Safety Stocks in Manufacturing Systems

by

Stephen C. Graves*

WP 1894-87

January 1987
revised June 1987

A.P. Sloan School of Management
Massachusetts Institute of Technology
Room E53-390
Cambridge, MA 02139

SAFETY STOCKS IN MANUFACTURING SYSTEMS

Stephen C. Graves
A. P. Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

January 1987
revised, June 1987
Draft 2.0

The author would like to express his thanks to Cheena Srinivasan for his help with the literature review, and to Professor Ken Baker, who provided some very useful feedback on an earlier draft.

ABSTRACT

Within manufacturing systems, inventories perform various functions and occur for various reasons. We define safety stocks in manufacturing systems as all inventory that is needed because the manufacturing environment is not deterministic and is not uncapacitated. In effect, we include all inventories except for cycle stocks that result due to batch production, and pipeline stocks due to processing and transfer times. This paper provides a critical review of the literature on safety stocks in manufacturing systems. Based on this review, the paper proposes a new modelling approach for thinking about safety stock issues within a manufacturing system. A key feature of the proposed model is that it highlights the tradeoff between the flexibility of a manufacturing system both to change rate and mix, and the investment in inventory.

INTRODUCTION

All manufacturing systems operate with significant investments in inventory. This inventory consists of raw material and parts stock, work-in process inventory, and end-item inventory. These inventories are needed for many reasons. A certain portion, called pipeline stock, is due to processing or transit times. Another portion, cycle stock, is due to the fact that production and material handling activities occur in batches. These two components of the inventory are completely predictable and explainable: the average pipeline stock depends only on the production volumes and processing/transit times; the average cycle stock depends only on the production volumes and production batch sizes. Furthermore, it is clear how to affect these inventories: to reduce the pipeline stock (cycle stock), we need reduce the processing/transit times (batch sizes) for a given production volume. If the manufacturing system operated in a deterministic world, and if there were always adequate capacity, this would be the only inventory needed by the manufacturing system. Needless to say, this is anything but the case. Indeed, for most manufacturing systems, inventory in excess of the pipeline and cycle stocks is very significant. This excess inventory, which we will call safety stocks, is needed in a manufacturing system due to uncertainties in the requirements, production, and supply processes, and due to the inflexibility of the manufacturing system. A manufacturing system uses safety stocks to maintain satisfactory performance, in terms of customer service and production costs, in the face of the various sources of uncertainty and in light of its own inability to respond adequately. Safety stocks are "excess" inventories held beyond the minimum inventory level that would be possible in a deterministic and uncapacitated world.

This definition of safety stocks is much broader than usual. It includes both stocks that explicitly protect against various types of uncertainty, and stocks that perform either a smoothing or decoupling function within a manufacturing operation. The reason for this broader definition is that it is neither possible nor desirable to separate these stocks by category or function in most instances. Indeed, most manufacturing operations do not admit to having any stock that is labeled as safety stock; rather, they just have large work-in-process inventories, which serve multiple purposes: protect against various uncertainties and disruptions, permit production smoothing, and provide some decoupling across multiple production stages. Furthermore, this could be the best policy since explicitly categorizing the manufacturing stock by function would lead to inefficiencies from redundant stocks.

Our understanding of safety stocks as they exist in manufacturing systems is nowhere near that for pipeline and cycle stocks. We have neither a predictive nor prescriptive theory for assessing safety stock levels in manufacturing systems. We can, though, describe some of the reasons that these stocks occur in manufacturing systems. Foremost is the presence of stochastic variability in various forms. On the requirements side, we may have to base production decisions on forecasts of requirements since firm customer orders do not cover the full production lead time. Since these forecasts will change over time as orders are realized, we may need excess inventory across the manufacturing system to be able to provide satisfactory service. On the production side, a particular production process may not be totally reliable; for instance, there may be yield uncertainty or uncertainty in the process duration. Similarly, on the supply side, a vendor may be unreliable with uncertainties either in the replenishment time or quantity. In both cases, excess inventory is required to protect the production schedule against some degree of variability.

The need for safety stocks is also due to the inflexibility of manufacturing systems. Manufacturing systems typically consist of multiple production stages, requiring multiple resources and producing multiple products. A particular product may require processing at several stages, and must compete for production resources at each stage with other products. Since these resources are limited, the manufacturing system does not have full flexibility to respond to schedule changes or recover from process disruptions. In addition, certain stages will perform assembly operations in which component parts are brought together into an assembly. Since a component may be common to several products, a product must also compete for components at an assembly stage. Since the availability of components may also be limited by resource availability, this is another source of inflexibility in the manufacturing system.

The intent of this paper is twofold. First, I provide a review and critique of the research literature on safety stocks for manufacturing systems. To my knowledge, this has not been done before. I hope that this review will be a useful reference for researchers and will stimulate new activity in this area. Second, based on my assessment of the literature, I suggest a new modeling approach for safety stock policy. This approach permits the explicit examination of the tradeoff between safety stocks and production flexibility.

In the next section I describe the primary research paradigm that appears in the literature. I then summarize the major research accomplishments that have come from this paradigm, and follow this with a critical assessment of the progress to date. The key shortfalling is the inability to model the inflexibility of a manufacturing system. Whereas the paradigm permits a wide variety of uncertainties to enter the manufacturing system, it effectively assumes that the system has full flexibility to change its production rate and mix in response to disturbances or disruptions. Based on this assessment, I then describe an alternate model that permits some characterization of the inflexibility of a manufacturing system. This model consists of an aggregate component, in which we represent the (in)ability to change the aggregate production rate, and a disaggregation component for representing mix flexibility. I present first the model for a single production stage. I then show how to use this as a building block for modeling a network of production stages, as would exist in most manufacturing systems. I describe how to use the model not only for sizing and locating safety stocks, but also for examining the tradeoff between inventory and increased production flexibility. I finish with a discussion of the limitations of this model and point out important issues that remain to be addressed.

RESEARCH PARADIGM

Most of the research literature on safety stocks in manufacturing systems uses a common model for representing the behavior of the manufacturing system. The assumptions for this model are effectively the same that underlie the logic for Material Requirements Planning (MRP) systems. The model is a discrete-time model, in which events occur only at the start (or end) of a period. The structure of each product is given by its bill-of-materials. The manufacturing system is represented as a network of production stages or sectors. The processing requirements for a product or a component part are given by a routing sheet which indicates the series of production stages that a product or part must pass through to complete its processing. Associated with each production stage is a known, constant lead time. The assumed behavior of each production stage is given by this lead time: namely, whatever is released into the production stage in time period t, completes processing and is ready for the next production stage in time period t+n, where n is the lead time. This lead time is assumed to be given and inviolable. As a consequence, we treat each production stage as a black box that imposes a fixed delay on any work released to it.

Within this context, the research paradigm has been to introduce some form(s) of uncertainty and then to explore how to deal with it. The most common assumption is that there is uncertainty in the requirements process, i.e. forecast errors. Then, the focus of the research has been to decide how much inventory to keep between various production stages in order to provide satisfactory customer service. To do this also requires the determination of how much work to release into each production stage on a period by period basis.

While this paradigm does not cover all of the relevant research, it does apply to the vast majority. It is an attractive model since viewing the production stages as black boxes not only simplifies the problem but also is consistent with an MRP viewpoint. In the following review of the literature, we will note how specific studies either fit within this paradigm or deviate from it.

LITERATURE REVIEW

While there is not a large literature on safety stocks in manufacturing systems, there are several distinct approaches that have been proposed and studied. The vast majority of these approaches start from the paradigm given above, and can be roughly categorized into exact analyses that attempt to characterize rigorously the optimal inventory policies, and approximate models that attempt to provide good and implementable solutions. In addition, there are several other studies that do not fit neatly into either category, but that are worthy of note.

By no means do I provide an exhaustive survey of the literature. But I have tried to be thorough in terms of giving a representative and balanced view of the field. To the extent that there is bias, I have focused on the modelling literature, with particular emphasis on works I deem to be important. Nevertheless, if the favorite paper of the reader is not included here, it may just be because I missed it in my review efforts.

Exact Analyses

The work of Clark and Scarf(1960) is noteworthy in that it characterizes the optimal inventory policies for a multistage, serial inventory system with stochastic demand (see Figure 1). They use a discrete-time model and assume a single product that is processed through a series of N stages. Each stage has a constant and known lead time, and has a linear processing cost and a linear inventory holding cost. (The raw material stage may also have a fixed ordering or production cost.) Demand that cannot be met from inventory at the final stage is backordered at a linear cost. The objective is to minimize the expected discounted costs. Clark and Scarf show that for a stationary demand process the optimal inventory policy for a serial multistage system is a function of the echelon inventories at each stage and is given by a critical number for each stage. Each period each stage places a replenishment order to bring its echelon inventory position back to its critical number. Furthermore, they show that this policy can be computed by solving a series of one-stage inventory problems. Their solution procedure computes first the optimal policy for the end-item stage (stage 1), assuming that sufficient input is always available from stage 2. From this optimal policy, they then determine the costs imputed upon stage 1 by a stockout by stage 2. This cost is used as the shortage cost for finding next the optimal policy for stage 2 under the assumption that stage 3 never stocks out. The procedure can be repeated for stage 3 and so on.

Schmidt and Nahmias (1985) study a scenario similar to that of Clark and Scarf, except that the single product is an assembly of two components (see Figure 1). They assume three production stages: one each for the procurement or fabrication of each component, and one stage for the assembly of the two components into the end item. Otherwise, all of the assumptions are the same as Clark and Scarf. This modest change to the product structure, however, makes the analysis much harder. And while they are able to characterize the optimal inventory policy for each component and for the end item, it is not clear how their work could be extended to more complex product structures. Nevertheless, the exact analysis of this two-component assembly does provide useful insight into the complexity of managing stocks for components with differing lead times.

Approximate Models: Without Lot-Sizing

Given the great difficulty of deriving optimal inventory policies, there has been surprisingly little work on approximate models for determining safety stocks in manufacturing systems. What work there is, though, is quite interesting. This work falls into two categories based on whether or not lot sizing is considered.

We will first discuss the literature that does not include lot sizing. Rather, this work addresses the safety stock issues without regard for how lot sizing is done. In effect, it assumes a lot-for-lot policy where each production stage reorders each period. As such, this should result in conservative safety stock policies since less frequent ordering (larger lot sizes) implies less exposure to stockout occasions, and hence less need for safety stocks. Since much of this work assumes some form of a base-stock control policy (e.g., Silver and Peterson, 1985, pp 476-480), we first present the base-stock model and its analysis.

Consider a single production stage with a fixed lead time of n time periods where n is a positive integer. Assume a single product that is processed by this production stage and that has a stochastic demand process with D_t being the demand in time period t. Each period a decision is made as to how much work to release to the production stage. We assume that sufficient raw stock is available so that the input to the production stage is never delayed. Since the production lead time is n time periods, work released in period t is completed and put into inventory in period t+n, and is available to satisfy demand in that period. One can view the production stage as a black box with an infinite-capacity conveyor that moves the work through the box at a constant rate; regardless of the load placed on the production stage, it takes n time periods for the conveyor to move a unit of work from start to finish. Demand that cannot be satisfied by inventory, is backordered. To analyze this inventory system, we define the following random variables:

W_t: the work-in-process inventory within the production stage at the beginning of period t;

I_t: the end-item inventory at the beginning of period t;

R_t: the amount of work released to the production stage at the beginning of period t;

P_t: the amount of production completed during period t.

To specify the relationship between the production and inventory variables and to clarify the timing of events, we write the balance equations for this system:

$$W_t = W_{t-1} + R_t - P_{t-1} \qquad (1),$$

$$I_t = I_{t-1} + P_{t-1} - D_t \qquad (2).$$

We refer to W_t as the intrastage inventory and I_t as the interstage inventory. Then, W_t is the intrastage inventory just after the work release at the start of period t, and I_t is the interstage inventory just after satisfying the demand at the start of period t. The production during period t, P_t, converts intrastage inventory available at the start of period t into interstage inventory that is available for satisfying demand at the start of period t+1. The convention of defining inventories at the start of the period is just a matter of taste, and can be changed without any loss. I prefer this convention, though, since I view the inventories as the state variables for time period t, and will express the control variables (the release and production decisions) as functions of these state variables.

Now, a base-stock control policy is a pull system: we initiate in each period a one-for-one replenishment of the observed demand in that period. In the given context, we set the release quantity equal to the demand, i.e., $R_t = D_t$. This is appropriate when we have no forecast of future demand, except to believe that the demand process is stationary. Combining this rule ($R_t = D_t$) with the above balance equations, we see that for all values of t, $W_t + I_t$ is a constant, which we define to be the base stock B. The level of base stock is a decision variable that we need to set to provide the best customer service with the least amount of inventory.

To determine the base stock level, we need to specify the production random variable. For the convention of viewing inventories at the start of a time period, production during time period t becomes available at the start of time period t+1 and can satisfy demand in that period. By assumption, we have a fixed lead time of n periods (n≥1 and integer), which implies that work released in period t (R_t) is available to meet demand in period t+n. That is, $P_{t+n-1} = R_t$ or equivalently, $P_t = R_{t-n+1}$. For t > n, if we assume $W_0 = 0$, and $I_0 = B$, we can substitute for P_t and R_t in (1) and (2) to obtain

$$W_t = D(t-n+1, t) \qquad (3),$$

$$I_t = B - D(t-n+1, t) \qquad (4),$$

where $D(t-n+1, t) = D_{t-n+1} + \dots + D_t$. When I_t is negative, the current period's demand cannot be completely satisfied from inventory, and a backorder results. We set B so that the probability of a backorder condition does not exceed a given service level. If we assume that D_t is an i.i.d. normally-distributed random variable with mean μ and variance σ^2, then we set B by

$$B = n\mu + k\,\sigma\sqrt{n} \qquad (5).$$

k is a service factor that is set to provide a guaranteed service level (e.g., k=1.65 yields a .95 probability that I_t is nonnegative). Using this specification of B, we find the expected inventory levels to be

$$E[W_t] = n\mu \ ,$$

$$E[I_t] = k\,\sigma\sqrt{n} \ .$$

Note that $E[I_t] = k\,\sigma\sqrt{n}$ is the excess inventory that is needed here to provide customer service in the face of demand uncertainty. When inventory shortages result in lost sales rather than backorders, the analysis is much harder since the total inventory ($W_t + I_t$) does not remain constant. In this case simple results such as (5) are not possible.

We are now in a position to describe approximate models for setting safety stocks in manufacturing systems. The earliest work is that of Simpson (1958), who studied a serial production system with base-stock control (see Figure 1). Simpson assumes that each stage observes the end-item (stage 1) demand process D_t, and each stage initiates in each period a one-for-one replenishment of the observed demand in that period; that is, we set the release quantity for stage i equal to the end-item demand, i.e., $R_{it} = D_t$. However, we no longer assume that sufficient input material is immediately available to accomplish the desired release, except for the raw material. Rather, between every pair of adjacent stages we specify a service time, which is a policy or decision parameter. The upstream stage will satisfy the release requests of its downstream stage within the service time. If we set m to be this service time, then the upstream stage must supply to the downstream stage at the start of period t the amount requested m periods ago, namely D_{t-m}. As a consequence, the lead time to replenish the inventory at stage i is the sum of the service time of the upstream stage to supply the input material, call it m_{i+1}, plus the fixed lead time within stage i, n_i. If we assume that the upstream stages are totally reliable, then we can apply a similar analysis to that given for the single-stage system. At the start of each period t, stage i must supply $D(t-m_i)$, which was requested m_i periods ago by its downstream stage. At the start of period t, stage i completes production of $D(t-n_i-m_{i+1})$, since its replenishment lead time is n_i+m_{i+1} periods. If B_i is the base stock for stage i, we can express the inventory after stage i as

$$I_{it} = B_i - D(t-n_i-m_{i+1}+1, t-m_i) \qquad (6),$$

where we define $D(a,b)=0$ for $a>b$. If $0 \le m_i < n_i+m_{i+1}$, the interstage inventory at time t is the base stock minus the demand history that has been supplied to the downstream stage, but has not yet been replenished by the upstream stage, namely the demand history from $t-n_i-m_{i+1}+1$ to $t-m_i$. If $m_i = n_i+m_{i+1}$, then the service time promised by stage i is equal to the time for stage i to replenish its inventory; hence in this case, the stage produces to order and the interstage inventory I_{it} should be constant (and equal to zero). The case when $m_i > n_i+m_{i+1}$ is not considered, since in this context there is no reason to promise a service time strictly greater than the replenishment time.

The derivation of (6) assumes that each stage is always able to fill within its service time a request by its downstream stage. In terms of (6), this equates to I_{it} being nonnegative with probability one. Thus, we would seem to have to set B_i to be greater than the maximum possible demand over an interval of length

$n_i+m_{i+1}-m_i$ time periods. It is here that Simpson makes an approximation. He assumes that at stage i the base stock level B_i is set to ensure routine coverage of a given "maximum reasonable demand" over an interval of length $n_i+m_{i+1}-m_i$ time periods. Implicitly, he seems to assume that when actual demand exceeds the maximum reasonable demand, the production stage will perform the extraordinary actions(e.g., expediting) necessary to fulfill the service time commitment. For instance, the maximum reasonable demand might be defined by a percentile of the demand distribution, where this percentile would reflect the frequency with which the production stage is willing to go into an expediting or overtime mode. Then we would set B_i as

$$B_i = \tau\mu + k\,\sigma\sqrt{\tau} \qquad (7),$$

where $\tau = n_i+m_{i+1}-m_i$ and k is the service factor for the required percentile for the standard normal distribution. From (6) and (7) we see that the expected inventory beyond stage i, the excess inventory, is

$$E[I_{it}] = k\,\sigma\sqrt{\tau}$$

where $\tau = n_i+m_{i+1}-m_i$.

Simpson assumes that the maximum reasonable demand has been preset, and then specifies an optimization problem to find the service times that minimize the total holding costs for the excess inventory. He then shows that an optimal choice for the service times satisfies an extreme point property, namely m_i either equals 0 or equals n_i+m_{i+1}. The significance of this observation is that the optimal policy is an 'all-or-nothing' policy: between any two stages either there is no inventory ($m_i = n_i+m_{i+1}$) or there is sufficient inventory to decouple completely the two stages ($m_i = 0$). Based on this result, the determination of the optimal policy reduces to a simple dynamic program over the stages of the production system.

Hanssmann (1959) considers a very similar scenario to Simpson, but with some significant differences in assumptions. He examines a serial system operating with a base-stock policy. Each stage observes the end-item demand in each period and sets its release quantity equal to the demand, $R_{it} = D_t$. The upstream stage is normally expected to provide the input material in the period of the release; that is, the service time between every two stages is expected to be zero. However, Hanssmann now assumes that this service time can be violated. When an upstream stage has insufficient stock, it does not take extraordinary actions to satisfy the downstream stage. Rather, the delivery of the shortfall is delayed until the upstream stage has sufficient stock. Although the length of this delay is a random variable, Hanssmann approximates it as a deterministic delay equal to its expected value. This deterministic delay from an upstream stage is added to the fixed lead time for the downstream stage. Hence, the poorer is the service provided by the upstream stage, the longer will be the replenishment lead time for the downstream stage and the more excess inventory it will need. For the end-item stage, however, this delay is imposed not upon another stage, but upon the customer. Hanssmann assumes that the demand process is a function of the expected delivery delay seen by the customer; in particular, the level of lost sales and lost customers increases with the length of the delay. For this model of system performance, Hanssmann formulates an optimization problem to find the base stock levels (and expected interstage delays) that maximize sales revenues minus inventory holding costs. This optimization problem can be solved as a dynamic program over the production stages. Unlike Simpson's model, though, the solutions from this dynamic program need not result in an all-or-nothing stocking policy.

In comparing the models of Simpson and Hanssmann, it is interesting to note the difference in the assumed behavior of the production system, particularly when a stage stocks out. Hanssmann assumes that when an upstream stage stocks out, the releases into the downstream stage are delayed. He then approximates this stochastic delay by its expected value to simplify the analysis of his model. As such, Hanssmann's model of system behavior is mathematically well-defined, and his approximation is quite testable by means of a Monte Carlo simulation. Simpson's model, however, is not as rigidly specified, and is more subtle. Simpson assumes that there is no delay on the downstream stage when its upstream stage stocks out; in effect, he avoids the consequences of interstage shortages. His justification for this seems to be the supposition that the purpose of safety stocks is to protect against normal variability, i.e., the maximum reasonable demand; safety stocks permit the system to function routinely in the face of normal variability. Safety stocks should not be held for protection against abnormal variability; rather, the organization maintains some slack capability to respond to abnormal variability. That is, the organization will switch from a routine operating mode to an emergency mode as needed. The specification of what is normal versus abnormal variability depends upon the frequency with which the organization is willing to revert to its emergency mode. But given this specification, then Simpson's model finds the best allocation of safety stock to deal with normal variability.

It is not clear how to choose between these two models. On the one hand, Hanssmann's model is appealing in that its mathematics can be fully specified and the effectiveness of his approximation can be quantified. On the other hand, Simpson's model seems to be more descriptive of how many organizations work.

The question seems to rest on the role of safety stocks: should we plan these stocks to account for all contingencies assuming that the production system is inflexible, or only to account for reasonable contingencies assuming that the production system will always be able to bend for the remaining cases. We comment upon this point again in my critique of the literature.

Miller (1979) introduced the concept of "hedging" as a means to provide safety stocks within a manufacturing system. He describes the approach in terms of a Materials Requirement Planning (MRP) system. In the face of demand uncertainty or forecast errors, he suggests that the master schedule (the production schedule for the end-item stage) be inflated to reflect the uncertainty over time in the end-item demand. The amount by which the schedule is inflated is called the hedge. While this notion has some intuitive appeal, the specific implementation suggested by Miller is not compatible with the earlier models of Simpson and Hannsmann, and seems to be without analytical support.

To explain the concept of hedging, consider a serial system (Figure 1) with lot-for-lot scheduling. Suppose the end-item demand process is stationary, i.i.d., and normally distributed with mean μ and variance σ^2. In MRP terminology, the demand forecast for each period is μ, and the single period forecast error is σ. Then Miller suggests setting the master schedule for the end item so that the cumulative planned production over the next τ time periods is $\tau\mu + k\sigma\sqrt{\tau}$ for all values of τ for some service factor k; that is, the production schedule is set to cover some desired percentile of the possible demand realizations, e.g., k=1.65 for 95% service. The cumulative hedge over the next τ time periods is $k\,\sigma\sqrt{\tau}$, and is realized as safety stock spread across the production pipeline.

In the notation of the current paper, we can interpret this hedging policy as a base-stock system. For a serial system with lot-for-lot scheduling, the hedging policy would set planned orders (or releases) for stage i at the start of time period t such that

$$(W_{1t} + I_{1t}) + \ldots + (W_{it} + I_{it}) = \tau_i\mu + k\sigma\sqrt{\tau_i}$$

where $\tau_i = n_i + n_{i-1} + \ldots + n_1$ is the cumulative lead time from stage i to the completion of the end item. Here, W_{it} denotes the planned orders for stage i that are in process at time t, and I_{it} is the on-hand inventory at time t. Thus, at each stage i, we set the planned orders so that the planned production over the next τ_i periods can cover a cumulative demand of $\tau_i\mu + k\sigma\sqrt{\tau_i}$. This is equivalent to the base-stock system with zero service times where (for $\tau_0 = 0$)

$$B_i = W_{it} + I_{it} = n_i\mu + k\,\sigma(\sqrt{\tau_i} - \sqrt{\tau_{i-1}}) \qquad (8a).$$

Each period each stage will observe the end-item demand D_t, and will set its planned orders (releases) equal to this demand, i.e., $R_{it} = D_t$. From (6), we can write the inventory after stage i as

$$I_{it} = B_i - D(t-n_i+1, t)$$

$$= n_i\mu + k\,\sigma(\sqrt{\tau_i} - \sqrt{\tau_{i-1}}) - D(t-n_i+1, t) \qquad (8b),$$

since the service times m_i are all zero. From (8b) we see that the expected inventory beyond stage i, the excess inventory, is

$$E[I_{it}] = k\,\sigma(\sqrt{\tau_i} - \sqrt{\tau_{i-1}}),$$

as found by Miller. However, this specification of the base stocks (8a) will not provide the service levels implied by Miller for either Simpson's model of system behavior or that of Hanssmann. For Simpson's model, the frequency with which each stage stocks out would be much greater than is implied by the service factor k in (8a). Using (7), we can show that the actual service factor for the suggested base stock (or hedge) would be $k' = k\,(\sqrt{\tau_i} - \sqrt{\tau_{i-1}}\,) / \sqrt{\tau_i}$, which is strictly less than k. Indeed, for Simpson's model, the base-stock levels given by (8) would ensure that the quantity

$$I_{it} + B_{i-1} + B_{i-2} + \dots + B_1$$

is nonnegative with the probability associated with the service factor k (e.g., probability .95 for k=1.65). To see why this is true, we can use (8a) and (8b) to obtain

$$I_{it} + B_{i-1} + B_{i-2} + \dots + B_1 = \tau_i\mu + k\sigma\sqrt{\tau_i} - D(t-n_i+1, t) ,$$

from which this observation follows. But it is not clear why the above quantity is of any interest. For Hanssmann's model, there would be additional replenishment delays due to stockouts that are not reflected in (8a) or (8b). Hence, although the qualitative ideas in Miller's paper are of interest, I cannot identify a model that supports the explicit suggestions for the safety stock levels.

Wijngaard and Wortmann (1985) provide a thoughtful review paper on inventories within MRP systems. Their primary focus is on prescribing interstage inventories under the standard research paradigm described earlier. They examine not only serial systems, but also simple assembly and distribution structures. Unfortunately, though, they use the same result as Miller did, namely that the safety stock required by stage i is given by $k\,\sigma(\sqrt{\tau_i} - \sqrt{\tau_{i-1}}\,)$, where $\tau_i = n_i + n_{i-1} + \dots + n_1$ is the cumulative lead time for stage i.

Approximate Models: With Lot-Sizing

The second category of approximate models considers lot sizing along with safety stocks. The earliest work is that of Clark and Scarf (1962) who extend their 1960 work to allow a fixed ordering cost at each stage. Again, they assume a serial system with periodic review and a stationary but uncertain demand process for the end item. Each stage has a linear inventory holding cost, a linear production cost and a fixed cost for initiating a replenishment. End-item demand that cannot be met from stock is backordered with a linear penalty cost. Their solution method successively computes the optimal (s,S) policy for each stage, where s is the reorder point, S is the order-up-to level, and both parameters are in terms of echelon inventory. Successive stages are linked by a penalty cost that represents the cost on the downstream stage of a stockout by the upstream stage. While this solution procedure does not guarantee the optimal multistage policy, it does provide both upper and lower bounds on the cost of the optimal policy.

Lambrecht et al. (1984) extend the Clark-Scarf procedure in two ways. First, they point out the ineffectiveness of the Clark-Scarf procedure when the natural order quantity (EOQ using echelon costs) for a downstream stage is greater than that for its upstream stage. For this case, they suggest collapsing the two stages into one stage before applying the Clark-Scarf procedure; in effect, they impose a constraint that forces the two stages to order concurrently with the same order quantity. Second, they show how to extend the Clark-Scarf approach to an assembly structure. In essence, the modification is to recognize the need to coordinate the replenishment policies for components for the same assembly. In addition, Lambrecht et al. provide experimental results that show the effectiveness of policies from their approximate procedure compared with the optimal policies from solving a Markov decision problem. These experimental results also provide some insight into the general form of optimal or near optimal policies. For two-stage serial systems, they find that these policies maintain a safety stock for the end item (stage 1); for the component (stage 2), the optimal policies plan the component replenishments to arrive a bit before these components are needed by stage 1, on average. The authors interpret the optimal policy for the components in terms of a safety time, where the safety time is defined as the expected time between when a replenishment quantity becomes first available and when there is the first usage of any of this replenishment quantity.

Lambrecht et al. (1985) extend their previous work to permit capacity restrictions υn production by the end-item stage. They recognize that the optimal policy can again be obtained in theory by solving a Markov decision problem, and provide experimental results on a series of test problems. These experiments illustrate the form of the optimal policy, and indicate the impact of the capacity constraint on the inventory policy.

Carlson and Yano (1984, 1986) consider a two-level assembly system with stationary but uncertain demand for the end item. They assume that the timing of the production replenishments for the end item has been planned in advance, and is cyclic; that is, the end item will be replenished every T periods, where T is prespecified. However, the amount replenished can vary and will reflect the recent demand history. They call this "fixed scheduling." In the 1984 paper they assume that the timing of component replenishments is also fixed in advance and cyclic, where the cycle length is an integer multiple of that for the end item. In the 1986 paper they assume "flexible scheduling" for the components; that is, component replenishments are replanned each period, and emergency replenishments are scheduled whenever a component runs short. In both cases, Carlson and Yano develop an algorithm for setting component and end-item safety stocks based on an approximate marginal analysis. For the case of fixed scheduling for the components, they find that the best allocation has no safety stock for the components; for the case of flexible scheduling for the components, they find that there are benefits from having safety stock at both the component and end-item level. In Yano and Carlson (1985, 1987), they compare via simulation the performance of a fixed scheduling policy with that for a flexible scheduling policy. For a two-level assembly system, they find that a fixed scheduling policy for both components and the end item dominates any other policy. This finding implies that if fixed scheduling is possible, there may be little value for component safety stocks.

De Bodt and Graves (1985) consider virtually the same scenario as Clark and Scarf (1962), but with a continuous-review policy. They restrict attention to policies specified by a reorder point and order quantity for each stage, where each parameter is expressed in terms of echelon inventory. This is in contrast to having an (s,S) policy in the echelon inventory for each stage, as assumed by Clark and Scarf for a periodic review system. Furthermore, De Bodt and Graves assume a nested policy; whenever a stage reorders, all downstream stages also reorder. In order for the policies to be stationary, the order quantity at each stage must be an integral multiple of the order quantity of its downstream stage. They then give an approximate cost model as a function of the policy parameters, and show experimentally the accuracy of the approximate cost model. For this cost model, the best choice of policy parameters can be found analytically.

It is interesting to note that the reorder policy assumed by De Bodt and Graves is similar in spirit to the fixed scheduling policy of Yano and Carlson (1984). Both policies are nested in that when a stage reorders, its downstream stage also reorders. De Bodt and Graves assume that the order quantities remain fixed, but allow the timing between replenishments to vary with the demand; Yano and Carlson fix the timing between replenishments, but allow the order quantities to vary according to the demand realization. Furthermore, the policy form assumed by De Bodt and Graves necessarily results in only safety time for the component stages, but safety stock for the end item. As such, it is consistent with the findings of Lambrecht et al. (1984) and Yano and Carlson (1984, 1985).

Other Studies

There have been several other research efforts that are worthy of note, but that do not fit cleanly into the material reviewed above. In particular, there are four sub-categories that we comment upon here, namely (i) studies that use simulation as an exploratory tool to identify possible principles for setting safety stock policy; (ii) papers that describe the relevant issues and tradeoffs, and propose operational guidelines for establishing safety stock levels; (iii) papers that focus on understanding the role of component commonality; and (iv) papers that study process time variability and how to prescribe safety times.

The best known simulation study is that of Whybark and Williams (1976). They identify four types of uncertainty in a production system: uncertainty in supply timing, in demand timing, in supply quantity, and in demand quantity. They then show, via a simulation study of a single-item, single-stage system, that safety stocks are the best mechanism for protecting against uncertainty in the supply or demand quantity, while safety times are preferred for timing uncertainties in either the supply or demand processes. Other simulation studies have been performed by Grasso and Taylor (1984), Schmitt(1984) and Guerrero et al. (1986). Grasso and Taylor simulate an MRP system with three end items, each with a multi-level product structure. They examine the performance of various buffering policies and lot-sizing policies in the face of timing uncertainty in the resupply of purchased parts. Their findings are not consistent with those of Whybark and Williams; for their simulation experiments, Grasso and Taylor find that safety stock is preferred over safety time to buffer against supply-timing uncertainty. Schmitt simulates an MRP system with an assembly and a fabrication production stage that produce four end items and twelve components, respectively. He allows uncertainty in end-item demand and in process time, and considers the effect of end-item safety stock, slack capacity at the production stages, and more frequent rescheduling on system performance. The major finding seems to be to identify the significance of the tradeoff between the use of safety stocks, excess capacity, and more frequent rescheduling (e.g., more setups). Guerrero et al. simulate Miller's hedging policy for a three-stage serial system where end-item shortages result in lost sales. They find that this hedging policy provides the desired service level (fill rate), but with a slight bias due to the fact that the development of the hedging policy assumes backorders rather than lost sales. They also simulate a hedging policy in which the replenishment of the safety stocks is smoothed; this policy is similar in spirit to the model proposed later in the current paper.

There are innumerable papers, particularly in trade journals, that discuss the need for safety stocks and propose general guidelines or operational schemes for establishing safety stocks in manufacturing systems. Much of this literature is in reaction to the proposition that safety stocks are only needed for the end item in a properly-implemented MRP system. This proposition is a byproduct of the fact that most implementations of MRP systems make no explicit provision for dealing with uncertainties. Nevertheless, this proposition is not uniformly accepted, as indicated by the outpouring of papers on safety stocks in MRP systems. We will not try to survey these papers here, but refer the interested reader to the excellent and comprehensive papers by New (1975), Berry and Whybark (1977), and Meal (1979). In particular, we note the paper of Meal, who suggests how to modify an MRP system first to measure the relevant uncertainties in its environment, and second to use these measurements to set safety stocks or times within the MRP framework.

Component commonality occurs when a component is used by more than one end item. While it is widely recognized that commonality should require less safety stock because of the opportunity to pool risks, we only have approximate models for determining the actuals benefits. Recently, there have been several efforts at understanding better the role and importance of component commonality. Both Collier (1982) and Baker (1985) point out the safety-stock reductions possible from component commonality. Baker also shows the difficulty of predicting the service level for a set of end items from the safety stock levels for their components when commonality exists; that is, the service levels anticipated for the components (e.g., 95% service for a service factor k=1.65) do not directly translate into a service level for the end items when commonality is present. Baker et al. (1986) analyze a simple single-period model with two end items, and three components, where one of the components is common to both end items. Their intent is to provide a framework for thinking about the relationship between service level and safety stocks, as well as to derive qualitative guidelines for setting inventory policy. They show that while there are inventory savings for the common component, the inventory levels for the unique (non-common) components actually increase in the presence of commonality; nevertheless, component commonality still results in a net inventory reduction. Gerchak et al. (1986) and Gerchak and Henig (1986) extend the model and findings of Baker et al. to more general settings with less restrictive assumptions.

Although the previous work that we cite seems to focus nearly exclusively on uncertainty in the demand process, there are other sources of uncertainty to consider. Yano (1987) considers a multistage serial system in which the lead times for each stage are stochastic. She assumes a control policy based on having a planned lead times for each production stage. For instance, in a two-stage system, a job is released to the upstream stage (stage 2) exactly n_1+n_2 time periods before its due date, where n_i is the planned lead time for stage i. The job is released to the downstream stage (stage 1) either n_1 time periods before its due date or when it completes processing at stage 2, whichever occurs later. Assuming that the realized lead times are independent of the planned lead times, Yano formulates an optimization problem to choose the planned lead times that minimize expected tardiness and earliness costs. In effect, she finds optimal safety times, where the safety time at a stage is the difference between its planned lead time and its expected realized lead time.

Finally, we note two bodies of literature that we have not reviewed, but that may have relevance to safety stocks in manufacturing systems. First, there is the research on inventory policies for distribution systems (e.g., Eppen and Schrage, 1981; Schwarz, Deuermeyer and Badinelli, 1985). A central theme of this research is the sizing and positioning of safety stock, but for distribution systems rather than production systems. Second is the research on the design of transfer lines (e.g., Buzacott and Hanifin, 1978; Gershwin and Schick, 1983). The primary focus of this area has been to determine the size of physical buffers needed within a transfer line whose stations are unreliable. The determination of these physical buffers, in effect, establishes safety stocks between the stations on the transfer line. We mention these two areas because there is a similarity of interests and there may be an opportunity to transfer results from one problem area to another. However, at this time these areas remain fairly distinct and we believe a review of these literatures would be premature.

CRITIQUE OF LITERATURE

Without question there has been substantial progress over the past thirty years. We have good models, both approximate and exact, for determining safety stock policy for serial systems with stochastic demand for a range of cost and operational assumptions. From these models, we have identified some general findings about good policies. In particular, I would cite the introduction of an echelon stock perspective as promulgated by Clark and Scarf (1960), and the all-or-nothing policy for interstage safety stocks derived by Simpson (1958). Also of significance is the work on safety stocks in the presence of lot sizing, in which general findings are the effectiveness of nested reordering policies, and the preference for safety times over safety stocks for the upstream stages (Lambrecht et al. 1984, Yano and Carlson 1984, 1985, De Bodt and Graves 1985).

There has been less progress for more complex product structures. A two-level assembly structure seems to be the richest product structure that has been dealt with in any depth (Schmidt and Nahmias 1985, Carlson and Yano 1986). Yet, even here general guidelines are hard to come by. There has also been progress made on component commonality, but mainly in terms of establishing its importance and the analytic difficulty it presents (Baker at al. 1986). Great opportunities exist for developing good models to study these more complex product structures.

In assessing the research to date, two aspects require greater comment, namely the research paradigm and the presumed role for safety stocks.

The research paradigm is based on the assumption that production stages can be modeled as black boxes. That is, we can describe the operation of a production stage by a fixed and deterministic lead time: what goes into the stage in period t, comes out in period t+n for n being the lead time. There are several consequences from this assumption. The first consequence is that all models based on this paradigm are essentially for a single product. Since the production lead times are given and fixed, the production and inventory policy of one product has no influence on any other product; hence, the paradigm implies that safety-stock planning separates by product, since there is no competition for scarce processing resources or components. Furthermore, the models do not distinguish between a production stage and an element in the product structure (i.e., bill-of-material). For a single product, each element in the bill-of-material (a component or subassembly, say) has a replenishment lead time, and as such, can be viewed as having a dedicated production stage that just produces that element according to the stated lead time. A second consequence is that fixing the lead times determines the intrastage inventory: a lead time of n time periods results in an average intrastage inventory of $n\mu$ for μ being the expected output per period. The intrastage inventory provides no protection against variability in the manufacturing system, and the models are restricted to using only interstage inventories for this function. A third consequence is that the models cannot consider the (in)flexibility of the manufacturing system in setting safety stock policy. The research paradigm assumes that the manufacturing system is completely flexible with respect to its ability to change production rate and mix. As a result, it rules out consideration of the tradeoff between safety stocks and increasing the flexibility of the manufacturing system.

The second comment concerns the question raised in the discussion of the papers by Hannsmann and by Simpson, namely the role of safety stocks. Are safety stocks suppose to cover all possible variability in a manufacturing system or just normal variability, e.g. the maximum reasonable demand? Another way of putting this is, under extreme circumstances do manufacturing systems behave routinely or do they respond in an equally extreme manner? Most of the literature has assumed the former. One consequence of this is that we then have a complete description of how the system operates. But most manufacturing systems do not behave in this fashion; rather, these systems rely on various slack resources or capabilities, in addition to safety stocks, to respond to variability. In particular, managers will take extreme actions when faced with extreme circumstances. But from a modelling standpoint, this view poses a dilemma of how to represent the possible responses the system will make when subject to abnormal variability. We no longer have a clear description of how the system operates. One approach is to assume that for the purposes of setting safety stocks, only normal variability is to be considered; hence, no attempt is made to model system behavior when extreme circumstances persist, other than to assume that it brings the system out of this condition. One attraction of this approach is the greater analytical tractability that seems possible for this reduced role of safety stocks (e.g., Simpson). But this approach creates a new problem of having to specify what is normal and what is not. And, although we may contend that this view is more consistent with actual practice, there is no proof other than anecdotal evidence. Finally, the question remains as to whether this viewpoint will lead to more useful models for setting safety stock policy.

In the next section we present a safety stock model that attempts to address the comments and concerns raised above. In this model we take the viewpoint that we plan safety stocks for protection against normal variability. We allow multiple products where the products share production resources. Furthermore, the model treats both interstage and intrastage inventory, and permits consideration of the flexibility for a production stage to change rate and mix. As such, the model allows the examination of the tradeoff between safety stock and increased flexibility. Needless to say, we require some additional assumptions about system behavior that we will comment upon as we go along.

A SAFETY STOCK MODEL

We present a modelling approach that entails an aggregate component and a detailed component. The aggregate component models the aggregate (multi-item) behavior of the manufacturing system, while the detailed component provides a characterization of item behavior via a disaggregation of the aggregate model. The primary assumptions for the modelling approach are the same as for the research paradigm, with the exception of the assumed behavior of a production stage. We do not assume that each production stage behaves as a black box that delays work by a prespecified lead time. Rather, we assume we can specify the behavior of each stage by a control rule, which is parameterized by a planned lead time. This planned lead time acts as a target for the production stage, and dictates how the production stage performs. Furthermore, the planned lead time is the parameter by which we introduce into the model the flexibility of a production stage to change its aggregate production rate.

The modelling approach is quite general in its ability to model multiple products with complex product structures, complex production networks, and various sources of uncertainty. However, we will present the model in its simplest realization, and then point out how the model might extend to include various complicating factors. The simple setting has one production stage that processes multiple items. The only uncertainty is in the demand process, which is assumed to be stationary. We use a base-stock (or pull) policy for inventory control. We first describe the aggregate model and then indicate how to disaggregate the results from this model to set safety stocks for individual products. This one-stage model will be a building block for constructing a model for a multistage system.

The modelling approach is related to the model developed and tested by Graves, Meal et al. (1986) for aggregate production planning in a two-stage system. However, there are significant differences with regard to the specifics and the intent of the model presented in the current paper.

Aggregate Model

To describe the aggregate model we use the same notation as previously introduced for the base-stock system. I_t and W_t are the interstage and intrastage inventory, respectively; P_t and R_t are production and release quantities, respectively. Now, however, the random variables I_t, W_t, P_t and R_t are aggregate entities. For instance,

$$I_t = \sum I_{it} ,$$

where I_{it} is the interstage inventory for product i at the start of time period t. D_t is the aggregate demand in period t. The balance equations are the same as for the base-stock system:

$$W_t = W_{t-1} + R_t - P_{t-1} \qquad (9),$$

$$I_t = I_{t-1} + P_{t-1} - D_t \qquad (10).$$

And the release rule remains the same, namely $R_t = D_t$; thus, the total aggregate inventory, $W_t + I_t$, remains constant and equals a base stock B.

We no longer assume that what was released into the production stage in period t-n becomes available to meet demand in period t, for n equal to the given lead time. Rather, we assume that the aggregate production output is determined by a production control rule that attempts to smooth the aggregate output and is parameterized by a planned lead time. The planned lead time for a stage is a decision variable and represents the total time that we plan for an item to spend at the production stage. The actual time, however, may deviate from the planned lead time due to production smoothing. Our intent is to prescribe a simple smoothing rule for which the actual lead times correspond closely to the planned lead time. In particular, we assume

$$P_t = W_t/n \qquad (11)$$

where n ≥1 is the planned lead time for the stage (which we do not require to be integer). This control rule outputs 1/n of the in-process inventory each period. Whereas we clearly need to do this on average to achieve an average lead time of n periods, the proposed control rule does this exactly each period.

With this specification of the control rule, we no longer view the production stage as a black box through which work moves at a constant rate, as if on a fixed speed conveyor. Rather we view each production stage as a filter. Each production stage sees an input process, given by the time series of demands, and converts this input process into an output process, namely the production time series. We envision the production stage acting as a filter that smoothes or damps the input process. Smoothing is achieved at a production stage by using the work-in-process inventory to average out fluctuations in the input (demand) process. Increasing the work-in-process inventory provides more smoothing since it permits more opportunity to average the peaks and valleys of the input process. The degree of smoothing needed depends on the noisiness or variability in the input (demand) process relative to the production capability at that stage. As will be seen, the control rule (11) is a smoothing rule in which the degree of smoothing is parameterized by the planned lead time n: increasing the lead time results in a larger work-in-process inventory and thus, a greater damping of the input process.

We can justify or defend this rule on several counts. This control rule is the simplest of a family of linear control rules that minimize a quadratic objective function in the production and in-process inventory levels; see the appendix for this derivation. Furthermore, Graves (1986) shows for a different context that the actual lead times from this rule are surprisingly reliable for a stationary demand process. Finally, we argue that this rule is an improvement over the typical rule, $P_t = R_{t-n+1}$, in that it views the planned lead time as a decision parameter. As a consequence, we need model the raison d'etre for these lead times, and include consideration of these lead times when setting safety stocks. In particular, it forces one to consider the use of both interstage and intrastage inventories. See Graves (1986) for additional discussion of this control rule.

By substituting (9) into (11) and D_t for R_t, we can reexpress (11) as a simple smoothing equation:

$$P_t = D_t/n + (n-1)P_{t-1}/n \qquad (12).$$

From (12), the effect of the planned lead time is clear: the larger is the planned lead time, the more we smooth the aggregate output. With smoother production, the production process requires less flexibility to change the production rate. But, as we will show, larger planned lead times require more interstage and intrastage inventory. Hence, to set the planned lead time, we must examine the tradeoff between inventory and production flexibility.

Suppose the aggregate demand process D_t is i.i.d. and is normally distributed with mean and variance given by μ and σ^2. Then, if we assume an infinite history, we obtain from (12) that

$$E[P_t] = \mu \qquad \text{Var}\,[P_t] = \sigma^2/(2n-1) \qquad (13)$$

where E[ ] denotes expectation and Var[ ] denotes the variance. From (11) and (13), we have that

$$E[W_t] = n\mu \qquad \mathrm{Var}\,[W_t] = n^2\sigma^2/(2n-1) \qquad (14).$$

Since $W_t + I_t = B$, we find from (14) that

$$E[I_t] = B - n\mu \qquad \mathrm{Var}\,[I_t] = n^2\sigma^2/(2n-1) \qquad (15).$$

Furthermore, since D_t is normally distributed, P_t, W_t, and I_t are also normally distributed with means and variances given above.

We use (13) to quantify the amount of production smoothing as a function of n, and the amount of production flexibility needed by the production stage. In particular, the production stage needs the capability to output P_t where P_t is a normal random variable with parameters given by (13). Presumably, we choose n so that P_t is consistent with the given capability of the stage. Suppose that we express the capability of a production stage in terms of a maximum reasonable output level, given by $\mu + \chi$. Since μ is the average output level, then χ denotes the slack that is normally available at the production stage. We say that P_t is consistent with the capability of the stage if the probability that P_t exceeds $\mu + \chi$ is acceptably small. Thus, we would set n such that $k\sqrt{\mathrm{Var}[P_t]}$ equals χ, where k is a service factor. From (13) this implies we set n by

$$n = (k^2\sigma^2 + \chi^2)/2\chi^2 ,$$

where we have assumed that $\chi \le k\sigma$; otherwise, we would set $n = 1$.

As an example, suppose we have a production stage with an aggregate demand process characterized by $\mu = 100$ and $\sigma = 25$. Suppose the capability of the production stage (maximum reasonble output level) is 130 ($\chi = 30$), and we specify a service factor $k = 1.65$ to ensure a probability of .95 that the output P_t is within the stated capability. Then we need set the planned lead time $n = 1.45$.

We define the flexibility to change the production rate by a dimensionless parameter $F = \chi/k\sigma$; then n is given by

$$n = (1 + F^2)/2F^2 \qquad (16).$$

We use F to characterize the amount of rate flexibility in a production stage. F is the ratio of a measure of the slack available to vary the aggregate production rate (χ) to a measure of the variability of the demand process, namely $k\sigma$. In the example above, the available slack is $\chi = 30$, and the demand variability is given by $k\sigma = (1.65)(25) = 40.25$; hence, $F = .75$, which indicates that the ability of the production stage to vary its production rate is only 75% of that needed by the demand process. When $F = 1$, the production stage has sufficient flexibility to vary its production rate to match the demand process. At the other extreme, $F = 0$, the production stage has no flexibility and can only produce at the expected demand rate (i.e., $P_t = \mu$); this is unrealistic since it impies $n = \infty$, and leads to infinite in-process inventory. For $0<F<1$, the production stage has some flexibility but not enough to match the demand process; thus, the stage will set its production to smooth the demand process, where more smoothing is needed for smaller values of F. This smoothing is made operational by the control rule (11), for which we need specify a planned lead time n.

We use (15) to specify the amount of safety stock necessary to provide acceptable service. However, this cannot be done immediately from (15) since I_t corresponds to the aggregate interstage inventory. Rather, we first need to characterize the "make-up" of this aggregate inventory since the service provided by this inventory is on an item level. (That is, we cannot use inventory of item j to satisfy a requirement for item i.) We describe two extreme disaggregations of the one-stage model that differ according to the assumed degree of mix flexibility. By "disaggregation," we mean to characterize the production and inventory random variables for individual items in a way that is consistent with the aggregate model. For instance, for P_{it} being the production output for item i in period t, we seek to characterize P_{it} given that

$$P_t = \sum P_{it} \qquad (17),$$

and given that the aggregate production P_t is specified by (11). Throughout this discussion of disaggregation, we assume that all variables are expressed in common units, namely in units of capacity of the critical resource for the production stage.

The first disaggregation assumes limited mix flexibility. By this, we mean that the production stage has limited ability to expedite or de-expedite the processing of individual items within the production stage. More specifically, we assume that the production decisions for item i are given by

$$R_{it} = D_{it} \qquad\qquad P_{it} = W_{it}/n \ .$$

That is, each item releases work to the stage according to a base-stock control system, and outputs work from the stage according to the same smoothing rule as for aggregate production. Since $W_t = \Sigma\, W_{it}$, this specification for P_{it} satisfies (17). By repeating the analysis for the aggregate model, we find that

$$E[W_{it}] = n\mu_i \qquad\qquad Var\ [W_{it}] = n^2\sigma_i^2/(2n-1)$$

$$E[I_{it}] = B_i - n\mu_i \qquad\qquad Var\ [I_{it}] = n^2\sigma_i^2/(2n-1) \qquad (18),$$

where we now assume the item demand process D_{it} is i.i.d. and is normally distributed with mean and variance given by μ_i and σ_i^2. In order to assure satisfactory service, we use (18) to set the total intrastage and interstage inventory B_i by

$$B_i = E[W_{it}] + E[I_{it}] = n\mu_i + k\ [n\sigma_i/\sqrt{(2n-1)}]\ ,$$

where k is the service factor and is assumed to be the same for all items. In terms of the parameter for rate flexibility, we can write B_i as

$$B_i = \{(1+F^2)/2F^2\}\ \{\mu_i + kF\sigma_i\ \},$$

and by summing over all of the items, we have

$$B = \{(1+F^2)/2F^2\}\ \{\mu + kF \sum \sigma_i\ \} \qquad (19).$$

This expression gives the total inventory as a function of the production rate flexibility, under the assumption of limited mix flexibility. (Note that we have used a generic k to denote a service factor both for specifying the base-stock levels B_i and for specifying the planned lead time n and flexibility parameter F. These service factors need not be the same, and could be distinguished via additional notation.)

The second disaggregation assumes complete mix flexibility, in which the production stage can completely alter the make-up of the aggregate output on a period by period basis. We again assume that each item releases work to the stage according to a base-stock (or pull) control system, i.e., $R_{it} = D_{it}$. However, each period we now set the output for each item so that the aggregate constraint (17) is satisfied, and so that all items have the same protection against stockout in the upcoming period. That is, we set P_{it} so that

$$(I_{it} + P_{it} - \mu_i)/\sigma_i = K_t \qquad (20)$$

where K_t is the same for all items. A realization of K_t corresponds to the common service factor for the time period t+1. For complete mix flexibility, we assume that (20) is always feasible; that is, we can expedite out of the stage whatever is needed to get equal protection across all items. By multiplying (20) by σ_i and summing over all i, we obtain

$$K_t \sum \sigma_i = \sum (I_{it} + P_{it} - \mu_i) = I_t + P_t - \mu \qquad (21).$$

From (21) we see that K_t is a random variable that is fully determined by the aggregate model. For instance, from (13) and (15), we find that

$$E[K_t] = (B - n\mu)/ \sum \sigma_i \qquad (22).$$

Suppose we now set the aggregate base stock B so that the expected value of K equals the desired service factor k. Thus, we set

$$B = n\mu + k \sum \sigma_i$$

$$= \{(1+F^2)/2F^2\}\mu + k\sum\sigma_i \qquad (23).$$

This expression gives the total inventory as a function of the production rate flexibility, under the assumption of complete mix flexibility.

We propose to use (19) and (23) as an upper and lower bound on the total inventory required by the production stage with rate flexibility F. These bounds differ according to the mix flexibility: (23) assumes complete flexibility to alter the production mix, whereas (19) assumes virtually no ability to change the production mix. In other words, (19) corresponds to a FCFS processing discipline within the production stage, while (23) corresponds to having no restrictions on the order in which items are processed. Both (19) and (23) assume that the inventory is to provide a guaranteed service level specified by the service factor

k. By subtracting (23) from (19) we obtain

$$k\,\{(1\text{-}F)^2/2F\}\sum\sigma_i\,,$$

which is the maximum inventory reduction from mix flexibility; we might view this as the benefits possible from an expediting activity within the production stage. In Figure 2 we plot (19) and (23) as a function of F for k=2, and $\mu = \sum\sigma_i = 1$. This figure shows the relationship between the required inventory for a production stage and the flexibility of the production stage, both in terms of rate and mix. We see here that the amount of mix flexibility (e.g., expediting) becomes an important determinant of the inventory only when the rate flexibility is low, say $F < .3$ $(n > 6)$.

We discuss next how we might apply this single-stage model to plan safety stocks in a multistage system. We descr'be one way to make this extension and then point out open issues and limitations of this approach.

Extension to Multistage System

A multistage system consists of a network of production stages that produce multiple items, where an item corresponds to a single processing task at a single production stage. An item may be an end item for customer sale, or a component for another item, or both. The interrelationship between items is specified by a "goes-into" matrix $A = \{a_{ij}\}$, where a_{ij} denotes the amount of item i required per unit of item j. That is, associated with each item j is a production stage, call it S_j, and a set of inputs given by $\{$ i: $a_{ij}>0\}$. The production of item j requires that all inputs be available in the right proportions and requires processing by stage S_j. We assume that the units of item j are given in terms of the capacity consumed in processing at stage S_j.

We assume that each production stage operates via a base-stock control system. That is, each period each stage initiates a replenishment to replace the items that have been consumed by external demand. If $\underline{E}_t$ denotes the random vector of external demand in period t, then $(\mathbf{I}-A)^{-1}\underline{E}_t$, where $\mathbf{I}$ is the identity matrix, gives the vector of induced demand. A base-stock control system would set R_{jt}, the release quantity for item j in period t, equal to the induced demand for item j. Thus, for $\underline{R}_t$ being the release vector with elements R_{jt}, we set $\underline{R}_t = (\mathbf{I}-A)^{-1}\underline{E}_t$. We also assume that each stage has a planned lead time determined by its degree of rate flexibility, and operates according to the linear control rule (11).

We take the viewpoint that we plan safety stocks to protect against normal variability, i.e., some maximum reasonable demand. Whenever safety stocks are inadequate, we assume the production system will take the extraordinary actions necessary to cover the shortfall; however, we position and size the safety stocks so that the frequency of this occurrence is acceptable. For each item we specify a service time, as used by Simpson, that is the duration by which any replenishment request will be satisfied. If the service time for an item is zero, then we need to be able to meet any reasonable replenishment request by a customer or by another production stage at the time of the request. If the service time is non-zero and equal to an integer m, then any reasonable replenishment request at time t must be met by time t+m. We equate the planning of safety stocks in a multistage system with setting these service times for all items.

The previous analysis for a single-stage system assumes that the service time for each item is zero, and that all inputs are available with no delay. We show here how to modify this analysis when the service times are non-zero. We present here only the case when the production stage has limited mix flexibility.

Consider an item j that is produced at stage S. Suppose its service time is m and that the maximum service time over the set of input items for j, { i: $a_{ij}>0$}, is m'. Without loss of generality we assume that the service time for all of the items in the input set { i: $a_{ij}>0$} is the same, since there is no value from having shorter service times for a subset of the input items. Let D_{jt} be the induced demand for item j in period t; that is, D_{jt} is the j^{th} element of the vector of induced demand $(\mathbf{I}-A)^{-1}\underline{E}_t$. If we assume that the $\underline{E}_t$ is i.i.d. and has a multivariate normal distribution, then D_{jt} is an i.i.d. normal random variable, say with mean μ_j and variance σ_j^2. In period t, we initiate a release request for item j to replenish the demand D_{jt}; but since the components for j have a service time of m', the components necessary to release D_{jt} become available in time period t+m'. Hence, D_{jt} enters the production stage in time period t+m', and correspondingly, the actual release in period t is the release request from period t-m', namely $D_{j,t-m'}$. Furthermore, since the service time for item j is m, the production stage must supply in period t $D_{j,t-m}$ from its inventory. More specifically, the balance equations for the inventories of item j are

$$W_{jt} = W_{j,t-1} + D_{j,t-m'} - P_{j,t-1} \qquad (24),$$

$$I_{jt} = I_{j,t-1} + P_{j,t-1} - D_{j,t-m} \qquad (25).$$

By combining (24) and (25), and back-substituting, we obtain

$$W_{jt} + I_{jt} = B_j + D_j(t\text{-}m\text{+}1,\ t\text{-}m') \qquad \text{if } m' < m \qquad (26),$$

$$= B_j - D_j(t\text{-}m'\text{+}1,\ t\text{-}m) \qquad \text{if } m' \geq m \qquad (27).$$

$D_j(s, t) = D_{js} + ... + D_{jt}$ for $s \leq t$, and $= 0$ otherwise. B_j is the base stock level.

Now, since we assume limited mix flexibility, we set the production output for j in period t by

$$P_{jt} = W_{jt}/n \qquad (28),$$

where n is the planned lead time for stage S. We assume that the aggregate production for stage S is set by (11), and that n has been determined based on the available slack at the production stage, i.e., by (17). Hence, (28) is consistent with the aggregate production rule.

Assuming an infinite history, we find from (24) and (28) that

$$W_{jt} = \sum (1 - 1/n)^s D_{j,t\text{-}m'\text{-}s} \qquad (29),$$

where the summation is from s=0 to s=∞. From (29) we obtain

$$E[W_{jt}] = n\mu_j \qquad \text{Var}\ [W_{jt}] = n^2\sigma_j^2/(2n\text{-}1) \qquad (30).$$

We can also use (29) in (26) and (27) to write I_{jt} as a weighted average of the demand history starting from time t-m' if m' < m, and from t-m if m' ≥ m. To do this, we need the boundary condition, $W_{j0} + I_{j0} = B_j$. Then we obtain if m' < m

$$E[I_{jt}] = B_j - (n\text{-}m\text{+}m')\mu_j$$

$$\text{Var}\ [I_{jt}] = \sigma_j^2\ [(m\text{-}m')\ \text{-}2n\{1 - (1 - 1/n)^{m\text{-}m'}\} + (n^2/(2n\text{-}1))] \qquad (31),$$

and if m' ≥ m

$$E[I_{jt}] = B_j - (n\text{-}m\text{+}m')\mu_j$$

$$\text{Var}\ [I_{jt}] = \sigma_j^2\ [(m'\text{-}m) + (n^2/(2n\text{-}1))] \qquad (32).$$

To ensure that we cover the maximum reasonable demand, we need set the base stock B_j to achieve a desired service level defined by the probability that I_{jt} is nonnegative. Thus, we set B_j by

$$B_j = (n-m+m')\mu_j + k\sqrt{Var[I_{jt}]}$$

where $Var[I_{jt}]$ is given by (31) or (32), and k is an appropriate service factor. Then the expected intrastage and interstage inventory is given by

$$E[W_{jt}] + E[I_{jt}] = n\mu_j + k\sqrt{Var[I_{jt}]} \qquad (33).$$

Thus, we are able to specify the expected inventory levels for item j as a function of the planned lead time n, and the service times m and m'.

The next step is to determine how to set the service times to minimize the investment in intrastage and interstage inventory. Each item has a service time and each stage has a planned lead time. We can use (33) to determine the expected inventory for each item. Increasing the service time for an item reduces its inventory, but will increase the inventory needed by an assembly or end item that uses it as a component. Hence, we cannot set the service times one at a time, but must consider the entire product structure as given by the goes-into matrix A; in particular, we must incorporate the definition that the replenishment service time for item j (termed m' above) is the maximum of the service times for the components fo j. Furthermore, to the extent that the planned lead times are decision variables, we need consider them simultaneously with the service times. This would be the case if we can modify the flexibility of a production stage either by adding or deleting resources.

This suggests an optimization problem for setting the service times and possibly, the planned lead times. We have not formally explored this optimization problem, but leave it to future research to do so. However, we have discovered that the expected inventory function given by (33) is not concave in the service times. As a consequence, Simpson's result, namely an all-or-nothing policy for the interstage inventory, does not apply here.

DISCUSSION

In this paper we have provided a critical review of the safety stock literature, and have suggested a new approach for modelling safety stocks in manufacturing systems based on this appraisal. One observation from the literature review is that previous research has relied on a paradigm that ignores the role of production flexibility in planning safety stocks. Rather, this paradigm assumes a rigid specification of the behavior of the production system. We present a model which includes consideration of the flexibility of a production stage. A second observation concerns the modelling philosophy that is most appropriate for planning safety stocks. Most of the previous research views safety stocks as the only mechanism available for responding to variability in a manufacturing system. We propose an alternate viewpoint in which we plan safety stocks for protection against normal variability, and assume that the remaining variability is dealt with by other means.

There remain many limitations and questions concerning the model presented here. With regard to the model assumptions we note that we considered only demand uncertainty and only for a stationary demand process without forecasts. We effectively ignore lot sizing by assuming lot-for-lot scheduling, and assume a very specific linear control rule for setting the production output of a production stage. We have specified two models of mix flexibility for a single production stage, but have only been able to show how to extend one version (limited mix flexibility) to a multistage setting. We have outlined an optimization problem for setting safety stocks in a multistage system, but have not explored the algorithmic implications for this problem.

All of these issues deserve further examination. Yet, a more fundamental issue may be to determine the validity and/or appropriateness of the proposed approach. Is the proposed model descriptive of the behavior of any manufacturing system? Does the model need to be descriptive in order for it to be useful in prescribing safety stocks? What is an appropriate role for safety stocks? Discussion of these questions is largely absent from the existing literature. While I wish I had answers to these questions, I can only hope that over time we will couple our modeling efforts with supporting empirical work to resolve these issues.

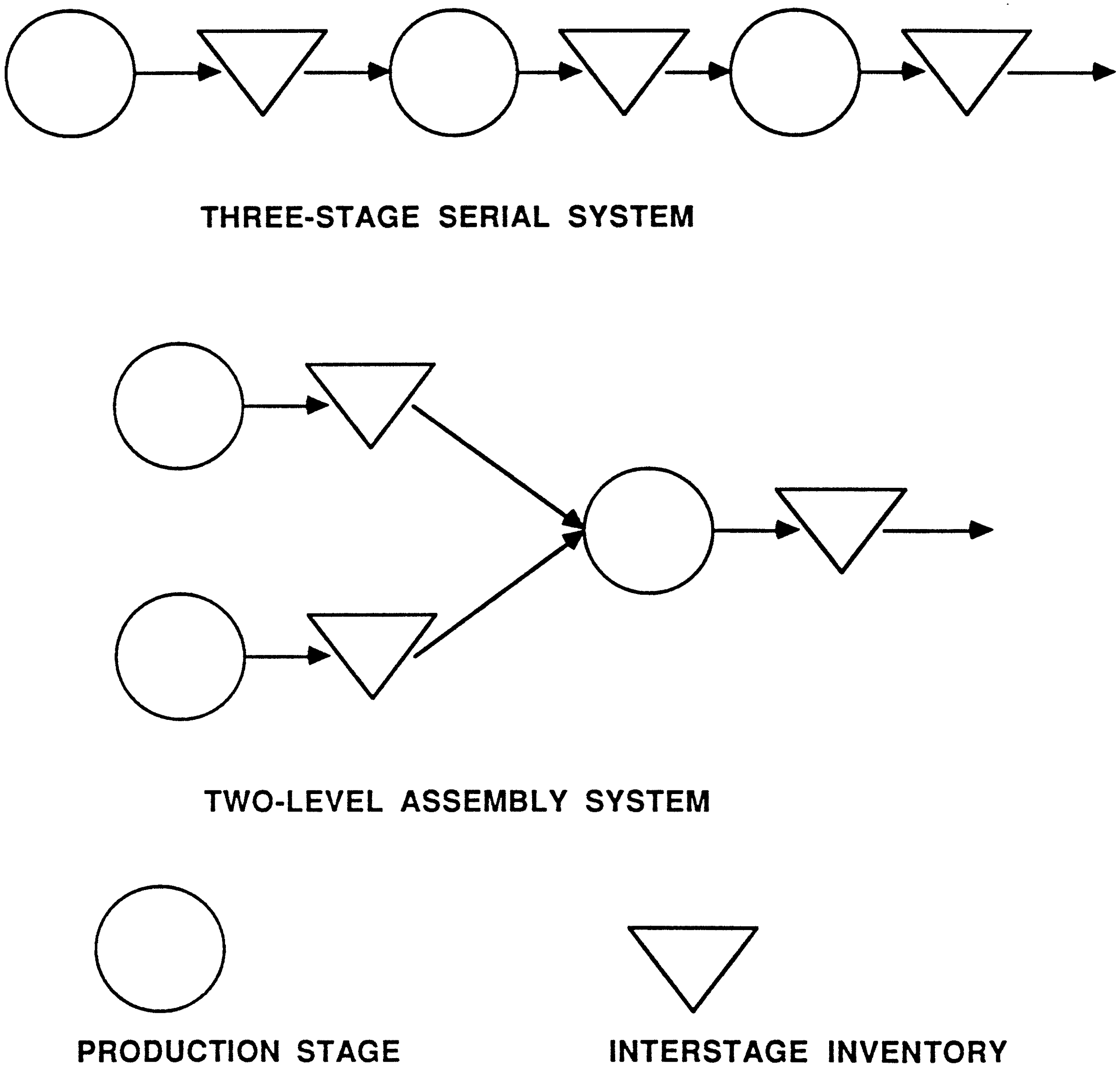
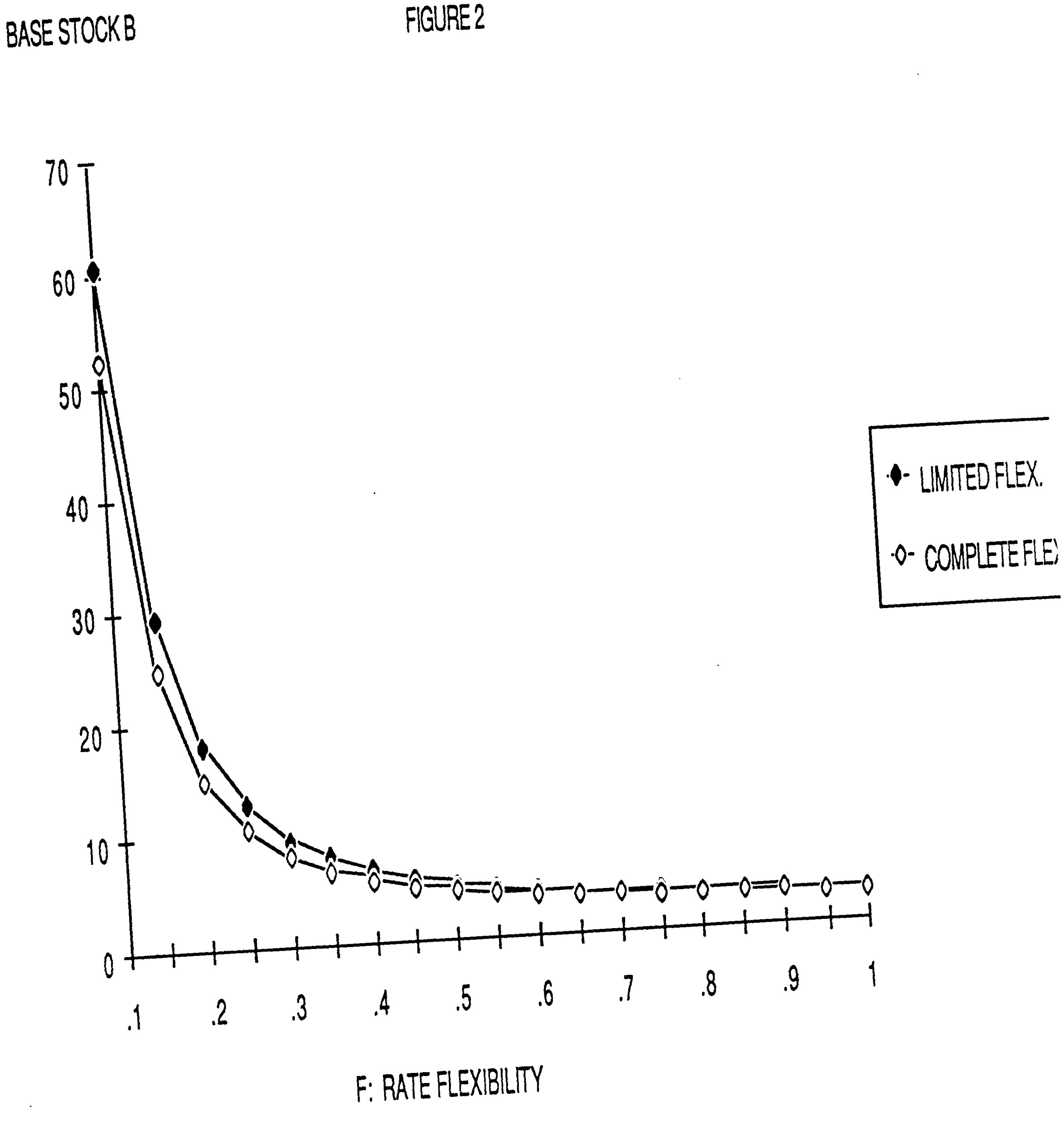

THREE-STAGE SERIAL SYSTEM
TWO-LEVEL ASSEMBLY SYSTEM
PRODUCTION STAGE
INTERSTAGE INVENTORY

FIGURE 1

FIGURE 2


BASE STOCK B
70
60
50
40
30
20
10
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
F: RATE FLEXIBILITY
LIMITED FLEX.
COMPLETE FLE

References

Baker, K. R. (1985), "Safety Stocks and Component Commonality," Journal of Operations Management, Vol. 6, No. 1, pp 13-22.

Baker, K. R., M. J. Magazine and H. L. W. Nuttle (1986), "The Effect of Commonality on Safety Stock in a Simple Inventory Model," Management Science, Vol. 32, No. 8, pp 982-988.

Berry, W. L. and D. C. Whybark (1977), "Buffering against Uncertainty in Material Requirements Planning Systems," Discussion Paper #82, Division of Research, Graduate School of Business, Indiana University, Bloomington Indiana.

Buzacott, J. A. and L. E. Hanifin (1978), "Models of Automatic Transfer Lines with Inventory Banks - A Review and Comparison," AIIE Transactions, Vol. 10, No. 2, pp 197-207.

Carlson, R. C. and C. A. Yano (1986), "Safety Stocks in MRP - Systems with Emergency Setups for Components," Management Science, Vol. 32, No. 4, pp 403-412.

Clark A. J. and H. Scarf (1960), "Optimal Policies for a Multi-Echelon Inventory Problem," Management Science, Vol. 6, No. 4, pp 475-490.

Clark, A. J. and H. Scarf (1962), "Approximate Solutions to a Simple Multi-Echelon Inventory Problem," in Arrow, K. J. et al. (Eds.), Studies in Applied Probability and Management Science, Stanford University Press, Stanford CA, pp 88-100.

Collier, D. A. (1982), "Aggregate Safety Stock Levels and Component Part Commonality," Management Science, Vol. 28, No. 11, pp 1296-1303.

De Bodt, M. A. and S. C. Graves (1985), "Continuous Review Policies for a Multi-Echelon Inventory Problem with Stochastic Demand," Management Science, Vol. 31, No. 10, pp 1286-1299.

Eppen, G. and L. Schrage (1981), "Centralized Ordering Policies in a Multi-Warehouse System with Lead Times and Random Demand," in Schwarz, L. B. (Ed.), Multi-Level Production/Inventory Control Systems: Theory and Practice, TIMS Studies in the Management Sciences, Vol. 16, North-Holland, Amsterdam, pp

51-67.

Gerchak, Y., M. J. Magazine and A. B. Gamble (1986), "Component Commonality with Service Level Requirements," working paper.

Gerchak, Y., and M. Henig (1986), "An Inventory Model with Component Commonality," Operations Research Letters, Vol. 5, No. 3, pp 157-160.

Gershwin, S. B. and I. C. Schick (1983), "Modeling and Analysis of Three-Stage Transfer Lines with Unreliable Machines and Finite Buffers," Operations Research, Vol. 31, No. 2, pp 354-380.

Grasso, E. T. and B. W. Taylor III (1984), "A Simulation-Based Experimental Investigation of Supply/Timing Uncertainty in MRP Systems," International Journal of Production Research, Vol. 22, No. 3, pp 485-497.

Graves, S. C., H. C. Meal, S. Dasu and Y. Qiu (1986), "Two-Stage Production Planning in a Dynamic Environment," in Axsater, S., C. Schneeweiss and E. Silver, (Eds.), Multi-Stage Production Planning and Inventory Control, Lecture Notes in Economics and Mathematical Systems, Vol. 266, Springer-Verlag, Berlin, pp 9-43.

Graves, S. C. (1986), "A Tactical Planning Model for a Job Shop," Operations Research, Vol. 34, No. 4, pp 522-533.

Guerrero, H. H., K. R. Baker and M. H. Southard (1986), "The Dynamics of Hedging the Master Schedule," International Journal of Production Research, Vol. 24, No. 6, pp 1475-1483.

Hanssmann, F. (1959), "Optimal Inventory Location and Control in Production and Distribution Networks," Operations Research, Vol. 7, No. 9, pp 483-498.

Lambrecht, M. R., J. A. Muckstadt, and R. Luyten (1984), "Protective Stocks in Multi-Stage Production Systems," International Journal of Production Research, Vol. 22, No. 6, pp1001-1025.

Lambrecht, M. R., R. Luyten and J. Vander Eecken (1984), "Protective Inventories and Bottlenecks in Production Systems," European Journal of Operational Research, Vol. 22, pp 319-328.

Meal, H. C. (1979), "Safety Stocks in MRP Systems," Technical Report #166, Operations Research Center, MIT, Cambridge MA.

Miller, J. G. (1979), "Hedging the Master Schedule," in L. P. Ritzman et al. (eds.), Disaggregation Problems in Manufacturing and Service Organizations, Martinus Nifhoff, Boston MA, pp 237-256.

New, C. (1975), "Safety Stocks for Requirements Planning," Production & Inventory Management, Vol. 12, No. 2, pp 1-18.

Schmidt, C. P. and S. Nahmias (1985), "Optimal Policy for a Two-Stage Assembly System under Random Demand," Operations Research, Vol. 33, No. 5, pp 1130-1145.

Schmitt, T. G. (1984), "Resolving Uncertainty in Manufacturing Systems," Journal of Operations Management, Vol. 4, No. 4, pp 331-345.

Schwarz, L. B., B. L. Deuermeyer and R. D. Badinelli (1985), "Fill-Rate Optimization in a One-Warehouse, N-Identical Retailer Distribution System," Management Science, Vol. 31, No. 4, pp 488-498.

Silver, E. A. and R. Peterson (1985), Decision Systems for Inventory Management and Production Planning, 2nd edition, John Wiley & Sons, New York.

Simpson, K. F. (1958), "In-Process Inventories," Operations Research, Vol. 6, pp 863-873.

Whybark, C. D. and J. G. Williams (1976), "Material Requirements Under Uncertainty," Decision Sciences, Vol.7, No. 4, pp 595-606.

Wijngaard, J. and J. C. Wortmann (1985), "MRP and Inventories," European Journal of Operational Research, Vol. 20, pp 281-293.

Yano, C. A. and R. C. Carlson (1984), "Buffering against Demand Uncertainty in Material Requirements Planning Systems - Systems with No Emergency Setups," Technical Report 84-13, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI.

Yano, C. A. and R. C. Carlson (1985), "An Analysis of Scheduling Policies in Multiechelon Production Systems," IIE Transactions, Vol. 17, No. 4, pp 370-377.

Yano, C. A. and R. C. Carlson (1987), "Interaction Between Frequency of Rescheduling and the Role of Safety Stock in Material Requirements Planning Systems," International Journal of Production Research, Vol. 25, No. 2, pp 221-232.

Yano, C. A. (1987), "Setting Planned Leadtimes in Serial Production Systems with Tardiness Costs," Management Science, Vol. 33, No. 1, pp 95-106.

APPENDIX

In this section we provide some justification for the linear control rule (11). As might be expected, we can derive a linear control rule from the minimization of a quadratic cost function. In particular, consider the following dynamic minimization problem:

$$\text{MIN } E\{\sum \beta^t[(P_t - \mu)^2 + \alpha(W_{t+1} - n\mu)^2]\}, \tag{A1}$$

subject to $W_t = W_{t-1} + D_t - P_{t-1}$.

The summation runs from t = 0 to t = T, β is a discount factor, and α is a relative (positive) cost factor. D_t is demand in period t, and is assumed to be an i.i.d. random variable with mean μ and variance σ^2. The problem is to minimize the expected cost, where at the start of each time period we know W_t and must set P_t.

The cost function in each time period consists of a production smoothing cost and an inventory-related cost. The production smoothing cost is proportional to the squared deviation of the production variable from its mean. The inventory-related cost is proportional to the squared deviation of the in-process inventory from its mean, where we have preset $n\mu$ as the target in-process inventory. This corresponds to a planned lead time of n periods. The objective function is then the discounted sum of these cost terms over the relevant time interval [0, T].

We can solve this minimization problem by dynamic programming. The general form of the optimal policy is given by

$$P_t = \mu + a_t(W_t - n\mu) \tag{A2}$$

where $a_T = \alpha/(1+\alpha)$, and $a_t = (\alpha + \beta a_{t+1})/ (1 + \alpha + \beta a_{t+1})$. For $\beta < 1$, the parameter a_t is less than 1 for all t, and converges to a constant, call it a, as T increases to infinity. In this case, we can use the inventory balance equation to rewrite (A2) as a simple smoothing equation:

$$P_t = aD_t + (1-a)P_{t-1} \qquad \text{(A3).}$$

(A3) is the same as (12), where a replaces 1/n. Thus, the control rule given by (11) [or equivalently (12)] is a special instance of the solution to (A1) where a=1/n. Indeed, we obtain the equivalent solution if the parameters α and β are such that

$$\alpha = 1/(n-1) - \beta/n \qquad \text{(A4).}$$

In this case a_t converges to 1/n and (A2) is the same as (11).

For general problem parameters, however, the optimal solution to (A1) is a linear control rule given by (A2) that will differ from (11). Nevertheless, the qualitative behavior of the production and inventory random variables remains essentially the same as that derived from the specific instance given by (11).