

Process Flexibility in Supply Chains

Stephen C. Graves • Brian T. Tomlin

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307

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

Chapel Hill, North Carolina 27599-3490

sgraves@mit.edu • brian_tomlin@unc.edu

Process flexibility, whereby a production facility can produce multiple products, is a critical design consideration in multiproduct supply chains facing uncertain demand. The challenge is to determine a cost-effective flexibility configuration that is able to meet the demand with high likelihood. In this paper, we present a framework for analyzing the benefits from flexibility in multistage supply chains. We find two phenomena, stage-spanning bottlenecks and floating bottlenecks, neither of which are present in single-stage supply chains, which reduce the effectiveness of a flexibility configuration. We develop a flexibility measure g and show that increasing this measure results in greater protection from these supply-chain inefficiencies. We also identify flexibility guidelines that perform very well for multistage supply chains. These guidelines employ and adapt the single-stage chaining strategy of Jordan and Graves (1995) to multistage supply chains.

(Supply Chain; Flexibility; Capacity; Product Allocation)

1. Introduction

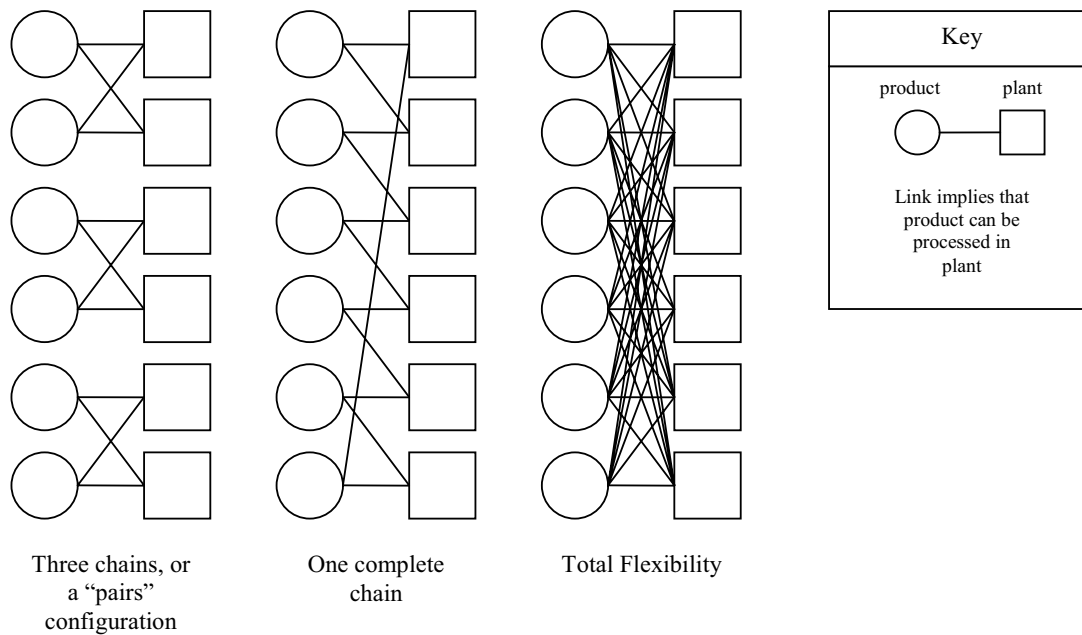
Manufacturing firms invest in plant capacity in anticipation of future product demand. At the time of capacity commitment, a firm has only a forecast of the unknown product demand. One approach to addressing forecast uncertainty is to build dedicated plants with sufficient capacity to cover the maximum possible demand. This strategy is expensive, and the expected capacity utilization is low. Flexibility provides an alternative means of coping with demand uncertainty. By enabling plants to process multiple products, a firm can allocate products to plants so as to meet realized demand most effectively.

A number of authors (e.g., Fine and Freund 1990; Gupta et al. 1992; Li and Tirupati 1994, 1995, 1997; Van Mieghem 1998) have examined investments in dedicated plants versus totally flexible plants, where a totally flexible plant can process all products. Partial flexibility, whereby a plant can produce a subset of products, has received less attention (Jordan and Graves 1995).

Jordan and Graves (J-G) investigate process flexibility in a single-stage manufacturing system with multiple products and plants. Flexibility is represented by a bi-partite graph, as shown for three configurations in Figure 1. Flexibility investments are then investments in these product-plant links.

J-G introduce the concept of chaining: "A chain is a group of products and plants which are all connected, directly or indirectly, by product assignment decisions. In terms of graph theory, a chain is a connected graph. Within a chain, a path can be traced from any product or plant to any other product or plant via the product assignment links. No product in a chain is built by a plant from outside that chain; no plant in a chain builds a product from outside that chain" (p. 580). In Figure 1, both the first and second configurations contain chains in which exactly two plants can process each product and each plant can process exactly two products. However, their performance is quite different. J-G demonstrate that the complete chain configuration, in which all products and plants are contained in one chain, and the chain is "closed," significantly outperforms the configuration

Figure 1 Examples of Flexibility Configurations



that has numerous distinct chains. In fact, the complete chain configuration performs remarkably like total flexibility in terms of expected throughput, even though it has far fewer product-plant links.

J-G develop three principles for guiding flexibility investments: (1) Try to equalize the capacity to which each product is directly connected, (2) try to equalize the total expected demand to which each plant is directly connected, and (3) try to create a chain(s) that encompasses as many plants and products as possible. These guidelines have been widely deployed in industry, including General Motors (J-G) and Ford (Kidd 1998).

Gavish (1994) extends the work of J-G on single-stage supply chains to investigate a specific seven-product, two-stage system, representing the component and assembly stages of an automotive supply chain. He shows that chaining is effective for this particular two-stage supply chain. However, his results depend on the supply chain chosen for study, and the extension to more complex multistage supply chains is not obvious. In this paper, we aim to understand the role of process flexibility in general

multistage supply chains and to develop insights into strategies for the deployment of process flexibility.

In §2, we develop the supply-chain framework used to evaluate flexibility. In §3, we show why inefficiencies cause multistage supply chains to differ from single-stage supply chains. Furthermore, we define and provide metrics for the inefficiencies. In §4, we introduce a flexibility measure, g , that can be used to classify supply chains. We also use analytic measures to show that increasing this flexibility measure reduces the supply-chain inefficiencies. In §5, we use simulation to further investigate the performance of flexibility configurations, and develop effective multistage flexibility guidelines. Conclusions are presented in §6.

2. The Model

We consider a multiproduct, multistage supply chain, consisting of K stages and I different products, with each product requiring processing at each stage. We do not assume any specific network structure for the supply chain but allow any general multistage production system in which each stage performs a distinct operation and requires its own processing resources.

An automotive supply chain might be modeled as a four-stage supply chain, comprising the component, engine, body, and final assembly operations.

Stage k , $k = 1, \dots, K$, has J_k different plants, where we use the term plant to refer to any capacitated processing resource. Product-plant links (i, j) at stage k are represented by an arc set A_k . At stage k , plant j can process product i iff $(i, j) \in A_k$. $P^k(i)$ defines the set of plants of stage k that can process i , i.e., $j \in P^k(i)$ iff $(i, j) \in A_k$. Similarly, we define the set of plants of stage k that can process one or more of the products in set M as $P^k(M) = \bigcup_{i \in M} P^k(i)$. To enable analytical tractability and simplify the presentation, we assume that all products i , such that $(i, j) \in A_k$, require the same amount of plant j 's capacity per unit processed. Thus, we define the capacity of plant j of stage k , c_j^k , to be the number of product units that can be processed in the planning horizon.

As is common in the flexibility and capacity planning literature (Eppen et al. 1989, Jordan and Graves 1995, Harrison and Van Mieghem 1999), we assume a two-stage sequential decision process. In the first stage, one determines the flexibility configuration for the supply chain, namely which products can be processed in each of the plants. In the second stage, demand is realized, and one allocates production capacity to meet demand. Thus, we choose the flexibility configuration when demand is uncertain and plan production after demand is realized.

To evaluate a flexibility configuration, we define a single-period production-planning problem that minimizes the amount of demand that cannot be met by the supply chain. For a given demand realization, $\mathbf{d} = \{d_1, \dots, d_I\}$, and flexibility configuration, $\mathbf{A} = \{A_1, \dots, A_K\}$, the production planning problem is the following linear program, $\mathbf{P1}(\mathbf{d}, \mathbf{A})$:

$$sf(\mathbf{d}, \mathbf{A}) = \text{Min} \left\{ \sum_{i=1}^I s_i \right\}$$

subject to

$$\begin{aligned} \sum_{(i,j) \in A_k} x_{ij}^k + s_i &\geq d_i, & i = 1, \dots, I, \quad k = 1, \dots, K, \\ \sum_{(i,j) \in A_k} x_{ij}^k &\leq c_j^k, & j = 1, \dots, J_k, \quad k = 1, \dots, K, \\ x_{ij}^k, s_i &\geq 0, \end{aligned}$$

where $sf(\mathbf{d}, \mathbf{A})$ is the total shortfall, s_i is the shortfall for product i , x_{ij}^k is the amount of product i processed in plant j at stage k over the planning horizon, and the other parameters are defined above. As noted earlier, to meet one unit of demand for product i , one needs one unit of capacity from each stage. For this model, we ignore temporal considerations in production planning.

When we determine the flexibility configuration, demand is a random vector denoted by $\mathbf{D} = \{D_1, \dots, D_I\}$ with a known distribution. The shortfall is a random variable, denoted as $SF(\mathbf{D}, \mathbf{A})$, that depends on the demand distribution and the flexibility configuration. For a given demand realization \mathbf{d} , the shortfall is $sf(\mathbf{d}, \mathbf{A})$, as found by solving $\mathbf{P1}(\mathbf{d}, \mathbf{A})$. We evaluate a given configuration \mathbf{A} by the expected total shortfall, $E[SF(\mathbf{D}, \mathbf{A})]$, where the expectation is over the demand random vector \mathbf{D} .

Although this framework suggests the formulation of an integer stochastic program to identify an optimal flexibility configuration, this is not the focus of this paper (see Birge and Louveaux 1997 for a stochastic program formulation for a single-stage flexibility problem). From our work with both GM and Ford, we have found that the final choice of supply-chain configuration is often influenced by strategic imperatives that are difficult to codify in a model. We have also learned that it can be challenging to accurately capture flexibility investment costs. Hence, industry practitioners are interested in tools that quickly identify a number of promising flexibility configurations that can then be further analyzed and modified. Therefore, in this paper we develop insights into what drives multistage supply-chain performance and then provide guidelines for the effective deployment of process flexibility in supply chains.

2.1. A Lower Bound for the Minimum Shortfall

We develop a lower bound for the minimum shortfall obtained in $\mathbf{P1}(\mathbf{d}, \mathbf{A})$, which we will use to understand how flexibility drives supply-chain performance.

THEOREM 1. (i) *A lower bound for the minimum shortfall in the production planning problem, $\mathbf{P1}(\mathbf{d}, \mathbf{A})$, is given*

by problem **P2**(\mathbf{d}, \mathbf{A}):

$$\begin{aligned} & \text{Max}_M \left\{ \sum_{i \in M} d_i - \min_{L_1, \dots, L_K} \left\{ \sum_{k=1}^K \sum_{j \in P^k(L_k)} c_j^k \right\} \right\} \\ \text{subject to} \quad & M \subseteq \{1, \dots, I\}, \\ & L_k \cap L_{k'} = \emptyset \quad \forall k \neq k', \\ & \bigcup_{k=1}^K L_k = M. \end{aligned}$$

(ii) If either the number of stages or the number of products is less than three, then the lower bound is exact, i.e., the minimum shortfall in **P1**(\mathbf{d}, \mathbf{A}) equals the optimum value for **P2**(\mathbf{d}, \mathbf{A}).

For a proof of the theorem, see Tomlin (2000). As an explanation, we note that the shortfall equals total demand minus total production. An upper bound on the total production is given by

$$\sum_{i \notin M} d_i + \min_{L_1, \dots, L_K} \left\{ \sum_{k=1}^K \sum_{j \in P^k(L_k)} c_j^k \right\},$$

where M is any subset of products and the L_k 's partition M . The second term is an upper bound (due to capacity) on the total production for the set of products M . The first term is the total demand for the remaining products, and hence an upper bound on the production for these products. By subtracting the upper bound on total production from the total demand, we obtain a lower bound on the total shortfall, equal to the objective function in **P2**(\mathbf{d}, \mathbf{A}); solving **P2**(\mathbf{d}, \mathbf{A}) provides the largest such lower bound. This lower bound is a multistage generalization of the shortfall expression, $V(A)$, of J-G.

In general, the expression in Theorem 1 provides a strict lower bound and not the actual shortfall. Tomlin (2000) shows that if the dual solution to **P1**(\mathbf{d}, \mathbf{A}) is integral, then the lower bound is exact. As a consequence, the lower bound is exact for the following supply-chain types: Supply chains with less than three stages, totally flexible supply chains, and totally dedicated supply chains. Experimental results indicate that it is also exact for a much wider class of supply chains. We use Theorem 1 in developing the analytical results in §4. For the simulations in §5, we solve **P1**(\mathbf{d}, \mathbf{A}) exactly.

3. Supply-Chain Inefficiencies

Consider a single-stage supply chain in which stage k is the only stage. The shortfall for such a supply chain is termed the stand-alone shortfall, which we denote as $E[SF_k(\mathbf{D}, \mathbf{A}_k)]$, for stage k . Without loss of generality, suppose that stage 1 is the *stand-alone bottleneck*, i.e., it has the greatest expected stand-alone shortfall:

$$E[SF_1(\mathbf{D}, \mathbf{A}_1)] = \text{Max}_{k=1, \dots, K} \{E[SF_k(\mathbf{D}, \mathbf{A}_k)]\}.$$

How does the supply chain perform overall relative to the stand-alone bottleneck? In this section, we present two multistage supply-chain phenomena that lead to inefficiencies by which the multistage supply chain performs worse than the stand-alone bottleneck; that is, $E[SF(\mathbf{D}, \mathbf{A})] \geq E[SF_1(\mathbf{D}, \mathbf{A}_1)]$. We define and measure this configuration inefficiency, CI, as follows:

$$\text{CI} = 100 \times \left(\frac{E[SF(\mathbf{D}, \mathbf{A})] - E[SF_1(\mathbf{D}, \mathbf{A}_1)]}{E[SF_1(\mathbf{D}, \mathbf{A}_1)]} \right).$$

The CI is the relative increase in expected shortfall resulting from the interaction of the multiple stages in the supply chain.

One way to avoid this inefficiency is to make stage 1 the bottleneck for all products for all possible demand outcomes. In effect, we set the capacity at the other stages sufficiently high so that these stages are never a constraint. However, this is likely to be very expensive due to the cost of the excess capacity. Alternatively, we might configure every stage to be identical so that each stage is an exact replica of the other stages. In this case, the production-planning problem collapses to single-stage problem; the shortfall is always given by the shortfall of stage 1, and the inefficiency is zero. However, such a policy may be prohibitively difficult, if not impossible, to employ for reasons of cost, technical feasibility, and/or challenges in interstage design coordination.

The supply-chain CI is caused by two phenomena: floating bottlenecks and stage-spanning bottlenecks.

The floating bottleneck is a direct result of demand uncertainty. If demand were certain, then the bottleneck for the supply chain is the stage with maximum stand-alone shortfall, namely stage 1. However, for uncertain demand, the fact that

$$\text{Max}_{k=1, \dots, K} \{E[SF_k(\mathbf{D}, \mathbf{A}_k)]\} = E[SF_1(\mathbf{D}, \mathbf{A}_1)]$$

does not imply that for every demand realization

$$\text{Max}_{k=1,\dots,K} \{sf_k(\mathbf{d}, \mathbf{A}_k)\} = sf_1(\mathbf{d}, \mathbf{A}_1).$$

In other words, for any demand realization, the stand-alone bottleneck need not be stage 1, but can float from one stage to another. Therefore,

$$E\left[\text{Max}_{k=1,\dots,K} \{SF_k(\mathbf{D}, \mathbf{A}_k)\}\right] \geq \text{Max}_{k=1,\dots,K} \{E[SF_k(\mathbf{D}, \mathbf{A}_k)]\},$$

where we say there is a floating inefficiency in the supply chain if the inequality is strict. We define and measure the inefficiency from floating bottlenecks as the relative increase in the expected maximum stand-alone shortfall over the expected shortfall for the stand-alone bottleneck:

$$\text{CFI}=100 \times \left(\frac{E[\text{Max}_{k=1,\dots,K} \{SF_k(\mathbf{D}, \mathbf{A}_k)\}] - \text{Max}_{k=1,\dots,K} \{E[SF_k(\mathbf{D}, \mathbf{A}_k)]\}}{\text{Max}_{k=1,\dots,K} \{E[SF_k(\mathbf{D}, \mathbf{A}_k)]\}} \right).$$

In §5, we use simulation to measure the protection various flexibility configurations provide against this floating inefficiency. Another measure of this inefficiency is the probability that stage 1 is the bottleneck stage. We use this measure in §4 to develop analytic measures of the protection that flexibility provides.

The notion of a floating bottleneck has previously been noted in the context of machine shops (see Hopp and Spearman 1996, p. 515). In the machine shop context, floating bottlenecks arise due to machines having different processing rates for products. In the supply-chain context here, it arises due to supply chains having partial flexibility. We also extend the literature

by providing a measure of this inefficiency in supply chains, quantifying its effect via simulation and developing flexibility strategies that protect against this inefficiency.

Care was taken in the preceding paragraphs to discuss the “stand-alone” bottleneck stage rather than simply the bottleneck stage. The reason lies in the second cause of inefficiency, the stage-spanning bottleneck. Floating bottlenecks only arise if demand is uncertain; the stage-spanning bottleneck can manifest itself even if demand is certain.

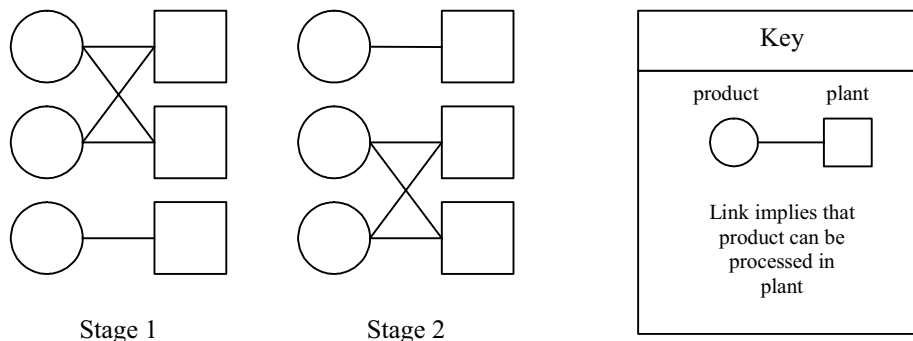
For a given demand realization, we define a stage-spanning bottleneck to occur whenever

$$sf(d, \mathbf{A}) > \text{Max}_{k=1,\dots,K} \{sf_k(d, \mathbf{A}_k)\};$$

that is, the supply-chain shortfall is strictly greater than the maximum stand-alone shortfall. Consider the set-based formulation, $\mathbf{P2}(\mathbf{d}, \mathbf{A})$, for the shortfall. A stage-spanning bottleneck occurs if the solution to $\mathbf{P2}(\mathbf{d}, \mathbf{A})$ has more than one nonempty L_k . The L_k sets identify the plants that limit production. If more than one set is nonempty, the plants that limit production span multiple stages, hence the term *stage-spanning bottleneck*.

As an example, consider the two-stage three-product supply chain shown in Figure 2. Recall that $\mathbf{P2}(\mathbf{d}, \mathbf{A})$ gives the exact shortfall for two-stage supply chains. Let the demands for products 1, 2, and 3 be 150, 50, and 150 units, respectively, and the capacity be 100 units for all plants. The shortfall of either stage on a stand-alone basis is 50 units. For $\mathbf{P2}(\mathbf{d}, \mathbf{A})$, the optimal product sets are $M^* = \{1, 3\}$, $L_1^* = \{3\}$, and

Figure 2 A Supply Chain That Suffers from Stage-Spanning Bottlenecks



$L_2^* = \{1\}$, and the supply-chain shortfall is 100 units. The bottleneck plants are those plants that can process the products in L_k^* . Thus, plant 3 in stage 1 is a bottleneck, as is plant 1 in stage 2.

This example demonstrates that the bottleneck for the supply chain need not be a single stage. We show in §5, that even reasonable flexibility designs can result in stage-spanning bottlenecks occurring with a high probability. This then increases the overall supply-chain inefficiency. We define and measure the spanning inefficiency as follows:

$$CSI = 100 \times \left(\frac{E[SF(\mathbf{D}, \mathbf{A})] - E[\text{Max}_{k=1, \dots, K}\{SF_k(\mathbf{D}, \mathbf{A}_k)\}]}{\text{Max}_{k=1, \dots, K}\{E[SF_k(\mathbf{D}, \mathbf{A}_k)]\}} \right).$$

In §5, we use simulation to measure the protection various flexibility configurations provide against this spanning inefficiency. Another measure of this inefficiency is the probability of occurrence of stage-spanning bottlenecks. We use this in §4 to develop analytic measures of the protection that flexibility provides.

Note that the overall inefficiency, CI, is the sum of the floating and spanning inefficiencies, i.e., $CI = CFI + CSI$. Therefore, we can measure the relative importance of both to the overall inefficiency. By developing flexibility policies that offer protection against both the floating and spanning inefficiencies, we provide protection against the overall inefficiency.

Zero inefficiency does not guarantee a well-designed supply chain. For instance, consider a supply chain with equally-sized dedicated plants at each stage. This results in zero inefficiency, because the performance of the supply chain is the same as that for a single stage. But the performance might be quite poor if the single-stage dedicated supply chain performs poorly. To assess the relative performance of a flexibility configuration, we define the configuration loss, CL:

$$CL = 100 \times \left(\frac{E[SF(\mathbf{D}, \mathbf{A})] - E[SF(\mathbf{D}, TF)]}{E[SF(\mathbf{D}, TF)]} \right),$$

where $E[SF(\mathbf{D}, TF)]$ is the expected shortfall for a totally flexible supply chain. The configuration loss is the relative increase in expected shortfall resulting from the supply chain not being totally flexible.

4. A Flexibility Measure for Supply Chains

We increase the flexibility of a supply chain by adding product-plant links. For a subset of products M , adding links to plants that cannot currently process products in M , increases the capacity available to the subset M , and decreases the probability that demand for the subset M cannot be met. We propose a flexibility measure based on the excess capacity available to any subset of products, relative to an equal-share allocation of the capacity. Consider a single-stage supply chain that processes I products and has J plants with a total stage capacity C_T , equal to the sum of plant capacities. Thus, an equal allocation of the capacity would provide each product with capacity $\bar{c} = C_T/I$. For any subset of products M , the difference in available capacity between the supply chain and an equal allocation, is given by $\sum_{j \in P(M)} c_j - |M|\bar{c}$. Expressing this excess capacity in units of the equal allocation, we define the excess capacity as

$$g(M) = \frac{\sum_{j \in P(M)} c_j - |M|\bar{c}}{\bar{c}} = \frac{\sum_{j \in P(M)} c_j}{\bar{c}} - |M|.$$

In words, $g(M)$ is the excess capacity (over its equal allocation) available to M , as measured in units of the equal allocation. This measure $g(M)$ is increasing in the process flexibility, namely, it increases as we add product-plant links.

We measure the flexibility of a single-stage supply chain by

$$g = \text{Min}_M \{g(M) : |P(M)| < J\};$$

that is, we take the minimum value over all product subsets that do not have access to the total stage capacity. This restriction is put in place as the excess capacity is bounded above by the total stage capacity. For the case of total flexibility, where all product subsets have access to the total stage capacity, we use the convention that $g = I - 1$.

The value of g provides a lower bound on the amount of excess capacity, measured in units of the equal allocation, which is available to any subset of products that does not have access to the total stage capacity. We note the relationship between flexibility and capacity from the following expression:

$$\sum_{j \in P(M)} c_j \geq \text{Min}\{(|M| + g)\bar{c}, C_T\}.$$

A larger value of g indicates that a larger fraction of the stage capacity is available to product subsets.

We developed the flexibility measure, g , for a single-stage supply chain. For multistage supply chains, we define g to be the minimum value of the g values for the individual stages. We now proceed to show the relationship between this measure and the supply-chain inefficiencies.

4.1. An Analytic Measure for the Spanning Inefficiency Based on the g -Value

Consider a set of products M , with subsets L_1, \dots, L_K , as defined in **P2(d, A)**. For set M , we define the set of plants that span multiple stages to be $\{j: j \in P^k(L_k), k = 1, \dots, K\}$.

THEOREM 2. *Consider a supply chain processing I products, in which each stage has a total capacity at least as large as the total expected demand and in which the I product demands are iid normally distributed. For the set of plants that span multiple stages, denoted by (M, L_1, \dots, L_k) , the probability that this set is the bottleneck, i.e., limits the production, is bounded above by*

$$\Omega_s(I, g) = \left(\Phi \left[\frac{-2g}{CV\sqrt{I/2}} \right] \right)^2,$$

where g is the flexibility measure for the supply-chain configuration, CV is the coefficient of variation for the individual product demands, and Φ is the standard normal cumulative distribution function.

For the proof, see Tomlin (2000). Note that this is not an upper bound on the probability of a stage-spanning bottleneck occurring, but rather an upper

bound on the probability of any stage-spanning set of plants being the bottleneck. However, if this upper bound is small, we conjecture that the probability of occurrence of a stage-spanning bottleneck is also low, and hence the spanning inefficiency should be small. The upper bound measure increases in the number of products and the coefficient of variation of demand. It decreases in g , but with diminishing returns.

We present selected numeric values for the upper bound in Table 1 (where we set values less than 1.0E-10 to zero). As can be seen, the upper bound decreases rapidly as g increases. For $g = 1$, the measure is less than 0.001 if the number of products is less than 25, and there would appear to be little benefit to increasing the flexibility measure g beyond 1. There may be a benefit if the number of products is large, e.g., above 25. We examine this observation by means of simulation in §5.

4.2. An Analytic Measure for the Floating Inefficiency Based on the g -Value

As noted in §3, a floating bottleneck occurs when stage 1 (the stage with the greatest expected shortfall) is not the bottleneck stage for a given demand realization. In this section, we develop a measure for the floating inefficiency, under the assumption that each stage has the same total capacity.

For any demand realization, the stand-alone shortfall at stage 1 is at least as large as the stand-alone shortfall for stage 1 under total flexibility. Therefore, if the stand-alone shortfalls for all other stages do not exceed the stage 1 shortfall under total flexibility, then stage 1 is a bottleneck stage. This is a suf-

Table 1 Values for Ω_s as the Flexibility Measure g , Increases from 0 to 3 for a CV of 0.3

Number of products	g										
	0	0.1	0.2	0.4	0.6	0.8	1	1.5	2	2.5	3
5	0.250	0.113	0.040	0.002	3.3E-05	1.4E-07	1.5E-10	0	0	0	0
10	0.250	0.147	0.076	0.014	0.001	7.3E-05	2.1E-06	0	0	0	0
15	0.250	0.163	0.098	0.027	0.005	0.001	5.6E-05	1.7E-08	0	0	0
20	0.250	0.173	0.113	0.040	0.011	0.002	3.1E-04	6.1E-07	1.5E-10	0	0
25	0.250	0.181	0.125	0.051	0.017	0.004	8.8E-04	5.5E-06	6.6E-09	0	0
30	0.250	0.186	0.133	0.060	0.023	0.007	0.002	2.4E-05	8.3E-08	0	0
35	0.250	0.191	0.141	0.069	0.029	0.010	0.003	7.1E-05	5.2E-07	1.1E-09	0
40	0.250	0.194	0.147	0.076	0.034	0.014	0.005	1.6E-04	2.1E-06	9.4E-09	0

ficient, but not necessary, condition for stage 1 to be the bottleneck. So a lower bound on the probability that a floating bottleneck does not occur is

$$\text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) \leq SF(\mathbf{D}, TF_1), \forall k = 2, \dots, K],$$

where $SF_k(\mathbf{D}, TF_1)$ equals the stand-alone shortfall for stage 1 when it is totally flexible.

Due to the assumption that each stage has the same total capacity, we can express the lower bound as

$$\text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) = SF(\mathbf{D}, TF_k), \forall k = 2, \dots, K],$$

where we replace the inequality by an equality as the shortfall is never strictly less than the total flexibility shortfall.

Simulation evidence indicates that the probabilities of stand-alone shortfalls equaling total flexibility shortfalls are positively correlated. Assuming this to be true, we can restate the lower bound on the probability of no bottleneck as

$$\prod_{k=2}^K (\text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) = SF(\mathbf{D}, TF_k)])$$

or

$$\prod_{k=2}^K (1 - \text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) > SF(\mathbf{D}, TF_k)]).$$

Therefore,

$$1 - \prod_{k=2}^K (1 - \text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) > SF(\mathbf{D}, TF_k)])$$

is an upper bound on the probability of a floating bottleneck.

To quantify this bound, we need an approximation for

$$\text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) > SF(\mathbf{D}, TF_k)],$$

as we have not found a closed-form expression. Using Theorem 1(ii), we find

$$\begin{aligned} & \text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) > SF(\mathbf{D}, TF_k)] \\ &= \text{Prob}\left[\text{Max}_M \left\{ \sum_{i \in M} D_i - \sum_{j \in P^k(M)} c_j^k \right\} \right. \\ & \quad \left. > \text{Max} \left\{ \sum_{i=1}^I D_i - \sum_{j=1}^{J_k} c_j^k, 0 \right\} \right] \end{aligned}$$

$$\begin{aligned} & \approx \text{Max}_M \left\{ \text{Prob} \left[\sum_{i \in M} D_i - \sum_{j \in P^k(M)} c_j^k \right. \right. \\ & \quad \left. \left. > \text{Max} \left\{ \sum_{i=1}^I D_i - \sum_{j=1}^{J_k} c_j^k, 0 \right\} \right] \right\}, \end{aligned}$$

where we use the same approximation as J-G. Namely, we use the maximum probability that the shortfall induced by a set of products M exceeds that for total flexibility. If this probability is small, then we expect that

$$\text{Prob}[SF_k(\mathbf{D}, \mathbf{A}_k) > SF(\mathbf{D}, TF_k)]$$

is also small. We now define a measure for the floating inefficiency:

$$\begin{aligned} \Omega_F &= 1 - \prod_{k=2}^K \left(1 - \text{Max}_M \left\{ \text{Prob} \left[\sum_{i \in M} D_i - \sum_{j \in P^k(M)} c_j^k \right. \right. \right. \\ & \quad \left. \left. \left. > \text{Max} \left\{ \sum_{i=1}^I D_i - \sum_{j=1}^{J_k} c_j^k, 0 \right\} \right] \right\} \right), \end{aligned}$$

for which we have the following result.

THEOREM 3. Consider a K stage supply chain that processes I products, in which each stage has a total capacity at least as large as the total expected demand and in which the I product demands are iid normally distributed. Then,

$$\Omega_F(I, g) = 1 - \left[1 - \left(\Phi \left[\frac{-g}{CV\sqrt{I/2}} \right] \right)^2 \right]^{K-1},$$

where g is the flexibility measure for the supply-chain configuration, CV is the coefficient of variation for the individual products, and Φ is the standard normal cumulative distribution function.

For the proof, see Tomlin (2000). Ω_F increases in the number of stages, in the number of products, and in the coefficient of variation of demand. This suggests that floating bottlenecks are more likely in supply chains as the number of stages, the number of products, and/or the coefficients of variation increase. It decreases in the flexibility measure g .

Table 2 shows Ω_F as the value of g increases for a ten-stage supply chain for various numbers of products. As can be seen it decreases rapidly as g increases.

Table 2 Values for Ω_F as the Flexibility Measure, g , Increases from 0 to 3 for a CV of 0.3 and $K = 10$

Number of products	g										
	0	0.2	0.5	0.75	1	1.25	1.5	1.75	2	2.5	3
5	0.925	0.661	0.176	0.029	0.003	0.000	0.000	0.000	0.000	0.000	0.000
10	0.925	0.760	0.382	0.146	0.041	0.009	0.001	0.000	0.000	0.000	0.000
15	0.925	0.799	0.498	0.258	0.107	0.036	0.010	0.002	0.001	0.000	0.000
20	0.925	0.820	0.570	0.346	0.176	0.076	0.029	0.009	0.003	0.000	0.000
25	0.925	0.834	0.619	0.413	0.239	0.121	0.054	0.022	0.008	0.001	0.000
30	0.925	0.844	0.654	0.466	0.294	0.165	0.084	0.039	0.016	0.002	0.000
35	0.925	0.851	0.681	0.507	0.341	0.207	0.115	0.058	0.027	0.005	0.001
40	0.925	0.857	0.702	0.541	0.382	0.246	0.146	0.080	0.041	0.009	0.001

For $g = 1$, it is less than 0.1 if the number of products is less than 15. There would appear to be little benefit, in terms of decreasing the probability of a floating bottleneck, to increasing g beyond 1, unless the number of stages or products is large, e.g. above 15.

5. Evaluation of Flexibility Policies Using Simulation

Having developed a flexibility measure, g , and provided evidence that increasing the flexibility measure from 0 to 1 dramatically improves performance, we use simulation to confirm this hypothesis. We also test the hypothesis that supply chains with a large number of products or stages may need extra flexibility, that is the g -value should be greater than 1.

For the experiments, we assume that the supply chain has I equal capacity plants (100 units) at each stage, where I is the number of products. $I = 10$ unless otherwise stated. We assume that product demands are iid normal, $N(\mu, \sigma)$, truncated at $\mu \pm 2\sigma$, with $\mu = 100$ and $\sigma = 30$, unless otherwise stated. For a given supply-chain configuration, the product demand vector \mathbf{d} is randomly generated. For each demand realization, we determine the shortfall by solving **P1**. We used 10,000 demand realizations to generate the estimates for the expected shortfall values. The 95% confidence intervals for the expected shortfall estimates were calculated (Law and Kelton 1991) and found to be within $\pm 3\%$ of the estimates.

Consider the pairs configuration shown in Figure 1. The g -value for this configuration equals 0. To see

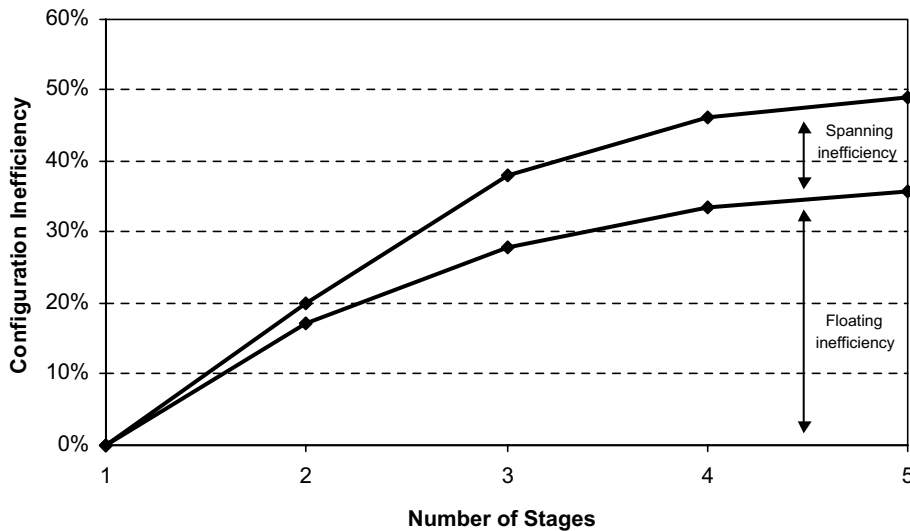
this, let M equal the first two products. In the simulation experiments, we use a pairs configuration policy to create supply chains with $g = 0$.

Consider the complete chain configuration depicted in Figure 1. It has two plants connected to each product. For equal-capacity plants, the g -value equals 1 for this chain configuration. As shown in J-G, one can have chain configurations with more than two plants connected to each product. For I products and I equal-capacity plants, if the number of plants connected to each product equals h , then the g -value equals $h - 1$ as $g(M) \geq h - 1$ for all M such that $|P(M)| < I$. We note that for a single-stage supply chain with I products and I equal-capacity plants, any configuration with a g -value equal to $h - 1$ must have at least hI product-plant links and thus has at least as many links as a complete chain in which each product is connected to h plants (such a configuration has hI links). The chaining configuration of J-G is therefore very efficient in terms of the number of product-plant links required for given g -values. In the simulation experiments, we use a chaining policy to create supply chains with $g = 1$ and 2.

By a configuration policy, we mean that each stage in the supply chain is configured according to that policy, e.g., a complete chain with $g = 1$. However, the policy is not one of replication. Rather, the particular pairs or chain configuration can differ for each stage and is chosen randomly in the simulation tests.

In §4, we provided analytical evidence that $g = 0$ supply chains would suffer from both floating and spanning inefficiencies. Simulation of the performance of the $g = 0$ policy confirms this, as can be seen

Figure 3 Both Floating and Spanning Inefficiencies Are Significant for $g = 0$ Configurations



from Figure 3. The numerical data for all figures can be obtained from the second author.

The configuration inefficiency for a five-stage supply chain is 49%: the expected shortfall for the five-stage supply chain is 49% higher than a single-stage supply chain. Both types of inefficiencies are present. The floating inefficiency contributes four-fifths of the total inefficiency while the spanning inefficiency contributes one-fifth. A stage-spanning bottleneck occurs with a frequency of 20% in two-stage supply chains and 68% in five-stage supply chains. This

confirms the analytic evidence in §4 that $g = 0$ supply chains have a high likelihood of stage-spanning bottlenecks.

From §4, we expect that increasing the flexibility measure to $g = 1$ should reduce the supply-chain inefficiency. Simulation results confirm this. Figure 4 shows the inefficiency for $g = 1$ and $g = 2$ supply chains. For a five-stage supply chain, the inefficiency is 6% for a $g = 1$ policy and 0% for a $g = 2$ policy. This compares to 49% for a $g = 0$ policy. Therefore, as conjectured from analytic evidence in §4, increasing

Figure 4 The Inefficiency Is Much Lower for $g = 1$ and $g = 2$ Supply Chains than for $g = 0$ Supply Chains

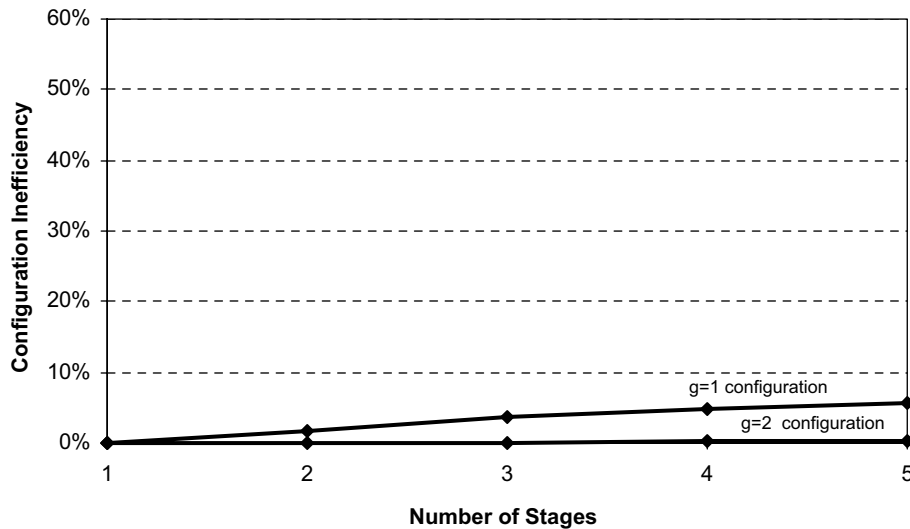
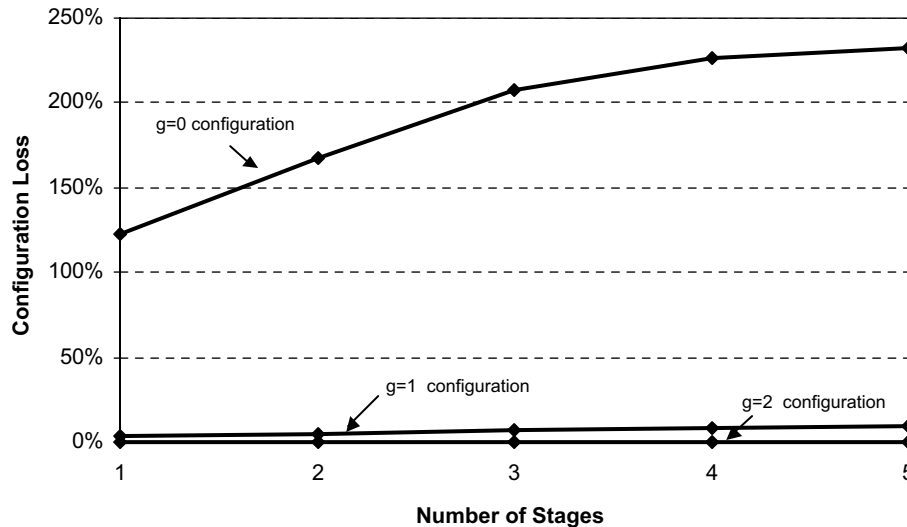


Figure 5 Supply Chains with g -Values = 1 and 2 Significantly Outperform Supply Chains with g -Values = 0



the flexibility measure, g , from 0 to 1 has dramatic results. Increasing it beyond 1 has limited value.

We note that stage-spanning bottlenecks did not occur once in any of the simulations for $g = 1$ and 2 supply chains: the inefficiency is entirely due to floating bottlenecks.

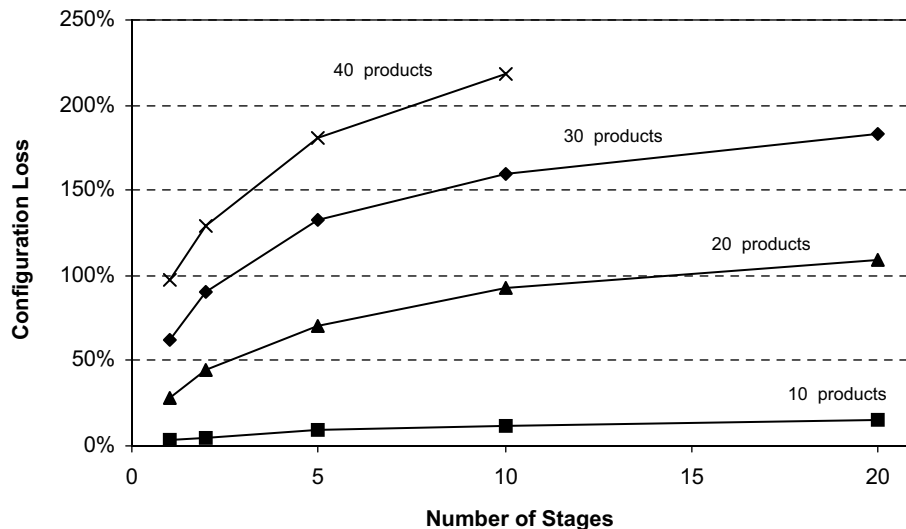
In §4, the analytic measure used for the probability of stage-spanning bottlenecks increases in the coefficient of variation. We tested this by increasing the coefficient of variation from 0.3 to 0.5. In this case, stage-spanning bottlenecks did occur for $g = 1$ supply chains, with a frequency of 0.67% but once again did not occur for $g = 2$ supply chains. This confirms that the coefficient of variation impacts the likelihood of stage-spanning bottlenecks. It also shows that supply chains with g -values of 1 or greater offer great protection against this inefficiency.

As noted earlier, inefficiency is only one measure of the supply-chain performance. It measures the impact due to the interaction of multiple stages in the supply chain. The configuration loss, introduced earlier, provides another measure of supply-chain performance, namely the relative increase in expected shortfall (resulting from partial flexibility) over total flexibility. Figure 5 shows the configuration loss for $g = 0, 1$, and 2 configurations. The $g = 0$ configuration performs extremely poorly. For a five-stage supply chain, the configuration loss is 232%: the expected

shortfall is 232% higher than for a totally flexible supply chain. There are two reasons for the poor performance. First, a single-stage supply chain configured using a $g = 0$ (pairs) policy chain does not perform well. Second, this is compounded in multistage supply chains by the fact that the configuration inefficiency is large for a $g = 0$ policy. In comparison, the $g = 1$ and 2 policies work extremely well. For a five-stage supply chain, the configuration loss is 9% for $g = 1$ and 0.3% for $g = 2$. The performance of supply chains with a g -value greater than or equal to 1 approaches that of total flexibility.

In §4, we suggested that supply chains with large numbers of products or stages might need extra flexibility above and beyond $g = 1$. Figure 6 confirms this. For $g = 1$ supply chains, the configuration loss becomes significant as the number of products or stages increases. However, a similar simulation for $g = 2$ supply chains resulted in a maximum configuration loss of 1% over all test cases, indicating that limited flexibility configured correctly offers the benefits of total flexibility even in supply chains with 30 products and 20 stages. We note that a $g = 2$ supply chain with 30 products requires 90 product-plant links per stage, whereas a totally flexible supply chain would require 900 product-plant links per stage.

Figure 6 Supply Chains with $g = 1$ Perform Less Well as the Number of Products Increase



5.1. The Recommended Flexibility Policy

From the analytic and simulation results, we see that supply chains with g -values greater than or equal to 1 perform very well. Furthermore, the chaining strategy of J-G provides an efficient way, in terms of the number of product-plant links, to achieve g -values of 1 or greater. Another important benefit from the chaining policy is that the supply chain performs well even when each stage is not identically configured. As long as each stage uses a chaining policy, there is no need to coordinate the exact chain design between stages. To see this, look again at Figures 4 and 6. A replicating strategy where each stage has an identical configuration, performs equivalent to a single-stage supply chain. For chains, because the inefficiency is low, the loss in performance from allowing a random chain at each stage is low. This is not true for a pairs configuration.

We therefore propose the following guidelines for designing flexibility in supply chains. In supply chains with a moderate number of stages and products (based on the analytical measure and our simulation experience we would suggest 10 or fewer stages and 20 or fewer products), let each stage follow the guidelines in J-G. That is, (1) create chains that encompass as many plants and products as possible, with ideally all plants and products as part of one single chain; (2) equalize the number of plants, measured in

total units of capacity, to which each product in the chain is directly connected; and (3) equalize the number of products, measured in total units of expected demand, to which each plant in the chain is directly connected. For supply chains with more products or stages, consider increasing the flexibility by directly connecting products to more plants but keeping a chain configuration. Note that the particular instance of flexibility policy implemented does not need to be coordinated between stages.

Results in this paper assume equal capacity usage of products in plants. We tested the robustness of the policy with unequal capacity usages and found that the guidelines remain valid; see Tomlin (2000) for details.

6. Conclusions

We identify and provide metrics for two inefficiencies that impact the performance of multistage supply chains, termed floating and stage-spanning bottlenecks. A general flexibility measure, g , is developed. We show analytically that flexibility configurations with values of g greater than or equal to one provide effective protection against these inefficiencies. We use simulation to confirm this and to show that lower values of g are associated with high inefficiencies. We show that a chaining policy is very efficient at delivering g -values greater than or equal to one.

The chaining policy, augmented in the case of supply chains with large numbers of products or stages, is shown to be an effective flexibility policy for multi-stage supply chains. Importantly, we demonstrate that each stage in the supply chain can design its particular chain in isolation and the overall supply chain will still perform well.

Acknowledgments

The authors acknowledge the input from Bill Jordan of General Motors in helping provide context and direction for this research. We would also like to thank the referees, the associate editor, and the departmental editor for helpful and constructive suggestions that led to a substantial improvement over a prior version of this paper. This research was carried out with the support of the MIT Leaders for Manufacturing Program. The first author also acknowledges the support provided by the Singapore-MIT Alliance in completing this work. The second author also acknowledges the support provided by the Burrell Faculty Development and Support Fund in completing this work.

References

- Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer, New York.
- Eppen, G. D., R. K. Martin, L. Schrage. 1989. A scenario approach to capacity planning. *Oper. Res.* **37**(4) 517–527.
- Fine, C. H., R. M. Freund. 1990. Optimal investment in product flexible manufacturing capacity. *Management Sci.* **36**(4) 449–466.
- Gavish, R. 1994. Flexibility in multi-product multi-assembly multi-component systems. Masters thesis, MIT, Cambridge, MA.
- Gupta D., Y. Gerchak, J. A. Buzacott. 1992. The optimal mix of flexible and dedicated manufacturing capacities: Hedging against demand uncertainty. *Internat. J. Prod. Econom.* **28** 309–319.
- Harrison, J. M., J. A. Van Mieghem. 1999. Multi-resource investment strategies: Operational hedging under demand uncertainty. *Eur. J. Oper. Res.* **113** 17–29.
- Hopp, W. J., M. L. Spearman. 1996. *Factory Physics: Foundations of Manufacturing Management*. Irwin, Chicago, IL.
- Jordan, W. C., S. C. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management Sci.* **41**(4) 577–594.
- Kidd, S. L. 1998. A systematic method for valuing a product platform strategy. Thesis, Leaders for Manufacturing Program, MIT, Cambridge, MA.
- Law, A. M., W. D. Kelton. 1991. *Simulation Modeling and Analysis*. McGraw-Hill, New York.
- Li, S., D. Tirupati. 1994. Dynamic capacity expansion problem with multiple products: technology selection and timing of capacity additions. *Oper. Res.* **42**(5) 958–976.
- , ———. 1995. Technology choice with stochastic demands and dynamic capacity allocation: A two product analysis. *J. Oper. Management* **12** 239–258.
- , ———. 1997. Impact of product mix flexibility and allocation policies on technology. *Comput. Oper. Res.* **24**(7) 611–626.
- Tomlin, B. 2000. Supply chain design: Capacity, flexibility and wholesale price strategies. Ph. D. thesis, Sloan School of Management, MIT, Cambridge, MA.
- Van Mieghem, J. A. 1998. Investment strategies for flexible resources. *Management Sci.* **44**(8) 1071–1078.

Accepted by Christopher Tang, former department editor; received March 2000. This paper was with the authors 9 months for 2 revisions.