Design for Postponement

Jayashankar M. Swaminathan

Kenan-Flagler Business School
University of North Carolina
Chapel Hill, NC-27599

Hau L. Lee

Graduate School of Business
Stanford University
Stanford, CA 94305

August, 2001; Revised August 2002, January 2003

To appear as a chapter in the Handbook of OR/MS on Supply Chain Management edited by Graves and de Kok. The work of the first author was supported in part by NSF CAREER Award #0296081.
1. Introduction

In this age of increasing globalization and shortening of product life cycles, companies are faced with the demand for escalating product variety to meet the diverse needs of global customers. Indeed, mass customization has become a business requirement for many high technology companies. However, the provision of product variety comes with a price. With forecasting becomes more difficult, overhead for product support is higher, inventory control is more difficult, manufacturing complexity increases, and after-sales support is more complex. One solution that innovative companies have exploited is the power of product and process design, by integrating design with their supply chain operations to gain control of product variety proliferation.

Design has always been viewed as a key driver of manufacturing cost. Past research has indicated that as much as 80% of the manufacturing cost of the product is determined by the design of the product or the process in which the product is to be manufactured. Design can also be leveraged to address the problem of mass customization (Martin et al. 1998). By properly designing the product structure and the manufacturing and supply chain process, one can delay the point in which the final personality of the product is to be configured, thereby increasing the flexibility to handle the changing demand for the multiple products. This approach is termed postponement\(^1\).

Alderson (1950) appears to be the first who coined this term, and identified it as a means of reducing marketing costs. Alderson held that “the most general method which can be applied in promoting the efficiency of a marketing system is the postponement of differentiation, ..., postpone changes in form and identity to the latest possible point in the marketing flow; postpone change in inventory location to the latest possible point in the

---

\(^1\) This approach has also been termed as delayed product differentiation or late customization.
time”. He believed that this approach could reduce the amount of uncertainty related to marketing operations. Bucklin (1965) provided arguments as to how postponement as identified by Alderson could be a useful concept but would be difficult to implement through the channel particularly in manufacturing environments predominantly operating on a “make-to-stock” basis. He argued that some entity in the channel would have to bear the risks associated with product variety, and postponement only helped in shifting this risk to some other partner in the channel. However, as manufacturing firms started to move away from the traditional make-to-stock environment, postponement has become an attractive alternative.

Zinn and Bowersox (1988) describe different types of postponement that could be implemented. These included labeling postponement, packaging postponement, assembly postponement, manufacturing postponement and time postponement. Labeling postponement is a situation where a standard product is stocked and labeled differently based on the realized demand. In packaging postponement products are not packaged into individual packs until final orders are received. Assembly and manufacturing postponement refer to situations where additional assembly or manufacturing may be performed at the assembly facility or at a warehouse before shipping the product to the customer after demand is realized. Finally, time postponement refers to the concept that products are not shipped to the retail warehouses but are held at a central warehouse and are shipped to customers directly.

Clearly, different types of postponement strategies have different costs and benefits associated with them. For example, with packaging postponement, inventory costs are reduced due to stocking of the standard product, whereas the packaging costs
are higher since it is not done in one big batch thereby losing economies of scale. Similarly, in manufacturing and assembly postponement, component costs may increase, and in some cases, a more complex process may have to be used. Moreover, there are multiple ways in which postponement can be pursued, each with different cost and service performance impacts.

Fundamentally, there are three types of factors that affect the benefits and costs associated with postponement - market factors, process factors and product factors. Market factors are those related to customer demand and service requirements. These parameters include demand fluctuations or variance, correlation in demand across the different products, lead time and service requirements for customization (which affect the penalty cost for stock-outs or late deliveries). Process factors are those manufacturing and distribution processes under the control of the firm. These include the sequence of operations performed to customize the product, the network structure of the supply chain (manufacturing and distribution sites), whether the product is made to order or made to stock as well as how much and at which location inventories (components, subassemblies, and finished products) are stored in the supply chain. Product factors are related to the design of the product or product lines. These include the degree of standardization that is present in the components and the costs associated with standardizing components, modularity in the product design, as well as the degree to which end products can be substituted for each other’s demand.

The ability of a firm to implement a successful postponement strategy depends on how well the firm can tailor its process and product characteristics to the market requirements. Primarily these relate to the changes in the design of the product or
process so that implementing postponement strategies becomes easier and more cost effective. There are mainly two types of changes -those related to process design changes, termed *process postponement*, and those related to product design changes, termed *product postponement*. Process postponement usually requires (1) process standardization, i.e., making some part of the process standard so that the different product variants share that process; and (2) process re-sequencing, i.e., changing the sequence of customization steps in which the product attains distinct functionalities and characteristics. Product postponement often requires standardizing some key components, or introducing parts commonality in the product structure.

In this chapter, we discuss analytical models for evaluating postponement alternatives. Earlier survey articles on similar areas include Garg and Lee (1998) and Swaminathan and Tayur (1998). The rest of this chapter is organized as follows. In Section 2, we introduce the three key postponement enablers: process standardization, process re-sequencing and component standardization. These three enablers and associated performance evaluation models are described in greater details in Sections 3-5. We also describe industry applications utilizing these enablers. In Section 6, we discuss other techniques for managing product variety such as modularity and downward substitution and explore the additional benefits of postponement in pricing and information processing. In Section 7, we provide our concluding remarks.

**2. Postponement Enablers**

Postponement can be enabled through changes in the manufacturing-distribution process or the product architecture. In this section we introduce three enablers of postponement -
process standardization, process re-sequencing and component standardization. Process standardization refers to standardizing the initial steps in the process across the product line so that products are not differentiated at these steps, and distinct personalities of the products are added at a later stage. All the products in the product line (or a subset of it) are processed through these standard steps. A complementary approach to process standardization is process re-sequencing. Here, the sequence is changed so that more common components are added in the beginning of the process. The components or features that create product differentiation are added later. The key benefit from both of the above approaches is that the initial stages in the process are less differentiated, leading to partially completed products at the end of these common stages. This enables the firm to pool the risk across the different product demands and to effect lower inventory requirements. Clearly the success of the above approaches depends to a great extent on how modular in structure the process is. Process modularity is the same as the product modularity concept applied to a process. If a process can be divided into separate sub-steps so that these sub-steps can be performed in either parallel or in different sequence then it is classified as a modular process. For example, the testing process of a product may require multiple tests and burn-ins. In some cases, the whole test process may have to be carried out in a continuous fashion, while in other cases, the process can be broken up into sub-tests. Process modularity is closely related to the flexibility of the process in that processes that are more flexible are likely to be modular processes as well (see chapter 5 for models on flexibility). In addition to process modularity, the feasibility of process re-sequencing depends on common or standard components in the product line. Indeed, the third enabler to postponement is component standardization.
We should note that all three enablers could be used individually or in any combination to achieve postponement. In the next sections, we describe the models that have been developed for each of these enablers. Some notation that is used throughout the paper is given in table 1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_i )</td>
<td>Realized demand for product ( i )</td>
</tr>
<tr>
<td>( \mu_i )</td>
<td>Mean demand for product ( i )</td>
</tr>
<tr>
<td>( \sigma_i )</td>
<td>Standard deviation of demand for product ( i )</td>
</tr>
<tr>
<td>( \rho_{ij} )</td>
<td>Correlation of demand for product ( i ) and product ( j )</td>
</tr>
<tr>
<td>( S_i )</td>
<td>Base stock inventory level for ( i )</td>
</tr>
<tr>
<td>( s_i )</td>
<td>Safety stock for ( i )</td>
</tr>
<tr>
<td>( E(x) )</td>
<td>Expected value of ( x )</td>
</tr>
<tr>
<td>( \text{Var}(x) )</td>
<td>Variance of ( x )</td>
</tr>
<tr>
<td>( z )</td>
<td>Safety factor</td>
</tr>
<tr>
<td>( h )</td>
<td>Per unit holding cost</td>
</tr>
<tr>
<td>( U )</td>
<td>A vanilla box configuration in terms of the components</td>
</tr>
</tbody>
</table>

### Table 1

3. **Process Standardization**

In process standardization approaches, inventory may be carried at the intermediate stage after the common steps in the process (known as *the point of differentiation*) as well as at the final product level. The models developed here differ in terms of consideration of single versus multiple points of differentiation.
3.1 Single Point of Differentiation

Lee (1996) describes the most basic version of this model where there are $M$ products and inventories are carried in finished form. All the products are customized from the inventory available at the end of the standard steps of the process. It is assumed that negligible stock of inventory of the generic product is stored, in that upon arrival from the generic production process it is allocated for customization. The basic assumption is that the standard part of the process takes $t$ time periods to complete and the remaining $T-t$ periods correspond to time for the customization step (see Figure 1). This is analogous to the warehouse lead time of $t$ periods and $T-t$ periods of transportation time from the warehouse to the retailers. All the products are assumed to have independent and normal demands with mean $\mu_i$ per period and standard deviation $\sigma_i$ per period and that the system follows a periodic review policy with a period length of one, and complete backlogging for unmet demands. Eppen and Schrage (1981) assumed that the inventory allocation at the intermediate stage to the end-products follows the equal fractile allocation rule, i.e. after allocation, the inventory position for each end product should be the sum of the
mean demand for that end product over $T-t$ time periods and a common safety stock factor multiplied by the standard deviation of demand for the end product over the $T-t$ time periods. It was assumed that the probability of stock imbalance, i.e., that the stocks for the different products cannot be re-allocated to satisfy the equal fractile rule after allocation, is negligible. They showed that when the costs are identical at each site it is optimal to operate the end product inventory stockpiles in an order-up-to manner with a base stock level of $S_i$. Erkip et al. (1990) extended the analysis to allow item demands to be correlated both across warehouses and also correlated in time. Lee (1996) studied the case where demand across products $(j, k)$ may be correlated in each period ($\rho_{jk}$). For such a system the steady state end of period inventory level for product $i$ ($I_i$) is given by:

$$E(I_i) = A_i - R_iT\sum_{j} \mu_j$$  \hspace{1cm} (1)

$$Var(I_i) = R_i^2 t\{\sum_{j} \sigma_j^2 + \sum_{j\neq k} \rho_{jk}\} + (T-t)\sigma_i^2$$  \hspace{1cm} (2)

where $A_i$ is a function of $S_i$ and $\mu_i$, but is independent of $t$ and $R_i = \sigma_i / \sum_j \sigma_j$. Based on these two moments, service measures such as fill rate can be derived. The value of $S_i$ can be determined to satisfy the target service level. Lee (1996) analyzes the above system with process standardization and addresses the impact of postponement which is reflected in the parameter $t$. Clearly, $E(I_i)$ is independent of $t$, but $Var(I_i)$ is decreasing in $t$ for a given $S_i$.

$$\frac{\partial Var(I_i)}{\partial t} = R_i^2 \{\sum_{j} \sigma_j^2 + \sum_{j\neq k} \rho_{jk}\} - \sigma_i^2$$

$$= \sigma_i^2 \{(\sum_{j} \sigma_j^2 + \sum_{j\neq k} \rho_{jk}) / (\sum_{j} \sigma_j^2)^2 - 1\}$$  \hspace{1cm} (3)
For all \( i \) and \( j \), \( \rho_{ij} \leq \sigma_i \sigma_j \) and therefore \( \sum_j \sigma_j^2 + \sum_{j \neq k} \rho_{jk} \leq (\sum_j \sigma_j^2)^2 \). Hence, the variance of the steady state end-of-period inventory for product \( i \) is decreasing in the degree of postponement. Thus, postponement will lead to reduction of inventory of finished products. Further, the reduction in inventory is greater when the end product demands are negatively correlated. For identical and independent demand for products, the expression for \( \text{Var}(I_i) \) simplifies to:

\[
\text{Var}(I_i) = \left[ \frac{t}{M} + (T - t) \right] \sigma^2
\]

Clearly one can see in this case (from the derivative with respect to \( t \)) that the reduction in variance is greater when the number of products is larger.

Lee and Whang (1998) explore this model further by assuming that demands are not IID (independent and identically distributed) over time. With non-IID demand, the value of postponement is more than just being able to make product commitments at a later point in time when realized demands have been revealed. In addition, the progression of demands may also help to improve the forecast of the future demands. Lee and Whang term these two values as the value of uncertainty resolutions and the value of forecast improvement. To illustrate these two different values, they used a random walk demand model with the characteristics that the variance of future demand increases as we look further out into the future. Hence, if we let demand for end product \( i \) after \( t \) periods from today be \( D_i(t) \), then

\[
D_i(t) = \mu_i + \sum_{k=1}^{t'} \varepsilon_{ik}
\]

where \( \varepsilon_{ik} \) is normally distributed with mean 0 and standard deviation \( \sigma_{ik} \). When we have products with identical means and variances of demands, and when \( \sigma_{ik} = \sigma \) for all \( i \)'s and
k’s, then the safety stock that needs to be carried at the end product level for the original
Lee (1996) model becomes:

\[
s_i^*(t) = z \cdot \sigma \sqrt{\frac{M}{6} (T+1)(T+2)(2T+3) + \frac{M(M-1)}{6} (T+1-t)(T+2-t)(2T+3-2t)}
\] (6)

where \(z\) is the safety factor. In that case, the percentage savings obtained due to
postponement as compared to no postponement (the case where \(t = 0\)) is given by

\[
V_i(t) = 1 - \frac{s_i^*(t)}{s_i^*(0)} = 1 - \frac{1}{M + \left(1 - \frac{1}{M}\right) \frac{(T+1-t)(T+2-t)(2T+3-2t)}{(T+1)(T+2)(2T+3)}}
\] (7)

One can observe that the safety stock required is decreasing in \(t\) and the percentage
savings due to postponement are increasing and convex in \(t\). The reduction in safety
stock with postponement in this case is greater than that when demand is stationary. The
reason is that, with stationary demand, postponement allows the allocation decision (to
the multiple end-products from the common intermediate product) be made after demand
realizations of the end products have been revealed during the time when the common
process was performed. When demand is a random walk process in which future
demands are more variable into the future time, then there is an added value of
postponement – by delaying the point when allocation to end-products has to be made,
the demand variability of the end-product is reduced, since we are now closer to that
future demand period than when we begin the total production process (forecast
improvement). Thus, postponement with time-dependent demands may be even more
valuable.

In the above models there is an implicit assumption that the production-
distribution process is continuous and inventory can be stored only in finished product
form. In general, manufacturing environments involve a discrete set of operations, and
inventory can be stored immediately following any one of these stages. Furthermore, the costs associated with delaying differentiation have not been considered in the above models. There is a stream of research that extended the standard postponement model by allowing for multiple manufacturing steps with intermediate buffer inventory, and by explicitly modeling the costs of resequencing the process steps.

Lee and Tang (1997) consider a model where there are two products that require $N$ sequential tasks for completion (see Figure 2). Inventory can be stored in a buffer after each task with the buffer after the $N$th task being finished goods. The first $k$ tasks are assumed to have been standardized, i.e., the inventory in the buffer after the $k$th operation can be used for customization into either products. The tasks $k+1$ to $N$ are distinct for the two final products. Thus, the point of differentiation is right after the $k$th step. Under normal demand assumption for the two products and a discrete time setting, they consider the costs associated with standardizing stages. Let $Z_i$ denote the average investment cost per period (amortized) if task $i$ is changed into a common operation for both products. It is possible that $Z_i < 0$, e.g., when standardizing that task leads to overall reduction in costs. Let $L_i(k)$, $p_i(k)$ and $h_i(k)$ denote the lead time for task $i$, unit processing cost for task $i$ and the per unit inventory holding cost for items in the buffer following task $i$. 

![Figure 2](image_url)
respectively, when the first \( k \) tasks are standardized. Further, they assume that the same safety factor \( z \) is used at all stages in the process, and a base stock policy is followed. Then, the average buffer inventory at any stage is given by

\[
\mu / 2 + z\sigma \sqrt{(L + 1)}
\]  

(8)

where \( \mu \) and \( \sigma \) are the mean and standard deviation of demand faced at that stage. The relevant cost per period for the case when the first \( k \) operations are standardized is given by:

\[
C(k) = \sum_{i=1}^{k} Z_i + \sum_{i=1}^{N} p_i(k)(\mu_1 + \mu_2) + \sum_{i=1}^{N} h_i(k)[L_i(k)(\mu_1 + \mu_2)] + \sum_{i=1}^{k} h_i(k)[(\mu_1 + \mu_2)/2 + z\sigma_{i2} \sqrt{L_i(k) + 1}] 
\]  

(9)

\[
+ \sum_{i=k+1}^{N} h_i(k)[(\mu_1 + \mu_2)/2 + z(\sigma_1 + \sigma_2) \sqrt{L_i(k) + 1}]
\]

This includes the average investments, processing costs, in-transit inventory (WIP) and buffer inventory holding costs.

Consider the special case when the lead time and holding costs at different stages are not affected by the point of differentiation, and when \( p_i(k) = p_i + \beta_i \), i.e., \( \beta_i \) represents the additional processing cost for standardizing an operation. Then conditions under which \( C(k) \) may be convex or concave can be derived. For the case when \( C(k) \) is convex in \( k \), the optimal \( k^* \) is decreasing in demand correlation among the two products, \( Z_i \), \( \beta_{ii} \), and mean demand, but is increasing in \( h_i \sqrt{L_i} + 1 \). As the demand correlation decreases, the resulting savings in inventory cost increase. In order to take advantage of the savings it is desirable to defer the common operation. When the cost of standardizing an operation \( Z_i \) or the incremental processing cost associated with delayed differentiation...
\( \beta_i \) increase then it is not as desirable to delay differentiation. As mean demand increases (while holding the variances constant), the resulting demand is less variable, therefore delaying differentiation is less attractive. Increase in \( h_i \sqrt{L_i + 1} \) leads to higher inventory savings due to delayed differentiation so the optimal delay point is further out. Furthermore, \( C(k) \) is concave in \( k \) if (1) \( Z_i, \beta_i \) and \( h_i \) are identical; (2) \( S_i \) and \( h_i \) are proportional to \( h_i \) and \( L_i \) is constant; or (3) \( Z_i, \beta_i \) are identical and \( h_i \) is linear in \( i \).

### 3.2 Multiple Points of Differentiation

![Figure 3](image-url)

So far, all the above models are restricted to only one point of differentiation. Garg and Tang (1997) consider a system with two points of differentiation, the first is the family differentiation point and the second the product differentiation point (see Figure 3). In their system, there are three stages in the process. At the first stage all the products are in their generic form, the family differentiation point occurs at the beginning of the second
stage where specific components are added to differentiate a generic product into different families. The product differentiation point occurs at the beginning of the third stage where specific components are used to customize semi-finished products into different end products of that family. Note that the points of differentiation emerge because of adding specific components. The lead times for the different stages are assumed to be $T_1$, $T_2$ and $T_3$, respectively. They assume that the manufacturing lead times for customizing the products into the different families $T_2$ are the same and that the manufacturing lead times for customizing different end products of different families $T_3$ are the same. Early postponement is defined as increasing $T_1$ to $T_1+1$ while reducing $T_2$ to $T_2-1$, and late postponement is defined as increasing $T_2$ to $T_2+1$ and reducing $T_3$ to $T_3-1$. For the above system, the authors consider two possible scenarios. In the first scenario, inventory is stored only in the finished goods form (called centralized system) and in the second scenario, inventory is stored at each point of differentiation as well as at the finished goods level (called decentralized system). The centralized system extends the model studied by Eppen and Schrage (1981) to three stages and correlated demand. The demand for the final products are assumed to be independent normal and identical across time periods. For each time period, the demand for the final products are correlated. The assumption is that the system operates under a base stock policy and periodic review. An equal fractile allocation is assumed at the first and second stages. For the centralized system, under an equal-fractile allocation policy and identical equivalent degree of correlation of demand at the family level (defined as the ratio of variance for perfectly correlated demand and the actual variance in demand for the family), they show that both early and late postponement lead to reduction in total
inventory. Further, they show that as the product demand across a family becomes more negatively correlated then late postponement becomes more preferable as compared to early postponement. In the decentralized model, they assume that inventory is stored at all locations and the service level at each of the stages is high enough that the system can be decoupled into independent single stage inventory systems. For such an environment, they analyze the inventory savings across the whole network due to early and late postponement. They show that if T1 > T2 > T3 then both early and later postponement are beneficial. Further, when T2 is sufficiently smaller (larger) than T1 and T3, then early (late) postponement is beneficial.

3.3 Vanilla Boxes

The above papers assume that the production distribution process does not have any capacity constraints. Swaminathan and Tayur (1998) analyze a final assembly process with production capacity where inventory is stored in the intermediate form (called vanilla boxes). In addition to the intermediate form, they allow the two extreme forms of vanilla boxes – as components and as finished products. Therefore, this model captures both assemble to order (where components are stocked and products assembled from the components after demand is realized) as well as make-to-stock (where inventory is carried in finished form only) as special cases. This approach allows for multiple points of differentiation, in that there is no restriction on the type of vanilla box that can be stored. For example, Figure 4 below shows a product line with three products p1, p2 and p3 made of components a, b, c and d. Vanilla box V1=(a, b) can support products p1 and p3 while vanilla box V3 = (b, c) can support p1 and p2. In general, every product i (1,...,
$M$) may be assembled either directly from its components, or from any vanilla box whose component set is a subset of those required by $i$, thus avoiding redundant components. A binary bill of material in terms of the components is assumed without loss of generality.

\[
p1 = a, b, c \quad \quad p2 = b, c, d \quad \quad p3 = a, b, d
\]

Demands for the final products are random but follow one of $L$ given scenarios, each with a given likelihood. They assume that the vanilla box inventory follows a base stock policy in that every period the inventory is brought up to that level, then demand is realized, products are assembled from vanilla boxes by adding other components within the production capacity. Unsatisfied demand is lost with a penalty and remaining inventory of vanilla boxes incurs holding cost. Clearly, the main benefit of having vanilla boxes is that the amount of lead time for customization is much lower than customizing from the component level. Also, the capacity for customization may be limited which makes the problem more challenging. Under the above setting, they develop a stochastic integer program to determine the optimal types of vanilla boxes as well as their inventory levels which minimize the expected holding and penalty costs in single and multi-period settings. The first stage variables determine which components
should be present in the different vanilla boxes and the base stock levels for those vanilla boxes. The second stage variables determine how those vanilla boxes should be allocated to the different products on realization of product demand.

Let $C$ denote the capacity available to assemble products from vanilla boxes or from basic components, $t_{i0}$ and $t_{ik}$ the per unit assembly time for product $i$ from components or from vanilla box $k$ ($t_{ik} = \infty$ if product $i$ cannot be made from vanilla box $k$), respectively, $\pi_i$ the per unit per period stock-out cost for product $i$, $h_k$ the per unit per period holding cost for vanilla box $k$, $S = (S_1, \ldots, S_K)$ the vector of base stock levels of vanilla boxes $k$ ($1, \ldots, K$), $D_i$ a realization $(D_{i1}, \ldots, D_{iM})$ of product demands in scenario $l$ where $D_{i1}, \ldots, D_{iM}$ have a joint distribution $F$, and $r_{ikl}$ the quantity of product $i$ made using vanilla box $k$ in scenario $l$ ($k = 0$ implies that product $i$ is assembled directly from components). Then the two-stage stochastic program corresponding to a vanilla box configuration $U$ can be formulated as follows.

$$P_l(U) = \min_S E_l Q(S, U, D_i),$$

where

$$Q(S, U, D_i) = \min \sum_{r_i} \left( \pi_i \left( D_{i1} - \sum_{k=0}^{K} r_{ikl} \right) \right) + \sum_{k=1}^{K} \left( h_k \left( S_k - \sum_{i=1}^{M} r_{ikl} \right) \right)$$ (10)

s.t. $\sum_{i=1}^{M} \sum_{k=0}^{K} t_{ikl} r_{ikl} \leq C \ \forall l,$ (11)

$\sum_{i=1}^{M} r_{ikl} \leq S_k \ \forall k \geq 1, \ \forall l,$ (12)

$\sum_{k=0}^{K} r_{ikl} \leq D_{il} \ \forall i, \ \forall l,$ (13)

$r_{ikl}, S_j \in R_+.$ (14)
Utilizing the above framework and through the development of an efficient simulation based algorithm, the authors explore the benefits of postponement through vanilla boxes under various settings. Among other results, they show that postponement using vanilla boxes outperforms both assemble to order and make to stock systems when the assembly capacity available is neither too slack nor too tight (representative of most real environments). Further, they find that the vanilla box approach is extremely powerful under high variance and negative correlation among product demands. Finally, they provide examples where stocking two types of vanilla boxes may be sufficient for a product family with ten products and the performance may be better than a make-to-stock approach (with all the ten products).

Graman and Magazine (2000) consider a postponement model with capacity constraints where inventory can be stored in an intermediate form. On realization of demand all the finished goods are used first, and then the semi-finished product is used to satisfy the demand subject to a capacity constraint. This problem can be viewed as a special case of the vanilla box problem with only one type of vanilla box. For this model, they derive analytical expressions for service measure and also inventory calculations and through a numerical study show that very little postponement capacity can actually provide all the benefits related to inventory reduction.

Benjafer and Gupta (2000) present models that utilize queuing approximations to analyze a system where both make-to-stock and make-to-order environments are utilized while delaying differentiation of the product. There are two stages in the production process, the first stage produces products to stock while the second stage produces
products to order. The two stages are separated by a buffer that holds semi-finished inventory. The authors utilize queuing approximations by decoupling the two stage and assume each of them will behave similar to a M/M/1 queue. For the above approximation, solutions for inventory and service are available which the authors utilize to develop an optimization problem. The objective is to minimize the total costs subject to service level constraints by changing the stocking level of the intermediate product and the degree of differentiation. Further, the authors present several computational insights and show the impact of congestion effects on the postponement decision.

3.4. Process Standardization Applications

Lee, Billington and Carter (1993) describe the process standardization efforts at Hewlett Packard DeskJet Printer business. The printer line had three distribution centers in Europe, the US and the Far East and needed localization for the different countries in terms of power supply module with correct voltage, power cord terminators and a manual in the appropriate language. The existing operation was one where the products were “localized” at the US factory before being shipped to the respective distribution centers. The manufacturing in the US was done through a pull system based on the target safety levels set for the different distribution centers while taking into account the one month lead time in transit to the overseas distribution centers. As a result, high levels of safety stock are needed in the overseas distribution centers. The re-engineering of the distribution process involved re-sequencing the transportation and localization steps so that localization would now be done at the distribution centers. This was accomplished by making changes to the product design so that the power supply and the manuals could
be added later at the distribution centers. There were also additional investments in the form of product re-design, package re-design and enhancement to distribution center capabilities which were offset by the inventory savings that resulted from postponement. Additional benefits included lower capital investment for in-transit inventory, lower freight costs (due to the use of bulk packaging of the generic printers as opposed to packaging finished printers) and local presence of final assembly in the overseas markets. Based on a detailed modeling and analysis, Hewlett-Packard adopted process standardization in their inkjet business, and was rewarded with huge costs savings and improvements in customer service.

Swaminathan and Tayur (1998) analyze the final assembly stage of RS6000 server machines produced by IBM. Each model in the product line had 50 to 75 end products mainly differentiated by ten main features or components. A component is defined to be a part that is directly used in final assembly, so a component may be a subassembly in itself, e.g., a planar card. Different end products across the product line showed a high degree of component commonality. Since demand for end products were highly stochastic and correlated, the existing mode of operation was to start final assembly only after a firm customer order had been received. The typical steps in final assembly involved getting components together (kitting), putting them in the right place (assembly), testing, loading software (preloading) and packing the final product. At the time of the research this process often finished later than the customer requested arrival date leading to a sizeable percentage of late orders. This order delay problem was becoming increasingly acute as customers who once were satisfied with delivery within a month were now demanding products to be shipped within seven to ten days after the
orders were placed. The change in customer requirements was due primarily to competition in the industry and increase in service expectations. The vanilla assembly process based on delayed differentiation stocked vanilla boxes (semi-finished inventory). Clearly, there were additional costs in terms of redesigning the line to enable vanilla boxes, including workforce training and having inventory of vanilla boxes in the process that tied up capital. However, the benefits of such an approach were that the lead time experienced by the customer was only limited to the customization time starting from the vanilla box, hence most of the orders could be satisfied on time. After a thorough analysis of the costs and benefits as well as the change process involved, the vanilla assembly process was introduced in one of the two assembly plants (which had a satellite plant that was redesigned to produce vanilla boxes).

Brown et al. (2000) describe the postponement approaches at Xilinx which involved process standardization. As a leader in the field-programmable logic business, Xilinx made use of the postponement practice to achieve significant cost savings and service improvements. The manufacturing of integrated circuits consisted of two major steps: a front-end wafer fabrication at their outsourced manufacturer in Taiwan; and a backend assembly and test at their outsourced assembly sites in the Philippines and other Asian sites. The front-end process was standardized so that multiple devices share the same process. This way, the product does not have to be highly differentiated at the end of the front-end process. Fabricated wafers are then stored as intermediate inventory, known as the die bank, and they would go through the backend process that customize the products into the exact end device, only after the customer orders have come in. This way, the lead time to the customers is only the backend process time, which is much
shorter than the sum of front-end and backend process times (the lead time when a totally build-to-order process is used); but the flexibility to customer orders is much greater than if finished goods inventory is stored under a build-to-stock process is used.

Zara (one of the famous brands of Inditex) utilizes process standardization and vanilla boxes in the design phase of the product life (Harle et al. 2001, and Fraiman and Singh, 2003). Zara introduces new products at a rapid rate; in fact 70% of the products change every two weeks in a typical retail outlet. In order to create large variety and quick response to customers, the firm employs several strategies including standardization of the design modules. At the beginning of each selling season, the designers create a library of models that serve as platforms for the models that will be eventually launched. Twenty designers walk the streets and go to discos in order to get a feel of the latest fashion trends. After carefully watching the latest in fashion trends, Zara designers give adaptation (or customization) to the models from the library (which are vanilla boxes) and create 5 to 8 new designs every day! In total about 12000 new products and designs are created every year.

4. Process Resequencing

Process resequencing is another approach for enabling postponement and making it more effective. The basic idea is that it may be possible to change the sequence of operations in a process so that products get differentiated later. However, there may be costs associated with changing the sequence of operations, and hence it is important to have models that provide insights on these costs and benefits.
4.1. Linear Process Sequencing

Lee and Tang (1998) consider a two-stage system where at each stage a distinct feature is introduced into the product. They consider the knitting and dyeing tasks for garments as representative of the two stages. Each feature may have multiple options, for example, garment could be knitted under different settings or dyed with different colors. They analyze the case where each feature has two alternative options (see Figure 5). Thus there are four possible products available to a customer. Figure 5a represents the case where the garments are dyed first and are knitted later, while Figure 5b represents the case where the garments are knitted first and dyed later.

![Figure 5](image)

In such a system, changing the sequence of operations (which determines the feature that should be introduced first into the product) does not affect the ends of the process (raw material and finished products) but only affects the inventory that is stored at the end of the first stage. The objective is to minimize the total variance for the two intermediate buffers since the variance influences the inventory requirements for the system. The use
of the total variance as an objective function is of course a stylized assumption. The authors argued that the cost of manufacturing, such as the use of overtime or expediting, is often directly linked to the variances of production requirements (which, in this case, the same as the variances of the intermediate buffers). Indeed, as we see below, the use of a different objective function can lead to different results. The total demand in any period (across all the four products) is assumed to be a random variable with mean $\mu$ and standard deviation $\sigma$. The demands for the end products are modeled as a multi-variate normal distribution with parameters $(N, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ where $\theta_{11}$ represents the fraction of customers buying the first option in both the features. They show that it is optimal to have feature $A$ sequenced before feature $B$, if

$$
(\mu - \sigma^2)[p(1-p) - q(1-q)] < 0
$$

where $p$ is the probability of a customer buying option 1 on feature $A$ and $q$ is the probability of a customer buying option 1 on feature $B$. Clearly if the variance associated with feature $A$, given by $p(1-p)$, is smaller than the variance associated with feature $B$, given by $q(1-q)$, then one expects that feature $A$ should be sequenced first. However if $\sigma^2 > \mu$ then the reverse result is true, which is counterintuitive. They also show that when more options are available on the two features and each of these options are equally likely, then it is better to sequence the operation with fewer options first when $\mu > \sigma^2$ and vice versa otherwise. Kapuscinski and Tayur (1999) show that if the objective is to minimize the sum of standard deviations rather than the sum of variances at the
intermediate stage, then for the two feature - two option case, the counterintuitive result corresponding to the case $\mu < \sigma^2$ vanishes.

4.2. Assembly Sequence Design

It is clear from the above models that the sequence of tasks could play an important role in enabling postponement and thereby reducing inventory requirements. However, the physical assembly sequence is often defined through a complex set of precedence relationships among the different tasks. The general assembly design sequence problem has been primarily studied by researchers in engineering (see Nevins and Whitney 1989). For example, Figure 6 shows a product line with four products and six components. The assembly sequences $FAS_1$ and $FAS_2$ represent possible sequences for the product line. $FAS_1$ represents an assembly sequence where component $a$ needs to be introduced first followed by $b$ which can be followed by $c$ or $d$. Once component $c$ has been added, $e$ or $f$ can be added to the subassembly. $FAS_2$ represents another assembly sequence where component $a$ still needs to be added first at which point either $b$ or $c$ can be added. Once $c$ has been added, $e$ and $f$ can be added in any sequence and once $b$ has been added, component $d$ can be added to the subassembly. Note that in $FAS_1$, component $b$ needs to precede component $c$ in the assembly whereas in $FAS_2$ that precedence has been relaxed.
Gupta and Krishnan (1998) introduce the notion of a “generic sub-assembly (GSA)” which is similar to the vanilla box concept. GSA is a sub-assembly that satisfies all the precedence relationships among its components and is a feasible sub-assembly. In the above example \((a,c,f)\) is a GSA for \(FAS2\) but is not a GSA for \(FAS1\) because in \(FAS1\) component \(b\) has to be in place before component \(c\) can be introduced. A MGSA is a maximal generic sub-assembly according to criteria such as number of components in the
assembly or number of final products that can be supported by it. In the above example, 
\((a,b,c)\) which is a generic sub-assembly (GSA) covers 3 products P1, P2 and P4 and has 
three components. It is a MGSA in terms of number of components. On the other hand, 
\((a,b)\) which covers P1, P2, P3 and P4 (all the four products) is a MGSA in terms of 
number of products covered. For a given feasible assembly sequence, Gupta and 
Krishnan (1998) present an algorithm that generates the MGSA for a product family with 
criterion of maximizing the number of end products supported. Although a useful step in 
the right direction, the above model does not provide a cost benefit analysis related to 
assembly sequence design.

Swaminathan and Tayur (1999) utilize the vanilla box model described earlier, 
along with an assembly sequence design model to generate useful managerial insights. In 
the assembly sequence design problem (ASDP) they develop a mathematical 
programming model that generates the best sequence taking into account costs associated 
with the design of components to make such an assembly sequence. Thus, they model the 
situation where components could be designed in a flexible manner to satisfy alternative 
precedence conditions during execution. In combination with the vanilla box model, they 
consider two approaches to task re-sequencing: (1) where the best vanilla boxes are 
determined first and the sequence design is generated to enable assembly of the vanilla 
boxes and finished products with minimal design costs; and (2) where the most efficient 
assembly sequence is determined for the set of finished products and then the best vanilla 
boxes are found while taking into account the assembly constraints. As opposed to 
earlier work, this approach integrates the assembly sequence decisions with the
postponement decisions, and hence enables analysis of various “what if” questions pertaining to process re-sequencing.

Their notation is as follows. Products are indexed by \( i = 1 \ldots M \), components by \( j = 1 \ldots n \) and vanilla boxes by \( k = 1 \ldots K \). Let \( u_{kj} \) denote the content of the \( k \)th vanilla box in terms of components, \( U \) the vanilla configuration (matrix of \( u_{kj} \)) and \( U_k \) the configuration of the \( k \)th vanilla box. Let \( a_{ij} \) denote the bill of materials for the products in terms of the components, \( g_{pq} \) the cost of assembling component \( p \) before component \( q \), \( e_{pq} \) the cost of allowing independence between components \( p \) and \( q \), and \( Y \) the assembly sequence defined through the Boolean variables \( y_{pq} \) (set to 1 if component \( p \) is assembled before component \( q \) and to 0 otherwise). The difference in design costs between a fixed and an independent precedence relationship is given by \( c_{pq} = g_{pq} - e_{pq} \leq 0 \) and the objective is to minimize the total cost incurred. For a particular vanilla box configuration \( U \) the problem can be formulated as given below.

\[
\text{ASDP}(U): \quad \min \sum_{p=1}^{n} \sum_{q=1}^{n} c_{pq} y_{pq}
\]

s.t. \[1-y_{qp} \geq u_{kp}(1-u_{kq}) \quad \forall p,q,k, \quad (16)\]
\[1-y_{qp} \geq a_{ip}(1-a_{iq}) \quad \forall i,p,q, \quad (17)\]
\[y_{pq} + y_{qp} \leq 1 \quad \forall p,q, \quad (18)\]
\[y_{pq} + y_{qr} - y_{pr} \leq 1 \quad \forall p,q,r, \quad (19)\]
\[y_{rp} = 0 \quad \forall p, \quad (20)\]
\[y_{pq} \in \{0,1\} \quad \forall p,q. \quad (21)\]
In the above formulation, constraint (16) represents that if a component \( q \) is not present in a vanilla box \( (u_{kq} = 0) \) then it cannot be a predecessor of any component \( p \) in that vanilla box \( (u_{kp} = 1) \). Constraint (17) represents that if a component \( q \) is not present in a product \( (a_{iq} = 0) \) then it cannot be a predecessor of any component \( p \) in that product \( (a_{ip} = 1) \). These constraints assure that all vanilla boxes and products can be assembled using the assembly sequence \( Y \). Constraint (18) indicates that two components are either unordered in the assembly sequence \( (y_{pq} = y_{qp} = 0) \) or there exists a unique ordering of these components in the assembly sequence \( (y_{pq} = 1 \text{ or } y_{qp} = 1, \text{ but not both}) \). Constraint (19) maintains the transitivity relationship between components and constraint (20) indicates that all components of the same type are at the same level in the assembly sequence.

The authors conducted an extensive computational study which, in addition to validating earlier observations on the role of demand variance and correlation, provides additional insights on issues such as: it is better to sequence features with higher degree of variance later in the process; when the total amount of options provided across all features is kept constant, it is better to provide more options in a restricted number of features.

### 4.3. Process Re-sequencing Applications

Benetton (described in Dapiran 1992) is the earliest reported application of process re-sequencing that the authors are aware of. Traditionally sweaters were manufactured by
first dyeing the yarn into different colors, and then knitting the garments out of the colored yarns. The garments were stored in the form of finished goods to be shipped to the retailers. Dapiran describes how Benetton interchanged the knit and dye operations when they realized that most of the demand variability was due to the uncertainty of the customers’ preference of colors in a particular season. The interchange of the knit and dye operations enabled the firm to stock inventory of “greige” (uncolored) knit garments that could be dyed once the seasonal demand was known, enabling postponement and reducing inventory. Benetton had to invest in improving the dyeing technology so that the quality of the garments would not deteriorate due to the process changes.

Swaminathan and Tayur (1999) describe the assembly sequencing problem for US Filter, a manufacturer of reverse osmosis pumps. The sequence of operations at the final assembly was altered to enable faster response to customers. Costs related to product-process redesign as well as worker retraining at the final assembly had to be taken into account. The process sequencing approach has also been successfully applied by Garg (1999) to a large electronic manufacturer whose products are tailored for the telecommunication industry. The process involved board insertion and assembly, a station “marrying” different modules together, and the packaging of accessories and other components to make the final product. Alternative sequences of the process would result in different inventory and waiting times for the manufacturing of the product. Garg employed a queueing network model to evaluate these alternatives. More examples of application of postponement can be found in Lee et al. (1997) and Lee (1998).
5. Component Standardization

5.1. Commonality and Inventory Management

Component commonality has traditionally been studied in the context of make to stock or assemble to order systems. Traditional research on component commonality in the operations management stream has been mainly focused on reduction in component inventory due to commonality. Collier (1982) introduced the notion of commonality index - a measure of degree of commonality in a product line. Gerchak and Henig (1986) showed that when components are combined (standardized), the inventory requirements for product specific components always increase. Further, they also showed that a myopic inventory policy is optimal for a dynamic multi-period inventory problem with component commonality. Baker et al. (1986) and later Gerchak et al. (1988) explored the benefits of having common components in terms of reduced inventory or increased service. In particular they considered two products each with two components and analyzed the impact of standardizing one of the components. Since then, several authors including Eynan and Rosenblatt (1996) and Thonemann and Brandeau (2000) among others have explored the benefits associated with component commonality under different settings. Fisher et al. (1999) study commonality issues in the automotive industry.

5.2. Commonality and Postponement

Lee (1996) points out that in order to perform a complete analysis regarding the benefits of postponement due to component standardization, one needs a model that takes into account the following aspects: (1) inventory savings for the part; (2) increase in material
costs for the common parts; (3) additional costs for the engineering change; and (4) inventory savings for the finished goods. Most of the analysis on postponement has focused on (4) and the analysis on commonality has focused on (1). Lee and Tang (1998) present a model where they consider the costs associated with standardizing a process step. In order to standardize a process step one needs to standardize the part associated with that step and associate a cost for that. They incorporate the above cost in their model while analyzing the degree of postponement that is optimal.

The value of component standardization may be different at different stages of the product life cycle. This is due to the dynamic changes of demand uncertainty, shortage costs, inventory holding costs, and rework costs (to convert one version of end product to another). For example, the uncertainty of demand may be much higher in the product introduction and end-of-life phase than the mature phase. The shortage cost may be the highest during the product introduction phase, while the inventory holding cost may be the highest during the end-of-life phase, when left-over inventory may have to be written off as the product becomes obsolete. Consequently, models capturing these dynamics are needed to assess the value of component standardization. Lee and Sasser (1995) describe one simple model that shows that, given all the dynamics of the demand and cost characteristics, the value of standardization for postponement is high in the product introduction phase, low in the mature phase, and high in the end-of-life phase.

Swaminathan (1999) considers the problem related to optimizing the level of commonality while simultaneously considering the costs of commonality as well as the benefits due to inventory savings obtained due to higher levels of commonality. In particular, the author considers a two-product system with one common subassembly and
two product specific subassemblies. The parameter to be optimized is the size of this common subassembly in terms of the degree of commonality. The author assumes that the cost of the common component increases in a convex fashion with respect to the degree of commonality while the costs of the product specific components decrease in a linear fashion. Inventory of both common and product specific components are stored in anticipation of demand. Under the above assumptions and standard inventory assumptions related to holding and penalty costs and linear costs of commonality, the author shows that the two products have either complete commonality or no commonality. Further the optimal level of commonality is lower in product lines where the costs of introducing commonality are higher. Based on a computational study, the author shows that the optimal level of commonality is always higher in postponement as compared to the optimal level of commonality when product inventories are managed independently. Moreover, the cost of commonality affects the impact of operational factors on optimal commonality and inventory under postponement. That impact is limited when the cost of providing commonality is either very high or very low.

Van Mieghem (2002) analyzes a model with two products similar to Swaminathan (1999), where each product is assembled from two components. However, Van Mieghem (2002) assumes that both common and product specific components are stocked and derives conditions under which commonality should be adopted. This condition is stated in terms of a maximal commonality threshold cost that depends on the demand forecast only through its demand correlation as well as on financial data. For high commonality cost, neither commonality nor postponement is optimal. A pure commonality strategy where each product is assembled using a common component,
however, is never optimal unless complexity costs are introduced. Finally, the author shows that while the value of the commonality strategy decreases in the correlation between product demands, commonality is optimal even when the product demands move in lockstep (perfectly correlated) if there is a sufficient profit differential between the two products.

5.3. **Component Standardization Applications**

Lee (1996) provides the example of a large computer printer manufacturer which produced both mono as well as color printers. The manufacturing process for the two products are very similar except for the materials used. There are two key stages in the production, printed circuit board assembly and final assembly. At each of these stages a distinct component (print mechanism interface or head driver board) is inserted to differentiate the two types of products. The product differentiation began as soon as the head driver board was inserted at the printed circuit board stage. The demand was highly uncertain for both products and often correlated which led to high forecast errors. The firm evaluated the option of standardizing the head driver board or both the head driver board and print interface mechanism, which would lead to postponement. However, the costs of designing the additional functionality in the common components needed consideration.

Brown *et al.* (2000) describe the component standardization approach utilized at Xilinx where the final end product was actually designed in such a way such that customization could be done through software deployment at the customer site. The result is an integrated circuit that is field-programmable. In this case, we have an
extreme case of commonality standardization, since the final product has been standardized.

Thonemann and Brandeau (2000) describe a model that determines the optimal degree of standardization of components in a multi-product environment. This model was successfully utilized by a large automobile manufacturer to determine the optimal degree of commonality of wire harnesses that go into a product family. The model provided decision support for the design of future generation of components. Hewlett-Packard Company had also used this approach to redesign their network printer (Lee, 1997). The network printer was made in Japan and used to have two distinct engines, one for 110 voltage countries (e.g., North America), and one for 220 voltage counties (e.g., Europe and Asia). The printers with different engines are thus not inter-changeable to meet the changing demands in the different continents. An alternative design called for using universal power supply and fuser, which would result in a universal printer. This way, the printers made in Japan could be used for any market. At the end of life of the product, this could be particularly beneficial, as excess inventory in one continent does not have to be written off, but instead can be transshipped to another continent if there are any demand/supply imbalances.

Most recently, Lucent Technologies in Spain has utilized the component standardization strategy to achieve great success (Hoyt and Lopez-Tello, 2001). In 1998, the Tres Cantos plant, which builds telecommunications-switching systems, was faced with a great sales potential in Saudi Arabia that was worth millions of dollars. However, the lead time required was much shorter than usual, due to the Saudi government’s desire to have all systems implemented prior to Y2K. The specific configurations required,
however, could only be known after detailed site engineering work had been performed. The complete manufacturing time for the build-to-order process was far greater than what the Saudi customers wanted. In addition, the Tres Cantos actually did not have enough capacity to meet this big order from Saudi. By redesigning the product so that they could have common building blocks, Lucent was able to (1) pre-build the common building blocks before detailed site engineering tasks were completed; and (2) utilize the US plant in Oklahoma to help solve the capacity limitation problem. The result was that the company was able to win the contract and delivered the products on time.

6. Related Strategies and Other Benefits

Thus far we have considered models that explore mostly the inventory-related benefits of postponement. However, postponement decisions may be tightly linked to pricing decisions and information flow decisions. We will explore models for those and other strategies related to product variety management in this section.

6.1. Postponement, Information and Pricing

Although the key benefit identified with postponement has been inventory reduction, there are other issues related to postponement that are important. The main benefit related to postponement stems from the fact that one can delay the decision point for differentiation so that one can get more demand information before making a final commitment. Benefits of postponement can also be due to better forecasts generated in cases where the future forecasts are improved as one gets closer to the period when the demand occurs. As described earlier, Lee and Whang (1998) differentiated the value of
postponement as “uncertainty resolution” and “forecast improvement.” Anand and Mendelson (1998) study the increased flexibility and pooling benefits of delayed production in a multi-product supply chain with a noisy information system on a binary demand distribution. Aviv and Federgruen (1999) provide a detailed characterization for the benefits related to postponement under unknown demand distributions.

Another benefit of postponement beyond risk pooling relates to more effective usage of existing capacity. Swaminathan and Tayur (1998) show that the benefit of postponement using vanilla boxes is extremely high when the available capacity is medium (neither too high or too low). This is because, under that condition, postponement also benefits capacity utilization, which in turn affects the total costs. Gavirneni and Tayur (1999) explore this issue further by considering a two-item system where the manufacturer has the option to produce them in separate facilities or postpone the differentiation to the distribution stage. The authors explore the benefits of postponement under varying assumptions about the information at the manufacturer about the ordering policies being utilized.

Although we have only focused on the flexibilities provided by postponement strategies related to production of products, one could also envision situations where a firm tries to obtain a similar flexibility through pricing. Van Mieghem and Dada (1999) present a comparative analysis related to price, production and capacity decisions. They show that competition, uncertainty and timing of operational decisions can influence investment decisions of the firm related to capacity and inventory. Using a simple model for uncertainty in demand (captured by a random shock) they show that, in contrast to
production postponement, price postponement can make the investment decisions related to capacity and inventory relatively insensitive to uncertainty.

6.2. Postponement, Modularity and Substitution

Postponement strategy is often affected to a great extent by the product architecture decisions. One such decision relates to the degree of modularity that is present in the product architecture. Recently, several firms have started designing their products in “families” where the individual products are distinguished by the alternative combinations that are given to the modular components. Examples include the personal computer, electronics and automobile industry. Although this is an important concept from a product design perspective, analytical models studying its relationship with postponement have been very limited.

Another concept related to postponement for managing product variety is substitution. Substitution is a strategy where the manufacturer is able to provide a customer with an alternative product when the product ordered is not available in stock, and in the process incurs some kind of goodwill cost (or a real cost related to providing a better product to the customer or providing a gift voucher of some sort). This is a powerful strategy that has been used by firms over the years in several industries. Researchers have studied several versions of this problem (see Jordan and Graves 1995, Bassok et al. 1999, Rao et al. 2002) following the initial characterization of the optimal policy by Ignall and Veinott (1969). Some firms have also utilized the substitution strategy along with the postponement strategy. Swaminathan and Kukukyavuz (2000) analyze one such an environment from the biotech industry, and show the comparative
benefits of these strategies. An alternative situation related to substitution where a customer automatically buys a substitute product, has also been studied (see Mahajan and Van Ryzin 1999 for details). Swaminathan (2001) presents a managerial framework that relates product variety strategies such as postponement, substitution and commonality to the product and process modularity and identifies the most appropriate approach under different conditions.

7. Conclusions

Global markets with diverse needs and dramatic shortening of product life cycles put a great premium on effective product variety management. As a result, designing products for postponement is a high priority. The inter-disciplinary and complicated nature of the problem has generated a need for models that can provide multiple perspectives on the costs and values of postponement. In this chapter, we have provided a summary of such research conducted from the operations management perspective. In particular, we focused on the three enablers of postponement; namely, process standardization, process re-sequencing and component standardization. We presented models which provided insights on their benefits as well as industrial application of these strategies.

Clearly, there is a need for research and models along two additional dimensions. First, we need models that can be incorporated into decision support systems that allow managers to benefit from model based decision support. To achieve this, more emphasis is needed on large-scale models that capture the essential characteristics of the real environment and development of algorithms to solve those models in a fast and efficient manner. Second, most of the models developed this far relate to product postponement.
As a larger set of firms move towards service-oriented businesses, models that can capture postponement benefits in those environments are going to be extremely useful.

As more and more firms adopt the Internet technology to conduct their business on-line as well as have more interactions with their customers over the Internet, they are beginning to gather richer and detailed information about customer preferences. This has provided firms with an opportunity to tailor their products and services around customer preferences, i.e., mass customization. Postponement provides a powerful way for firms to pursue mass customization without incurring the usual huge operational costs associated with proliferating product variety. Indeed, as Feitzinger and Lee (1997) indicated, postponement is a strategy that allows firms to implement cost-effective mass customization. As a result, it is all the more important that new models related to postponement and other strategies for effective product variety management be studied and analyzed by future researchers.
References


