Annals of Operations Research 126, 71–101, 2004

On the Effect of Product Variety in Production–Inventory Systems

SAIF BENJAAFAR and JOON-SEOK KIM {saif,jkim}@me.umn.edu
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455, USA

N. VISHWANADHAM mpenv@nus.eng
Logistics Institute – Asia Pacific, National University of Singapore, Singapore-119260

Abstract. In this paper, we examine the effect of product variety on inventory costs in a production–inventory system with finite capacity where products are made to stock and share the same manufacturing facility. The facility incurs a setup time whenever it switches from producing one product type to another. The production facility has a finite production rate and stochastic production times. In order to mitigate the effect of setups, products are produced in batches. In contrast to inventory systems with exogenous lead times, we show that inventory costs increase almost linearly in the number of products. More importantly, we show that the rate of increase is sensitive to system parameters including demand and process variability, demand and capacity levels, and setup times. The effect of these parameters can be counterintuitive. For example, we show that the relative increase in cost due to higher product variety is decreasing in demand and process variability. We also show that it is decreasing in expected production time. On the other hand, we find that the relative cost is increasing in expected setup time, setup time variability and aggregate demand rate. Furthermore, we show that the effect of product variety on optimal base stock levels is not monotonic. We use the model to draw several managerial insights regarding the value of variety-reducing strategies such as product consolidation and delayed differentiation.

Keywords: production/inventory systems, product variety, delayed differentiation, queueing systems

1. Introduction

Determining how much product variety to offer is central to the strategy of most manufacturing firms. Intuitively, the costs and benefits of product variety are well understood. Increased product variety allows a closer match between customer preferences and offered products, which then has the potential of increasing or maintaining market share and/or yielding higher prices. On the other hand, higher product variety could lead to operational inefficiencies incurred whenever the production system switches from making one item to another or in increased costs of raw material, component procurement, and storage and distribution of finished goods. Additional costs may include higher costs in product development, marketing, and customer service. Although the cost–benefits tradeoffs are qualitatively clear, there has been an ongoing debate, in both industry and academia, regarding the true magnitude of these costs and benefits. A sample of recent articles include (Zipkin, 2001; Agrawal, Kumaresh, and Mercer, 2001; Randall and Ulrich, 2001; Fisher and It-

tner, 1999; MacDuffie, Sethuraman, and Fisher, 1998; Quelch and Kenny, 1994; Kekre and Srinivasan, 1990).

In this paper, we contribute to this debate by examining the costs of product variety in a specific context. We consider an integrated production–inventory system with finite capacity where products are made to stock and share the same manufacturing facility. The facility incurs a setup time whenever it switches from producing one product type to another. The manufacturing facility has a finite production rate and stochastic production times. In order to mitigate the effect of setups, products are produced in batches. Demand for each item occurs one unit a time according to an independent renewal process with stochastic order inter-arrival times. Finished inventory of each product is managed according to a continuous review base-stock policy. If available, an order is always satisfied from stock, otherwise it is backlogged with the production system. The system incurs a holding cost per unit of inventory per unit time and a backordering cost per backordered unit per unit time. Base-stock levels and batch sizes are chosen to minimize the long run average of the sum of holding and backordering costs. Note that, in contrast to a conventional inventory system where replenishment lead times are exogenous, inventory replenishment lead times in our system are endogenous and depend on the current level of congestion in the production system.

In our setting, higher product variety affects cost in two ways. Increasing the number of products increases batch sizes, which leads to longer supply leadtimes and, consequently, to higher inventory and backorder levels. Increasing the number of products also leads to an increase in the number of distinct inventoried items which induces, even when setup times are negligible, higher inventory costs. In contrast to inventory systems with exogenous lead times, we show that cost increases almost linearly in the number of products. More importantly, we show that the rate of increase is highly sensitive to various system parameters, including demand and process variability, demand and capacity levels, and setup times. Surprisingly, the effect of these parameters is not always in line with intuition. In particular, we show that the relative increase in cost due to higher product variety is decreasing in demand and process variability. It is also decreasing in expected production time. On the other hand, we find that the relative cost is increasing in expected setup time, setup time variability and aggregate demand rate. Furthermore, we show that the effect of product variety on optimal base stock levels is not monotonic.

Literature that is closely related to our work include papers by Thonemann and Bradley (2002), Federgruen, Gallego, and Katalan (2000), Zipkin (1995), and Benjaafar, Cooper, and Kim (2003). Thoneman and Bradley consider a decentralized system consisting of a manufacturer and several retailers with multiple products. They use an $M/G/1$ queueing model to approximate expected manufacturing lead-time, which they then use to approximate lead-time demand using a normal distribution. By treating leadtimes as i.i.d. random variables, they are able to decouple the analysis of the inventory and production systems. A related problem is studied by Federgruen, Gallego, and Katalan (2000) who develop upper and lower bounds on the optimal cost. They show that, under a periodic base stock policy, both the upper and lower bounds grow linearly in the number of products. The studies of Thonemann and Bradley and Federgruen, Gallego,

and Katalan are in part motivated by an earlier simulation study by DeGroote, Yucesan, and Kavadias (1999). Zipkin (1995) considers the case of a perfectly flexible production system with no setups between items. He shows that the sum of the standard deviations of lead-time demand for the different items increases proportionally to the square root of the number of products. He also shows that this sum is increasing in capacity utilization and production time variability. This analysis is extended in Benjaafar, Cooper, and Kim (2003) who examine the value of inventory pooling (product consolidation) and show that the benefits of pooling in a production–inventory setting diminish with increases in the loading of the production system. Benjaafar and Kim (2001) examine the effect of demand variability on the benefits of pooling.

The model we present in this paper is distinct from the above studies in that we do not assume specific distributions for demand, production time and setup times. Also, in our model, we do not decouple the production and the inventory systems. Instead we explicitly relate the distribution of inventory and backorder levels to the distribution of order queue size at the production system. This allows us then to directly estimate various performance measures of interest, including expected inventory and backorder levels, order queue size, and supply lead time. More importantly, it allows us to capture important effects due to characteristics of the distributions of various parameters such as demand, production times and setup times. Finally, in our model, we jointly optimize batch size and base-stock levels.

We do, however, share some important assumptions with previous literature. In particular, we assume that aggregate demand is not affected by our choice of variety level. Instead, higher variety leads to a segmentation of the existing demand among a larger number of items. This allows us to more readily isolate the effect of product variety and does correspond to situations of mature industries, where increasing product variety is needed to hold on to existing market share. Although pricing decisions are beyond the scope of our model, our assessment of how variety affects costs can be easily integrated into a *manufacturing–marketing* model that trades-off operational costs with higher prices.

Our work is of course related to the vast literature on product variety that spans the fields of economics (Lancaster, 1990), marketing science (Green and Krieger, 1985), operations management (Ho and Tang, 1998), and inventory theory (Garg and Lee, 1999). In the inventory theory literature, the focus has been on systems with exogenous lead times. An important result from this literature is the statistical economies of scale associated with consolidating multiple items into fewer ones, the so-called risk pooling effect. In particular, for symmetric systems it has been shown that inventory costs increase proportionally to the square root of the number of items (Eppen, 1979). In this paper, we show that in a production–inventory system the increase is linear in the number of items. Recent examples from the inventory literature include (Dobson and Yano, 2001; Aviv and Federgruen, 1999; Van Ryzin and Mahajan, 1999; Alfaro and Corbett, 1999).

The remainder of the paper is organized as follows. In section 2, we describe our model and our main analytical results. In section 3, we extend our analysis to asymmetric systems. In section 4, we provide computational results and various managerial insights.

In section 5, we include simulation results in support of our analysis. In section 6, we offer discussion and concluding comments.

2. Model description

We consider a multi-item production–inventory system where a single production facility produces K items and separate inventory buffers are kept for each item. The demand for item i occurs one unit at a time according to an independent renewal process with rate λ_i. The inter-arrival time between orders is a random variable denoted by X_i, with $E(X_i) = 1/\lambda_i$ and coefficient of variation C_{a_i}. Separate inventory buffers are kept for each item. If available, an order is satisfied from buffer stock. If not, the demand is backordered. The system incurs a holding cost h_i per unit of inventory of type i per unit time and a backordering cost b_i per unit of type i backordered per unit time. The inventory buffer of each item i is managed according to a continuous review base-stock policy with base-stock level s_i. This means that the arrival of each new order triggers the placement of a replenishment order with the production facility. Replenishment orders at the production facility are processed in batches of size Q_i on a first-come first-served basis. This means that replenishment orders for each item are accumulated until the number of orders reaches Q_i. Once the number of orders reaches Q_i, the batch is allowed to queue up for processing at the production facility. The production facility can process only one order at a time. We assume that unit production times for product i are i.i.d. generally distributed random variables, denoted Y_i, with $t_i \equiv E(Y_i)$ and $\eta_i \equiv \mathrm{Var}(Y_i)$. The production facility incurs a setup time when two consecutive batches are of different type. Setup times for product i are i.i.d. generally distributed random variables, denoted by Z_i, with $\tau_i \equiv E(Z_i)$ and $\theta_i \equiv \mathrm{Var}(Z_i)$.

As shown in figure 1, the production facility has two stages: a batching stage and a processing stage. In the batching stage, customer orders are accumulated until a batch of size Q_i is reached. In the processing stage, batches queue up in front of the production facility and processed in the order they arrive. Units within a batch are processed one at a time. Once completed, a batch is delivered to the corresponding inventory buffer. Since the demand occurs according to a renewal process, the distribution of number of customers for item i in the batching stage is uniformly distributed with

$$\Pr\left(N_i^b = n\right) = \frac{1}{Q_i}, \quad n = 0, 1, \ldots, Q_i - 1, \tag{1}$$

where N_i^b is the number of orders of type i in the batching stage.

In our model, we assume that a base-stock level s_i and batch size Q_i is chosen so that the long run expected total cost per unit time is minimized. We denote this expected total cost by

$$z(\mathbf{s}, \mathbf{Q}) = \sum_{i=1}^{K} E(h_i I_i + b_i B_i), \tag{2}$$

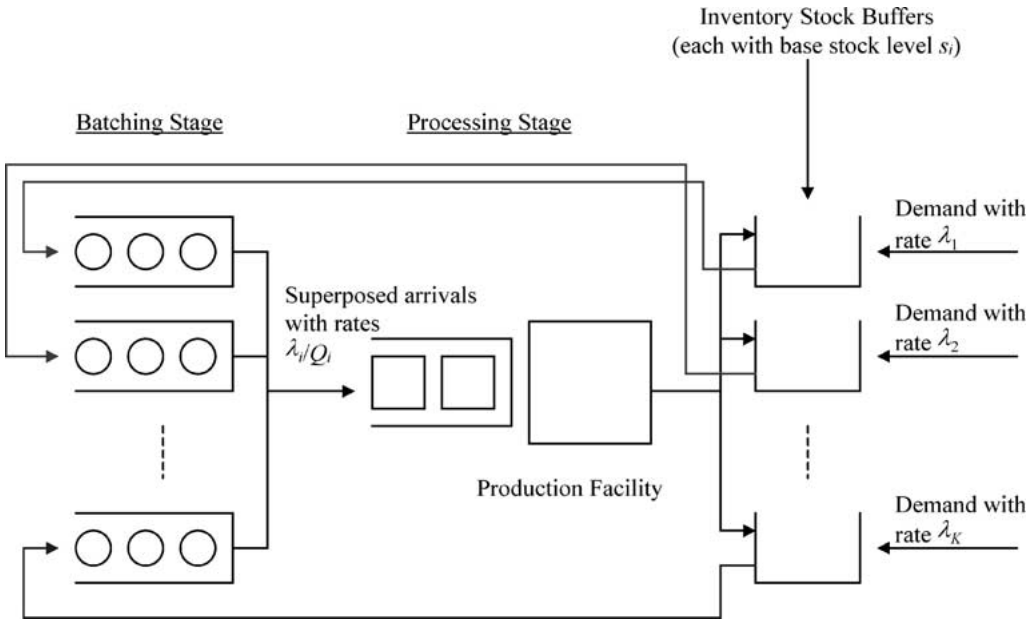

Figure 1. The production–inventory system.

where I_i and B_i are random variables equal in distribution to, respectively, the steady-state inventory and backorder levels for item i and $\mathbf{s}$ and $\mathbf{Q}$ are vectors representing product base stock levels and batch sizes. The use of a base-stock policy is in part justified by the fact that in most production–inventory systems, the transaction cost of communicating an order to the production system is small. Although not always optimal, base-stock policies are easy to implement and analyze and have well-known properties. They have been shown to be nearly optimal for a variety of conditions and are useful as benchmark for the amount of inventory needed if other system parameters are varied. Similarly, the use of a fixed batch is motivated by the prevalence of batching in practice and by the ease of implementing batching policies.

In order to streamline the presentation and isolate the effect of product variety, we shall momentarily assume that we have a symmetric system. In particular, we let $\lambda_i = \lambda/K$, $s_i = s$, $Q_i = Q$, $b_i = b$ and $h_i = h$ for $i = 1, \ldots, K$. Similarly, we assume identical distributions of demand inter-arrival times, setup times and production times among all products. Therefore we drop the index i from all associated notation i. The symmetry assumption allows us to explicitly examine the effect of K (the number of products) on various system characteristics. In section 3, we show how the analysis can be extended to general systems with asymmetric products.

When viewed in isolation, the processing stage of the production system forms a $GI/G/1$ queue with K batch arrival streams with rate λ/Q. Characterizing the probability distribution of queue size in a $GI/G/1$ is in general difficult. Therefore, we use a development described in Buzacott and Shanthikumar (1993) to approximate the probability distribution of batches in the processing stage using a geometric distribution of the

following form:

$$\Pr\left(N^p = n\right) \approx \begin{cases} 1-\rho, & n = 0, \\ \rho(1-\sigma)\sigma^{n-1}, & n = 1, 2, \ldots, \end{cases} \tag{3}$$

where N^p is the number of batches in the processing stage, $\sigma = (\hat{N} - \rho)/\hat{N}$, $\hat{N}$ is an approximation for the expected number of customers in a $GI/G/1$ queue, and ρ is the steady state utilization of the production facility. From Buzacott and Shanthikumar (1993), we also borrow the following approximation of the $GI/G/1$ expected waiting time in queue:

$$\hat{W} = \left(\frac{\rho^2(1+C_{s,p}^2)}{1+\rho^2 C_{s,p}^2}\right)\left(\frac{C_{a,p}^2 + \rho^2 C_{s,p}^2}{2\hat{\lambda}(1-\rho)}\right), \tag{4}$$

where $C_{a,p}^2$ is the squared coefficient of variation i inter-arrival times for the superposed arrival process at the processing stage and $C_{s,p}^2$ is the squared coefficient of variation for the effective service time. Then, by invoking Little's law, we obtain $\hat{N} = \hat{\lambda}\hat{W} + \rho$, where $\hat{\lambda} = \lambda/Q$.

The coefficient of variation of the arrival process to the production facility can be obtained by noting that it consists of the superposition of K renewal processes corresponding to the departure process from each batching stream. Since the departure of a batch coincides with the arrival of Q orders with i.i.d. inter-arrival times and coefficient of variation C_a (the coefficient of variation in the inter-arrival time of demand orders of each type), the distribution of batch inter-departure times for each stream has a coefficient of variation C_a/Q. Following the *asymptotic approach* suggested by Whitt (1982), we approximate the superposition of K identical renewal processes by a renewal process with the same coefficient variation, which leads to

$$C_{a,p} \approx \frac{C_a}{Q}. \tag{5}$$

Note that the superposition of renewal processes is not, in general, itself a renewal process; therefore, the arrival process of batches at the production stage will typically not be a renewal process. Nevertheless, we shall provide simulation results that help validate the above approximation in our production–inventory setting (see section 5 and tables 1–4).

To obtain $C_{s,p}^2$, we need to obtain the mean and variance of the effective service time of a batch. The processing of a batch has two components a setup time component and a service time component. A setup is incurred only when two consecutive batches are of a different type. Given the symmetry and the independence assumption with regard to the arrival process, the probability that the next batch to be processed is of the same type as the one that preceded it is $1/K$ (independent of the sequence of batches

already processed). Let U and W be random variables that denote respectively the setup time and process time experienced by a batch. Then, it is not difficult to see that

$$E(U) = \frac{(K-1)\tau}{K}, \qquad \mathrm{Var}(U) = \left(\frac{K-1}{K}\right)\left(\eta + \frac{\tau^2}{K}\right) \tag{6}$$

and

$$E(W) = Qt, \qquad \mathrm{Var}(W) = Q\theta, \tag{7}$$

where η and θ refer to the variance of production time and setup time, respectively. Letting S be a random variable that denotes batch service time, then $S = U + W$, and both mean and variance can be readily obtained, which can then be used to approximate the coefficient of variation in batch service time:

$$C_{s,p}^2 \approx \frac{\mathrm{Var}(S)}{E^2(S)} = \frac{(K-1)(K\eta + \tau^2) + K^2 Q\theta}{((K-1)\tau + QtK)^2}. \tag{8}$$

Note that in our treatment, we assume that batch service times are i.i.d., which is clearly not the case since only a fraction of the batches actually experiences a setup. However, as we show using simulation, this approximation still yields a reasonable estimate of steady-state performance measures of interest (see section 4 and tables 1–4).

Equations (6) and (7) can also be used to obtain the steady state utilization of the production system, which we denote by ρ, as follows

$$\rho = \frac{\lambda}{Q}\big(E(U) + E(W)\big) = \frac{\lambda(K-1)\tau}{QK} + \lambda t. \tag{9}$$

Since for stability, we must have $\rho < 1$, there is a minimum feasible batch size given by

$$Q > Q_{\min} = \frac{\lambda(K-1)\tau}{K(1-\lambda t)}. \tag{10}$$

We are now ready to state our first result.

Result 1. The distribution of the number of orders in the production system (including both the batching and processing stages) of product type i can be approximated as follows:

$$\Pr(N_i = n) \approx \begin{cases} \dfrac{1}{Q}\left(1 - \left(\dfrac{\rho}{\sigma}\right)r\right), & n = 0, 1, \ldots, Q-1, \\ \dfrac{1}{Q}\left(\dfrac{\rho}{\sigma}\right)(1-r)r^{\lfloor n/Q \rfloor}, & n = Q, Q+1, \ldots, \end{cases} \tag{11}$$

where $r = \sigma/(K(1-\sigma)+\sigma)$.

Proof. First, we obtain the probability distribution of the number of batches of type i in queue. The conditional probability $\Pr(N_i^p = n_i^p | N^p = n^p) \equiv p_i(n_i^p | n^p)$, where

$N^p = N_1^p + N_2^p + \cdots + N_N^p$, has a binomial distribution with parameter $1/K$. Hence, we have

$$p_i\left(n_i^p \middle| n^p\right) = \frac{n^p!}{n_i^p!(n^p - n_i^p)!}\left(\frac{1}{K}\right)^{n_i^p}\left(1 - \frac{1}{K}\right)^{n^p - n_i^p} \quad \forall n^p \geqslant n_i^p,$$

and

$$\begin{aligned} p_i\left(n_i^p\right) &= \sum_{n^p = n_i^p}^{\infty} p_i\left(n_i^p \middle| n^p\right) p\left(n^p\right) \\ &= \sum_{n^p = n_i^p}^{\infty} \frac{n^p!}{n_i^p!(n^p - n_i^p)!}\left(\frac{1}{K}\right)^{n_i^p}\left(1 - \frac{1}{K}\right)^{n^p - n_i^p} \rho(1-\sigma)\sigma^{n^p - 1}, \quad \text{for } n_i^p \geqslant 1. \end{aligned}$$

which can be rewritten as

$$p_i\left(n_i^p\right) = \frac{\rho(1-\sigma)}{\sigma}\left(\frac{1}{K}\right)^{n_i^p}\left(\frac{K}{K-1}\right)^{n_i^p} \sum_{n^p = n_i^p}^{\infty} \frac{n^p!}{n_i^p!(n^p - n_i^p)!}\left(\frac{(K-1)\sigma}{K}\right)^{n^p}.$$

Using the fact that

$$\sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} a^n = \frac{a^k}{(1-a)^{k+1}}$$

leads to

$$p_i\left(n_i^p\right) = \left(\frac{\rho}{\sigma}\right)\left(\frac{K(1-\sigma)}{K(1-\sigma)+\sigma}\right)\left(\frac{\sigma}{K(1-\sigma)+\sigma}\right)^{n_i^p}.$$

Letting

$$r = \frac{\sigma}{K(1-\sigma)+\sigma}$$

leads to

$$p_i\left(n_i^p\right) = \left(\frac{\rho}{\sigma}\right)(1-r)r^{n_i^p}.$$

For $n_i^p = 0$, we have

$$p_i\left(n_i^p = 0\right) = 1 - \sum_{n_i^p = 1}^{\infty} p_i\left(n_i^p\right) = 1 - \left(\frac{\rho}{\sigma}\right)r.$$

□

In order to relate the total number of orders in the system to the number of units in the batching and production stages, we note that when there is less than Q orders of type i in the production system, then these orders are in the batching stage. If there are $N_i > Q$ orders of type i, then $Q\lfloor N_i/Q \rfloor$ are in the production stage and the remaining

are in the batching stage. Assuming independence between the batching and production stages, we obtain:

$$\begin{aligned}\Pr(N_i = n) &\approx \Pr\big(N_i^b = n\big)\Pr\big(N_i^p = 0\big)\\ &= \frac{1}{Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right), \quad n = 0, 1, \ldots, Q-1,\end{aligned}$$

and

$$\begin{aligned}\Pr(N_i = n) &\approx \Pr\left(N_i^b = n - \left\lfloor \frac{n}{Q} \right\rfloor Q\right)\Pr\left(N_i^p = \left\lfloor \frac{n}{Q} \right\rfloor\right)\\ &= \frac{1}{Q}\left(\frac{\rho}{\sigma}\right)(1-r)r^{\lfloor n/Q \rfloor}, \quad n = Q, Q+1, \ldots.\end{aligned}$$

The independence assumption is appropriate under heavy traffic ($\rho \to 1$) where the number of orders in the batching stage represents only a small fraction of total number of orders in the system. Simulation suggests that the amount of error introduced is also small for low and moderate utilization (see section 5 and tables 1–3).

Using the estimated distribution of number of units in the system, we can now obtain performance measures of interest including average inventory and average backorders for each product type.

Result 2. Expected inventory and backorders levels for each product i, $E(I_i)$ and $E(B_i)$, can be approximated as follows:

$$E(I_i) \approx E\big(\tilde{I}_i\big) = \begin{cases} \dfrac{s(s+1)}{2Q}\left(1 - \left(\dfrac{\rho}{\sigma}\right)r\right), & \text{if } Q \geqslant s,\\ s - \dfrac{Q-1}{2} - \left(\dfrac{\rho}{\sigma}\right)\left(\dfrac{Qr}{1-r}\right)\\ \quad + \left(\dfrac{\rho}{\sigma}\right)r^k\left[Q\left(k + \dfrac{r}{1-r} + \dfrac{1}{2}\right) - \left(s + \dfrac{1}{2}\right)\right.\\ \qquad \left. + \left(\dfrac{s^2+s}{2Q} + \dfrac{k^2Q - 2ks - k}{2}\right)(1-r)\right], & \text{if } Q < s, \end{cases}$$

and

$$E(B_i) \approx E\big(\tilde{B}_i\big) = \begin{cases} \dfrac{(Q-s-1)(Q-s)}{2Q} - \left(\dfrac{\rho}{\sigma}\right)r\left(\dfrac{s^2+s}{2Q} - \dfrac{Q}{1-r}\right), & \text{if } Q > s,\\ \left(\dfrac{\rho}{\sigma}\right)r^k\left[\dfrac{((k+1)Q - 1 - s)((k+1)Q - s)(1-r)}{2Q}\right.\\ \qquad \left. + \left(kQ + \dfrac{Q-1}{2} - s + \dfrac{Q}{1-r}\right)r\right], & \text{if } Q \leqslant s, \end{cases}$$

where $k = \lfloor s/Q \rfloor$.

Proof. For $Q > S$,

$$\begin{aligned} E(I_i) &= \sum_{n=0}^{s}(s-n)\Pr(N_i = n) \\ &\approx \big[s + (s-1) + (s-2) + \cdots + 1\big]\frac{1}{Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right) \\ &= \frac{S(S+1)}{2Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right), \end{aligned}$$

and

$$\begin{aligned} E(B_i) &= \sum_{n=s}^{\infty}(n-s)\Pr(N_i = n) \\ &\approx \big[1 + 2 + \cdots + (Q-1-s)\big]\frac{1}{Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right) \\ &\quad + \big[(Q-s) + \cdots + (2Q-1-s)\big]\frac{1}{Q}\left(\frac{\rho}{\sigma}\right)(1-r)r \\ &\quad + \big[(2Q-s) + \cdots + (3Q-1-s)\big]\frac{1}{Q}\left(\frac{\rho}{\sigma}\right)(1-r)r^2 + \cdots \\ &= \frac{(Q-S-1)(Q-s)}{2Q} - \left(\frac{\rho}{\sigma}\right)r\left(\frac{s^2+s}{2Q} - \frac{Q}{1-r}\right). \end{aligned}$$

For $Q < S$, let $k = \lfloor s/Q \rfloor$,

$$\begin{aligned} E(I_i) &= \sum_{n=0}^{s}(s-n)\Pr(N_i = n) \\ &= \sum_{n=0}^{Q-1}(s-n)\Pr(N_i = n) + \sum_{n=Q}^{s}(s-n)P(N_i = n) \\ &\approx \big[s + (s-1) + \cdots + (s-Q+1)\big]\frac{1}{Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right) \\ &\quad + \big[(s-Q) + (s-Q-1) + \cdots + (s-2Q+1)\big]\frac{1}{Q}\left(\frac{\rho}{\sigma}\right)(1-r)r \\ &\quad + \big[(s-2Q) + (s-2Q-1) + \cdots + (s-3Q+1)\big]\frac{1}{Q}\left(\frac{\rho}{\sigma}\right)(1-r)r^2 \\ &\quad + \cdots \\ &\quad + \big[(s-kQ) + (s-kQ-1) + \cdots + 1\big]\frac{1}{Q}\left(\frac{\rho}{\sigma}\right)(1-r)r^k \\ &= \left[s - \frac{Q-1}{2}\right] - \left(\frac{\rho}{\sigma}\right)\left(\frac{Qr}{1-r}\right) \end{aligned}$$

$$+\left(\frac{\rho}{\sigma}\right)r^k\left[Q\left(k+\frac{r}{1-r}+\frac{1}{2}\right)-\left(s+\frac{1}{2}\right)\right.$$
$$\left.+\left(\frac{s^2+s}{2Q}+\frac{k^2Q-2ks-k}{2}\right)(1-r)\right],$$

and

$$E(B_i)=\sum_{n=s}^{\infty}(n-s)P(N_i=n)\approx\left(\frac{\rho}{\sigma}\right)r^k\left[\frac{((k+1)Q-1-s)((k+1)Q-s)(1-r)}{2Q}\right.$$
$$\left.+\left(kQ+\frac{Q-1}{2}-s+\frac{Q}{1-r}\right)r\right]$$

Finally, for $Q=s$,

$$\begin{aligned}E(I_i)&=\sum_{n=0}^{s}(s-n)P(N_i=n)\\&\approx\left[s+(s-1)+(s-2)+\cdots+1\right]\frac{1}{Q}\left(1-\left(\frac{\rho}{\sigma}\right)r\right)\\&=\frac{s(s+1)}{2Q}\left(1-\left(\frac{\rho}{\sigma}\right)r\right),\quad\text{and}\\E(B_i)&=\sum_{n=s}^{\infty}(n-s)P(n_i=s)\\&=P(n_i=s+1)+2P(n_i=S+2)+3P(n_i=S+3)+\cdots\\&\approx\left(\frac{\rho}{\sigma}\right)r^k\left[\frac{((k+1)Q-1-S)((k+1)Q-S)(1-r)}{2Q}\right.\\&\qquad\left.+\left(kQ+\frac{Q-1}{2}-S+\frac{Q}{1-r}\right)r\right].\end{aligned}$$

□

Result 3. Let $\tilde{z}(s,Q)=\sum_{i=1}^{K}\{hE(I_i)+bE(B_i)\}$ refer to the estimated expected total cost, then for fixed Q, $\tilde{z}$ is convex in s.

Proof. For $Q>s$,

$$\frac{\tilde{z}(s+1,Q)+\tilde{z}(s-1,Q)}{2}-\tilde{z}(s,Q)=\frac{K(h+b)}{2Q}\left(1-\left(\frac{\rho}{\sigma}\right)r\right)\geqslant 0.$$

For $Q<s$,

$$\frac{\tilde{z}(s+1,Q)+\tilde{z}(s-1,Q)}{2}-\tilde{z}(s,Q)=\frac{K(h+b)}{2Q}\left(\frac{\rho}{\sigma}\right)(1-r)r^k\geqslant 0.$$

For $Q = s$,

$$\frac{\tilde{z}(s+1, Q) + \tilde{z}(s-1, Q)}{2} - \tilde{z}(s, Q) = \frac{K}{Q}\left[hs\left(1 - \frac{\rho r}{\sigma}\right) + \frac{b}{2}\left(\frac{\rho}{\sigma}\right)(1-r)r^k\right] \geqslant 0.$$

Hence proved. □

Result 3 ensures that a search for the base-stock level that minimizes $\tilde{z}$ is computationally efficient. In general, a closed form solution for this optimal base stock level is difficult to obtain. However, this is possible when the condition $s < Q$ applies.

Result 4. For $s < Q$, a base-stock level that minimizes $\tilde{z}$ is given by

$$s^* = \begin{cases} \left\lfloor \left(\frac{Qb}{h+b}\right)\left(\frac{\sigma}{\sigma - \rho r}\right) \right\rfloor & \text{if } \gamma < 1 - \left(\frac{\rho}{\sigma}\right)r, \\ Q - 1 & \text{otherwise,} \end{cases} \tag{12}$$

where $\gamma = b/(b+h)$.

Proof. Let $\Delta\tilde{z} = \tilde{z}(s, Q) - \tilde{z}(s-1, Q)$, then

$$\begin{aligned} \Delta\tilde{z} = K\Bigg[\frac{s(s+1)}{2Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right)h + \frac{(Q-s-1)(Q-s)}{2Q}b \\ + \left(\frac{\rho r}{\sigma}\right)\left(\frac{Q}{1-r} - \frac{s(s+1)}{2Q}\right)b\Bigg] \\ - K\Bigg[\frac{s(s-1)}{2Q}\left(1 - \left(\frac{\rho}{\sigma}\right)r\right)h + \frac{(Q-s+1)(Q-s)}{2Q}b \\ + \left(\frac{\rho r}{\sigma}\right)\left(\frac{Q}{1-r} - \frac{s(s-1)}{2Q}\right)b\Bigg] \end{aligned}$$

which can be rewritten as

$$\Delta\tilde{z} = K\left[\left(\frac{s}{Q}\right)(h+b)\left(1 - \left(\frac{\rho}{\sigma}\right)r\right) - b\right].$$

Setting $\Delta\tilde{z} = 0$, leads to

$$s^* = \left(\frac{bQ}{h+b}\right) \Big/ \left(1 - \left(\frac{\rho}{\sigma}\right)r\right)$$

since the condition $s^* < Q$ is equivalent to

$$\gamma \equiv \frac{b}{h+b} < 1 - \left(\frac{\rho}{\sigma}\right)r$$

leads to the desired result. □

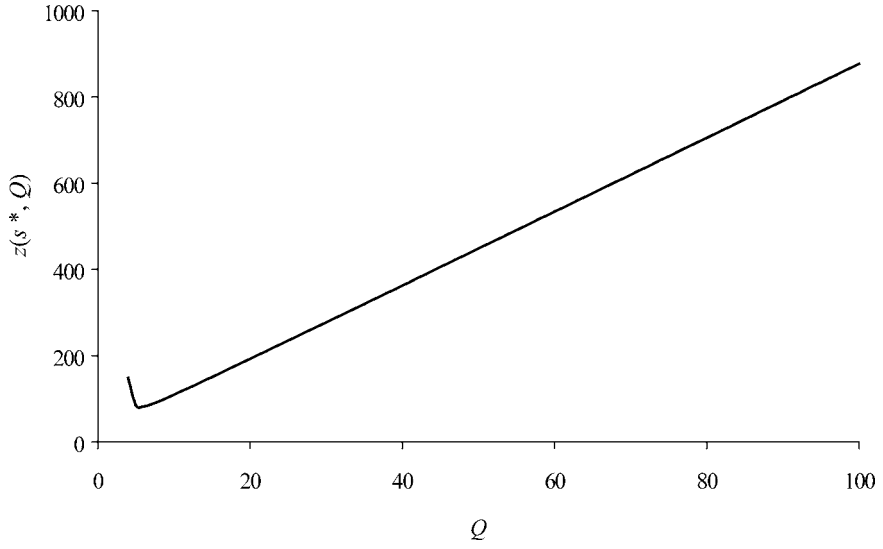

Figure 2. The effect of batch size on total cost ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$, $\tau = 1.0$, $K = 10$).

The cost function $\tilde{z}$ is generally not convex in Q, although we observe that it consistently assumes the convex-like U-shape of figure 2. As $Q \to \infty$, we can also show that $\tilde{z} \to \infty$. Given the convexity of $\tilde{z}$ in s, a search, even over a wide range of values of Q, can be carried out very efficiently.

We should note that in certain environments a separate direct cost may be associated with each setup. In that case, the cost function should be modified to include an additional component of the form $\nu(K-1)\lambda/KQ$, where ν is the cost per setup. All of our analysis remains valid, including results 4 and 5. Since the setup cost component is decreasing and convex in Q with a limiting value of zero, a search for Q can still be carried out efficiently. Finally, we note that in systems where there is a setup cost but no setup time, our model captures the dynamics of a production–inventory system where inventory is managed according to a (Q, r) policy with $r = s - Q$.

3. The general case

In this section, we show how the analysis can be extended to asymmetric systems. We allow products to have non-identical distributions of order inter-arrival, setup, and production times, and to have non-identical holding and backordering costs. Products may also have different base-stock levels and batch sizes. The production system can be viewed as a $GI/G/1$ queue whose arrival process is now the superposition of K non-identical renewal processes, each with an arrival rate λ_i/Q_i and squared coefficient of variation C_{a_i}/Q_i. Again, following Whitt (1982), we approximate the superposition of

K renewal processes by a renewal process whose squared coefficient of variation is the convex combination

$$C_{a,p}^2 = \sum_{i=1}^{K} p_i C_{a_i}^2, \quad \text{where } p_i = \frac{\lambda_i/Q_i}{\sum_{i=1}^{K} \lambda_i/Q_i} \tag{13}$$

represents the probability that an arrival is of type i. The approximation is exact when the individual renewal processes are Poisson and is asymptotically exact when the utilization of the queue is close to 1. Alternative approximations are discussed in Albin (1984, 1986) and could be used instead without affecting the subsequent analysis.

Noting that the probability that a batch of type i experiences a setup (a random variable with mean τ_i and variance η_i) is $1 - p_i$, we can obtain the first and second moments of setup time experienced by an arbitrary batch as

$$E(U) = \sum_{i=1}^{K} p_i(1-p_i)\tau_i \quad \text{and} \quad E\big(U^2\big) = \sum_{i=1}^{K} p_i(1-p_i)\big(\eta_i + \tau_i^2\big). \tag{14}$$

Similarly, we can obtain the first two moments of processing time for an arbitrary batch as

$$E(W) = \sum_{i=1}^{K} p_i Q_i t_i \quad \text{and} \quad E\big(W^2\big) = \sum_{i=1}^{K} p_i\big(Q_i\theta_i + Q_i^2 t_i^2\big). \tag{15}$$

From (14) and (15), we can now readily obtain the first two moments of the effective batch service time S ($S = U+W$) of an arbitrary batch, from which we can then compute the corresponding coefficient of variation $C_{s,p}$. The utilization of the production system is given by

$$\rho = \sum_{i=1}^{K}\left(\frac{\lambda_i}{Q_i}\right)E(S).$$

In order to characterize the distribution of the number of batches in the system, we retain the geometric approximation in (3), which we use to obtain the marginal distribution of the number of batches of each type in the system. First we note that the conditional probability $p_i(n_i^p|n^p)$ (the probability of having n_i^p batches of type i given a total of n^p batches in the system), has a binomial distribution with parameters p_i. Thus, we have

$$p_i\big(n_i^p\big|n^p\big) = \frac{n^p!}{n_i^p!(n^p - n_i^p)!} p_i^{n_i^p}(1-p_i)^{n^p - n_i^p} \quad \forall n^p \geqslant n_i^p \tag{16}$$

from which, after much algebra, we obtain

$$p_i\big(n_i^p\big) = \left(\frac{\rho}{\sigma}\right)(1-r_i)r_i^{n_i^p} \quad \text{for } n_i^p \geqslant 1, \tag{17}$$

and

$$p_i(0) = 1 - \left(\frac{\rho}{\sigma}\right) r_i, \tag{18}$$

where

$$r_i = \frac{p_i \sigma}{1 - \sigma(1 - p_i)}, \tag{19}$$

and σ is obtained from (4) as in the symmetric case. From (1), (17) and (18), we obtain the (approximate) distribution of the number of orders in the production system (including both the batching and processing stages) of product i as

$$\Pr(N_i = n) \approx \begin{cases} \dfrac{1}{Q_i}\left(1 - \left(\dfrac{\rho}{\sigma}\right) r_i\right), & n = 0, 1, \ldots, Q_i - 1, \\ \dfrac{1}{Q_i}\left(\dfrac{\rho}{\sigma}\right)(1 - r_i) r_i^{\lfloor n/Q \rfloor}, & n = Q_i, Q_i + 1, \ldots. \end{cases}$$

The rest of the analysis follows as in the symmetric case. Results similar to results 2–4 can be obtained by simply substituting Q_i, s_i and r_i for Q, s and r, respectively. For brevity the details are omitted.

4. Numerical results and managerial insights

In this section, we generate numerical results to examine the effect of product variety on system performance and study the relationship between product variety and various system parameters. We are particularly interested in examining the benefits of variety-reducing strategies, such as product standardization and delayed product differentiation. Unless otherwise specified, we shall assume a symmetric system with the following parameter values: $h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$, $\tau = 15.0$, $C_a = 1.0$, $C_{\text{process}} = 1.0$, and $C_{\text{setup}} = 1.0$, where C_{process} is the coefficient of variation in unit processing time and C_{setup} is the coefficient of variation in setup time.

Values for the estimated optimal total cost $\tilde{z}^*$ are obtained for different aggregate demand rates ($\lambda = 0.3, 0.6, 0.8, 0.9$), unit production times ($t = 0.2, 0.5, 0.8, 1.0$), setup times ($\tau = 1.0, 5.0, 10.0, 15.0$), coefficients of variation in demand inter-arrival time ($C_a = 0, 0.5, 1.0, 1.5, 2.0$), coefficients of variation in unit production time ($C_{\text{process}} = 0, 0.5, 1.0, 1.5, 2.0$), coefficient of variation in setup time ($C_{\text{setup}} = 0, 0.5, 1.0, 1.5, 2.0$), and number of items (1–50).

Observation 1. $\tilde{z}^*$ is increasing in K. For $K \gg 1$, the increase is almost linear in K. The rate of increase is increasing in λ, t, τ, C_a, C_{process}, and C_{setup}.

Supporting results are shown in figures 3–8. These results appear to support the argument that pure inventory models, where cost increases proportionally to the square root of K, would under-estimate the cost of greater variety. Pure inventory models would

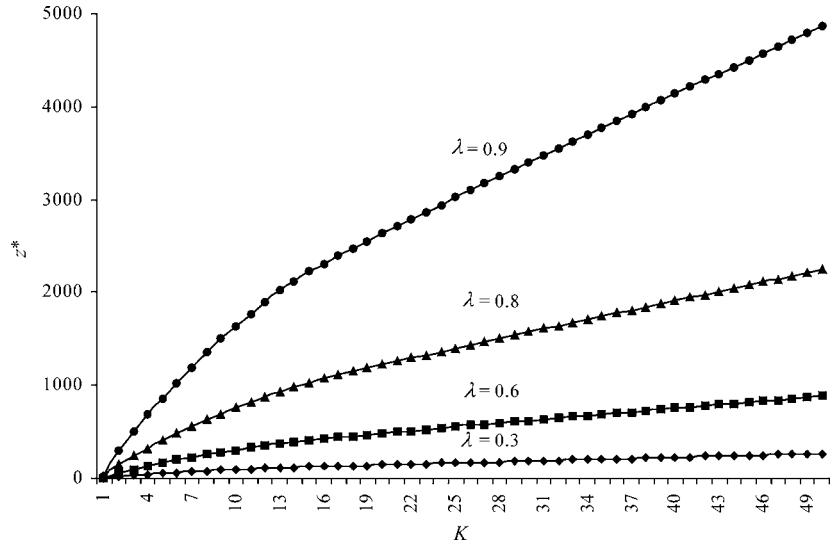

Figure 3. The impact of product variety on the optimal total cost for different demand rates ($h = 1$, $b = 10$, $t = 1.0$, $\tau = 15.0$).

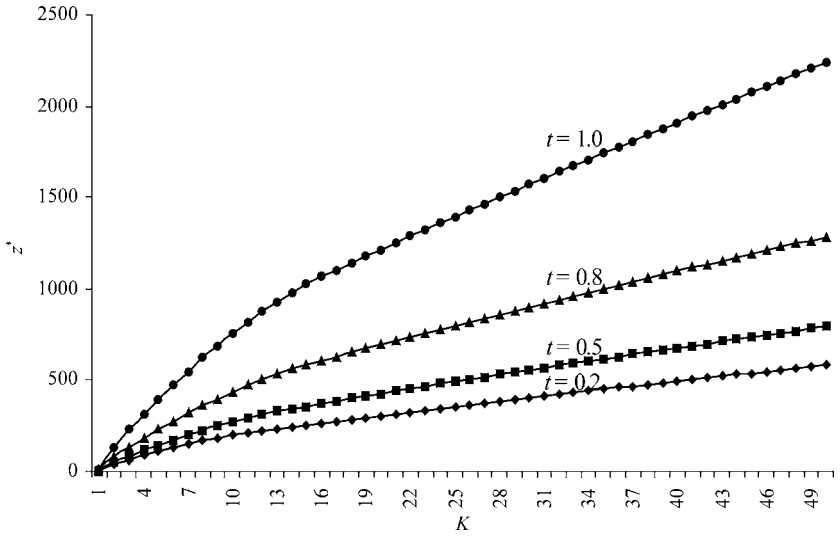

Figure 4. The impact of product variety on the optimal cost for different values of production time ($h = 1$, $b = 10$, $\lambda = 0.8$, $\tau = 15.0$).

also ignore the sensitivity of the rate of increase to system parameters, particularly those affecting system utilization and demand and process variability.

In order to assess the cost of variety relative to a strategy of product consolidation, we examine the ratio of the estimated optimal total cost in a system with K items to one with a single item while maintaining the same overall demand rate and use the

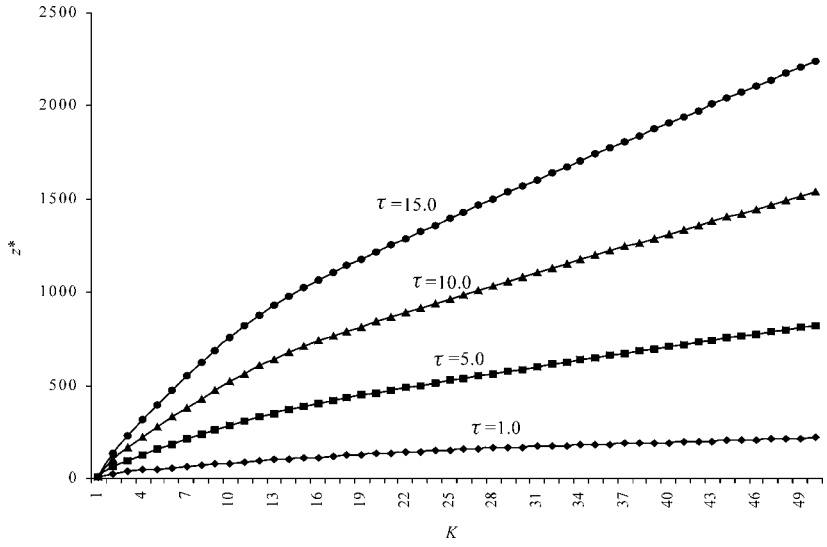

Figure 5. The impact of product variety on the optimal cost for different values of setup time ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$).

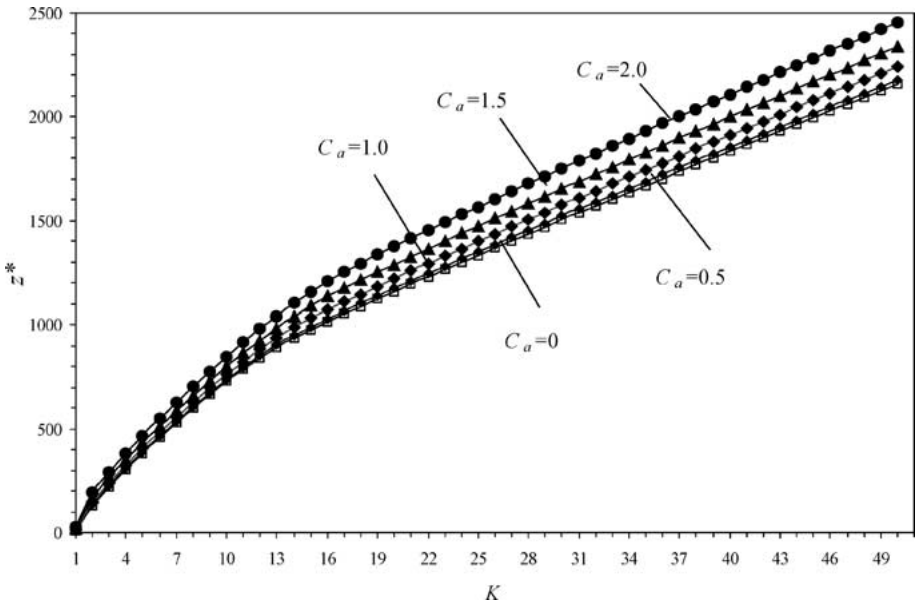

Figure 6. The impact of product variety on the total cost for different levels of demand variability ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$, $\tau = 15.0$).

notation $\delta = \tilde{z}^*_{(K)}/\tilde{z}^*_{(1)}$ to denote this ratio. We are especially interested in examining how different parameters affect the relative cost of offering greater variety; or equivalently the relative advantage of increasing product consolidation.

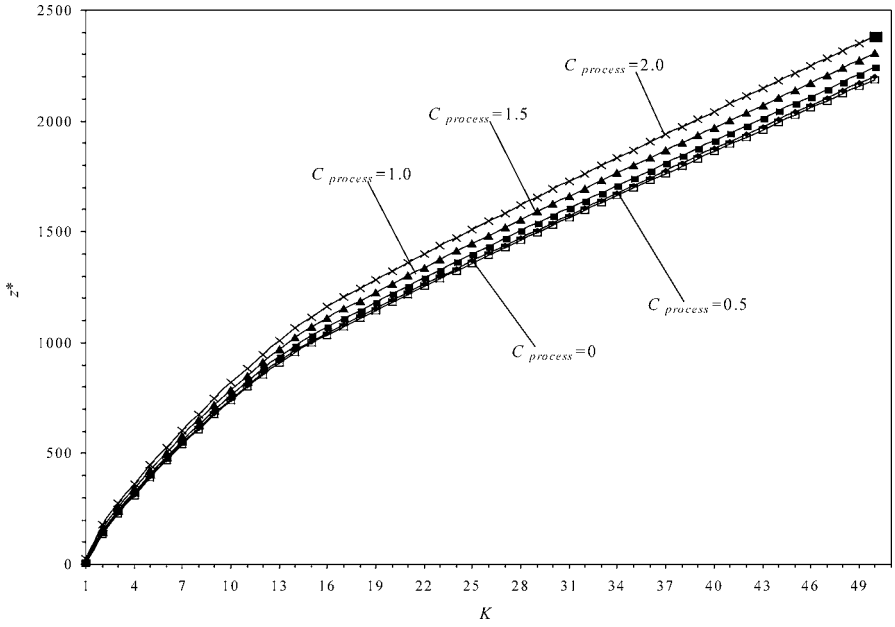

Figure 7. The impact of product variety on the total cost for different levels of processing time variability ($h = 1, b = 10, \lambda = 0.8, t = 1.0, \tau = 15.0$).

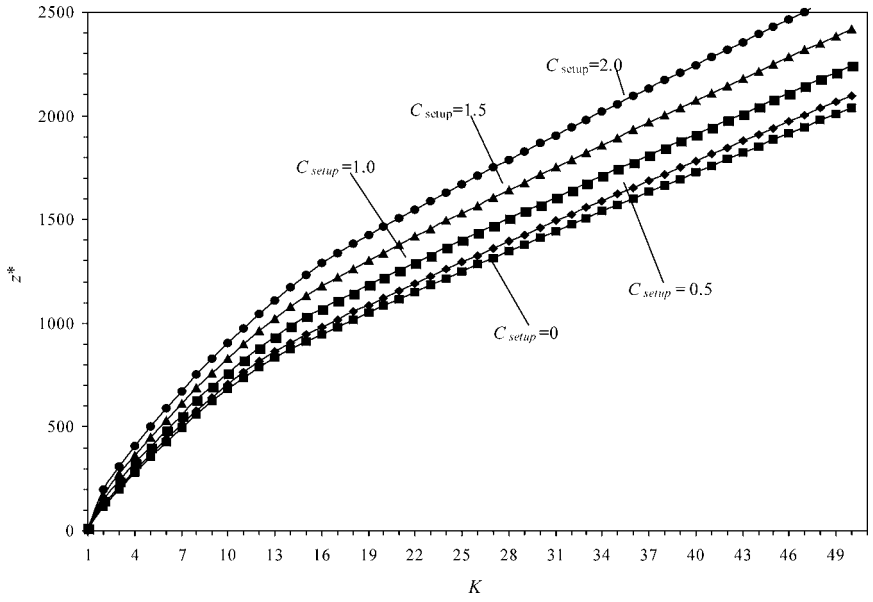

Figure 8. The impact of product variety on the total cost for different levels of setup variability ($h = 1$, $b = 10, \lambda = 0.8, t = 1.0, \tau = 15.0$).

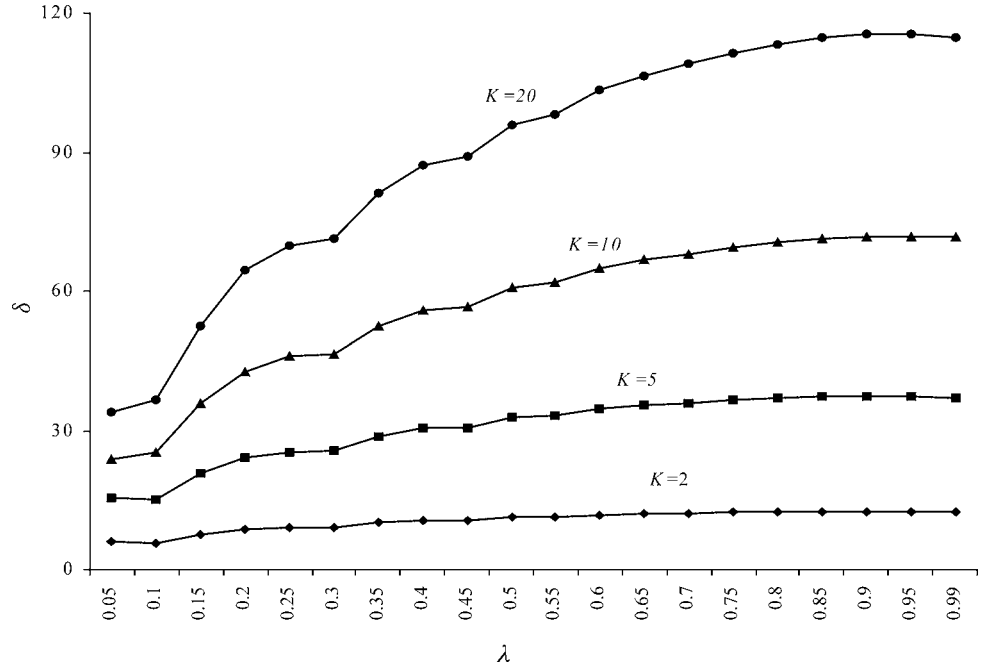

Figure 9. The effect of demand rate on the relative advantage of pooling ($h = 1$, $b = 10$, $\tau = 15.0$, $t = 1.0$).

Observation 2. The ratio δ is not monotonic in λ, although it is generally increasing with a finite limit when $\lambda t \to 1$.

Sample data illustrating observation 2 are shown in figure 9. Note that for relatively large λ, δ is relatively insensitive to changes in λ. Observation 2 suggests that, as demand increases, the relative advantage of product consolidation increases.

Surprisingly, the effect of expected production time t is different from that of the demand rate λ. In fact, as stated in observation 3 and illustrated in figure 10, δ is generally decreasing in t.

Observation 3. The ratio δ is not monotonic in t, although it is generally decreasing with a finite limit when $\lambda t \to 1$.

The above result suggests that increases in process efficiency make it relatively more desirable to consolidate products and offer less variety. Counter to intuition, this also means that the relative cost of offering variety is smaller in process-inefficient systems. The difference in the effect of λ and t can be, in part, explained by the difference in the way each parameter affects utilization of the production system. If we rewrite expression (9) as $\rho = \rho_{\text{setup}} + \rho_{\text{process}}$, where $\rho_{\text{setup}} = \lambda(K-1)\tau/QK$ and $\rho_{\text{process}} = \lambda t$, then we can see that an increase in λ affects both components of utilization, the one due to setups and the one due to processing, while an increase in t affects only the utiliza-

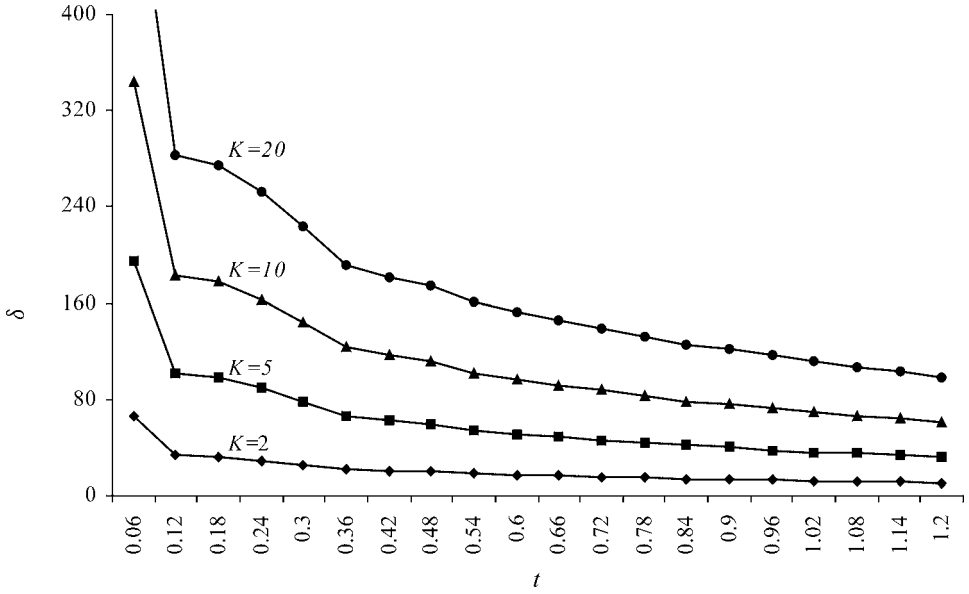

Figure 10. The effect of production time on the relative advantage of pooling ($h = 1$, $b = 10$, $\lambda = 0.8$, $\tau = 15.0$).

tion due to processing, which is independent of K. These results seem to support those observed in Benjaafar, Cooper, and Kim (2003) who study a perfectly flexible system with no setups and show that an increase in processing utilization tends to decrease the relative cost advantage of inventory pooling.

Observation 4. The ratio δ is increasing in τ, with the increase being approximately linear in τ.

Observation 4 is illustrated with sample data in figure 11. In line with intuition, the result suggests that variety becomes relatively more expensive as setup time increases. Together, observations 2–4 highlight the fact that parameters that impact utilization do not necessarily have the same effect on δ. In particular, the relative increase in cost due to higher variety is significant when setup times are high but can be relatively small when production times are long.

Observation 5. The ratio δ is decreasing in the variability of demand and production time but increasing in the variability of setup time.

Figures 12–14 provide supporting data for the result. The fact that δ is decreasing in the variability of demand and production time is surprising. It means that offering more variety is relatively less expensive when demand and process variability is high. It also means that strategies of product consolidation or delayed differentiation are relatively less valuable when demand variability is high. This is different from results

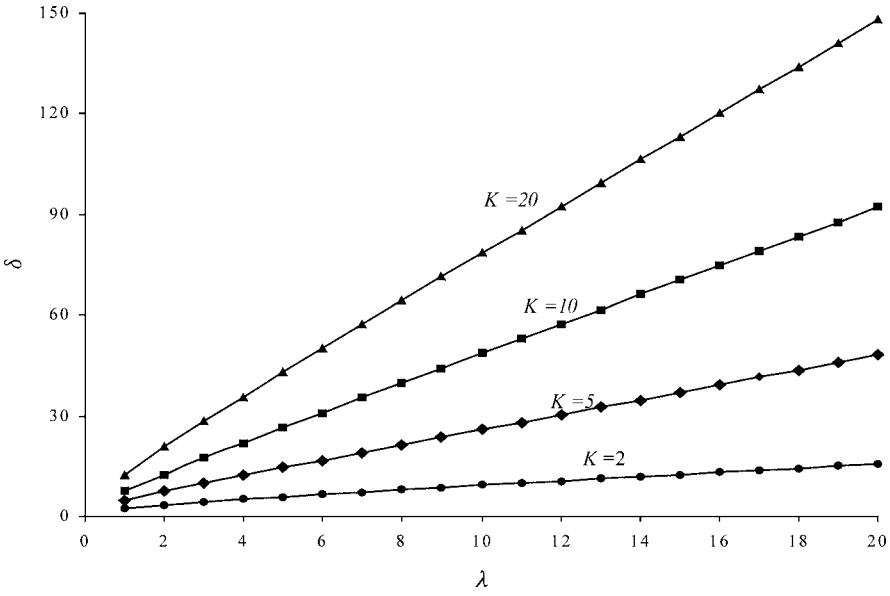

Figure 11. The effect of setup time on the relative advantage of pooling ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$).

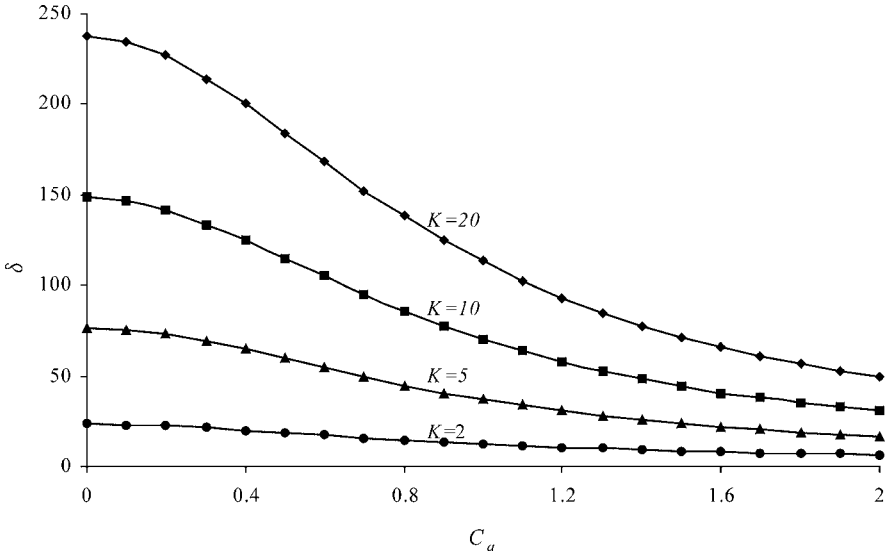

Figure 12. The effect of demand variability on the relative advantage of pooling ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$, $\tau = 15.0$).

obtained for inventory systems with exogenous lead times where it can be shown that the value of variety reduction (or demand pooling) increases when either demand or leadtime variability increases (Eppen, 1979). Interestingly, the effect of setup time variability is different from that of demand and process variability. Higher variability in this

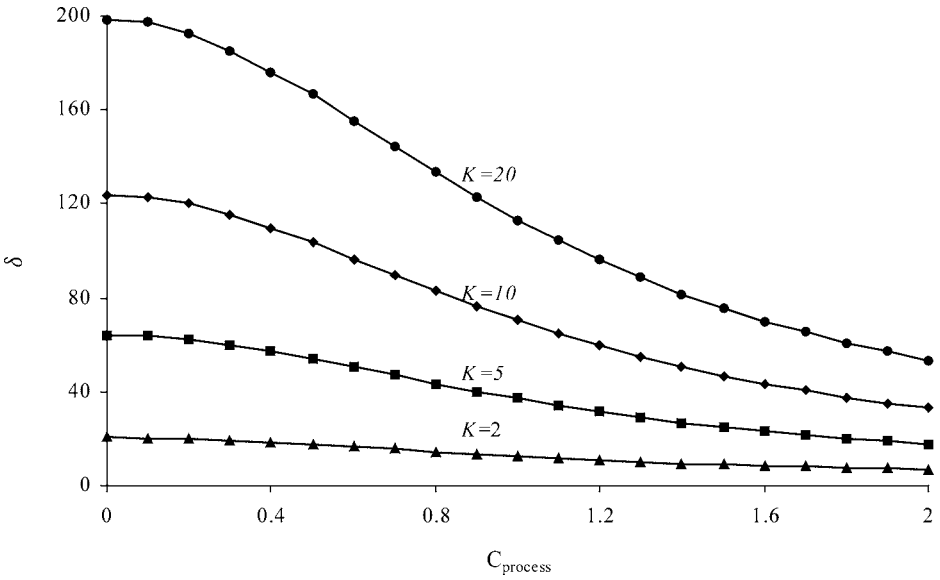

Figure 13. The effect of production time variability on the relative advantage of pooling ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$, $\tau = 15.0$).

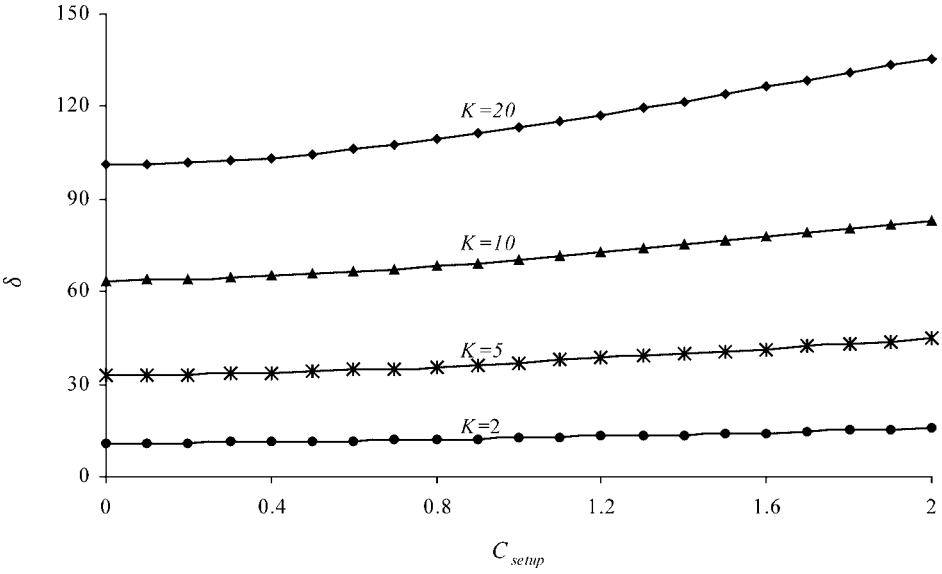

Figure 14. The effect of setup time variability on the relative advantage of pooling ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$, $\tau = 15.0$).

parameter tends to increase the relative cost of increased variety.

An intuitive explanation for these effects is difficult due to the complex relationship between various parameters. However, they appear to be related to the way the

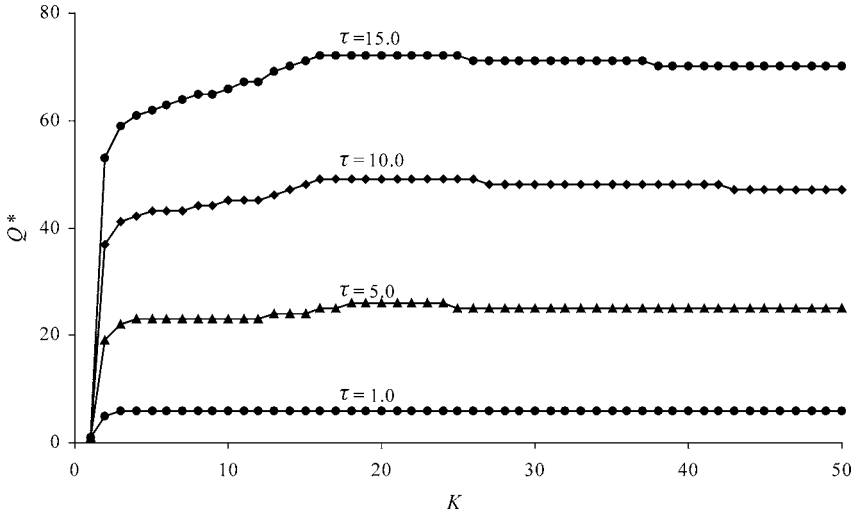

Figure 15. The impact of product variety on the optimal batch size ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$).

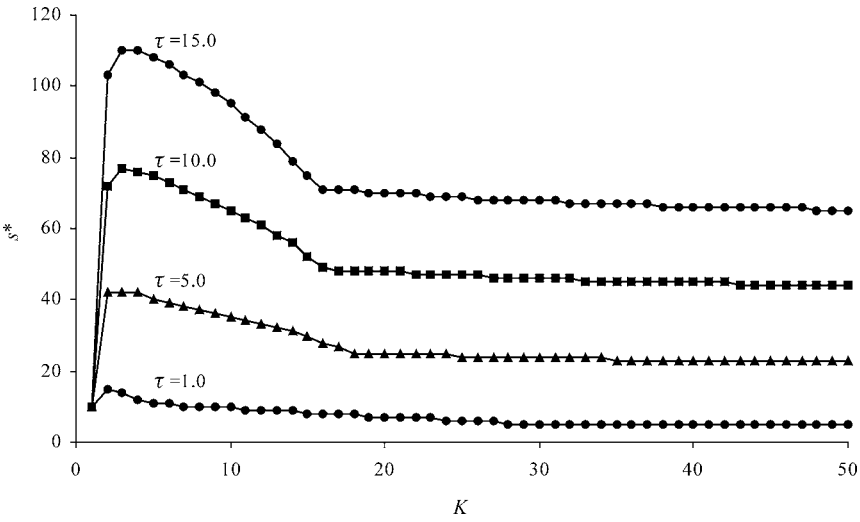

Figure 16. The impact of product variety on the optimal base stock level ($h = 1$, $b = 10$, $\lambda = 0.8$, $t = 1.0$).

distribution of *leadtime demand* for each product is affected by changes in these parameters. (Lead time demand is the amount of demand that arrives from the time an order is placed with the production system until it is delivered). Higher demand or process variability induces more congestion in the production system and more correlation in the lead times experienced by different products. Hence, the advantage of the statistical economies achieved by product consolidation diminishes when this variability is increased. A discussion of these effects in systems with no setups and no batching can be found in Benjaafar, Cooper, and Kim (2003).

Finally, we examine the effect of product variety on the control variables Q and s. The effect of K on Q and s is shown in figures 15 and 16. As we can see, initial increases in K, initially lead to an increase in Q. However, this eventually levels off and Q becomes relatively insensitive to increases in K. This effect can be explained by recalling that

$$\rho = \frac{\lambda(K-1)\tau}{QK} + \lambda t.$$

Clearly, when K is large, ρ becomes mostly independent of K. The effect of K on the optimal-base-stock level is less predictable and this is due to the complex interaction between Q and s. However, for sufficiently large K, s^* is generally decreasing in K.

We should note that the preceding results rely upon the assumption of first-come first-served (FCFS) sequencing of batches at the production process. Although FCFS is, in general, not optimal, it is widely used in practice for its simplicity and perceived fairness. Furthermore, batching is often used as a substitute for dynamic sequencing when real time control is not feasible or expensive to implement. Characterizing an optimal policy (especially in asymmetric systems) is a difficult problem that to date remains unresolved even for the simpler case of zero setup times; see de Vericourt, Karaesmen, and Dallery (2000) for results and references. Total cost under the FCFS policy can be viewed as an upper bound for total cost under an optimal policy, while a lower bound is given by the case of zero setup times and no batching. For the latter case, we can show that the relative advantage of production consolidation still diminishes with increases in process utilization and demand and process variability, with the cost ratios approaching one in the limit cases. Taken together, the behavior of the lower and upper bound cases seems to suggest that, at least qualitatively, the results presented here would not be fundamentally different under an optimal policy.

5. Simulation results

The numerical results of the previous section were validated using computer simulation. A small but representative sample of these simulation results is shown in tables 1–4. Comparisons with the analytical model are also provided. The simulation model is identical to the analytical model except that we do away with approximations regarding the distribution of batch arrivals and batch processing times at the production system. In particular, we let the arrival process of batches be determined by the arrival of individual orders. Similarly, we do not assume i.i.d. batch setup and processing times. Instead, we allow a setup time to be incurred only if the preceding batch on the production system is of a different type. Similarly, we simulate the individual production times for each item in the batch. We collect statistics directly on various performance measures including inventory and backorder levels. We use Gamma distributions to model inter-arrival, processing, and setup times. In order to vary the coefficient of variation while maintaining the same mean, we let the parameters of the Gamma distribution be α/m and β/m. This allows us to fix the mean at α/β and vary its coefficient of variation by varying

Table 1
Simulation results for systems with Gamma-distributed order inter-arrival times ($\alpha = 1$, $\beta = \lambda/10$) and exponentially-distributed processing times and setup times ($t = 1$, $\tau = 1$).

C_a^2	Simulated total cost $K = 1$			Simulated total cost $K = 10$			Ratio		
	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$
0.01	2.02	4.98	10.7	21	60.9	140.1	10.40	12.23	13.09
	[0.001]	[0.02]	[0.08]	[0.06]	[0.51]	[1.56]	12.17	13.15	13.42
0.20	2.72	6.01	12.8	23.5	68.7	146.8	8.64	11.43	11.47
	[0.006]	[0.05]	[0.11]	[0.05]	[0.78]	[2.56]	9.67	11.23	11.74
0.40	3.08	7.09	14.9	24.6	71.7	159.3	7.99	10.11	10.69
	[0.005]	[0.03]	[0.10]	[0.05]	[0.45]	[4.88]	8.57	9.96	10.41
0.60	3.6	8.2	17.1	25.6	73.3	166.2	7.11	8.94	9.72
	[0.005]	[0.03]	[0.20]	[0.08]	[0.43]	[2.75]	7.61	9.11	9.41
0.80	4.18	9.6	19.8	27.9	76.8	172.7	6.67	8.00	8.72
	[0.008]	[0.06]	[0.20]	[0.19]	[1.19]	[2.75]	7.05	8.30	8.67
1.00	4.6	10.7	22.4	30	79.7	179.5	6.52	7.45	8.01
	[0.010]	[0.08]	[0.18]	[0.15]	[1.25]	[2.84]	6.67	7.67	8.07
1.20	5.06	11.4	23.8	31.3	82.3	185.4	6.19	7.22	7.79
	[0.010]	[0.05]	[0.34]	[0.17]	[0.50]	[1.70]	6.35	7.12	7.54
1.40	5.48	12.8	27	33.3	86.2	190.1	6.08	6.73	7.04
	[0.020]	[0.25]	[0.41]	[0.22]	[1.04]	[2.25]	6.20	6.69	7.12
1.60	5.89	13.7	28.6	36.2	89.8	201.2	6.15	6.55	7.03
	[0.010]	[0.06]	[0.24]	[0.17]	[1.21]	[3.47]	5.92	6.34	6.77
1.80	6.33	15	30.6	37.3	95.9	205.4	5.89	6.39	6.71
	[0.020]	[0.14]	[0.40]	[0.19]	[1.65]	[1.66]	5.63	6.06	6.46
2.00	6.7	15.9	32.5	38.5	99.6	210.3	5.75	6.26	6.47
	[0.020]	[0.10]	[0.40]	[0.12]	[0.50]	[3.50]	5.35	5.83	6.18

m (since $C_a = \sqrt{m/\alpha}$). Each estimate from the simulation is based on simulating the system in question for 10^7 time units – 10 replications, each lasting 10^6 time units (not including a warm up period of 10^5 time units). The numbers shown in brackets are the half-widths of 95% confidence intervals for the corresponding expected total cost. In the columns under "Ratio", the first number in each cell is the ratio of the total cost estimates from the $K = 10$ and $K = 1$ columns. The second number in the cell is the value of the analytical approximation of δ. The simulation was carried out using the commercial software Arena (Kelton, Sadowski, and Sadowski, 1998).

6. Concluding comments

In this paper, we introduced a model for the analysis of production–inventory systems with multiple products. We used the model to examine the effect of product variety on inventory-related costs. We showed that total cost tends to increase linearly with the

Table 2
Simulation results for systems with Gamma-distributed processing times ($\alpha = 1$, $\beta = 1$), exponentially-distributed order inter-arrival times for each item and exponentially-distributed setup times ($\tau = 1$).

C^2_{process}	Simulated total cost $K = 1$			Simulated total cost $K = 10$			Ratio		
	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$
0.01	2.73 [0.004]	5.79 [0.01]	12.2 [0.07]	26.7 [0.09]	68.6 [0.90]	150.1 [2.65]	9.78 9.56	11.85 11.86	12.30 12.83
0.20	3.07 [0.005]	6.76 [0.02]	14.5 [0.08]	27.9 [0.08]	71.3 [0.53]	157 [1.12]	9.09 8.79	10.55 10.65	10.83 11.38
0.40	3.5 [0.008]	7.86 [0.02]	16.8 [0.08]	28.3 [0.07]	75.8 [0.87]	161.7 [1.02]	8.09 8.00	9.64 9.73	9.63 10.21
0.60	3.92 [0.007]	8.2 [0.03]	17.1 [0.20]	28.8 [0.09]	76.9 [0.16]	166.3 [2.50]	7.35 7.42	9.38 8.87	9.73 9.33
0.80	4.18 [0.008]	9.6 [0.06]	19.8 [0.20]	29.8 [0.09]	77.8 [0.08]	174.2 [0.48]	7.13 7.04	8.10 8.21	8.80 8.63
1.00	4.6 [0.010]	10.7 [0.08]	22.4 [0.18]	30 [0.15]	79.7 [1.25]	179.5 [2.84]	6.52 6.67	7.45 7.67	8.01 8.07
1.20	5.06 [0.010]	11.4 [0.05]	23.8 [0.34]	32.1 [0.14]	84.6 [0.34]	189 [1.50]	6.34 6.30	7.42 7.18	7.94 7.57
1.40	5.48 [0.020]	12.8 [0.25]	27 [0.41]	33.7 [0.31]	87.8 [0.15]	195.5 [2.47]	6.15 6.09	6.86 6.78	7.24 7.16
1.60	5.89 [0.010]	13.7 [0.06]	28.6 [0.24]	35 [0.04]	89.2 [0.26]	201.2 [1.97]	5.94 5.88	6.51 6.43	7.03 6.80
1.80	6.33 [0.020]	15 [0.14]	30.6 [0.40]	36.6 [0.05]	93.8 [1.65]	207.3 [1.35]	5.78 5.67	6.25 6.14	6.77 6.51
2.00	6.7 [0.020]	15.9 [0.10]	32.5 [0.40]	38.3 [0.22]	95.1 [1.50]	214.7 [2.51]	5.72 5.51	5.98 5.89	6.61 6.24

number of products. We found that the rate of increase is sensitive to system parameters such as demand and process variability, demand and capacity levels, and setup times. We found that the effect of these parameters can be counterintuitive. For example, increasing either demand or process variability decreases the relative cost of offering variety. A similar effect is observed with respect to expected production time. On the other hand an increase in either expected setup time, setup time variability, or demand rate increases the relative cost of offering variety.

These results highlight important differences between inventory systems with exogenous supply lead times and systems whose lead times are generated by a production facility with finite capacity and stochastic production times. They also highlight the impact of variety on manufacturing efficiency (via setups) and the resulting effect on inventory costs. More importantly, the results point to the need for managers to be aware that the relative costs of offering variety (or alternatively the relative advantage of reducing it) are sensitive to system parameters. For example, in industries where

Table 3
Simulation results for systems with Gamma-distributed setup times ($\alpha = 1$, $\beta = 1$), exponentially-distributed order inter-arrival times for each item and exponentially-distributed processing times.

C^2_{setup}	Simulated total cost $K = 1$			Simulated total cost $K = 10$			Ratio		
	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 0.9$
0.01				28.1 [0.03]	75.9 [0.11]	175.2 [1.22]	6.11 6.31	7.09 7.53	7.82 7.98
0.20				28.6 [0.02]	76.7 [0.08]	176.3 [1.28]	6.22 6.37	7.17 7.56	7.87 8.00
0.40				28.9 [0.02]	76.9 [0.14]	176.8 [0.54]	6.28 6.44	7.19 7.59	7.89 8.02
0.60				29.5 [0.07]	77.5 [0.13]	177.7 [1.25]	6.41 6.52	7.24 7.62	7.93 8.04
0.80				29.8 [0.09]	78.2 [0.59]	178.6 [1.45]	6.48 6.59	7.31 7.65	7.97 8.05
1.00	4.6 [0.010]	10.7 [0.08]	22.4 [0.18]	30 [0.15]	79.7 [1.25]	179.5 [2.84]	6.52 6.67	7.45 7.67	8.01 8.07
1.20				31.3 [0.11]	82.1 [0.67]	182.4 [1.67]	6.80 6.74	7.67 7.69	8.14 8.08
1.40				31.8 [0.12]	83.3 [1.11]	183.5 [1.24]	6.91 6.82	7.79 7.71	8.19 8.10
1.60				32.3 [0.16]	83.9 [1.32]	185.9 [1.38]	7.02 6.90	7.84 7.73	8.30 8.11
1.80				33.1 [0.09]	84.9 [1.51]	186.4 [1.17]	7.20 6.98	7.93 7.76	8.32 8.13
2.00				33.5 [0.18]	85.4 [1.48]	187.1 [1.55]	7.28 7.07	7.98 7.78	8.35 8.14

either demand or process variability is high, the relative penalty for offering variety is small. On the hand when expected setup times or setup time variability are high, the relative cost of variety can be significant. For firms that must offer variety, these results also suggest areas of improvement that would yield the largest percentage reduction in cost.

In this paper, we have focused on quantifying inventory-related costs due to increased variety. Clearly, these costs need to be traded off against possible benefits from potentially higher prices or increased market share. On the other hand, there might be additional costs associated with variety including higher costs of raw materials, product development, and marketing. Managers need to be aware of these additional costs and benefits when deciding when to increase product variety and by how much. Furthermore, in many applications, increasing product variety beyond a certain range would require investing in a different process technology. In that case, analyzing the impact of variety would need to be carried out for discrete ranges and for different choices of technology.

Table 4
Simulation results for systems with Gamma-distributed order inter-arrival times for each item ($\alpha = 1$, $\beta = \lambda/10$, individual arrival rate $= \lambda/10 = \rho/10$), exponentially-distributed processing times ($t = 1$) and zero setup times ($\tau = 0$).

C_a^2	Simulated total cost $K = 1$			Simulated total cost $K = 10$			Ratio		
	$\lambda = 0.70$	$\lambda = 0.90$	$\lambda = 0.95$	$\lambda = 0.70$	$\lambda = 0.90$	$\lambda = 0.95$	$\lambda = 0.70$	$\lambda = 0.90$	$\lambda = 0.95$
0.01	3.19	11.5	23.7	9.87	20.6	32.1	3.09	1.79	1.35
	[0.009]	[0.023]	[0.134]	[0.002]	[0.050]	[0.113]	3.09	1.76	1.37
0.20	3.86	13.6	27.9	10.2	21.7	38.1	2.64	1.60	1.37
	[0.009]	[0.052]	[0.467]	[0.002]	[0.075]	[0.484]	2.63	1.59	1.33
0.40	4.59	15.8	35	10.7	23.9	42.9	2.33	1.51	1.23
	[0.008]	[0.129]	[1.000]	[0.003]	[0.132]	[0.716]	2.32	1.50	1.28
0.60	5.25	18.1	38.8	11.2	26.6	47.6	2.13	1.47	1.23
	[0.005]	[0.051]	[0.745]	[0.012]	[0.213]	[0.762]	2.13	1.46	1.25
0.80	6.08	21.3	50.1	12.0	31.6	54.6	1.97	1.48	1.09
	[0.012]	[0.250]	[2.250]	[0.013]	[0.246]	[0.865]	2.04	1.45	1.22
1.00	6.78	23.5	52.0	12.6	32.8	66.4	1.86	1.40	1.28
	[0.012]	[0.263]	[0.496]	[0.012]	[0.334]	[1.650]	1.87	1.39	1.20
1.20	7.29	25.5	52.9	13.2	33.7	62.8	1.81	1.32	1.19
	[0.019]	[0.287]	[0.784]	[0.018]	[0.16]	[0.756]	1.80	1.34	1.19
1.40	8.07	29.4	57.2	14.0	36.3	67.6	1.73	1.23	1.18
	[0.017]	[0.344]	[1.270]	[0.011]	[0.611]	[1.030]	1.74	1.32	1.17
1.60	8.65	29.3	60.3	14.9	39.0	75.0	1.72	1.33	1.24
	[0.014]	[0.169]	[1.470]	[0.038]	[0.277]	[0.857]	1.70	1.31	1.16
1.80	9.44	32.8	73.0	15.7	42.1	78.9	1.66	1.28	1.08
	[0.062]	[0.450]	[1.580]	[0.063]	[0.234]	[0.630]	1.69	1.29	1.15
2.00	10.0	35.5	76.4	16.6	44.8	85.7	1.66	1.26	1.12
	[0.043]	[0.337]	[1.730]	[0.072]	[0.961]	[1.890]	1.66	1.26	1.14

Acknowledgments

This paper has benefited from helpful comments and suggestions by two anonymous referees. Research of the first author is supported, in part, by the National Science Foundation under grants DMI 9988721.

Appendix. Summary of notation

K: the number of products.
λ_i: the demand rate for product type i.
$\lambda = \sum_{i=1}^{K} \lambda_i$, the aggregate demand rate.
$C_{a_i}^2 = \mathrm{Var}(X_i)/E(X_i)^2$.

Q_i: batch size for product of type i.
$p_i = \lambda_i/Q_i/\sum_{i=1}^{K}(\lambda_i/Q_i)$, the probability that a batch in the processing stage is of type i.
s_i: base-stock level for product of type i.
X_i: a random variable denoting the inter-arrival time between orders of type i; $E(X_i) = 1/\lambda_i$.
Y_i: a random variable denoting unit production time for product i; $E(Y_i) = t_i$, $\mathrm{Var}(Y_i) = \eta_i$.
Z_i: a random variable denoting setup time for product i; $E(Z_i) = \tau_i$, $\mathrm{Var}(Z_i) = \theta_i$.
$U_i = Z_i$ if a batch of type i is processed after a batch of type j $(j \neq i)$; 0, otherwise.
$W_i = Q_i Y_i$.
$S_i = U_i + W_i$.
$U = \sum_{i=1}^{K} p_i U_i$.
$W = \sum_{i=1}^{K} p_i W_i$.
$S = \sum_{i=1}^{K} p_i S_i$.
$\rho = \sum_{i=1}^{K} (\lambda_i/Q_i)E(S)$, steady state utilization of the production facility.
$C_{a,p}^2 = \sum_{i=1}^{K} p_i C_{a_i}^2$.
$C_{s,p} = \mathrm{Var}(S)/E(S)^2$.
$\sigma = (\hat{N} - \rho)/\hat{N}$.
$r_i = p_i\sigma/(1 - \sigma(1 - p_i))$.
N_i^b: a random variable denoting the number of orders of type i in the batching stage.
N_i^p: a random variable denoting the number of batches of type i in the processing stage.
$N_i = N_i^b + Q_i N_i^p$.
$N^p = \sum_{i=1}^{K} N_i^p$.
$\hat{N}$: approximated number of customers in a $GI/G/1$ queue.
I_i: a random variable denoting inventory level for product i.
B_i: a random variable denoting backorder level for product i.
h_i: holding cost per unit of inventory of type i per unit time.
b_i: backordering cost per order of type i backordered per unit time.
$z(\mathbf{s}, \mathbf{Q}) = \sum_{i=1}^{K} E(h_i I_i + b_i B_i)$, the long run expected total cost per unit time given $\mathbf{s}$ and $\mathbf{Q}$, where $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ and $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_K)$.

References

Agrawal, M., T.V. Kumaresh, and G.A. Mercer. (2001). "The False Promise of Mass Customization." *The McKinsey Quaterly* 3.

Albin, S.L. (1984). "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues." *Operations Research* 32, 1133–1162.

Albin, S.L. (1986). "Delays for Customers from Different Arrival Streams to a Queue." *Management Science* 32, 329–340.

Alfaro, J.A. and C.J. Corbett. (1999). "The Value of SKU Rationalization in Practice: The Value of Inventory Pooling under Suboptimal Inventory Policies." Working Paper, Anderson School of Management, UCLA, CA.

Aviv, Y. and A. Federgruen. (1999). "The Benefits of Design for Postponement." In S. Tayur, R. Ganeshan, and M. Magazine (eds.), *Quantitative Models for Supply Chain Management*, pp. 553–584. Kluwer Academic.

Benjaafar, S., W. Cooper, and J.S. Kim. (2003). "On the Benefits of Inventory Pooling in Production–Inventory Systems." Working Paper, Department of Mechanical Engineering, University of Minnesota, Minneapolis.

Benjaafar, S. and J.S. Kim. (2001). "When does Higher Demand Variability Lead to Lower Safety Stocks?" Working Paper, Department of Mechanical Engineering, University of Minnesota, Minneapolis.

Buzacott, J.A. and J.G. Shanthikumar. (1993). *Stochastic Models of Manufacturing Systems*. Prentice Hall.

DeGroote, X., E. Yucesan, and S. Kavadias. (1999). "Product Variety and Leadtime Performance." Working Paper, INSEAD, Fontainebleau, France.

Dobson, G. and C.A. Yano. (2001). "Product Offering, Pricing, Make-to-Order versus Make-to-Stock and Cycle Time Decisions with Shared Manufacturing Capacity." Working Paper, University of Rochester.

Eppen, G.D. (1979). "Effects of Centralization on Expected Costs in Multi-Location Newsboy Problems." *Management Science* 25, 498–501.

Federgruen, A., G. Gallego, and Z. Katalan. (2000). "The Effect of Product Variety on Manufacturing Performance." Working Paper, School of Business, Columbia University.

Fisher, M.L. and C.D. Ittner. (1999). "The Impact of Product Variety on Automobile Assembly Operations: Empirical Evidence and Simulation Analysis." *Management Science* 45, 771–786.

Garg, A. and H.L. Lee. (1999). "Managing Product Variety: An Operations Perspective." In S. Tayur, R.M. Ganeshan, and M. Magazine (eds.), *Quantitative Models for Supply Chain Management*, pp. 467–490. Kluwer Academic.

Green, P. and A. Krieger. (1985). "Models and Heuristics for Product Line Selection." *Marketing Science* 4, 1–19.

Ho, T. and C. Tang (eds). (1998). *Managing Product Variety*. Boston, MA: Kluwer Academic.

Kekre, S. and K. Srinivasan. (1990). "Broader Product Line: A Necessity to Achieve Success?" *Management Science* 36, 240–251.

Kelton, D.W., R.P. Sadowski, and D.A. Sadowski. (1998). *Simulation with Arena*. Boston, MA: McGraw-Hill.

Lancaster, K. (1990). "The Economics of Product Variety: A Survey." *Marketing Science* 9, 189–210.

MacDuffie, J.P., K. Sethuraman, and M.L. Fisher. (1998). "Product Variety and Manufacturing Performance: Evidence from the International Automotive Assembly Plant Study." *Management Science* 42, 350–370.

Quelch, J.A. and D. Kenny. (1994). "Extend Profits, not Product Lines." *Harvard Business Review* 72(5), 153–160.

Randall, T. and K. Ulrich. (2001). "Product Variety, Supply Chain Structure, and Firm Performance: Analysis of the U.S. Bicycle Industry." *Management Science* 47, 1588–1604.

Thonemann, U.W. and J.R. Bradley. (2002). "The Effect of Product Variety on Supply Chain-Performance." *European Journal of Operational Research*, Forthcoming.

Van Ryzin, G. and S. Mahajan. (1999). "On the Relationship between Inventory Costs and Variety Benefits in Retail Assortments." *Management Science* 45, 1496–1509.

de Vericourt, F., F. Karaesmen, and Y. Dallery. (2000). "Dynamic Scheduling in a Make-to-Stock System: A Partial Characterization of Optimal Policies." *Operations Research* 40, 811–819.

Whitt, W. (1982). "Approximating a Point Process by a Renewal Process, I: Two Basic Approaches." *Operations Research* 30, 125–147.

Zipkin, P. (1995). "Performance Analysis of a Multi-Item Production–Inventory System under Alternative Policies." *Management Science* 41, 690–703.
Zipkin, P. (2001). "The Limits of Mass Customization." *Sloan Management Review* 42, 81–87.