Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make Schedule

William S. Lovejoy, Kannan Sethuraman,

To cite this article:
William S. Lovejoy, Kannan Sethuraman, (2000) Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make Schedule. Manufacturing & Service Operations Management 2(3):221-239. http://dx.doi.org/10.1287/msom.2.3.221.12348

Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make Schedule

William S. Lovejoy • Kannan Sethuraman
School of Business Administration, University of Michigan, 701 Tappan, Ann Arbor, Michigan 48109-1234
Indian Institute of Management, Vastrapur, Ahmedabad, 380015, India
wlovejoy@umich.edu • eskay@iimadh.ernet.in

In a firm that makes schedule, orders are always processed within a fixed time frame. In a congested facility, such a situation would be impossible using conventional queuing logic. We propose a conceptual model of a firm in which workers make schedule by rushing jobs, if necessary, with potential quality consequences. Hence, time and quality are substitutes, a feature that we recognize explicitly in our definition of the firm's capacity. Production yields are not exogenous parameters but endogenously determined by workers responding to schedule pressures. The plant manager can authorize overtime to relieve this pressure or live with the quality consequences of rushing. The model reveals the close relationships among the firm's workforce policies, the integrity of the inspection system, and the cost performance of the firm as its volume and/or product line expands. We consider pure congestion (driven by volume) and pure complexity (driven by product line breadth) effects in the context of our plant performance model. We consider two root causes of complexity costs: time and quality. The time effects of complexity will only have cost consequences in congested facilities, but quality effects are always present. Also in contrast to time effects, quality effects of complexity can be present in nonbottleneck workstations. Hence, the quality consequences of complexity can be as, or more important than the time consequences.
(*Product Line Complexity; Congestion; Capacity; Quality*)

This article proposes a conceptual model of congestion and complexity costs in a plant with fixed resources that strives to make schedule. "Fixed resources" means that productive resources (labor and capital equipment) are fixed in the near to intermediate term. "Making schedule" means that plant managers are expected to process all jobs released to their shop in a fixed amount of time. "Congestion costs" refers to the costs of imposing demand rates close to capacity on plant resources in a variable environment. Classical treatments of congestion costs focus on inventory and delay costs in queuing models of production systems. "Complexity costs" refers to the costs of performing a heterogeneous rather than homogeneous set of tasks. Classical treatments of complexity costs are discussed shortly.

Our model was developed to reflect reality in a client firm that makes large durables for predominantly business clients. Our charge was to provide recommendations regarding a potential initiative to redesign the firm's product line around common modules. Management felt that this would be a good way to reduce their complexity costs, which they believed to be very high. However, the firm is privately held, and data

1523-4614/00/0203/0221$05.00
1526-5498 electronic ISSN

records are absent or unclear in many important areas. The recommendations were to be made on a time scale that precluded a careful empirical study. Our input had to be in the form of a conceptual framing of what was going on and recommendations based on that framing. The alternative was to make decisions absent our input. We had a student team working with us on site.

The current literature on the costs of complexity is dominated by time-based logic. In essence, a heterogeneous product line requires the processing of an increased number of transactions, which require time on resources (especially in support functions, cf. Miller and Vollman 1985). Likewise, change-over losses between products impose nonproductive time on the system. As product variety increases, the plant handles higher variety with less effective capacity, and queuing delays (shipping delays or inventory costs) accrue. Indeed, to date the only analytical model of complexity that the authors are aware of (Banker et al. 1988) uses an M/G/1 queuing model of a bottleneck workstation to show how higher variety imposes higher delays and inventory costs. If additional time imposed on resources is the significant consequence of high product variety, then complexity phenomena reduce conceptually to the congestion phenomena familiar to us from queuing theory. We began our investigation in our client firm by looking for the imposition of nonproductive time due to product variety and the presence of delays.

Discussions with principals in the firm revealed some interesting features that could not be explained by this time-based logic or existing models. This firm has a long history of paternalistic human-resource practices; they do not lay off people. Also, their plant assets are relatively fixed in the intermediate term. This is a firm with fixed human and capital resources, and hence their capacity is relatively fixed. Also, plant managers are expected to "make schedule," meaning that all job orders released to the shop are to be completed and ready to ship in a constant amount of time (three days in most plants, four days in some). The precise statement from management would be something like, "97% of all orders through a plant must be complete in three days." This mandate is met. Finally, the firm operates in a volatile demand environment. How can a plant with fixed capacity, operating in a variable environment, get products through the plant in a fixed amount of time with high reliability? Queuing logic would suggest that this is impossible, unless the plant enjoys a lot of excess capacity. But, if the plant enjoys a lot of excess capacity, then there would be no time effect of complexity (extra time demanded from nonscarce resources is not costly). Does this plant suffer complexity costs? If so, by what mechanism?

Some answers were available on the shop floor. First, there is a possible appeal to overtime, which is a short-term expansion of capacity to meet demand. However, in this facility demand variability is not absorbed by overtime alone. Second, if the backlog of work endangers schedule completion, workers will, in their own words, "work like crazy" to get the products out on time. We call this behavior "rushing," and we model it as an implicit worker response to having too much to do in too little time. Managers who do not schedule overtime or allow delayed shipments will invite rushing when schedule completion is threatened. If a job is rushed, there can be task errors and/or omissions. Task errors or omissions will manifest themselves in higher probabilities of producing defectives, so that rushing will decrease the time invested in each unit but also increase the probability of producing defective products. Hence, production yields (percentage of nondefectives produced) in our model are not exogenous parameters but are endogenously determined by workers responding to schedule-completion pressures.

In our model, we explicitly recognize that time and quality can be traded off in job performance, and we represent the system "capability" as a time-quality pair. Different throughput rates are possible, with different quality levels. All output is filtered through an inspection process, the performance of which is represented by traditional type I and type II error probabilities. The quality of the output and the integrity of the inspection process will combine to release some product for downstream consumption and scrap or rework others.

We also recognize that task variety alone, even absent time pressures, can cause direct quality problems. It is recognized in the human factors literature that

managing a wider array of stimuli can decrease performance. These quality effects can impose rework that amplifies time pressures. Therefore, it can be difficult to untangle the effects of time-based and quality-based complexity phenomena, because behaviors will adjust to substitute one for the other.

We believe that our model has conceptual power in any system in which time and quality are substitutable and in which time is actively managed with quality consequences following. Our analysis suggests the following conceptual take-aways.

1. In a plant that makes schedule, classical congestion phenomena will not manifest themselves in inventory and delay as is usually assumed. Instead, the stress of congestion will find its way into a spectrum of cost categories (labor, internal failure costs, external failure costs) contingent upon the overtime policy in the plant and the integrity of the outgoing inspection process.

2. A plant that always makes schedule as volume is added in a variable environment either has a lot of excess capacity, uses overtime routinely to relieve congestion pressures, or has an imperfect inspection system and deteriorating outgoing quality levels. See Proposition 1.

3. The appropriate use of overtime to relieve schedule pressures depends essentially on the relative cost of labor and quality. See Corollary 2.1.

4. A plant manager whose performance is appraised on making schedule, but who is not accountable for external failure (e.g., warranty) costs, has an incentive to use a suboptimal amount of overtime. Incentives can be aligned either by including external failure costs in the plant report or by tightening up the outgoing inspection system. See Corollary 2.2.

5. Although it is possible to generate parameter values that favor rushing over overtime, numerical trials for a range of realistic parameter values indicate that if there are significant quality consequences for rushing, then overtime is almost always recommended as the preferred means to making schedule. If there are no significant quality consequences for rushing, then a firm should reexamine its notion of its own capacity, since there is no serious consequence for working faster.

6. Our model proposes two fundamental ways in which product variety (product line complexity) affects plant performance: time and quality. We reflect these phenomena, and investigate their influences, using separate time and quality parameters. Time-based complexity logic reduces, conceptually, to congestion intuition. Using time-based logic only, complexity costs can only be realized in congested environments and can only be generated at bottleneck resources. That is, excess capacity will mask complexity costs. Quality-based complexity costs, however, are independent of congestion and utilizations, and can be realized simultaneously throughout the plant at bottleneck and nonbottleneck resources.

7. A plant that is interested in addressing its costs of complexity with limited investigative time and resources may prefer to concentrate on quality issues rather than time issues. Quality issues are always there, independent of schedule pressures, labor policy, and/or whether or not a resource is a bottleneck. Also, in the case that schedule pressures are severe, time- and quality-complexity effects are similar in magnitude because one begets the other. Intuitively, unless a plant is fundamentally overloaded (there is simply too much to do in the available time even without confounding complexity effects), then quality-based complexity costs can have a greater impact than time-based complexity costs on total firm costs. This is discussed in more detail in § 5.

The next section reviews some of the literature relevant to issues of congestion and complexity. Section 2 describes our model. Section 3 considers the feasibility of making schedule, optimal labor policies, and the incentives facing a plant manager. Section 4 presents some cost signatures that calibrate one's intuition in this new context. Section 5 considers the implications of our conceptual model for firm investigations and decision making.

1. Existing Literature

There is an extensive queuing theory literature that relates facility utilizations and the variability in demand rates and processing times to inventory and/or delay consequences. See Heyman and Sobel (1982) for an entree into this body of work. Fundamental in this literature is the fact that idleness is imposed on a facility

by asynchronous variability between the arrival rate of jobs and the facility's processing capabilities. This idleness prevents a perfect match between the mean arrival rate of jobs and the facility's mean capacity to process jobs. The ratio of these two is the utilization of the system, and inventory and delays explode as the utilization approaches unity in a variable environment. This is the classical logic of congestion. One consequence of this logic is that anything that decreases the effective capacity of the system (such as the imposition of nonproductive time) will increase inventory and/or delays in the presence of variability.

Less extensive is the literature on the "cost of complexity," which (in the management literature) refers to the potential diseconomies that attend the performance of a heterogeneous, rather than a homogeneous, set of tasks. This is sometimes referred to as diseconomies of scope. Financial evidence of these diseconomies include costs rising more than proportionally as the firm adds products or services to its repertoire. Indeed, it is this type of financial indicator that attracts managerial attention to this topic.

As noted above, Banker et al. (1988) use an M/G/1 queuing model of the firm to show that increased product variety increases queuing delays and holding costs, explicitly linking the complexity and congestion literatures. Queuing intuition also underlies our notion of the time-related consequences of complexity, but when the firm makes schedule, the cost impact will not be in inventory or delays, as we shall see.

Economists have studied multiproduct firms from a theoretical (cf. Panzar and Willig 1981, Teece 1982) and empirical (Caves et al. 1979, Pulley and Braunstein 1992) perspective. In this literature, firms add products to better utilize indivisible shared resources. With such resources, a firm may choose to hold excess capacity until it develops new products, so that the conventional model of using an extra hour of labor costing the company the ambient wage rate is not necessarily operative. Utilizing slack resources incurs no immediate opportunity cost. Hence, there is an important relationship between congestion (utilization) and the opportunity cost for time on resources.

In the management literature, Miller and Vollman (1985) trace complexity costs to fundamental transactions (activities) that must be attended to for the daily operation of a firm. The basic logic is that adding products adds transactions (work) and hence will increase the utilization of support resources. This will result in either delays, via classical congestion logic, or increased capacity investment (not considered in our fixed resource model). The latter increases overhead expenditures. Hence, in Miller and Vollman's paper, complexity costs are closely related to classical congestion phenomena, but manifest themselves primarily in manufacturing overhead and not direct costs. The Miller and Vollman article initiated a series of empirical investigations into the existence of complexity costs (cf. Foster and Gupta 1990, Banker et al. 1990, Banker et al. 1995, Anderson 1995, MacDuffie et al. 1996, and Sethuraman 1994). The conclusions regarding the presence and magnitude of complexity costs are mixed in these papers. This is not surprising when one considers that these costs are contingent upon several root causes and potentially confounding circumstances, all subject to measurement error.

There is also an extensive literature on flexibility and focus (cf. Skinner 1974, Sethi and Sethi 1990) as generic strategies. The fundamental issue is trading off the market gains from variety and the costs due a system's potential loss of economic effectiveness. Undesirable cost or effectiveness consequences for high variety are implicit in that trade-off.

2. The Model

Our model is a mathematical reflection of the situation in our client firm. We model a plant as a single production station with inspection and rework. Since most production systems are networks of workstations, the model here is most appropriately applied to bottleneck workstations (that is, the single workstation that most impacts total system performance). Such bottlenecks can arise in any direct or indirect activity required to complete production. We will see, however, that some complexity costs do not rely on this bottleneck status. We assume that the firm is operating with fixed human and machine resources.

Our view of complexity is that higher task variety will impact one or both of two fundamental phenomena: time and quality. This assumption has support in the literature (time effects have been discussed, and

quality effects are discussed below). We model the time and quality effects of product variety using the model parameters $\gamma(n)$ and $h(n)$, where n is the number of products in the product line. A more complete description of how these functions are used appears below. By setting these to constants (independent of n), we can eliminate effects due to the number of products produced and focus on pure congestion phenomena. By allowing these terms to vary with the breadth of the product line, we can investigate the additional costs imposed by complexity phenomena.

In the next several pages we introduce our model and the notation we use to represent the model's parameters. A complete list of variable names and definitions appears in Table 1.

We denote the average demand rate for product i by λ_i units/day. The total demand rate over all products is therefore $\lambda := \sum_{i=1}^{n} \lambda_i$, and the long-run proportion of total demand for product i is λ_i/λ.

The system capability for product i is represented by a time-quality pair, (t_i^0, p_i^0), representing the nominal time devoted to processing each unit and the anticipated yield (fraction of output of acceptable quality) from that time investment. We assume these quantities are strictly positive.

The actual mean time it takes to process a unit of product i, t_i, may differ from t_i^0 for a variety of causes. First, the plant manager can schedule an overtime level, OT (days), effectively increasing the length of a working day. Then, the mean number of working days that each unit of product i requires is decreased by a factor of $1/(1 + OT)$. It is implicit in this formulation

Table 1 Variable Definitions

n	=	number of products in the product line.
λ_i	=	demand (units/day) for product i.
λ	=	$\sum_{i=1}^{n} \lambda_i$ = total units/day demanded.
t_i^0	=	base time (days/unit) to produce one unit of product i.
p_i^0	=	base yield (probability of a nondefective item) that results from an investment of t_i^0 in production of a unit of type i, before being adjusted for any complexity or rushing effects on quality.
θ	=	proportional change in processing time due to rushing. We always have $\theta \in (0,1]$, and θ decreases as workers rush more.
δ	=	parameter relating rushing to quality impacts. Yields are multiplied by the factor θ^δ to generate actual process yields.
$\gamma(n)$	=	proportional adjustment to time/unit in production for product i resulting from making n products. It is intuitive that $\gamma(n)$ is nondecreasing in n and that $\gamma(1) = 1$.
$h(n)$	=	proportional adjustment to yield/unit for product i resulting from making n products. It is intuitive that $h(n)$ is nonincreasing in n and that $h(1) = 1$.
OT	=	overtime (days/day) authorized.
α_i	=	probability of a type I error in inspection for product i.
β_i	=	probability of a type II error in inspection for product i.
t_i	=	$t_i^0\gamma_i(n)\theta/(1 + OT)$ = mean time invested per unit of production of product i, including rushing, overtime, and complexity adjustments.
p_i	=	$p_i^0 h_i(n)\theta^\delta$ = yield (probability of nondefective production) achieved for product i, including rushing and complexity effects.
π_i	=	$\beta_i + p_i(1 - \alpha_i - \beta_i)$ = probability that a unit of product i is released for consumption.
W_F	=	full-time wage bill ($/day) for fixed labor force.
ξ	=	overtime premium paid, so that one day of overtime costs $(1 + \xi)W_F$.
m_i	=	material cost per unit for product i.
s_i	=	fraction of product i output that, if judged defective, is scrapped instead of reworked.
ω_i	=	warranty and goodwill costs per defective item of product i released for consumption.
r_i	=	revenues per unit for product i.
$\bar{T}$	=	Long-run average production time per job in the plant.
$T^*(\lambda)$	=	Schedule of targets that $\bar{T}$ cannot exceed in order to make schedule. T^* will vary with total demand rate λ.
τ	=	desired throughput time in the plant to make schedule.
$P_D(\tau)$	=	probability that actual throughput time is less than or equal to τ, used to define making schedule.
$P^{\min}$	=	Parameter to define making schedule. Schedule is "made" if $P_D(\tau) \geq P^{\min}$
ρ	=	system utilization, which will equal $\lambda\bar{T}$.

that labor is the most constraining resource in production.

The addition of extra products can impose coordination system delays, for example, on scheduling systems not designed for high product variety. These include additional machine setups, information-system delays, and other impositions of nonproductive time. To model this we assume that the time required per unit of product i is multiplied by a "nonproductive time" function, $\gamma_i(n)$, of the number of products n. We assume that $\gamma_i > 0$ for all i and focus intuitively on the case where $\gamma_i(n)$ is greater than or equal to one and nondecreasing in n. This is most appropriate for build-to-order environments and is appropriate for our client firm because it essentially builds to order from the bottleneck work station.

Investing time t_i in production achieves a yield rate, p_i, which may also differ from p_i^0 for several reasons. Research in human factors and ergonomics (Sanders and McCormick 1993, Meister 1976, Cohen 1980) suggests that as the number of stimuli that need to be processed by an individual increases, performance decreases. However, there are a variety of contextual circumstances (e.g., how similar different stimuli are, the presentation rate, frequency of change, etc.) that modify the relationship between number of stimuli and performance. We assume that the process yield can deteriorate as the number of products that need to be processed by the system increases. Specifically, we use $h_i(n) \leq 1$ to represent the factor by which yield for product i is impacted due to the stimuli load associated with processing n different products. We focus on the case where $h_i(n)$ is decreasing in n. Based on our field application, we intuitively associate $h_i(n)$ with human error, but it can represent other system quality effects, such as information system errors driven by variety.

Workers may need to rush to make schedule, effectively decreasing the time they invest in each unit of product. Let $\theta \in (0, 1]$ denote the factor by which processing time per unit is decreased by rushing. Rushing, however, also decreases the quality of the output. We use a power model that allows for a range of strengths of impact, and curvatures, for the quality effects of rushing. Specifically, we assume that, as a result of rushing to a level θ, yields are decreased by a factor θ^δ for some $\delta > 0$. We will assume that these effects are the same for all products. δ less than (greater than) unity implies that yields decrease less than (more than) proportionally to the time saved by rushing.

In summary, the mean time invested to process a unit of product i is $t_i = t_i^0\gamma_i(n)\theta/(1 + OT)$ working days, reflecting the net effects of rushing, overtime, and the inefficiencies imposed by product variety. The yield achieved by this time investment is $p_i = p_i^0 h_i(n)\theta^\delta$, reflecting the net effects of rushing and yield reductions imposed by product variety.

After processing, each unit goes into an inspection process. The accuracy of this process is represented by parameters α_i and β_i. α_i is the probability that a quality item of type i is judged to be defective, and β_i is the probability that a defective item of type i is judged to be without defect. In the language of statistical process control, α_i and β_i are the probabilities of type I and type II errors, respectively, by the inspection process for units of product i. We assume that $\alpha_i + \beta_i \leq 1$ for all products i, and we disallow the case with $\alpha = 1$ and $\beta = 0$. This would be the situation in which we scrap or rework everything. In that case, the system would be unstable as jobs arrive but never leave.

A fraction s_i of the units of product i that are judged to be defective by the inspection system is scrapped. The remaining fraction, $(1 - s_i)$, of the output that is judged defective is reworked. Items that are reworked reenter the production process. If an item is scrapped, then another job must enter the production process to replace it. From the perspective of time demanded in the production system, there is no difference between replacement (due to being scrapped) and rework jobs. In either case, the system must initiate another processing iteration on the job. The most significant difference between scrap and rework is the material costs incurred.

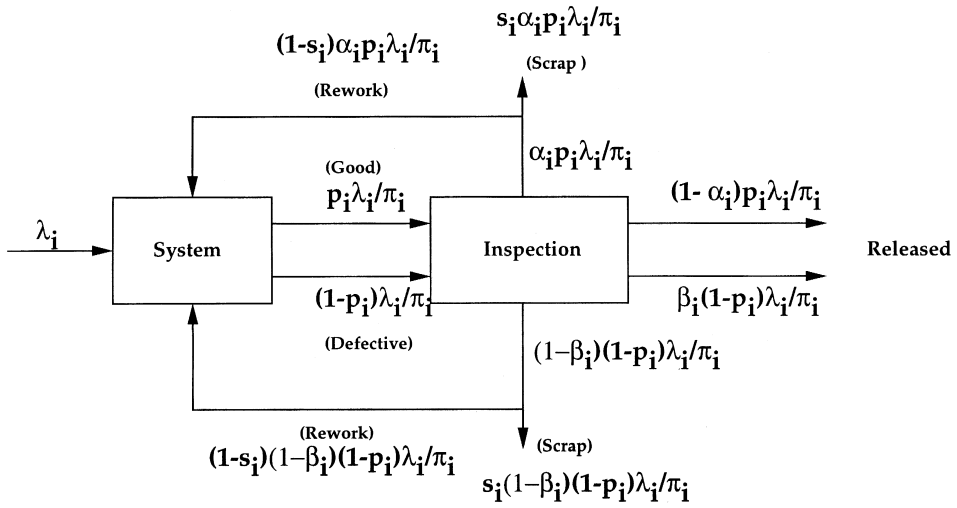
Let π_i be the probability that an item of type i, emerging from the production system, is released for consumption, and $1 - \pi_i$ the probability that it is scrapped or reworked. It is clear from the above that

$$\pi_i = p_i(1 - \alpha_i) + (1 - p_i)\beta_i = \beta_i + p_i(1 - \alpha_i - \beta_i).$$

See Figure 1 for a graphical representation of the material flow rates.

Our stylized model of the firm is, essentially, a single-station queue with scrap and rework deriving

Figure 1 Mathematical Model



λ_i = Demand for Product i

p_i = Probability of producing a good unit per processing iteration for product i

α_i, β_i = Inspection parameters

s_i = Scrap rate for product i

$\pi_i = p_i(1-\alpha_i) + \beta_i(1-p_i)$

from an imperfect inspection system, with the potential to reduce the mean service time and utilizations (reducing schedule completion pressures) by rushing and/or increasing the time available (by scheduling overtime). Yield rates are determined endogenously by rushing behaviors, which in turn depend on schedule completion pressures and overtime policies. Although stylized, we believe that the essential ingredients of this model are not uncommon and that its essential dynamics are appropriate for our client firm. This last belief derives from first-hand discussions with workers on the shop floor who described their rushing response to schedule pressures and subsequent presentations to management.

The material flows just described have attendant cash flows. We assume that the accounting system of the firm collects costs into cost categories, and that these cost figures comprise the bulk of the information that the plant manager has to assess plant performance. The specifics of how material flows translate into cash flows, and how the cash flows are accumulated into cost categories and are reported to management will vary from firm to firm. Here we assume one such translation. Our fundamental insights are not contingent on the details of these assumptions.

Here the cost categories are labor, material, indirect variable costs, and warranty costs. Revenues will be tracked, but the plant manager's actions do not impact that figure (we use warranty costs to capture any revenue loss resulting from poor quality). Inventory levels (and holding costs) can be tracked with more specific assumptions regarding the statistics of the demand and service time processes, but this will not play a role here (see below).

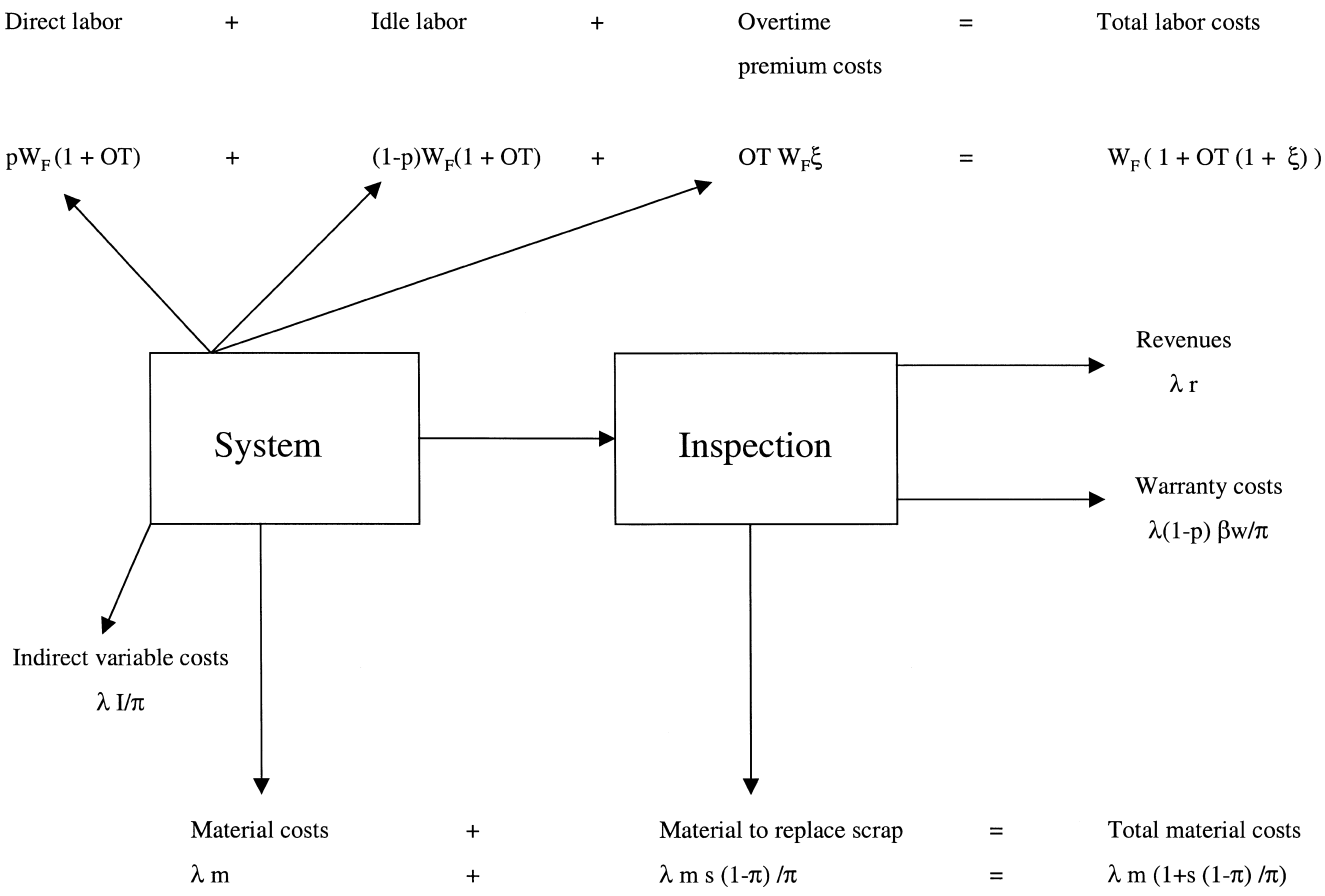
In the following we provide the long-run average cash flow rates (dollars/day) experienced by the firm for a fixed product portfolio. We assume that these are reasonable proxies for the intermediate-range performance that is the focus of this article (in the long run, all resource levels can be adjusted). These results can then be used to anticipate trends in the plant's cost signatures as we add volume or broaden the product

line. The cash flow expressions are expressed graphically in Figure 2 and are described in this section.

Labor Costs. With our fixed resource assumption, the labor force is fixed. Let W_F denote the full-time labor wage bill per working day. Direct labor will be charged when this force is productively busy; otherwise the wages are charged to an idle labor account. We assume that direct labor performs setups, etc., so that any time invested in production activities by direct labor personnel is charged to the direct labor account. Different cost taxonomies could be used without qualitatively changing our results. When there is no work to do, the labor costs are charged to an idle labor account. If ρ denotes the system utilization, the direct and idle labor charges (dollars per working day) are $\rho W_F(1 + OT)$ and $(1 - \rho)\, W_F(1 + OT)$, respectively. Overtime pay increases the base wage by a factor of ξ. The overtime premiums, paid in excess of the base wage, are tracked in a third account at a rate of $OT \times W_F \times \xi$ dollars per working day. Note that the total labor bill (direct plus idle plus overtime) will be $W_F(1 + OT(1 + \xi))$. Without overtime, the total labor bill is constant, and would remain so as we add products (however, costs would move from the idle labor to the direct labor account). The only labor cost consequence of adding products will be the potential use of overtime as increased time demands threaten schedule completion. This is a direct consequence of operating with fixed resources.

Direct Material and Indirect Variable Costs. One unit of product i requires m_i dollars in direct materials. A fraction, s_i, of items of product i that do not pass inspection are scrapped. These require more material to begin production anew. The remaining fraction (1

Figure 2 Cash Flows



$- s_i$) of items that do not pass inspection are reworked, incurring labor time and indirect variable costs but no direct material costs. Recall that $(1 - \pi_i)$ is the probability that a unit of product i finishing production will not pass inspection and that the total flow rate of product i through the production station (including original and rework flows) will be λ_i/π_i. Hence, considering the original product flows plus rework flows, the total cost rate for material for product i will be $\lambda_i m_i(1 + s_i (1/\pi_i - 1))$. We assume an indirect variable cost I_i per unit for product i that is incurred every time a unit (original or rework) passes through the production process. This captures any variable costs (e.g., supplies) not included in direct material. Hence, the indirect variable cost flow rate for product i will be $\lambda_i I_i/\pi_i$.

Revenues and Warranty and Goodwill Costs. We assume revenues of r_i per unit of product i released for downstream consumption. For each defective item of type i that we release for consumption, we incur a goodwill and warranty cost of ω_i (which may include returned revenues). The flow rate of defective products of type i to customers will be $(\lambda_i/\pi_i)(1 - p_i)\beta_i$, and the warranty cost flow rate will be ω_i times this number.

Inventory Holding Costs. Inventory will not exhibit the explosion that usually attends congestion in conventional queuing models. This is because the plant always makes schedule, which bounds the delay experienced by any job in the system. Indeed, in our model average inventory levels will increase at most linearly with sales, and hence inventory cost/sales figures will remain relatively flat. This is in contrast to the model in Banker et al. (1988) in which complexity costs are present only in inventory holding costs. The difference is the phenomenon of making schedule. When a plant always makes schedule, the costs that one intuitively associates with congestion phenomena will not be realized as inventory costs, but will diffuse into other cost categories (see below). Inventory costs will not be considered further.

3. Making Schedule

3.1. Definition and Feasibility

We assume that items failing inspection are reworked or restarted immediately. Hence, the number of iterations it takes to produce one good unit of type i is a geometrically distributed random variable, and the expected time between entering the system and being released for consumption will be t_i/π_i for products of type i. Since the long-run fraction of jobs that are type i is λ_i/λ, the long-run average service time per job (averaged over all jobs processed by the system) will be

$$\bar{T} = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \frac{t_i}{\pi_i}.$$

Recall that $\lambda = \Sigma_{i=1}^{n} \lambda_i$. With t expressed in "working days," the capacity of the system is $1/\bar{T}$ products/day, and the long-run utilization of the system is $\rho = \lambda\bar{T}$.

We model "making schedule" as follows. Define τ to be the desired maximal total delay between the arrival of a job and its release for consumption. Define $P_D(\tau)$ to be the probability that the delay is less than or equal to τ, and $P^{\min}$ to be a specified lower bound on this probability. We model the pressure to make schedule as the desire to keep $P_D(\tau) \geq P^{\min}$. For example, if plant managers are expected to have 97% of scheduled production complete within three days of its being released to the shop floor, we would have $P^{\min} = 0.97$ and $\tau = 3$ days.

Rather than relying on the statistics of specific demand streams and production processes, we will model the desire to make schedule in an indirect manner. Specifically, we assume that for any portfolio of products with an aggregate arrival rate of λ units/day, we want to adjust the mean processing time per unit ($\bar{T}$) such that the utilization $\rho = \lambda\bar{T}$ is less than or equal to some specified target $\rho^*(\lambda)$. This should be a generically applicable construction, because all we require is that the probability that an arriving job will be cleared from the system in τ working days or fewer can be increased to some specified target level by decreasing the utilization. It would be difficult to generate a reasonable queuing model of a firm that does not have this feature. The appendix demonstrates how one can estimate $\rho^*(\lambda)$ using an M/M/1 queuing approximation. We will assume henceforth that the schedule of target utilizations, $\rho^*(\lambda)$, is known.

The plant manager can decrease $\bar{T}$ and ρ by increasing the amount of time available in a working day (that is, use overtime). This would reduce schedule pressures on workers. If overtime is disallowed and schedule pressures loom, then workers will rush jobs. The

possibility is that rushing breeds poor quality and rework will render some schedules infeasible, a topic we address below. Since $\rho = \lambda\bar{T}$, it follows that, for any specific product portfolio, making schedule is equivalent to keeping $\bar{T} \leq T^*$, where $T^* = \rho^*(\lambda)/\lambda$ is a known quantity and $\bar{T}$ is adjusted by rushing and using overtime.

It is apparent that if unlimited overtime is available, then the plant can always make schedule. But, this is never the case. The following proposition articulates those cases in which the plant can make schedule for any fixed allocation of overtime. Recall that $\delta < 1$ means that rushing reduces product yields less than proportionally to the savings in labor time, and $\beta_i > 0$ means that some nonzero fraction of defective items will reach consumers. Basically, one or both of these inequalities needs to hold for the plant to make schedule.

Proposition 1.

a) *If $\delta < 1$ or $\beta_i > 0$ for all i, then for any $OT \geq 0$, the $\lim_{\theta\to 0}\rho = 0$. That is, starting with any level of overtime the plant can always make schedule by rushing.* b) *If $\delta \geq 1$ and $\beta_i = 0$ for some i, then it might not be possible to make schedule.*

Proof. Substituting the detailed expressions for π_i and t_i into the expression for ρ yields

$$\rho = \sum_{i=1}^{n} \lambda_i \frac{\theta t_i^0 \gamma_i(n)}{1 + OT} \frac{1}{\beta_i + p_i^0 h_i(n)\theta^\delta(1 - \alpha_i - \beta_i)}.$$

Define ρ_i as follows:

$$\rho_i = \lambda_i \frac{\theta t_i^0 \gamma_i(n)}{1 + OT} \frac{1}{\beta_i + p_i^0 h_i(n)\theta^\delta(1 - \alpha_i - \beta_i)}$$

so that $\rho = \Sigma_{i=1}^{n} \rho_i$. Define $S_\beta = \{i \mid \beta_i > 0\}$, then

$$\rho = \sum_{i\in S_\beta} \rho_i + \sum_{i\in S_\beta^c} \rho_i$$

where S_β^c denotes the complement of the set S_β. Note that regardless of δ, $\lim_{\theta\to 0}\rho_i = 0$ for $i \in S_\beta$. Now suppose that $\delta < 1$. Then for $i \in S_\beta$ the result holds as just shown. For $i \in S_\beta^c$ we have that ρ_i will be proportional to $\theta^{1-\delta}$ (recall that we disallow $\alpha_i + \beta_i = 1$ when $\beta_i = 0$ because that would characterize an unstable system). Hence, again the limit of ρ_i as we drive θ to zero is zero. This completes the proof of part (a) of Proposition 1. For part (b) of Proposition 1, note that if $\delta \geq 1$ and $S_\beta^c \neq \emptyset$, then for $i \in S_\beta^c$ the limit of ρ_i as θ goes to zero is proportional to $1/\theta^{\delta-1}$. This term is either constant (if $\delta = 1$) or increasing as we decrease θ, so that we might not be able to drive the utilization down to the target ρ^*. In particular, if $\delta > 1$ and $\beta_i = 0$ for all i, then ρ is increasing as we decrease θ, so that if we cannot make schedule using overtime only, we cannot make schedule at all. Q.E.D.

Proposition 1 generates some important intuition. If the time required to process each unit can be reduced without significant quality losses, then managers should revisit their current beliefs regarding firm capacity. They actually have more than they think. If, on the other hand, rushing can only be activated with significant quality effects ($\delta \geq 1$), then the ability to make schedule depends on having an imperfect quality control system. Essentially, there is a quality "vent" for congestion and complexity pressures. Schedule can be met as these pressures mount by ushering increasing levels of defective product out the door. If $\beta_i > 0$ for all i, then we release some bad product for consumption. The total production rate of poor quality product increases as we rush (decrease θ), and a constant fraction, β_i, of these leave the facility and are not reworked. Hence, we can make schedule by releasing an increasing quantity of poor quality goods from the facility. Goodwill and warranty costs will suffer, but we will not incur delays. Intuitively, if we observe a plant that always makes schedule as we add products, then we can reasonably suspect that either the plant enjoys a lot of excess capacity on average, makes regular use of overtime, or is allowing an increasing amount of poor quality items to reach market.

An inspection system that always detects and recycles defective output removes the potential use of the quality vent. In this case, if there are significant quality consequences for rushing ($\delta > 1$), the firm can increasingly miss schedule once maximal overtime is used. This is because rushing is counterproductive, making more work than it saves.

This emphasizes the point that in this model there is no one single capacity figure that limits the firm's output capabilities. Instead, there is a range of output rates available, each with a different quality level. The optimal cost level for any given level of output is

achieved by rationally trading off labor costs (incurred by increasing the working time per day) and quality costs (both internal rework costs and external failure costs). How should a plant manager behave (normatively) in this environment? In the following we analyze the optimal level of overtime required to meet schedule, given the anticipated reactions of workers. Recall that for a given λ and $T^* = T^*(\lambda)$, making schedule requires that

$$\rho = \sum_{i=1}^{n} \lambda_i \frac{t_i}{\pi_i} \leq \rho^*(\lambda) = T^*\lambda,$$

or equivalently

$$\bar{T}(OT, \theta) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \frac{\theta t_i^0 \gamma_i(n)}{1 + OT}$$

$$\frac{1}{\beta_i + p_i^0 h_i(n)\theta^{\gamma}(1 - \alpha_i - \beta_i)} \leq T^*.$$

Basically, the average processing time (in working days) per unit of production ($\bar{T}$) is a function of the overtime (OT) and rushing (θ) levels employed, and some combination of overtime and rushing must be used to make schedule, represented by keeping $\bar{T} \leq T^*$.

We assume that both the plant manager and the workers desire to make schedule, but that the plant manager wishes to minimize total costs and the workers wish to minimize their effort (subject to making schedule). We model this by assuming that the plant manager chooses an overtime level OT, and then the workers respond by rushing just enough to make schedule. That is, the workers choose

$$\theta = f(OT) := \max\{\theta \in (0,1] : \bar{T}(OT, \theta) \leq T^*\}.$$

Total costs as a function of overtime and rushing are

$$TC(OT, \theta) = \sum_{i=1}^{n} \left[\frac{I_i}{\pi_i} + \frac{1 - p_i}{\pi_i}\beta_i\omega_i + m_i\left(1 + \frac{1 - \pi_i}{\pi_i}s_i\right)\right]$$

$$\lambda_i + W_F[1 + OT(1 + \xi)]. \quad (1)$$

The plant manager's problem is to choose an overtime level (between zero and the maximal feasible amount of overtime, $OT_{\max}$) that minimizes TC, given worker reactions. That is, the plant manager wishes to solve the following problem, which we call problem *PM*:

$$\text{Minimize}_{OT \in [0, OT_{\max}]} TC(OT, f(OT)).$$

Problem *PM* is not easy to analyze directly, but in the following we generate an equivalent form of the problem that yields simpler results. The next lemma is a standard result in functional analysis (cf. Hoffman 1975).

Lemma 1. *If $f : R \to R$ is continuous and strictly monotone on an interval $I \subset R$, then $f(I)$ is an interval, f^{-1} exists on $f(I)$, and the set $\{(x, f(x)) : x \in I\} = \{(f^{-1}(y), y) : y \in f(I)\}$.*

Assume that either $\delta < 1$, or $\delta = 1$ and $\beta_i > 0$ for all i; then it follows from Proposition 1 that for any OT the workers have a feasible response, meaning that there is a $\theta \in (0, 1]$ such that schedule is made. Hence, $f(OT)$ is well defined. Under these same conditions, it can be verified that $\bar{T}(OT, \theta)$ is strictly monotone in $\theta > 0$ so that the workers' response to overtime level OT will be $f(OT) = 1$ if $\bar{T}(OT, \theta) < T^*$, and otherwise will be the unique θ such that $\bar{T}(OT, \theta) = T^*$.

Because overtime is costly, the plant manager would never choose a level of overtime greater than any OT for which $\bar{T}(OT, 1) \leq T^*$. To do so would cost money but would invite no change in worker response, which would remain at $\theta = 1$. Hence, if $\bar{T}(0, 1) \leq T^*$, then $(OT, f(OT)) = (0, 1)$ is the optimal solution to *PM*. This is the situation in which the plant has sufficient capacity to ensure schedule completion without rushing or overtime.

If $\bar{T}(0, 1) > T^*$, then define $\hat{OT}$ to be the unique $OT > 0$ such that $\bar{T}(OT, 1) = T^*$. Define $OT' = \min\{\hat{OT}, OT_{\max}\}$. Proposition 2 states that we can recast the plant manager's problem (problem *PM*) to be one of choosing an optimal level of rushing instead of choosing an optimal level of overtime. This is possible because the conditions stated ensure a one-to-one relationship between these two versions of the problem.

Proposition 2. *If $\delta < 1$, or $\delta = 1$ and $\beta_i > 0$ for all i, then*

a) *If $\bar{T}(0, 1) \leq T^*$, then the plant manager's problem (PM) is solved by $(OT, \theta) = (0, 1)$.*

b) *If $\bar{T}(0, 1) > T^*$, then the plant manager's problem (PM) is equivalent to*

$$\text{minimize}_{\theta \in J} TC(OT(\theta), \theta)$$

where

$$J = \text{the interval } f([0, OT']) \subset R^+$$

$$OT(\theta) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \frac{t_i^0 \gamma_i(n)\theta}{T^*} \frac{1}{\beta_i + p_i^0 h_i(n)\theta^\delta(1 - \alpha_i - \beta_i)} - 1.$$

Proof. Part (a) of Proposition 2 follows from the above comments. So, assume that $\bar{T}(0, 1) > T^*$. It also follows from the above comments that we can substitute $I := [0, OT']$ for $[0, OT_{\max}]$ as the search interval in problem *PM*. Under the stated assumptions, $f(OT)$ is strictly increasing and continous on I. Since for any $OT \in I$ there exists a $\theta > 0$ such that schedule is made (Proposition 1), it follows that $f(0) > 0$, so $I \subset R^+$. The result would follow from Lemma 1 if $f^{-1}(\theta) = OT(\theta)$. But for $OT \in I$, $f(OT)$ is the unique θ such that $\bar{T}(OT, \theta) = T^*$. Rearranging this equation so that OT is the dependent variable yields the appropriate expression for f^{-1}. Q.E.D.

We can write $TC(\theta) = TC(OT(\theta), \theta)$ by substituting the appropriate expression for $OT(\theta)$ into the total cost equation. For notational convenience define the following terms:

$$V_i := I_i + m_i s_i + \beta_i \omega_i.$$

$$Q_i := p_i^0 h_i(n) \beta_i \omega_i.$$

$$L_i := \frac{W_F(1 + \xi) t_i^0 \gamma_i(n)}{\lambda T^*}.$$

$$R_i = p_i^0 h_i(n)(1 - \alpha_i - \beta_i).$$

V_i, Q_i, L_i, and R_i are related, respectively, to variable costs, external quality costs, labor costs, and release probabilities for product i. Direct substitution generates

$$TC(\theta) = -W_F \xi + \sum_{i=1}^{n} \lambda_i m_i (1 - s_i) + \sum_{i=1}^{n} \lambda_i \frac{V_i - Q_i \theta^\delta + L_i \theta}{\beta_i + R_i \theta^\delta}.$$

Problem *PM* has an easily stated solution if $\delta = 1$, found by inspecting the derivative of $TC(\theta)$ with respect to θ. Corollary 2.1 shows that in this case intuitive notions of what should occur does occur. If variable costs and external failure costs are large relative to labor-related costs, overtime is preferred to induced rushing. If the reverse is true, then overtime is avoided, and it is economic for workers to rush to make schedule. But things are very interdependent. The integrity of the quality control system and the fraction of material judged defective that is scrapped versus reworked all play a role. For $\delta = 1$, Corollary 2.1 gives us an exact answer, using a comparison of costs that include all relevant parameters.

Corollary 2.1. *If $\delta = 1$ and $\beta_i L_i \geq R_i V_i + \beta_i Q_i$ for all i, then the plant manager will choose $\theta^*_{TC} = f(0)$, or equivalently $OT^*_{TC} = 0$. The reverse inequality will imply $\theta^*_{TC} = f(OT')$, or equivalently $OT^*_{TC} = OT'$.*

We can also say something about optimal labor policies when $\delta \neq 1$. As δ decreases from unity, the quality and cost consequences of rushing diminish. Eventually, rushing will become optimal regardless of the parameter values relevant to Corollary 2.1. Clearly, if $\delta = 0$ it is always optimal to rush, as there are no quality or cost consequences for working faster to reduce the per unit processing time. This is an unrealistic situation because capacity is essentially limitless and available on demand at no cost.

When $\delta > 1$, quality deterioration due to rushing is more than proportional to time savings. First, it may not be possible to make schedule if it cannot be made by overtime alone. So, problem *PM* may not be well defined. However, suppose that the plant can make schedule with some combination of overtime and rushing. It can be shown that $\bar{T}$ will be unimodal but not necessarily monotone in θ, so that a unique inverse will not exist. Problem *PM* is then well defined, but not as tractable analytically. Also, one is tempted to assume that if problem *PM* solves at $\theta = 1$ (no rushing) for $\delta = 1$, certainly it will not call for rushing at $\delta > 1$. That is, we would suspect that amplifying the quality deterioration that results from rushing should only make us more reluctant to rush. This is practically but not rigorously true.

That is, in most practical circumstances we will want to rush less as δ gets larger. But, when some combination of overtime or rushing is needed to make schedule, the limit as θ goes to zero of $dTC/d\theta$ is

$$\sum_{i=1}^{n} (\lambda_i / \beta_i) L_i$$

which is greater than zero. That is, for sufficiently small θ, total costs decrease by rushing more (decreasing θ still further). This is because rushing will save labor time proportionally, but for $\delta > 1$ rushing's quality effects approach their lower bound of zero more rapidly. Hence, for frenzied levels of rushing (resulting in extremely low quality) a marginal rushing increase saves labor time but does not decrease quality levels significantly. In the limit all product is defective, and the release probability is just that of committing a type II error (β_i for product i). The total flow of product i (new and rework) through the plant is λ_i/β_i, and rushing saves overtime costs for each unit but does not increase quality costs. Such a perverse situation may not be likely in practice, but this example does illustrate the potentially counterintuitive effects of δ and why Corollary 2.1 cannot easily be extended.

Certainly, specific parameter sets can be assumed that yield optima with any combination of overtime and rushing to make schedule. However, the authors have performed a large number of numerical trials with an array of parameter values (details of the structured experiment used are available from the authors). In each trial, optimal overtime and rushing levels were computed as products were added to plants, using the M/M/1 approximation for T^* described in the appendix. For most realistic parameter values overtime is preferred to rushing as long as δ is greater than about 0.5. This conclusion is especially credible in firms in which the labor content of the cost of sales is relatively small. Intuitively, overtime increases labor costs but does not affect quality. Rushing, however, compromises quality which can impact external failure costs (typically very costly) and/or scrap and rework. The latter incur internal failure costs and also impose additional time costs that exacerbate the need to rush. Overtime is typically preferred over either external failure costs or the self-reinforcing effects of internal quality costs.

3.2. The Plant Report and Incentives

It is common in many companies (including our client) for warranty costs to be accumulated at the corporate level but excluded from plant reports. In such companies, the plant manager may pursue a local agenda that differs from the minimization of total costs. If the plant manager perceives that the plant report is the most significant input to his or her performance appraisal, then omitting warranty costs on that report will reduce his or her attention to external failure costs. Define $C(\theta)$ to be $TC(\theta)$ with $\omega_i = 0$ for all products i. If warranty costs are not included on the plant report, then a plant manager's actions may be better anticipated by solving problem *PM* using C as his or her relevant cost function, and not TC.

We relate these two problems and provide some additional insights into optimal overtime allocations in the following corollary. Let θ^*_{TC} (OT^*_{TC}) denote the optimal θ (OT) in the problem *PM* using cost function TC (which includes all costs). Let θ^*_C (OT^*_C) denote those values using cost function C (which omits warranty costs). The following corollary follows directly from these definitions and the fact that

$$\frac{\partial TC}{\partial \theta} = \sum_{i=1}^{n} \frac{\lambda_i}{(\beta_i + R_i\theta^\delta)^2}\{\beta_i L_i - \delta(R_i V_i + \beta_i Q_i)\theta^{\delta-1} + R_i L_i \theta^\delta(1-\delta)\}.$$

Corollary 2.2. *If $\delta < 1$, or $\delta = 1$ and $\beta_i > 0$ for all i, then:*

a) $\partial TC/\partial\theta = \partial C/\partial\theta - \Sigma_{i=1}^n (\lambda_i/\pi_i^2)\,\delta\beta_i Q_i\theta^{\delta-1} < \partial C/\partial\theta$ *and hence* $\theta^*_C \le \theta^*_{TC}$ *and* $OT^*_C \le OT^*_{TC}$.

b) $Lim_{\beta_i \downarrow 0 \text{ for all } i} TC(\theta) = C(\theta)$.

Part (a) of Corollary 2.2 notes that if external failure costs are not included in plant reports, the plant manager will have an incentive to use less than an optimal level of overtime. One resolution to this incentive incompatibility, of course, is to include the external costs in the manager's performance reports. This is not done in many cases because these external costs are often not realized until long after production is complete, and matching the costs to specific plant schedules (while conceptually straightforward) is perceived as difficult. Another resolution to the incentive problem is suggested by part (b) of Corollary 2.2. If the inspection system is made more stringent, then all potential external costs will be internalized in plant costs, and C and TC become the same.

4. Cost Signatures and Intuition

4.1. Cost Trajectories for Pure Congestion Effects

Some intuition about how costs are affected by making schedule can be developed by looking at cost trajectories as a function of the number of products in the

facility. The data used in these numerical experiments, which are for illustration purposes only, are shown in Table 2. The intuition developed here is not dependent on these data assumptions.

We first set $h = \gamma = 1$ to eliminate complexity effects and concentrate on pure congestion. Figure 3 shows labor, internal (material and variable indirect), and external (warranty) cost/sales data as a function of number of products in the facility for four scenarios. The scenarios differ by the method used to make schedule (overtime without rushing or rushing without overtime), and the quality of the inspection system (no outgoing inspection, in which case $\alpha = 0$ and $\beta = 1$, and an excellent outgoing inspection system, in which case $\alpha = 0$ and $\beta = 0.02$). An "excellent" instead of "perfect" ($\beta = 0$) inspection system is used so that schedule completion will always be feasible.

Table 2 Parameter values for Figures 3 and 4

Revenue per unit:	$r = \$40$
Material cost per unit:	$m = \$8$
Indirect variable cost per unit:	$I = \$10$
Full time labor cost per day:	$W_F = \$40$
Warranty cost per defective released:	$\omega = \$40$
Fraction of defectives scrapped:	$s = 0.33$
Overtime premium:	$\xi = 0.75$
Rushing quality exponent:	$\delta = 1$
In Figure 3	$h = \gamma = 1$
In Figure 4	$h = 1 - 0.1\ (n - 1)$
(n denotes the number of products)	$\gamma = 1 + 0.1\ (n - 1)$

This table displays the data used for the illustrative cost trajectories in Figures 3 and 4. All products *i* have identical cost parameters to avoid introducing product mix effects into the calculations, so that the results focus only on the effects of congestion and complexity. The authors have performed a large number of numerical experiments with an array of parameter values, and the intuition developed by Figures 3 and 4 is not contingent upon the particular values used to generate them. We set the overtime premium to 0.75 to reflect a mix of regular (time-and-a-half) and weekend (double-time) rates. In the "excellent" inspection system we set a low but nonzero β (the probability of releasing defective products to customers) to ensure that it is always feasible to make schedule. Again, the qualitative insights discussed in the text are not dependent on these assumptions.

In Figure 4, we assume very significant *h* and γ effects, for illustration purposes. With these dramatic complexity effects, we need to keep a low base utilization. Specifically, we assume that the total flow rate $\lambda = 2$ units/day, which implies a base utilization of 2 (0.25)/0.98 = 51%. From the M/M/1 approximation this implies $T^* = 0.3155$, and $\rho^* = 0.631$.

To keep $\lambda = 2$ as we add products, we use the following flow rates per product (the same for all products):

n	1	2	3	4	5	6
λ_i	2	1	0.66667	0.5	0.4	0.33333

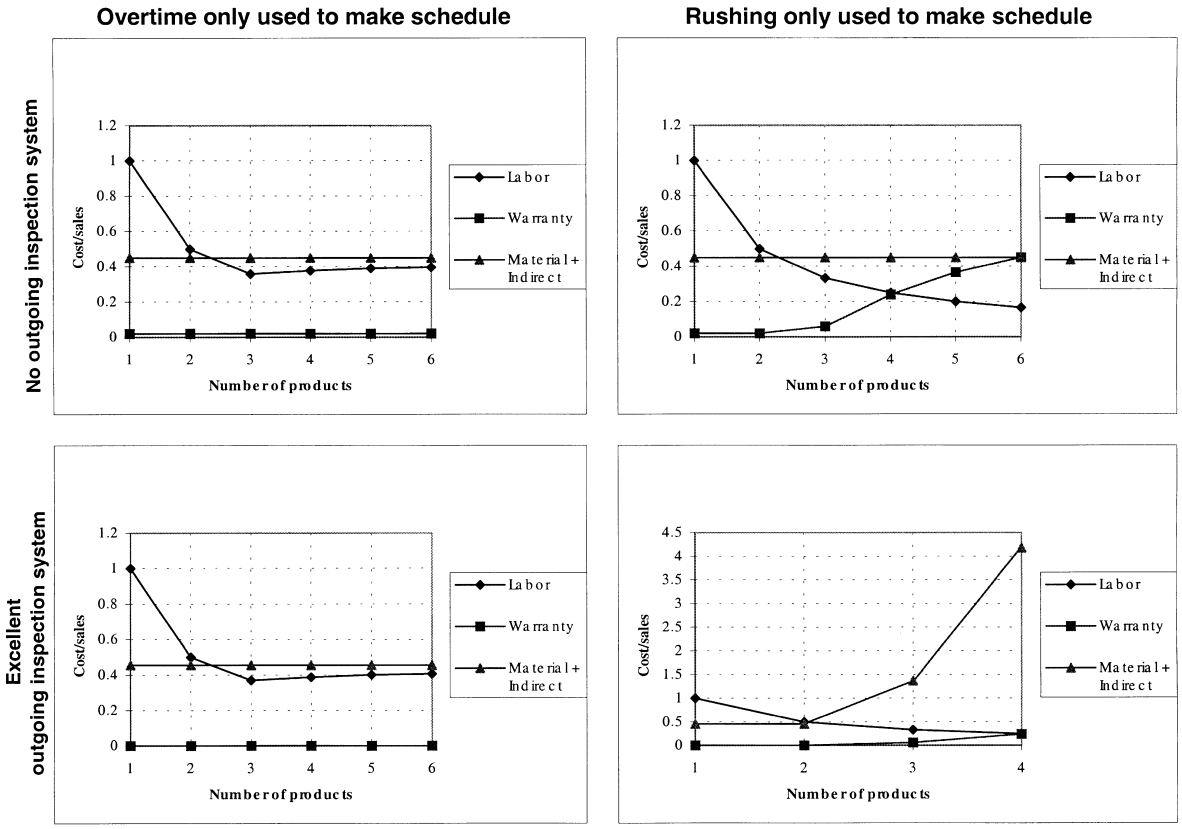
In all of the cases in Figure 4, it is optimal to use overtime to make schedule, and this policy is used in constructing the graphs. If rushing were used instead, then the labor costs would be shifted into internal and external failure costs as before.

The top two graphs show systems with no outgoing inspection system, so that failure costs will be felt externally and not internally (nothing is scrapped or reworked). Hence, the internal costs remain flat in both scenarios. Labor costs/sales decline as long as the fixed labor force is being more highly utilized, but when schedule completion is threatened additional costs are incurred. These are felt in the labor statistics if overtime is used to relieve schedule pressures and in external costs if rushing is used.

The bottom two graphs in the second row show systems with excellent outgoing inspection systems. Here the situation is similar, in that labor costs can be substituted for failure costs by using overtime versus rushing to make schedule. Here, however, failure costs are felt internally instead of externally due to the rigorous inspection system. Note the change of scale on the bottom right graph. Overtime and external failure costs are incurred once, and then they are history. However, rework imposes its own load on the system, which can increase schedule pressures and exacerbate the costs of rushing. That is, internal failure costs can explode in a self-reinforcing manner when schedule is threatened and both overtime and an external quality "vent" are disallowed.

In a plant that makes schedule, classical congestion phenomena will not manifest themselves in inventory and delay as is usually assumed. Instead, the stress of congestion will find its way into a spectrum of cost categories (labor, internal failure costs, external failure costs) contingent upon the overtime policy in the plant and the integrity of the outgoing inspection process. The firm's overtime policy determines whether schedule pressures will affect labor costs or failure (external or internal) costs. The integrity of the firm's inspection system determines whether failure costs are felt internally or externally. Also, internal failure costs have a self-reinforcing nature, exacerbating schedule pressures and all of their consequences.

Figure 3 **Cost Trajectories as a Function of Labor Policy and Inspection System Integrity**

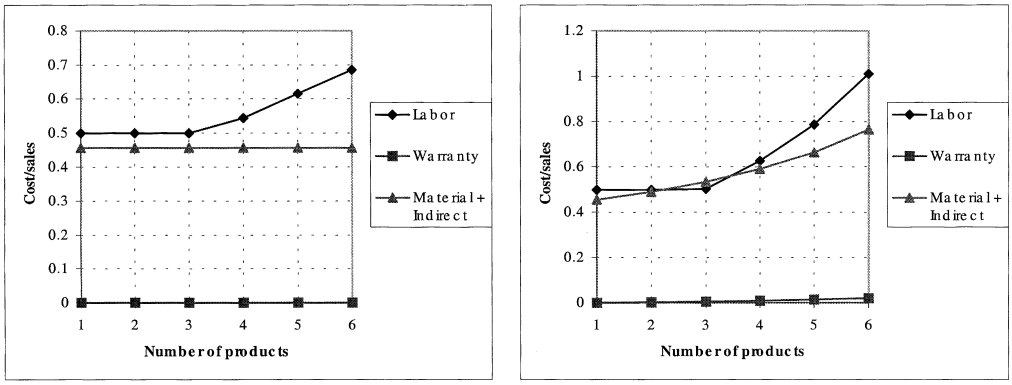


Visually, these graphs reinforce the intuition that overtime may be preferable to rushing as a means for making schedule because the consequences appear less severe. The labor/sales statistic will asymptotically approach the time cost of labor for one unit of production (at overtime rates) divided by revenue per unit, which in many instances is not large (labor costs as a percentage of sales is low in many firms, and remains so even when increased by an overtime premium). This reinforces the intuition that in most cases a firm should prefer overtime as the way to make schedule. To avoid contrary incentives, the plant manager should either be held accountable for external failure costs or the plant should have a rigorous outgoing inspection system (Corollary 2.2).

4.2. Cost Trajectories for Complexity Effects

True complexity costs are those that are a direct function of the heterogeneity of the task mix in the firm, and not the total volume of demand. Much of the intuition developed in the literature regarding the costs of complexity in firms can be considered to be variations on classical congestion themes. If a broader product line means more transactions must be executed by support functions, then the demand for "purchase orders" or "database updates" or other actions that these functions perform will increase. If demand levels approach a station's capacity, delay costs will result. If higher product variety increases utilizations and/or imposes greater variability on resources, then via classical queuing logic, inventory holding costs and/or delays in meeting customer needs will increase. These intuitive notions of how complexity will impact firm performance suggest that for a firm with fixed resources, complexity costs can only be felt at congested resources. Firms with excess capacity will never "see" complexity effects in their cost signatures because

Figure 4 **Cost Trajectories for Time versus Quality Effects of Complexity, with Overtime Used to Make Schedule**



Only time effects of complexity **Only quality effects of complexity**

extra time demanded of slack resources has no (short-term) cost.

In our model, complexity costs arise in two ways. First, higher task heterogeneity will impose nonproductive downtime on resources (as they change from one product to the next, for example). This effect is reflected in the $\gamma(n)$ term in our equations. If time is costly, then task heterogeneity has a cost. This is consistent with congestion logic in that this aspect of complexity costs can only be felt on scarce resources.

The second way in which task heterogeneity affects performance in our model is via quality effects. This aspect is reflected in the $h(n)$ term in our model and derives from the sort of relationship between stimuli complexity and performance assumed in the human factors literature. Because lower quality has a cost (internal and external failure costs) whether or not the firm has excess capacity, this aspect of complexity costs can be felt anywhere in the firm and simultaneously at all (bottleneck and nonbottleneck) resources. If, in addition, a resource is congested in the classical sense, then quality costs can impose time costs via rework.

To build an intuition regarding how complexity effects impact costs, we again look at cost trajectories as products are added to a facility. In this section, however, we control for volume effects by keeping the *base utilization*

$$\rho^0 = \sum_{i=1}^{n} \frac{\lambda_i t_i^0}{p_i^0}$$

constant as we add products. Now, as we add new products to the plant, the production of existing products must be cut back. Any cost effects will derive from product heterogeneity instead of pure congestion. The data used in these examples, which are for illustration purposes only, are shown in Table 2. In the first case we assume that there are only time effects of complexity ($\gamma(n)$ increases in n, but $h(n)$ remains constant at unity as products are added). In the other case, we assume that there are only quality effects of complexity ($h(n)$ decreases in n, but $\gamma(n)$ remains constant at unity as we add products). In both cases, we assume that overtime is used to make schedule (this is optimal). If overtime is denied, then some of the labor costs will be redirected into failure costs as noted above. The cost trajectories are shown in Figure 4 and are discussed in the next two subsections.

Time Effects ($\gamma(n)$) Only. In this case only time is impacted by complexity, and there will be no cost impact for adding products as long as total utilization is below that needed to make schedule. The plant can add up to three products without cost. After that point, overtime is needed and the labor/sales statistics deteriorate rapidly. The labor/sales cost statistics do not level out, as is the case with pure congestion. This is because as products are added each product produced becomes more costly because of the nonproductive time imposed by product heterogeneity. In the pure congestion case, higher volumes translate into higher revenues as well as costs, muting the labor cost to sales

ratio. Here, the total volume and revenues remain the same, but complexity effects demand more labor time for the same volume. Because in this example the nonproductive time imposed increases linearly with the number of products in the plant, labor costs will likewise increase linearly once schedule is threatened. Consistent with pure congestion logic, there is no cost consequence for adding products to the plant if the plant has excess capacity. Because all of the time effects are relieved using overtime, the internal and external failure costs are not affected by adding products.

Quality Effects ($h(n)$) Only. In this case only quality is impacted by complexity. There are two significant differences between this and the time-only graph. First, internal and external failure costs increase as the firm adds products, even though an overtime-only policy is used to make schedule. Second, the internal and external failure costs begin deteriorating immediately, even before schedule is threatened. Neither of these would happen with congestion or time-only effects. Third, once schedule is threatened labor costs rise more rapidly than in the time-only case because defective items are returned for rework with high probability ($\beta = 0.02$).

This emphasizes the point that quality-related complexity costs will be realized at bottleneck and nonbottleneck stations and in over- and underutilized facilities. Quality-related complexity costs will not be masked by surplus capacity. How and where these quality effects are realized depends on now-familiar contingencies. If the inspection system is poor, then some of the labor and internal failure costs shown in Figure 4 will be shifted to warranty costs. If a rushing policy is used to make schedule, then labor statistics will not suffer, but internal and external failure costs will suffer more dramatically.

5. Targeting Investigations and Process Improvements

Our conceptual model and the above discussions suggest the following strategy for targeting scarce investigative and process improvement efforts in a time-constrained and data-sparse firm. First, do you believe the firm is fundamentally congested (cannot make schedule reliably even at (t^0, p^0))? Prior beliefs on this point can be calibrated by plant tours and some rapid data assessment. If the plant is overloaded, then the common tactic of looking into time-based phenomena (e.g., change-over losses) will bear fruit in understanding the costs of complexity in the plant. If the plant is not overloaded, then the greatest return on invested effort will be had by looking into the quality effects of complexity (e.g., defectives, scrap, and rework issues).

If the firm is fundamentally congested, then remedies for the costs of complexity will be the same as those for congestion-related problems. Set up reductions, enhanced raw capacity, or demand reduction/prioritization are common tactics. If the firm is not congested, then focusing scarce effort on the quality-related costs of complexity is recommended. From the human factors literature, we can trace these costs to the complexity of stimuli that need to be processed by workers. Reducing them will require that stimuli be simplified or the effects prevented. Color coding to make product and activity selection simpler and fool proofing assembly and other tasks are common tactics.

A student team worked parallel to us and under our guidance in our client firm. Our initial instructions to the team, given before our understanding of complexity issues had matured to its current form, was to look for time drivers via a standard utilization study. They set about estimating, through employee interviews and observations, the theoretical throughput rate of the various workstations and indirect activities in a representative plant and the downtime imposed by changing tasks. They compared this with the actual load placed on the facility. The students concluded that the plant had excess theoretical capacity. A management accounting team working parallel to us, at the same time, reached the conclusion that complexity costs could not be statistically demonstrated in this firm, probably because they were masked by capacity. Quality data, unfortunately, were not available. However, we went back to the shop floor to get a sense for the potential for quality-related complexity costs. The bottleneck workstation in this plant is the paint cell. One quality-related complexity issue revealed to us is hanging the wrong part to be painted. This robs the paint cell of capacity, threatens schedule, and generates rework. But, simple color or bar coding can prevent this. Another issue raised was the many different

shades of grey offered in end products. It may not be detected that one part, for example, was painted the wrong shade of grey until it is assembled into the final product and its contrast with other parts is apparent. This can be remedied by color charting upsteam in the process or by bar coding paint guns. These stories were only revealed when we asked the right questions, prompted by our more complete model of complexity in this firm. Also, it is almost certain that a quality audit of current systems, with fool-proofing remedies, would be less expensive than a modular product redesign. We advised the firm's management that it is our opinion, based on the facts at hand, that if there are complexity costs in this firm with its current resources, they are quality based and can be reduced significantly with initiatives that are less costly than their modular initiative. That initiative cannot be justified based on the current resources and product line. Modular products may still be desirable for this firm, but managers will have to root the cost justification for this initiative in the capital cost savings for future tool sets and not cost savings with current resources.

The following summer, we advised a second student team to look carefully at quality issues. For a variety of administrative reasons, they began their work in the wood room, which was a nonbottleneck station. They found 15% rework and 7% scrap rates in the first section they investigated. About 20% of these problems could be traced directly to complexity issues (again, working on the wrong part or applying incorrect colors), and could be defused with more careful identifying procedures. Management had a sense that quality costs might be high in the firm, but were still surprised at these very low yield levels at a resource that was not commonly considered "critical" (that is, that commonly contributed to schedule anxieties). Quality issues are now a top priority in this firm. We cannot claim sole credit for this management perspective, but neither were our voices inconsequential.

Acknowledgments. The authors would like to thank University of Michigan professors Shannon Anderson, Bill Lanen, Peter Lenk, and Marty Young for helpful advice on estimation theory, the cost of complexity, and accounting systems. We also thank University of Michigan students Jennifer Brown, Charlie Choi, Bill Reeves, Kirk Schell, Ryan Schmidt, and Chris Spears for their on-site work and insights.

Appendix: Using an M/M/1 Approximation to Compute T* and ρ*(λ)

As described in the text, we model "making schedule" as keeping $P_D(\tau) \geq P^{\min}$, where τ is the desired maximal total delay between the arrival of a job and its release for consumption, $P_D(\tau)$ is the probability that the delay is less than or equal to τ, and $P^{\min}$ is a specified lower bound on this probability. At our industrial client's facility, plant managers were expected to have 97% of scheduled production complete within three days of it being released to the shop floor, which suggests $P^{\min} = 0.97$ and $\tau = 3$ days.

Recall that $\lambda = \Sigma_{i=1}^{n} \lambda_i$ is the aggregate demand arrival rate, and $\bar{T} = \Sigma_{i=1}^{n} (\lambda_i/\lambda)\ (t_i/\pi_i)$ is the expected total service time (including rework) experienced by the next arriving job. Hence, the system utilization is $\rho = \lambda\bar{T}$.

If we are willing to assume that the interarrival times and generic service times are exponentially distributed, then we know from M/M/1 queuing logic (cf. Heyman and Sobel 1982) that the long-run probability of there being k jobs in the system is $(1 - \rho)\rho^k$. The delay experienced by an arriving job, conditional on there being k jobs in the system, is the sum of $(k + 1)$ exponential service times (one for each of the k jobs in the system and the arriving job). This sum is gamma distributed, with

$$Pr\{Delay \leq \tau \mid k\} = 1 - e^{-\tau/\bar{T}} \sum_{j=0}^{k} \frac{(\tau/\bar{T})^j}{j!}.$$

So,

$$P_D(\tau) = \sum_{k=0}^{\infty} (1 - \rho)\rho^k \left[1 - e^{-\tau/\bar{T}} \sum_{j=0}^{k} \frac{\left(\frac{\tau}{T}\right)^j}{j!}\right],$$

which can be increased to unity by reducing $\bar{T}$ (and hence ρ) to zero.

In our formulation $P^{\min}$ and τ are parameters that remain fixed as we adjust demand and capacity. For a given aggregate arrival rate λ, define $T^*(\lambda)$ as the maximum $\bar{T}$ such that $P_D(\tau) = P^{\min}$. Making schedule is equivalent to ensuring that $\bar{T} \leq T^*(\lambda)$. The plant manager can decrease $\bar{T}$ by increasing the amount of time available in a working day (that is, use overtime) or decreasing the time needed to process a unit (by forcing the workers to rush jobs).

Table A.1 shows T^* and ρ^* as a function of λ using the above M/M/1 approximation with $\tau = 3$ days and $P^{\min} = 0.97$.

Table A.1 T*(λ) and ρ*(λ) from M/M/1 Approximation

λ	T*(λ)	ρ*(λ)
1	0.460	0.460
2	0.315	0.630
3	0.240	0.720
4	0.194	0.776
5	0.162	0.810
…	…	…
10	0.0895	0.895
…	…	…
20	0.0472	0.945

References

Anderson, Shannon. 1995. Measuring the impact of product mix heterogeneity on manufacturing overhead cost. *Accounting Rev.* **70** 363–387.

Banker, R., S. Datar, S. Kekre. 1988. Relevant costs, congestion and stochasticity in production environments. *J. Accounting Econom.* **10** 171–197.

———, ———, ———, T. Mukhopadhyay. 1990. Costs of product and process complexity. R. Kaplan, ed. *Measures of Manufacturing Excellence.* Harvard Business School Press, Boston, MA.

———, G. Potter, R. Schroeder. 1995. An empirical analysis of manufacturing overhead cost drivers. *J. Accounting Econom.* **19** 115–137.

Caves, D., L. Christiansen, M. Tretheway. 1979. Flexible cost functions for multiproduct firms. *Rev. Econom. Statist.* 477–481.

Cohen, S. 1980. Aftereffects of stress on human performance and social behavior: A review of research and theory. *Psych. Bull.* **88** 82–108.

Foster, G., M. Gupta. 1990. Manufacturing overhead cost driver analysis. *J. Accounting Econom.* **12** 309–337.

Heyman, D., M. Sobel. 1982. *Stochastic Models in Operations Research*, Vol. II. McGraw Hill, New York.

Hoffman, K. 1975. *Analysis in Euclidean Space.* Prentice-Hall, Englewood Cliffs, N.J.

Lovejoy, W., K. Sethuraman. 1998. A note on information gathering in complexity studies. Working paper, School of Business Administration, University of Michigan, Ann Arbor, MI.

MacDuffie, John Paul, Kannan Sethuraman, Marshall Fisher. 1996. Product variety and manufacturing performance: Evidence from the international automotive assembly plant study. *Management Sci.* **42** 350–368.

Meister, D. 1976. *Behavioral Foundations of System Development.* Wiley, New York.

Miller, J., T. Vollman. 1985. The hidden factory. *Harvard Bus. Rev.* (Sep–Oct).

Panzar, J., R. Willig. 1981. Economies of scope. *Amer. Econom. Rev.* **71** 268–272.

Pulley, L., Y. Braunstein. 1992. A composite cost function for multiproduct firms with an application to economies of scope in banking. *Rev. Econom. Statist.* **74** 221–230.

Sanders, M., E. McCormick. 1993. *Human Factors in Engineering and Design.* McGraw-Hill, New York.

Sethi, A., S. Sethi. 1990. Flexibility in manufacturing: A survey. *Internat. J. Flexible Manufacturing Systems* **2** 289–328.

Sethuraman, Kannan. 1994. The impact of product variety on manufacturing performance: An empirical investigation of the world automobile industry. Unpublished Ph.D. Thesis, Wharton School, University of Pennsylvania, Philadelphia, PA.

Skinner, W. 1974. The focused factory. *Harvard Bus. Rev.* 40–48.

Teece, D. 1982. Towards an economic theory of the multiproduct firm. *J. Econom. Behavior Organ.* **3** 39–63.

The consulting Senior Editor for this manuscript was Rajiv Banker. This manuscript was received on February 25, 1998, and was with the authors 297 days for 4 revisions. The average review cycle time was 61.6 days.