_0(t-1) - V^-(t-1) - Q \\ &\qquad - V^+(t+j-L) + (j-L+1)\cdot C \\ &\qquad - Z(t+j-L, t+j] < -V^+(t+j)\big) \\ &= \Pr\big(IP_0(t-1) - V^-(t-1) - Q + (j-L+1)\cdot C < 0\big) \\ &= \Pr\big(X + Z_n^{1-\tau} + Z_{-n}^{1-\tau} > S - Q + (j-L+1)\cdot C\big),\end{aligned}$$

which is evaluated by convolving X, $Z_n^{1-\tau}$, and $Z_{-n}^{1-\tau}$. (This is an approximation, as discussed above.)

We are now ready to determine the distribution of Δ:

$$\Pr(\Delta = 0) = 1 - \Pr(\Delta > 0),$$
$$\Pr(\Delta = j) = \Pr(\Delta > j-1) - \Pr(\Delta > j), \quad j = 1, 2, \ldots.$$

Note that the distribution of Δ is dependent on τ, which is uniformly distributed over $(0, 1)$. To make this dependency explicit, we write $\Pr(\Delta = j \mid \tau)$ for $\Pr(\Delta = j)$. The total expected holding and backorder costs associated with the specified demand is given by

$$\int_0^1 \sum_{j=0}^{\infty}\sum_{y=1}^{Q} c^y(\tau+j+l)\Pr(\Delta = j \mid \tau)\, d\tau,$$

which is a function of R, Q, and S. Because demand occurs on average λ units per period per retailer with an order being triggered every Q units, the long-run average holding and backorder costs per period at all retailers are

$$N\lambda\frac{1}{Q}\int_0^1 \sum_{j=0}^{\infty}\sum_{y=1}^{Q} c^y(\tau+j+l)\Pr(\Delta = j \mid \tau)\, d\tau. \tag{13}$$

Finally, note that the expected ordering costs per period are equal to

$$NK\frac{1}{Q}\sum_{y=R+1}^{R+Q} \Pr(y - D \leqslant R),$$

where D denotes the total demand at a retailer in a period. The long-run average total costs per period in the system are the sum of (9), (10), (13), and the above ordering costs.

4. Numerical Examples

We used numerical examples to compare the performance of (T, Y) policies with that of (R, Q) policies. Specifically, we were interested in the volatility of orders faced by the factory, the long-run average systemwide costs, and the impact of key system parameters on the relative cost performance of the two types of policies.

Table 1. System parameters.

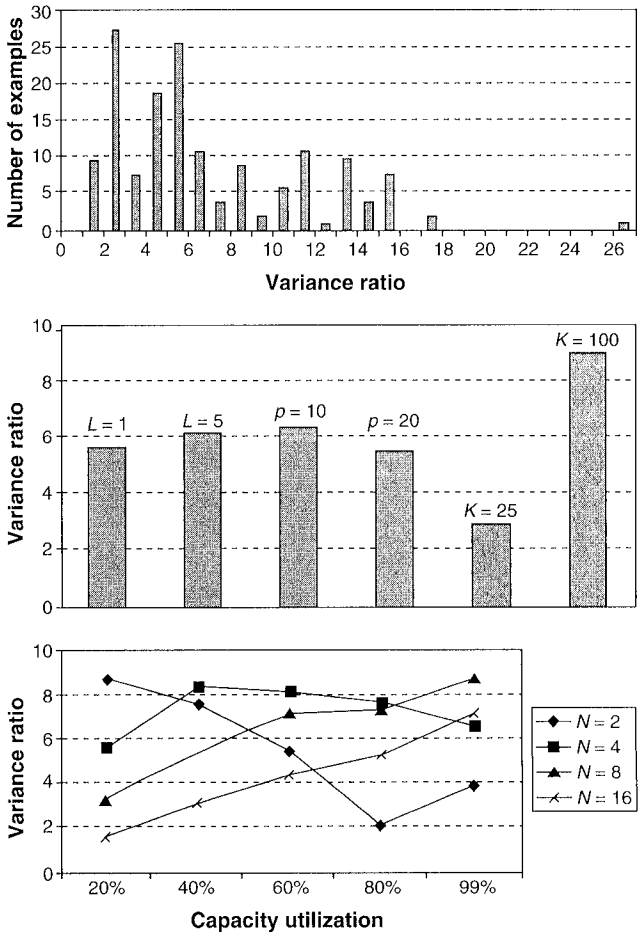
N	2, 4, 8, 16
ρ	20%, 40%, 60%, 80%, 99%
C	50
L	1, 5
H	1
K	25, 100
l	3
h	1
p	10, 20

We have chosen the sets of parameters in Table 1, where $\rho = N\lambda/C$ is the factory's capacity utilization, with the mean demand rate $\lambda = \rho C/N$. There are 160 examples.

For each example, we used the exact procedure developed in §3.1 to compute the long-run average costs of a (T, Y) policy and the approximate procedure developed in §3.2 for an (R, Q) policy. For both cases, we obtained the steady-state shortfall distribution via simulation and the optimal control parameters, i.e., the optimal values of (T, Y, S) and (R, Q, S), via a search. Once the optimal control parameters were obtained, a simulation of 50,000 periods was used to verify that the long-run average costs generated by the above algorithms were accurate for both (T, Y) and (R, Q). (For the (T, Y) case, even though the procedure was exact, there was a step where the random variable τ was discretized, introducing a potential error.) For (T, Y), the average absolute percentage difference between the evaluated and simulated costs was 0.1% and the standard deviation was 0.5%. For (R, Q), the average was 0.4% and the standard deviation was 0.7%. We henceforth use the simulated costs as our performance measures of the policies.

We also determined the variability of the orders seen by the factory. For the (R, Q) system, the variance of the total retailer orders in any period can be easily computed because the nominal inventory positions of the retailers are independent and uniformly distributed. (As we mentioned earlier, the orders by a retailer in different periods are typically correlated. Therefore, the variance of the total retailer orders in a period is, admittedly, a simplistic way to measure the uncertainty in the factory's demand process.) For (T, Y), we computed the variance of the total retailer orders received by the factory on any retailer order occasion. (It turns out that in all our examples, $T \leqslant N$. Therefore, the factory sees a retailer order(s) every period. In other words, every period is a retailer order occasion. Now if $T > N$, it may be necessary to consider alternative variance measures as the factory expects zero retailer orders in some periods.) Not surprisingly, the order variance under the optimal (R, Q) policy was found to be higher than or equal to that under the optimal (T, Y) policy in every example. The highest ratio of the order variances exceeds 25. Figure 2 shows the histogram of the variance ratio and how it depends on several system parameters. Note that the variance ratio increases as the transportation fixed cost

Figure 2. Ratio of order variance in the (R, Q) system to order variance in the (T, Y) system.



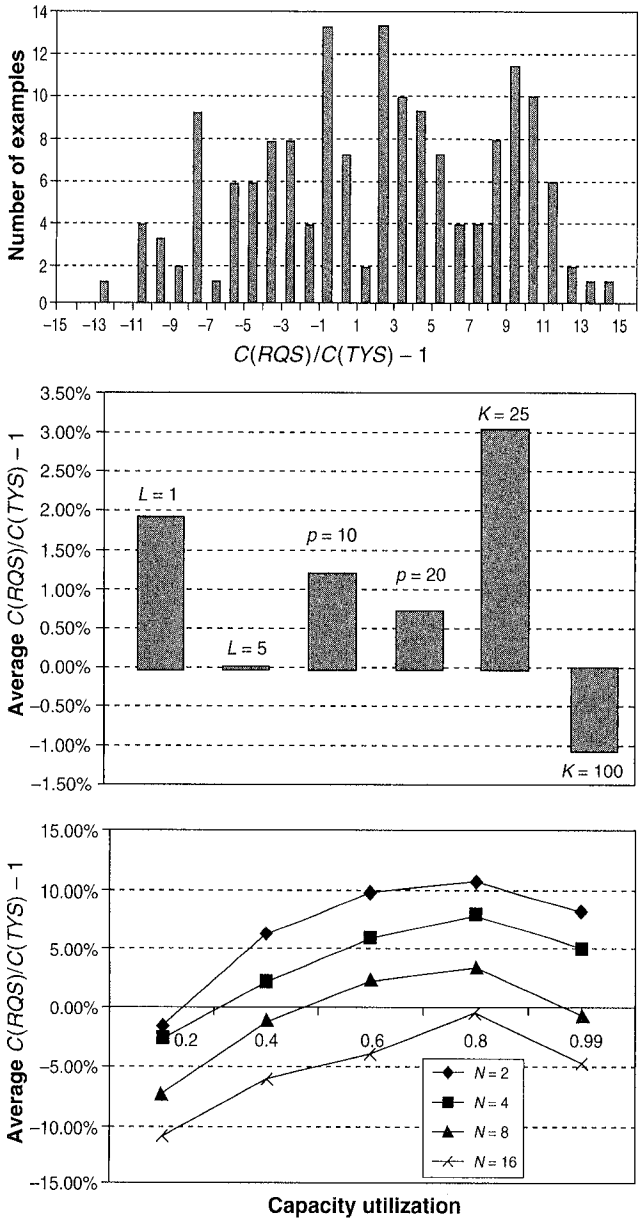
becomes larger, implying that the order variance in (R, Q) systems increases at a faster pace than the order variance in (T, Y) systems as the need for batching intensifies. Also, the nature of the relationship between the variance ratio and the capacity utilization is highly dependent on the number of retailers.

Unlike order variances, there is no clear dominance in terms of supply chain costs. In 45% of the examples, (R, Q) outperformed (T, Y). Figure 3 depicts a histogram of the relative cost difference between the two policies with a mean of 0.99%, as well as how this relative performance depends on several key system parameters. The figure suggests that (R, Q) policies are likely to outperform (T, Y) policies when one of the following is true. (Tentative explanations are in square brackets.)

(a) The factory production lead time is long. [As the factory production lead time lengthens, the demand variance at the factory becomes less harmful because of risk pooling of the factory's demands over its lead time (i.e., its lead time demand). This is in favor of (R, Q) policies, which have larger order variance at the factory.]

(b) The customer backorder cost is high. [As the customer backorder cost increases, the retailers carry more

Figure 3. Cost comparison between the (T, Y) system and (R, Q) system.



safety stock. As a result, the induced penalty for shipment delays at the factory decreases. This is again in favor of (R, Q) policies, which induce larger order variance at the factory.]

(c) The transportation fixed cost is large. [As the economies of scale for retailer replenishment increase, both T and Q are expected to increase. Thus, the responsiveness of the (T, Y) system diminishes greatly because each retailer is allowed to order only once every T periods, whereas the responsiveness of the (R, Q) system remains largely unchanged because each retailer can, if necessary, order in every period.]

(d) The number of retailers is large. [One possible explanation is risk pooling of retailer orders at the factory, which serves to again diminish the negative impact of the larger order variance induced by the (R, Q) policy.]

(e) The factory capacity utilization is either extremely low or extremely high. [When capacity utilization is either very low or very high, planning at the factory is less critical: either produce to demand or produce to capacity. Planning is important in the midrange, and that is where the (R, Q) policy is likely to do worse because of the larger order variance at the factory induced by it.]

5. Demand Information and Allocation Policy

Suppose that the factory has access to point-of-sale data on a real-time basis; i.e., the factory observes every customer demand as it occurs at the retail level. How can the factory make use of, and what is the value of, this centralized demand information? Moreover, is there an alternative allocation policy that the factory can use? We address these questions in the context of (T, Y) systems. (Recall that in (T, Y) systems with $T < N$, the current allocation policy at the factory allocates, if necessary, inventory among the retailers ordering at the same time by using the sequence of the individual demands contained in those retailers' orders. Therefore, strictly speaking, the factory's access to the point-of-sale demand information has already been implicitly assumed. However, the factory has so far not used that information in its production decisions.)

With access to centralized demand information, the factory is able to accumulate information about a future retailer order. For example, suppose $T = 8$. Consider a retailer order placed at time t. This order comprises eight periods' worth of demands at that retailer. In the previous (T, Y) system, all the factory knows about this retailer order before time t is that it is a Poisson random variable with mean λT. In the current system, the factory observes an increasing portion of the random variable as time approaches t. For example, at time $t - 4$, the factory has observed four periods' worth of demand that will be part of the order at time t. This demand information can be incorporated into the factory's production decision. Vendor-managed inventory (VMI) systems typically rely on such information to enable the supplier to anticipate the replenishment needs of its downstream partners. (Note that the factory is incorporating information about its future demands into its production decisions. Several papers have considered the impact of this type of demand information. A key difference is that the future-demand information here comes from within the supply chain, instead of from the end customers. As a result, the scenario we are considering here is also one of information sharing between supply chain stages. Many papers have studied this, but information sharing in the context of staggered (T, Y) policies is new. See Chen 2003 for a survey on the supply chain information-sharing literature.)

We assume that the factory follows a floating base-stock policy; i.e., in period t, the factory attempts to produce

enough to raise its inventory position to a base-stock level equal to $s + v(t)$, where s is a control parameter and $v(t)$ is a function of the observed demand data at time t.

Suppose we are at time t. To gain some intuition on the form of $v(t)$, let us consider a special case with $L = 2$. Let x_i be the random variable representing the unknown portion of the retailer order placed at time $t + i$, and let y_i be the observed part of this retailer order, $i = 1, 2$. Note that the total retailer orders received by the factory in the time interval $(t, t + L]$ is $x_1 + x_2 + y_1 + y_2$. Intuitively, the factory's target inventory position at time t, i.e., $s + v(t)$, should cover this lead-time demand. It seems reasonable to let s cover the unknown portion of the lead-time demand, i.e., $x_1 + x_2$, and let $v(t)$ cover the known portion, i.e., $y_1 + y_2$. In general, let $y_i(t)$ be the part of the retailer order that will be placed at time $t + i$ that has been observed by time t, $i = 1, \ldots, L$. It seems plausible to have $v(t) = \sum_{i=1}^{L} y_i(t)$.

Now let us consider the factory's allocation policy. Recall that under the FCFS allocation policy, the inventory allocated to a retailer's bin (at the factory) represents inventory committed to that retailer and thus cannot be used to satisfy another retailer's order. This may create a situation where there is inventory in one retailer's bin (to be shipped to the retailer on its next order occasion) while another retailer's order has to be backlogged. This motivates an alternative allocation policy that we call current order allocation (COA). Under this policy, the factory will give priority to the current retailer order. That is, in every period, the factory will attempt to fill the current retailer order as much as possible from its on-hand inventory. If the on-hand inventory is insufficient, the factory will create a backlog for the retailer for the unfilled portion. This backlog will be added to the next order placed by this retailer, and the total becomes the then-current retailer order. If in a period the factory has simultaneously received orders from multiple retailers, and if the factory on-hand inventory is insufficient to satisfy all these orders, inventory allocation is according to the sequence of demands that correspond to the individual units in the orders in an FCFS manner. The factory then creates a backlog for each of the retailers and retains the demand sequence for later allocation, and these backlogs will be added to the corresponding orders placed by this group of retailers on their next order occasion.

The original (T, Y) system can be modified in several ways, depending on whether or not the centralized demand information (CDI) is utilized and if the allocation policy is FCFS or COA. The value of CDI is defined to be the percentage reduction in systemwide costs resulting solely from the use of CDI (and the allocation policy remains FCFS). Similarly, the value of COA refers to the cost reduction resulting from COA alone (without CDI), and the value of CDI and COA combines the benefits of both the demand information and the new allocation policy.

To calculate the supply chain performance under the above three variations of the original (T, Y) system for the 160 examples described earlier, we fixed the values of T and Y found in the original system and searched for the optimal value of s based on simulation. This made computation manageable. Figures 4, 5, and 6 report the values of the various improvements and how they depend on key system parameters.

On average, CDI reduces costs by 0.9%, with a range from 0% to 4%. (This magnitude of improvement is comparable to what has been reported in the literature on the value of centralized demand information; see Chen 2002.) The value of CDI tends to be higher if one of the following is true. (Tentative explanations are in square brackets.)

(a) The factory production lead time is shorter. [If the factory production lead time is long, the factory would require high safety stock. Thus, the floating factor $v(t)$, due to CDI, is relatively small, implying that the value of CDI is small as well.]

(b) The customer backorder cost is lower. [When the customer backorder cost is low, the retailers carry low safety stock. Thus, the fill rate at the factory becomes important. CDI helps improve factory fill rate.]

Figure 4. Value of centralized demand information in (T, Y) system.

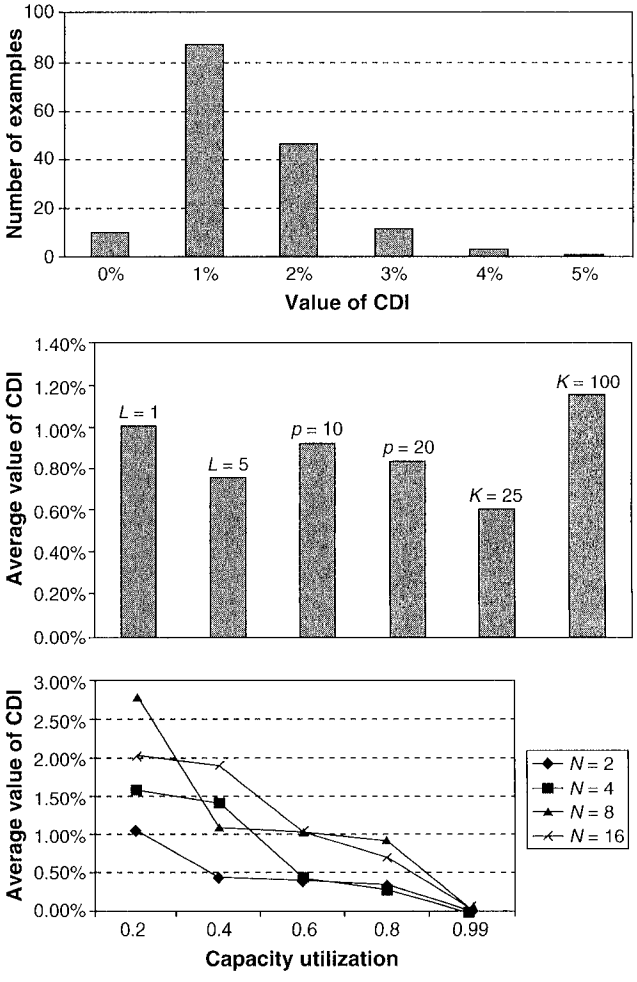
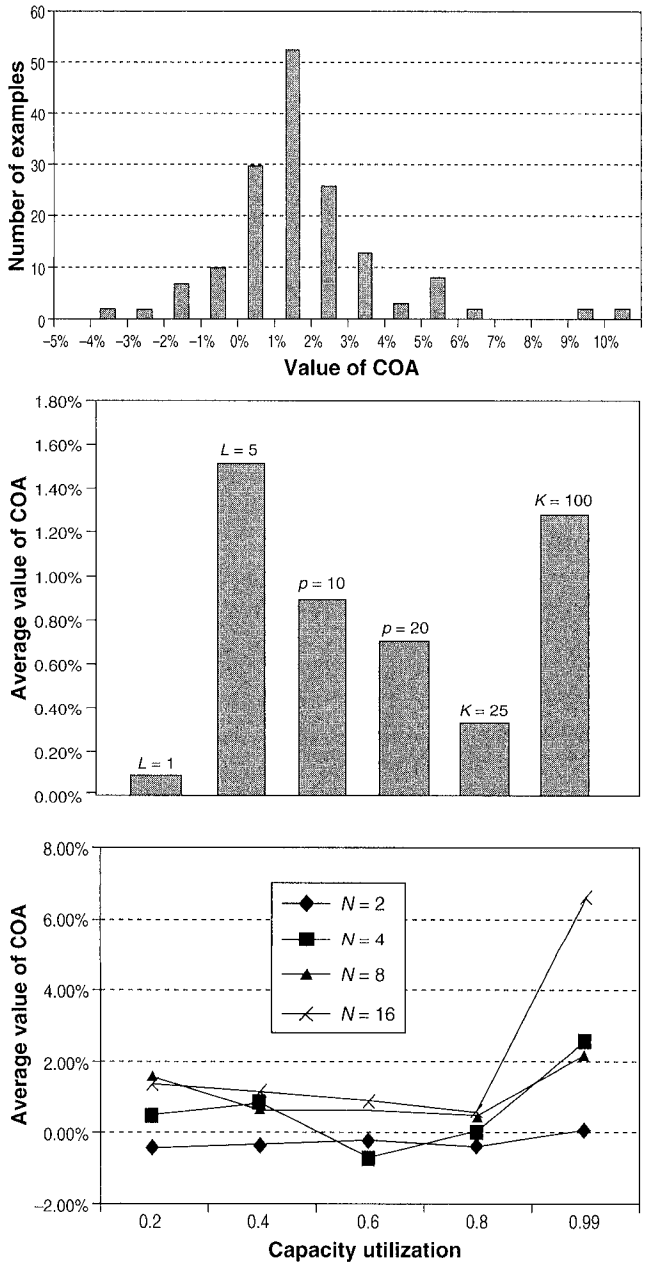


Figure 5. Value of current order allocation in the (T, Y) system.



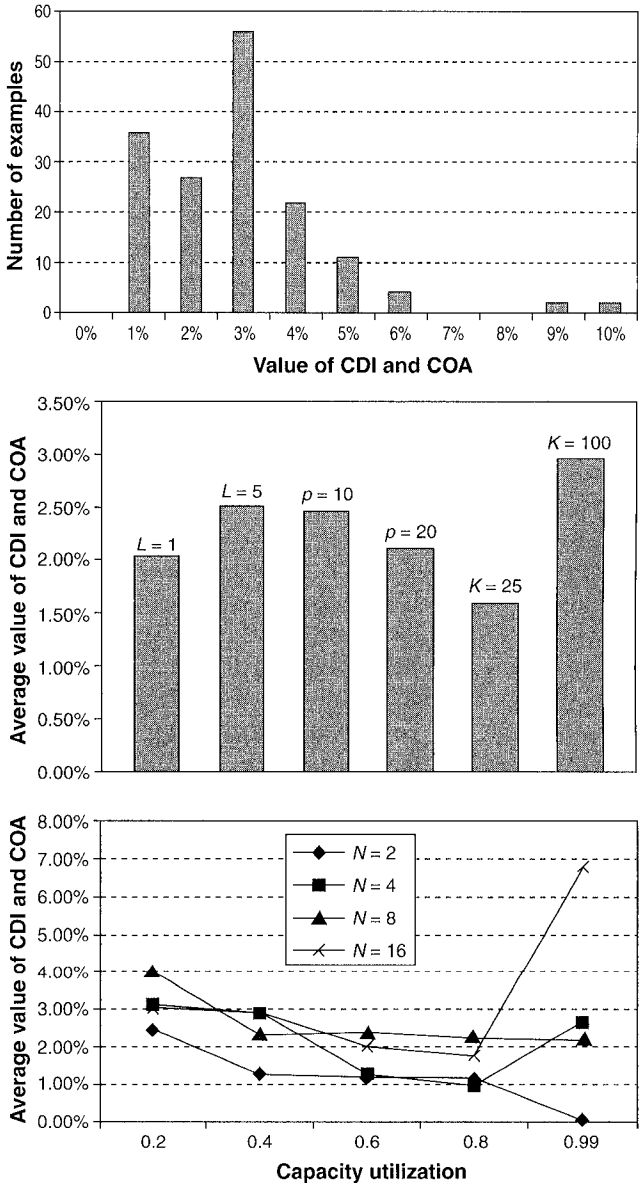
(c) The transportation fixed cost is higher. [As the fixed cost increases, T increases. The amount of information about future demands at the factory increases.]

(d) The number of retailers is larger. [As the number of retailers increases, while keeping capacity utilization constant, the average demand at each retailer decreases. Thus, the optimal T is likely to increase, implying that the factory has more future-demand information.]

(e) The factory capacity utilization is lower. [When the factory has ample capacity, it is in a better position to exploit the CDI.]

On the other hand, the new allocation policy is not always better than FCFS: On average, it reduces costs

Figure 6. The combined value of CDI and COA.



by 0.8%, with a range from −4.4% to 9.4%. One disadvantage of COA is that it does not anticipate future demands and hold-back inventory for an expected large retailer order in the future. The value of COA tends to be higher if one of the following is true. (Tentative explanations are in square brackets.)

(a) The factory production lead time is longer. [When the factory production lead time is long, the factory's ability to respond to demand changes is diminished. In this case, it is more important not to have any factory inventory "wasted," i.e., sitting idle in a retailer bin while another retailer's order is backlogged. This scenario may arise under FCFS, but it does not happen under COA.]

(b) The customer backorder cost is lower. [When the customer backorder cost is low, the retailers hold low safety stock. This implies that it is more important to have a high

fill rate at the factory. By not having idle inventory at the factory, COA tends to increase the factory fill rate.]

(c) The transportation fixed cost is higher. [When the fixed cost is higher, T is lengthened. As a result, under FCFS the idle time of inventory (sitting in a retailer bin) is also lengthened, making COA relatively better.]

(d) The number of retailers is larger. [As the number of retailers increases, while keeping capacity utilization constant, the average demand at each retailer decreases. Thus, the optimal T is likely to increase, implying that there is more "wasted" inventory at the factory under FCFS.]

(e) The factory capacity utilization is very high. [The effect of high factory capacity utilization is similar to a long factory lead time.]

Finally, by combining CDI with COA in the (T, Y) system, we have achieved an average cost reduction of 2.3%, with a range from 0% to 9.5%. This suggests "synergy" between CDI and COA: The two together are more than the sum of what they can achieve individually. Figure 6 suggests that the combined value tends to be higher if (a) the factory production lead time is longer, (b) the customer backorder cost is lower, (c) the transportation fixed cost is higher, (d) the factory capacity utilization is very low, or (e) the factory capacity utilization is very high and the number of retailers is large. These trends are, of course, a result of the relative strengths of the trends depicted in Figures 4 and 5.

6. Conclusions

The central message of this paper is that reducing or eliminating the bullwhip effect does not always improve supply chain efficiency. Therefore, for firms interested in improving their supply chains, measuring and controlling the bullwhip effect is no substitute for a sound economic analysis of the entire chain's operations. The paper has also found that in the considered supply chain model, the value of centralized demand information is comparable with the existing findings in the literature, and that this value tends to increase under an inventory allocation policy at the supplier that gives priority to the current retailer order rather than one that satisfies retailer orders on an FCFS basis.

Acknowledgments

The authors thank Gerard Cachon, the associate editor, and two referees for their helpful comments on earlier versions of this paper. The first author's research was supported in part by the National Science Foundation under grant SBR-97-0246.

References

Aviv, Y., A. Federgruen. 1997. The operational benefits of information sharing and vendor managed inventory (VMI) programs. Columbia University, New York.

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

Axsäter, S. 1993a. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Oper. Res.* **41** 777–785.

Axsäter, S. 1993b. Continuous review policies for multi-level inventory systems with stochastic demand. S. C. Graves, A. H. G. Rinnooy Kan, P. H. Zipkin, eds. *Logistics of Production and Inventory*, Vol. 4. *Handbook in Operations Research and Management Science*. North-Holland, Amsterdam, The Netherlands.

Blanchard, O. J. 1983. The production and inventory behavior of the American automobile industry. *J. Political Economy* **91** 365–400.

Blinder, A. S. 1986. Can the production smoothing model of inventory behavior be saved? *Quart. J. Econom.* **101** 431–454.

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

Cachon, G. 2001. Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review. *Oper. Res.* **49** 79–98.

Caplin, A. S. 1985. The variability of aggregate demand with (s, S) inventory policies. *Econometrica* **53** 1395–1409.

Chen, F. 2003. Information sharing and supply chain coordination. T. de Kok, S. Graves, eds. *Handbook in Operations Research and Management Science: Supply Chain Management*. North-Holland, Amsterdam, The Netherlands.

Chen, F., R. Samroengraja. 2000. A staggered ordering policy for one-warehouse, multi-retailer systems. *Oper. Res.* **48** 281–293.

Chen, F., Y.-S. Zheng. 1994. Evaluating echelon stock (R, nQ) policies in serial production/inventory systems with stochastic demand. *Management Sci.* **40** 1262–1275.

Chen, F., Y.-S. Zheng. 1997. One-warehouse multi-retailer systems with centralized stock information. *Oper. Res.* **45** 275–287.

De Bodt, M., S. Graves. 1985. Continuous review policies for a multi-echelon inventory problem with stochastic demand. *Management Sci.* **31** 1286–1295.

Deuermeyer, B., L. Schwarz. 1981. A model for the analysis of system service level in warehouse/retailer distribution systems: The identical retailer case. L. Schwarz, ed. *Studies in the Management Sciences: The Multi-Level Production/Inventory Control Systems*. North-Holland, Amsterdam, The Netherlands, 163–193.

EHCR (Efficient Healthcare Consumer Response: Improving the Efficiency of the Healthcare Supply Chain). 1996. Health Industry Distributors Association. www.hida.org.

Eppen, G., L. Schrage. 1981. Centralized ordering policies in a multi-warehouse system with leadtimes and random demand. L. Schwarz, ed. *Studies in the Management Sciences: The Multi-Level Production/Inventory Control Systems*. North-Holland, Amsterdam, The Netherlands, 51–69.

Federgruen, A. 1993. Centralized planning models for multi-echelon inventory systems under uncertainty. S. Graves, A. Rinnooy Kan, P. Zipkin, eds. *Logistics of Production and Inventory*, Vol. 4. *Handbook in Operations Research and Management Science*. North-Holland, Amsterdam, The Netherlands.

Federgruen, A., P. Zipkin. 1984. Approximation of dynamic, multi-location production and inventory problems. *Management Sci.* **30** 69–84.

Federgruen, A., P. Zipkin. 1986a. An inventory model with limited production capacity and uncertain demands I. The average cost criterion. *Math. Oper. Res.* **11** 193–207.

Federgruen, A., P. Zipkin. 1986b. An inventory model with limited production capacity and uncertain demands II. The discounted-cost criterion. *Math. Oper. Res.* **11** 208–215.

Forrester, J. 1961. *Industrial Dynamics*. MIT Press, Cambridge, MA.

Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.*, **45** 244–257.

Glasserman, P., S. Tayur. 1996. A simple approximation for a multi-state capacitated production-inventory system. *Naval Res. Logist.* **43** 41–58.

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

Jackson, P. 1988. Stock allocation in a two-echelon distribution system or "What to do until your ship comes in." *Management Sci.* **34** 880–895.

Kahn, J. A. 1987. Inventories and the volatility of production, *Amer. Econom. Rev.* **77**(4) 667–679.

Kurt Salmon Associates. 1993. Efficient consumer response: Enhancing consumer value in the grocery industry. The Joint Industry Project on Efficient Consumer Response, Washington, DC.

Lee, H. L., V. Padmanabhan, S. Whang. 1997. Information distortion in a supply chain: The bullwhip effect. *Management Sci.* **43** 546–558.

Roundy, R. O., J. A. Muckstadt. 1997. Coordinating production and inventory to improve service. *Management Sci.* **43** 1189–1197.

Svoronos, A., P. Zipkin. 1988. Estimating the performance of multi-level inventory systems. *Oper. Res.* **36** 57–72.

Tayur, S. 1992. Computing order-up-to levels in capacitated environments. *Stochastic Models* **9** 585–598.

Tayur, S., R. Ganeshan, M. Magazine. 1998. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Models for Supply Chain Management.* Kluwer Academic Publishers, Boston, MA.