Scheduling and Ordering Production Policies in a Limited Capacity Manufacturing System: The Multiple Replenishment Products Case

Share Your Story

Scheduling and Ordering Production Policies in a Limited Capacity Manufacturing System: The Multiple Replenishment Products Case

Soundouss Bouhia and Frederick H. Abernathy
Division of Engineering and Applied Sciences, Harvard University

July 10, 2003

Abstract

This paper formulates production policies to maximize the performance of an apparel manufacturing system that replenishes basic items characterized by a flat average demand. The simulation-based model compares a number of production strategies and chooses an ordering and scheduling policy that increases the overall performance of the supply chain. In doing so, different supplier configurations (i.e., different capacity resources and cost structures) are compared and the optimal strategy is selected to maximize the supplier's profit while maintaining a high order fill rate and minimizing the risk of carrying inventory. The system's performance is then compared to an upper limit, calculated using Linear Programming.
keywords: Ordering and Scheduling production; Heuristic Policies, Replenishment items;

1 Introduction

Before lean retailing, retailers placed orders with manufacturers many months in advance of the planned selling season; factory lead times[1] weren't important. Manufacturers organized the work in their factories to minimize the direct labor cost. Nowadays, with lean retailing and rapid replenishment a manufacturer of replenishment products wants to have just sufficient finished goods to satisfy the uncertain time varying demand but not so much that the inventory carrying costs diminish the profit [3].

Lean retailers require that their suppliers compete not only on price, but also on replenishment speed, flexibility and other services. The quick replenishment practices required by lean retailers were first adopted for basic apparel products and now are being introduced for fashion-basic products. Currently, very few fashion items are replenished on this basis.

These requirements have forced the channel players to adopt new strategies to cope with the lean retailing innovation. On one hand, the retailers must develop the capability to gather and analyze available information and to incorporate the results into internal forecasting, planning and decision-making processes. On the other hand, manufacturers supplying lean retailers with replenished items have to provide increased product variety and a high order fill rates. Two extreme strategies for the manufacturer to meet retailer's requirements would be:

1. Holding high levels of finished goods inventory to meet retailers' demands. This is not a sustainable policy in the long term as more retailers move toward lean retailing and the increased risks and costs associated with holding inventory, product proliferation and shorter product life cycles.

2. Alter its internal design, planning, procurement, manufacturing and distribution operations in order to respond rapidly to demand changes without increasing its own exposure to inventory risk.

In our model, manufacturers can better manage their own risk by explicitly taking into account different uncertainties introduced by lean retailing. In particular, we consider the uncertainty in demand

---

[1] The lead time is defined as the time interval between getting the product authorization to produce an item and receiving the finished good in the distribution center.

at the stock keeping unit (SKU) level. The model estimates the extent to which a manufacturer's finished goods inventory must increase as a function of greater demand uncertainty and higher order fulfillment rate(OFR)[2]requirements. This rate is a good proxy for the retailer's satisfaction with deliveries. Most items sold at retail are offered in a variety of sizes and colors for a given style. The pattern of weekly demand for each of these SKUs can be calculated and characterized by a coefficient of variation (CV) defined as the ratio of standard deviation of demand to the average demand. The CV of each SKU is one of the primary factors that determines the level of finished goods inventory a manufacturer needs to carry to achieve a given OFR. A SKU with a high CV (for which the weekly variation of demand departs much more from the average demand) requires relatively a larger number of weeks of average sales in finished goods inventory compared to a SKU with a lower CV. Empirical results and probability theory tell us that low sales volume SKUs will tend to have higher CVs than high sales volume SKUs.

2 Literature and Assumptions

When demand is deterministic and time varying, two main approaches have been considered [9]:

1. When variability of the demand pattern is very low, use the basic Economic Order Quantity (EOQ) based on the average demand rate.

2. Use the Wagner-Whitin dynamic lot sizing algorithm or one of its extension (e.g. the Silver-Meal algorithm) to solve the deterministic, time-varying demand problem. The Wagner-Whitin dynamic lot sizing model in [4] is considered to be the standard method to solve the deterministic demand models.

These methods assume: 1) a setup cost that determines the level of the EOQ and 2) no capacity constraint. The problem becomes more complex when a realistic production capacity constraint is introduced. Even for a single-item problem, the solution for the limited capacity case is complex and the lack of an analytical solution justifies the use of heuristic procedures to solve the problem. When demand is

random, a frequently used policy to determine the ordering level at each time period is the base-stock policy. This policy consists of building inventory up to a target level or safety stock level to deal with the demand variability. Extensive decision rules have been developed for finding the safety stock levels. Tayur and Glasserman [8] examine the stability of this policy in a limited production capacity model: "the system is stable if, on average, it can produce finished goods at a greater rate than they are demanded".

Apparel manufacturers, in our model, can better manage their own risk by explicitly taking into account different uncertainties introduced by lean retailing. In particular, the uncertainty in demand at the SKU level is the main driver of the manufacturer's inventory policy. The model estimates the extent to which a manufacturer's finished goods inventory must increase as a function of greater demand uncertainty and higher OFR requirements.

Some flexibility of the production capacity over time is a major requirement given the stochastic nature of the demand. The higher the upper bound of the production capacity, the higher the order fill rate for all SKUs and the higher the profit. This assumption is feasible when the production facility is a large plant that produces many similar styles and has the capability of stretching or reducing the capacity by reallocating the workers to different tasks.

3 Problem Formulation

In this problem, the supply chain is defined by a manufacturer having many different customers and a retailer having to make ordering decisions to satisfy customers' demand requirements on multiple SKUs. It is assumed that the demand is never backlogged[3]. The dynamics of the system are governed by the products' aggregate time-varying demand of their retailers and the production evolution from the raw materials to the finished goods in the distribution center. We refer to the system as SOPS (Scheduling and Ordering Production System).

Let there be P production lines in the system and each production line, identified by subscript p has 3 characteristics: the lead-time, LT_p; the production cost for SKU k, $PC_p(k)$ [4] and the produc-

---

[2]The order fill rate in the model is defined as being the fraction of customer demand for a particular SKU that can be immediately filled (i.e. from inventory) during a specified time period; Some call this a service level.

[3]The lost sales assumption is common in apparel industry. The customer could walk in to a store not find his size, and walk out without the product.

[4]includes the assembly cost, the labor cost, the transportation cost from the raw material suppliers to the plant and from the plant to the Distribution center and any tariffs, insurances or duty if applicable

tion capacity limitations, U_p^{min} and U_p^{max} [5]. There are K number of SKUs indexed by k, selling at price SP_k and an annual inventory carrying cost, ICC_k, expressed in annual percentage of inventory costs. The demand of each SKU is characterized by its mean μ_k and its coefficient of variation CV_k. Each time period is indexed by, t. T is the time horizon over which the profit is calculated. Δt is the time unit (e.g. week) for orders, demand etc...

$d_k(t)$: Demand at time t for each SKU k, defined on the interval $[t, t + \Delta t]$.

$x_k(t)$: Finished goods inventory level for each SKU k, at time t, at the end of the interval $[t, t + \Delta t]$.

$u_{p,k}(t)$: Total number of units of each SKU k, scheduled for production during $[t - \Delta t, t]$ in plant p.

$\gamma = [\gamma_p]$: Vector of dimension P of the proportion of total production allocated to each plant p, such that $\sum\limits_{m=1}^{P} \gamma_p = 1$.

The objective function is the expected profit over the projected time horizon:

$$E\{\text{Profit}\} = E\{\text{Revenue} - \text{Total Costs}\}$$

$$\begin{aligned} \text{Profit} \quad = \quad & \sum_t \sum_k P_k * \min[d_k(t), x_k(t)] \\ & - \sum_{t,k,p} \text{PC}_p(k) * u_{p,k}(t) \\ & - \sum_{t,p} \text{ICC}_k * [x_k(t) + \sum_p \sum_{s=1}^{L_p} u_{p,k}(t-s)], \end{aligned}$$

The constraints consist of the inventory dynamics and the capacity limits for each plant:

$$x_k(t+1) = \max[0, x_k(t) - d_k(t)] + \sum_p u_{p,k}(t - \text{LT}_p)$$

and

$$U_p^{\min}(t) \leq \sum_k u_{m,k}(t) \leq U_p^{\max}(t)$$

4 Solution Methodology

4.1 Ordinal Optimization

Simulation is one of the most broadly used techniques to analyze complex stochastic systems. When looking for the optimum or the best scenario that maximizes the system's performance, analytical solutions rarely exist. They exist primarily for simple system models based on the assumption that the objective function is smooth enough to use derivatives. The calculus-based methods to solve such problems include gradient and trajectory of steepest descent ("fall line") . These techniques, also referred to, metaphorically, as "skiing downhill in a fog" [5] are based on the improvement of the global performance of the system depending on its current stage.

Due to the presence of uncertainty, most real world problems are not solvable using the classical optimization techniques, unless strong assumptions are made to simplify the formulation of the problem. Most human-made systems are characterized by common criteria: a lack of structure, great uncertainty and a large search space that makes the problem NP-hard [6]. Analytical tools have a limited applicability due to the complexity of the systems. This complexity forces the goal of finding the absolute optimum to be softened to a goal of finding a solution that has a high probability of being good enough. This is the essence of Ordinal Optimization [6]. When the goal is softened, we can tolerate imprecise performance estimates, since a "good enough" solution can be obtained with high confidence from a selected set. Ordinal Optimization is a search-based method that matches the "selected subset" with the "good enough" subset. A "good enough" subset is predetermined. This method is used in SOPS to find the values of the parameters that maximize the system performance.

4.2 Simulation Based Solution and Target Inventory

The SOPS simulations mimic actual plant operations and model the non-deterministic demand. The SKUs are aggregated into clusters with common coefficient of variation (CV) to simplify the

---

[5]These values could be expressed as a percentage of the nominal production capacity or as absolute values of maximum and minimum units of production. If the nominal production capacity is 1000 units per week, then the capacity limitation could be expressed as 80% to 120% of that value or as a minimum production of 800 and a maximum production of 1200 units.

[6]A problem is NP-hard if an algorithm for solving it can be translated into one for solving any other NP-problem (Nondeterministic Polynomial time)

computation [7]. This decreases the search space dimension and improves the simulations' running time. The SKUs demand in a cluster are uncorrelated. During each simulated week, random demand levels are generated for each SKU. The demand is generated from a two-parameter probability distribution[7]. The two parameters are used to fit the two parameters characterizing the demand (mean and CV). We consider a make-to-stock manufacturing system in which inventory is managed through a base-stock policy which is called, in this study, a "target policy". The program finds the target policy that produces the best result using Ordinal Optimization. Operating profit is calculated weekly from sales income less production, material and inventory carrying costs.

The target inventory method is similar to the traditional approach in inventory problems, also known as the sS policy, where the supplier has an inventory policy of returning the SKU stock level to a target level S each time the inventory level of that SKU drops to a minimum level s. In this case, the manufacturer is trying to keep a certain target level for its inventory subject to plant capacity constraints.

4.3   Ordering and Scheduling Policies

In the multiple products, multiple plants production model, production cycle times (or leadtimes) of the plants range from LT_{min} to LT_{max}. Each plant has a limited production capacity. At each time period, an *ordering policy* determines the number of units of each SKU to be produced. At each time period, for each plant a *scheduling procedure* determines the number of units of each SKU to produce in each plant, consistent with production capacity constraints. Given the complexity of the system, there is no universal optimal ordering or scheduling policy that maximizes the manufacturer's profit under all demand conditions. Therefore, the heuristics approach is the only way to arrive at reasonable answers to the production problem. The heuristics used to define the scheduling and ordering policies depend on the demand and replenishment patterns.

In apparel manufacturing, three categories for demand patterns can be distinguished: 1. Replenishment items (basics) with a flat mean demand through out the selling season; 2. Fashionable items that have a short life cycle; 3. Replenishment items that have seasonal peaks in demand. For the first category, the scheduling and ordering policies are critical. The use of a short cycle production line can improve the profit and the inventory level[8]. As an illustration of this feature, see the HBR article [1] and the simulations at the end of this paper. For the second case, the manufacturer fixes a schedule ahead of the selling season to meet the forecast of demand by building up inventory in the least expensive long cycle plants[9]. The last two cases are treated separately in another study.

5   Different Ordering Policies

5.1   Classic Base Stock Policy or Lookback Policy

One ordering policy, called a *lookback* policy, is the classic base stock (order-up-to) level policy that determines at each time period the total inventory level position (Work In Process (WIP) + finished goods), compares it to the target total inventory level at that time and schedules the difference, if any.

5.2   Planning for the Near Future or the Lookforward Policy

Another ordering policy, the *lookforward* policy, is used in conjunction with a scheduling policy, projecting the finished goods inventory at LT_{max} units of time in the future before determining the ordering amount to be scheduled in the plants. The real demand per SKU is approximated[10] by the mean demand for that SKU. This requires some estimates of the order placed between the present time and LT_{max} units of time in the future. The difference between the finished goods inventory and the target inventory level for the given LT_{max} units of time determines the anticipated amount for each SKU to be scheduled in each plant.

The focus of this paper is the development of a multi-plant, multi-product algorithm to determine

---

[7]The density of the distribution takes on shapes similar to the gamma densities, intersects the origin and behave like a Gaussian towards infinity.

[8]This can be true even when the short cycle plants' production cost are higher than those of the long cycle time plant.

[9]The long cycle time is a bundle system where each worker is most efficient at one and only one operation - Taylorism

[10]One might better approximate the inventory level some time in the future by estimating the actual demand by the mean times the expected order service level. This order service level refinement does not appear to significantly change the results.

the profit. Each SKU has a nondeterministic demand pattern and can be produced in all plants. The objective function is the total profit; the maximization variables are the target inventory levels for each SKU. Each plant has a lead time, production capacity limitations and a cost structure. There is no uncertainty on the delivery time. The variation on lead time could be added easily into the algorithm by replacing LT_p by the realized LT_p^r.

We adopt the following definitions:

τ_k : Target inventory (order-up-to) level of finished goods for each SKU k.

$WIP_k^{\tau_k}(t)$: Work in process for SKU k at time t; does not include finished goods.

$FI_k^{\tau_k}(t)$: Finished goods inventory for SKU k at the beginning of time t.

$DL_{k,p}^{\tau_k}(t)$: Delivery from plant p of SKU k at the beginning of time t.

$O_{k,p}^{\tau_k}(t)$: The number of units of SKU k order at the start of period t to arrive at $t + LT_p$.

The evolution of the inventory system is described by the following equations:

$$FI_k^{\tau_k}(t) = \sum_p DL_{k,p}^{\tau_k}(t - LT_p) + \\ \text{Max}[0, FI_k^{\tau_k}(t-1) - D_k(t-1)],$$

and

$$O_{k,p}^{\tau_k}(t) = \text{Max}[0, \tau_k - FI_k^{\tau_k}(t)]. \qquad (1)$$

where

$$DL_{k,p}^{\tau_k}(t) \leq \sum_p O_{k,p}^{\tau_k}(t - LT_p), \qquad (2)$$

and

$$WIP_k^{\tau_k}(t) \leq \sum_p \sum_{i=1}^{LT_p - 1} O_{k,p}^{\tau_k}(t - i). \qquad (3)$$

The level of $WIP_k^{\tau_k}(t)$ depends on the scheduling policy: how much to order of each SKU in each plant when there is a limited production capacity in each plant.

6 Different Scheduling Policies

6.1 Scheduling Procedures

Each production plant is characterized not only by its cost structure (materials, production and transportation to the Distribution Center) and lead-time but also by its *scheduling procedure*. The long lead-time plants always have some flexibility over some minimum/maximum range. In a multiple plants, multiple products case, a collection of these scheduling procedures is called a "*scheduling policy*". We distinguish two main scheduling procedures: the "*HiCvFirst*" and "proportional" procedure. We assume that the order to be scheduled at a certain period, t in plant p for SKU k is expressed in units as $S_k^p(t)$. This amount is determined using an ordering policy.

In the "*HiCvFirst*" procedure, if the total orders[11] across all SKUs is between the minimum and maximum capacity, then all the orders are started. If the total order across all SKUs is higher than maximum capacity, then the order is scheduled, starting with the SKUs from the highest CV clusters to the lowest CV clusters until maximum capacity is reached. If the total orders across SKUs is lower than the minimum capacity, then the scheduler begins all orders and loads the plant with a scale factor until minimum capacity is reached. This scale factor for one SKU is equal to the ratio of the target level of that SKU to the total target levels of all SKUs. In this last case the following amount is scheduled for SKU, k:

$$u_{p,k}(t) = S_k^p(t) + \frac{\tau_k}{\sum_k \tau_k}(U_p^{min} - \sum_k S_k^p(t))$$

In the "*proportional*" procedure, the plant is loaded pro rata for each SKU. The proportional factor for a SKU is equal to the ratio of the mean demand of that SKU to the total average demand of all SKUs. If the total of the orders is between the minimum and maximum production capacity of the plant, i.e., $U_p^{\min} \leq \sum_k S_k^p \leq U_p^{\max}$, then for each SKu k, S_k^p units are scheduled in plant p:

$$u_{p,k}(t) = S_k^p(t)$$

At time t, if the total of orders is greater than the maximum capacity for that plant, i.e.,

---

[11]The total orders, equal to the amount needed to be scheduled at each time period for each SKU, is determined using an ordering policy.

$\sum_k S_k^p > U_p^{\max}$, then the scheduler starts in plant p the following amount for each SKU:

$$u_{p,k}(t) = \frac{S_k^p(t)}{\sum_k S_k^p(t)} * U_p^{\max}$$

At time t, if the total orders is less than the minimum capacity for that plant, i.e., $\sum_k S_k^p < U_p^{\min}$ then the scheduler in plant p begins production of the following amount for each SKU:

$$u_{p,k}(t) = S_k^p(t) + \frac{\mu_k}{\sum_k \mu_k}(U_p^{\min} - \sum_k S_k^p(t))$$

6.2 Scheduling Policies

One of the scheduling policies that has been used extensively is the *"HiCvFirst"* policy. This policy consists of loading all the plants, starting with the quickest cycle plant, using the *"HiCvFirst"* procedure for each plant. A second scheduling policy is a *"Proportional"* policy which loads all the plants, starting with the quickest cycle plant, using the *"Propportional"* procedure for each plant. A third policy is called a *"mixed"* policy that loads the quick cycle plants using the *"HiCvFirst"* procedure and the slow cycle plants using *"Propportional"* procedure. It also loads the quick cycle plants first. Intuitively, this procedure is appropriate when the shortest lead-time plant doesn't have much capacity flexibility given the highly skilled workers needed. Note that all these policies load the quickest lines to their maximum capacity.

According to [10], if an overall task is broken down into a series of separate tasks and a worker trained to efficiently and quickly do one task, then the workers productivity would be higher than if they have several tasks to do. In the apparel industry, this leads to the progressive bundle system. Not all tasks require the same time, and therefore, in an apparel factory a typical item might require 20 to 30 assembly and finishing operations. With some operations being much shorter than others, some tasks will require 3 or 4 operators to keep up with the production of a single operator doing a single task. These requirements for the balance of the production line leads to a plant minimum of 100 to 200 days worth of WIP between operators. This production system is called the progressive bundle system.

6.3 Linear Programming and the Profit Upper Limit

One can generate a deterministic demand situation by first generating the time varying demand, for each SKU, over the time horizon T. This demand is treated as a known demand. The profit generated from this pseudo-deterministic demand situation is a good benchmark for the profit calculated using the SOPS and constitutes an absolute upper limit. This maximum is calculated using a linear programming framework [2]. The objective function is the profit over the simulation time horizon, defined as the revenue minus the costs of production and inventory. Linear programming determines at each time period the production allocation for each SKU that maximizes the overall profit under the maximum and minimum production capacity constraints.

To assess the performance of the scheduling policies used by SOPS, the average profit[12] is compared to the profits generated by the deterministic demand situation. The profit can then be statistically compared using the profit standard deviation to estimate the performance of the ordering and scheduling policies.

7 Simulation Results

7.1 Ordering Policies Comparison

In a first attempt to compare the ordering policies, we considered the single SKU case to factor out the scheduling policy impact. The input parameters are inspired from the apparel industry. The inventory carrying cost is set to 18% and the plant capacity limitations are flexible enough not to interfere with the system performance. The two ordering policies are statistically equivalent for all CVs values and for all the plants' lead-times. This demonstrates that the two ordering policies are permutable when there is no forecasting algorithm involved. If there is a forecasting algorithm, the *look-Forward* policy seems to be advantageous.

7.2 Scheduling Policies Comparison

For comparing the scheduling policies, we have used 20 SKUs in a two plants setting. The first plant has a lead-time of 11 weeks and the second plant has a lead time of 2 weeks and has a higher

---

[12]the average is taken over a large number of simulation runs. This number of simulation is chosen such that incrementing this number changes the average profit by less than some predetermined small epsilon, for example 0.01%

production cost. All the SKUs are clustered into 3 groups: a high, a medium and a low CV group. The set of conditions for the input parameters are extracted from the apparel industry and mimic the values used in [1]. These values can be changed to accommodate other products. As it would be expected intuitively, the total inventory level (see [1]), decreases when the production capacity allocated to the quick line increases. The inventory is in the order of 6 weeks when all the items are produced in the quick plant and in the order of 16 weeks when all the items are produced in the slow plant. This emphasizes the idea of decreasing the cost and risk of carrying inventory by using quick production lines. The graph below shows the profit curve for the "proportional" and *"mixed"* scheduling policies, using the "lookforward" ordering policy. The profit of the two scheduling policies differ by less than one standard deviation of the profit. The highest curve in the graph is the upper bound of the average profit. The curve right below it is the average upper limit decreased by one standard deviation. The next curve is the *"mixed"* policy profit increased by one standard deviation. The two scheduling policies performance approaches the upper limit. We run the simulations for 4000 simulated years in order to achieve a stable profit. For this multiple products, two plant case, the SOPS's simulation and optimization running times[13] are respectively 30 seconds and 12 hours. This time decreases dramatically with the machine performance.

8 Conclusions

In this paper, we defined a list of ordering and scheduling policies in a multiple products, multiple plants case. This list is not exhaustive and other policies can be developed. However, we demonstrated that the scheduling policies' performances are within one standard deviation from one another and they approach the upper bound benchmark, extracted from the pseudo-deterministic case.
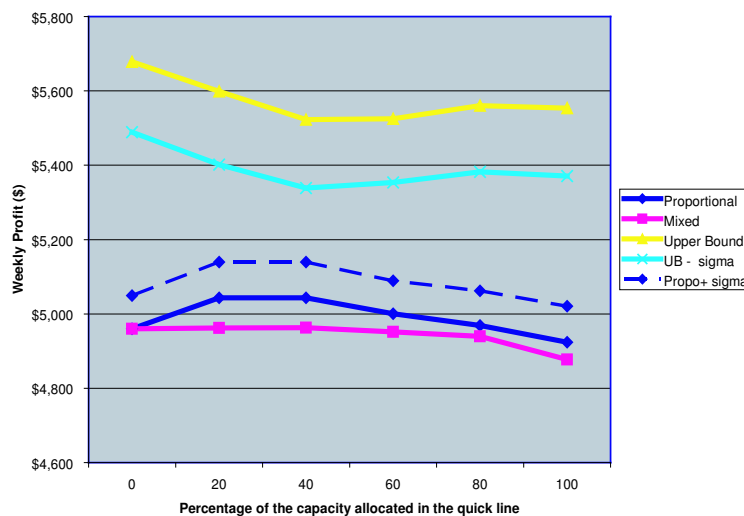
9 Acknowledgments

Figure 1: Comparing the scheduling policies to the profit's upper bound when the production capacity proportion in the quick production line, γ, varies from 0% to 100%.

---

[13]Based on a 1.6GHz machine running without disk calls

References

[1] F.H. Abernathy, J.T. Dunlop, J. Hammond, and D. Weil. Control your inventory in a world of lean retailing. *Harvard Business Review*, pages 169–176, November- December 2000.

[2] D.P. Bertsekas. *Dynamic Programming And Optimal Control, Vol.1-2.* Athena Scientific, Belmont, Massachussetts, 1995.

[3] F.H. Abernathy J.T. Dunlop, J. Hammond, and D. Weil. *A Stitch In Time: Lean Retailing and the Transformation of Manufacturing, Lessons from the Apparel and Textile Industries* . Oxford University Press, 1999.

[4] T.M. Whithin H.M. Wagner. Dynanmic version of the economic lot size model. *Management Science*, 5:89–96, 1958.

[5] Y.C. Ho. Heuristics rules of thumb and the 80/20 proposition. *Trans. on Automatic Control*, 39(5):1025–1027, May 1994.

[6] Y.C. Ho, L.H. Lee, and E.K. Lau. Explanation of goal softening in ordinal optimization. *IEEE Trans. on Automatic Control*, 44(1), January 1999.

[7] L.H. Lee. *Ordinal Optimization and Its Application in Apparel Manufacturing Systems.* PhD thesis, Harvard University, Cambridge, MA, May 1997.

[8] S. Tayur P. Glasserman. The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Operations Research*, 42(5):913–925, September-October 1994.

[9] E.A. Silver, D.F. Pyke, and R. Peterson. *Inventory Management and Production Planning and Scheduling.* John Wiley and Sons, 1998.

[10] F.W. Taylor. *The Principles of Scientific Management.* W. W. Norton and Comapany, 1911.