2005

Manufacturing lead-time rules: Customer retention versus tardiness costs

Susan A. Slotnick, *Cleveland State University*
Matthew J. Sobel, *Case Western Reserve University*

Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 163 (2005) 825–856
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Production, Manufacturing and Logistics

Manufacturing lead-time rules: Customer retention versus tardiness costs

Susan A. Slotnick [a], Matthew J. Sobel [b,*]

[a] *Department of Operations Management and Business Statistics, James J. Nance College of Business Administration, Cleveland State University, 2121 Euclid Avenue, Cleveland, OH 44115, USA*
[b] *Department of Operations, Weatherhead School of Management, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7235, USA*

Accepted 30 July 2003
Available online 25 January 2004

Abstract

Inaccurate production backlog information is a major cause of late deliveries, which can result in penalty fees and loss of reputation. We identify conditions when it is particularly worthwhile to improve an information system to provide good lead-time information. We first analyze a sequential decision process model of lead-time decisions at a firm which manufactures standard products to order, and has complete backlog information. There are Poisson arrivals, stochastic processing times, customers may balk in response to quoted delivery dates, and revenues are offset by tardiness penalties. We characterize an optimal policy and show how to accelerate computations. The second part of the paper is a computational comparison of this optimum (with full backlog information) with a lead-time quotation rule that is optimal with statistical shop-status information. This reveals when the partial-information method does well and when it is worth implementing measures to improve information transfer between operations and sales.

Keywords: Dynamic programming; Manufacturing; Markov decision processes; Due-date assignment; Marketing

1. Introduction

1.1. Background

As manufacturing firms work to increase the efficiency of their supply chains and reduce inventories, the importance of on-time delivery influences their efforts (Ansberry, 2002). Along with stringent quality standards, manufacturers are demanding that their suppliers meet contracted delivery dates or face substantial penalties. In the aerospace industry, for example, tardiness penalties as high as one million

* Corresponding author.
E-mail addresses: s.slotnick@csuohio.edu (S.A. Slotnick), mjs13@po.cwru.edu (M.J. Sobel).

0377-2217/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2003.07.023

dollars per day may be imposed on subcontractors of aircraft components (Stansbury, 2000). In order to avoid such penalties, along with the possibility of losing the business of important customers, supplier firms are under pressure to quote delivery dates that are attainable, as well as attractive to current and potential customers. While promises of short lead times may bring in business, it can be extremely embarrassing for a firm if it cannot deliver what it has advertised. For example, Allied Signal Plastics announced a 48-hour delivery time for some of its popular resins; it then had to rescind the promise within days of the announcement when it found that it could not live up to it, because of manufacturing and inventory constraints (Freeman, 1996). A major reason for inaccurate lead-time quotation is lack of information on manufacturing capacity (Wein, 1991; Bartholomew, 1996; APICS, 1991). Manufacturers are investing in ERP and other types of software systems that facilitate accurate delivery promises by using real-time production information in order to estimate lead times (Dwyer, 2000; Yeager, 1997; Leachman et al., 1996). Thus the calculation of accurate lead times is a prerequisite to the successful employment of a delivery-time guarantee (So and Song, 1998; Palaka et al., 1998; Chatterjee et al., 2002).

An important tradeoff in the assignment of lead times is that a relatively short lead-time quotation may attract and retain customers, but makes tardiness more likely, often with adverse effects on short- and long-term business. Longer lead-time quotations, on the other hand, are easier to fulfill, but customers may prefer competitors who promise shorter turnaround times. A good lead-time policy, then, is one that balances these two considerations to enhance profit. A sales department that has accurate information about the status of the factory is in a better position to make good lead-time decisions than one that has incomplete or dated information. How much better will those decisions be with improved information, that is, when sales has an accurate estimate of the current status and capabilities of the manufacturing facility? What is the value of manufacturing software that provides this information, when compared to the adverse effects of inaccurate lead-time decisions?

In order to answer these questions, we develop a model that maximizes the firm's profit, when sales has complete information about the manufacturing backlog. We compare the performance of a procedure based on this model with a previously developed method based on partial information. The larger the difference between the two profit figures, the more that it is worth to the firm to expend resources to enhance the information flow between the two departments.

1.2. Overview

We model the decision-making process of a firm that manufactures standard products to order. Such a firm may be using response time and customer service level to compete in a market in which products are standardized, but customers choose their vendors partially on the basis of delivery time (for example, circuit boards (Shapiro, 1988), lighting fixtures (Weng, 1999), floppy disks and CD's (Andel, 2002), printing (Boyaci and Ray, 2003), and other products (Rajagopalan, 2002)). The sales department solicits orders, customers present jobs with known processing requirements, and the sales department proposes contractual terms which the customer either accepts or takes her business elsewhere.

We restrict attention to processing requirements which vary only in processing time, and contractual terms which vary only in delivery date. If the job remains, it then proceeds to the factory. The net revenue from a job depends on its processing time (reflecting common pricing practices, as well as the concept that revenue should be related to the cost of resources that the firm expends in processing and related opportunity costs (for example, Johansen, 1991, 1994), and the firm pays a penalty if the job is delivered later than was contracted. The *lead-time quotation* is the difference between the quoted delivery time and the processing time, i.e., the number that sales adds to the processing time to come up with a delivery date. The *actual lead time* is the difference between the actual delivery time and the processing time. The *delivery time* is the time from when the order is placed until it is delivered to the customer.

The information which is available to salespeople who negotiate terms with customers should affect the modeling of due-date quotation. When a salesperson prepares a due-date quotation for a prospective job at a manufacturer of standard products to order, we assert that the routinely available information includes estimates of (a) the job's utilization of factory capacity, and (b) the aggregate claim on factory capacity of jobs that have been booked but not yet completed. Here, we summarize (a) with job processing time and (b) with the factory backlog in units of processing time. Our assertion is consistent with manufacturing control software (Palmatier and Shull, 1989), MRP and its extensions MRP II and ERP (Krajewski and Ritzman, 2002; Migliorelli and Swan, 1988), industry organizations, for example aerospace (Aerospace Industries Association, 2002), and government reporting requirements for companies in North America, Europe and Asia (Bureau of Census, 2002, 2000; Statistics Canada, 2002; Statisches Bundesamt Deutschland, 2002; Czech Statistical Office, 2002; Statistics Bureau and Statistics Center (Japan), 2002). The research related to due-date quotation is reviewed in Section 2 which notes that previous work either makes no use of (a) and (b) or uses a surrogate such as the number of jobs whose aggregate processing time comprises the backlog (Duenyas, 1995; Duenyas and Hopp, 1995). However, information on job processing times and factory backlogs is routinely available, and performance of the lead-time policy is degraded by ignoring this information and then inferring it.

In Sections 3 and 4 we analyze a model of due-date quotation in which the sales department of a manufacturer of standard products has current accurate information on factory conditions and on the job at hand. Customers arrive according to a Poisson process, bringing jobs whose processing times are independent and identically distributed positive random variables. The sales department learns a customer's job processing time upon arrival, and quotes a due date equal to a lead time plus processing time. The customer balks with a probability that depends on the lead-time quotation and, perhaps, job processing time. This reflects the situation where a customer with a larger order may be more sensitive to delivery times, and so is likely to get more favorable terms (Carter, 1993) (note that only Proposition 5 uses this assumption). If the customer permits the job to stay, the job goes to the factory which processes jobs on a first-come first-served basis.

Although the resulting sequential decision model is a semi-Markov decision process, its features permit the application of methods based on discrete-time Markov decision processes. The analytical results in Sections 3 and 4 use the criteria of expected present value of net profit and long-run average net profit per unit time. In Section 3.2 the structures of the revenue term, penalty cost, and balking probability are quite general. We characterize an optimal policy by analyzing the structure of the value function of an associated dynamic program. The characterizations yield managerial insights and would accelerate computations in embellished models which are tailored to particular applications. These are the primary results:

- An optimal due-date quotation balances the net profit of the job at hand with the increased tardiness penalties that acceptance of that job is likely to cause in the future (Proposition 1).
- When revenue is credited and tardiness penalties are charged at the time that a job goes to the factory (reflecting, for example, discounts to customers willing to take later deliveries (Cheung, 1998; Elimam and Dodin, 2001)), a higher backlog is only a detriment because it is likely to cause higher tardiness penalties in the future (Proposition 3). This valuable computational property would be lacking in the equivalent model in which revenues and penalties are posted when jobs complete their processing. This is a major difference between this paper and previous work.
- The change in the value function as processing time increases is bounded above by the rate at which the immediate revenue increases with processing time (Proposition 5). Because this bound involves essentially no computation, it is very useful for excluding suboptimal lead-time quotations in an algorithm.

- When a job's net revenue is proportional to its processing time, the tardiness penalty is proportional to lateness (the amount by which a job's completion date exceeds the due-date quotation), and the balking probability is an exponential function of the lead-time quotation, the optimal lead-time quotation increases with larger backlogs and decreases with longer processing times (Proposition 7).
- There are easily computed bounds on the maximal expected present value of net profits offset by tardiness penalties (Proposition 9). These bounds are useful for excluding suboptimal lead-time quotations in an algorithm.
- Most results are valid for a more general model which has multiple classes of customers and the extent to which maximizing the *gain rate*, i.e., the long-run average net profit per unit time, yields results that are analogous to those for the criterion of maximal expected present value of net profits.

A numerical study in Section 5 compares the profitability and lead-time quotations associated with the rule that is optimal with accurate current backlog information with those of the log-linear rule presented in a previous paper (Chatterjee et al., 2002). These results clarify the circumstances in which it is particularly advisable for the firm to have an information system that provides sales with accurate timely backlog information.

Section 6 presents the conclusions.

2. Related work

Actual lead times depend on the volume and attributes of jobs that the factory processes, on the sequence with which jobs are processed at work centers in the factory, and on other factors (availability of materials, accuracy and timeliness of information, reliability of mechanical devices, etc.). Since sales determines the volume and attributes of jobs arriving at the factory and current due-date quotations have a major impact on actual lead times in the future, due-date quotation is a complex sequential decision problem. All of the research on this problem makes draconian reductions of this complexity in order to achieve any results at all.

A major reduction of complexity is achieved by assuming that at the outset all job arrival times and processing times are known with certainty and the objective is minimization of costs. Under these assumptions, there is no risk of customers balking due to lead-time quotations, so these models are not directly useful for the negotiation of terms with prospective customers of manufacturers of standard products. Since the models do not yield simple heuristics which have been tested in a more complex setting, these models have not yet been shown to be indirectly useful either. The resulting scheduling literature focuses on selecting due dates, and sometimes, on selecting the sequence in which jobs should be processed. We note here that the deterministic literature concludes that using both job characteristics *and* shop status information such as congestion results in better due-date decisions (Baker and Bertrand, 1981; Baker, 1984; Ragatz and Mabert, 1984; Weeks, 1979). Cheng and Gupta (1989) survey this literature. For discussions of more recent work, including stochastic models, see Lawrence (1995), Moodie (1999), Moodie and Bobrowski (1999) and Chatterjee et al. (2002). In the remainder of this section we discuss those articles that are most relevant to the present paper.

Another reduction of complexity occurs by focusing on lead-time quotation while acknowledging imperfections in forecasts of actual arrival and processing times. That research makes uncertainty explicit but most of it suppresses the issue of job sequencing in the factory. The current paper follows this route and the remainder of this section discusses articles that are fellow travelers.

Nothing is known about the current customer in Duenyas (1995) except that the sales department knows the *type* of customer and it has up-to-date information on the number of jobs in the factory. It does not,

however, know the backlog or the numbers of each type of job that comprise the backlog. Sales does not know a current customer's processing time and the processing time distribution is the same for all types. However, customer type determines profitability and sensitivity to lead-time quotation. The number of jobs in the factory is used to infer the conditional distribution of the backlog. So the argument in Section 1.2 implies that the resulting policies are typically suboptimal because more precise backlog information is available than only the number of jobs in the backlog.

In Chatterjee et al. (2002), sales knows the current customer's processing time, profitability, and sensitivity to lead-time quotation, and the actual lead times are distributed as the queueing time in M/M/1. This leads to customer-specific quotations that are sketched in Section 1.2 and described in detail in Section 5.

The remainder of this section reviews papers whose models do not utilize customer-specific information at the time that a due-date quotation is made. Although they do not model due-date quotation in the manufacture of standard products, they employ methods that may be useful for that purpose. Consider a firm that uses experimental technologies to manufacture exotic prototype products. The sales department might not be able to estimate any differences among the processing times of successive job prospects. Therefore, in a model of such a setting, it may be appropriate to model processing times as independent random variables whose values are not revealed until actual processing is completed. This is a characteristic of the model in Duenyas and Hopp (1995) which also assumes that sales does not know the factory backlog but does know the number of jobs that comprise that backlog. The model includes revenue unrelated to processing time, a proportional tardiness penalty, and a balking probability that depends on the due-date quotation.

Palaka et al. (1998), So and Song (1998) and Weng (1999) assume that an exogenous probability distribution describes waiting time or time in the system, but nothing is known about the current customer's processing time, profitability or sensitivity to lead-time quotation. So the delivery-time quotation given to a customer is independent of the characteristics of the customer or the job. Palaka et al. (1998) and So and Song (1998) model system time (processing time plus waiting time) with the queueing time in M/M/1, and Weng (1999) models it with a phase-type distribution. Ogden and Turner (1996) present a decision-support model with empirically estimated parameters that aims to maximize customer satisfaction with delivery promise and performance; uncertainty of early, on-time or late delivery is modeled with probability distributions.

ElHafsi (2000) considers the problem of specifying a lead time for a lot and splitting the processing of the lot among several processing centers that are subject to failure. Keskinocak et al. (2001) study two deterministic models for coordinating scheduling with lead-time quotation.

There is related work which combines studies of delivery-time decisions with inventory and pricing. Boyaci and Ray (2003) consider two substitutable products which differ only in price and delivery time; the shorter delivery time, which attracts time-sensitive customers, bears a premium price. Here the guaranteed delivery times are based on industry standards, and the waiting time in the facility (corresponding to our lead time) is determined by the assumptions of Poisson arrivals, exponential service times and first-come–first-served processing. In two novel models (Li, 1992; Lederer and Li, 1997), firms make neither lead-time decisions nor delivery-time quotations. However, their production rate decisions affect delivery time delays and, therefore, influence the demand encountered by firms.

There is a sizable literature on input control that considers balking and reneging from the firm's point of view, and also from the customer's point of view (Stidham, 1985). Several papers deal with loss of customers due to waiting time. Boots and Tijms (1999) present exact and approximate formulas for the loss probability in a multiserver queueing system in which a customer may leave if service does not begin by a certain time. Whitt (1999) examines the benefits of supplying customers with information about waiting times when there is a fixed-capacity waiting room. Balking and reneging of customers is compared in systems with and without communication of anticipated delays.

3. Optimal due-date quotation

3.1. The model

Successive potential customers arrive at sales according to a Poisson process with intensity λ, with jobs whose processing times $S_1, S_2, \ldots$ are independent and identically distributed nonnegative random variables. When customer k arrives, sales learns S_k and quotes a due date $S_k + L_k$, where L_k is the lead time. The customer stays with probability $a(S_k, L_k)$ and departs with probability $1 - a(S_k, L_k)$ (which we interpret as a balking probability). Let B_k denote the backlog when customer k arrives. Throughout Section 3, we assume that sales knows B_k when it quotes L_k.

Let s, b, and L denote a generic processing time, backlog, and due-date quotation, respectively. The notation $\tau(s,b)z(b-L)$ unifies the treatment of fixed and proportional tardiness penalties. A proportional penalty corresponds to a unit cost $\tau(s,b)$ and $z(b-L) = (b-L)^+$ (where $(x)^+$ denotes $\max\{x,0\}$). That is, the penalty is zero if $L \geqslant b$ and it is $\tau(s,b)(b-L)$ if $L < b$. A fixed penalty corresponds to $z(b-L) = 1$ if $L < b$ and $z(b-L) = 0$ if $L \geqslant b$. That is, the penalty is zero if $L \geqslant b$ and it is $\tau(s,b)$ if $L < b$.

If the due date quotation results in customer k staying, the firm immediately books revenue $r(S_k, B_k)$ and tardiness penalty $\tau(S_k, B_k)z(B_k - L_k)$, the backlog grows to $B_k + S_k$, and the factory completes customer k's job at time $B_k + S_k$. So we assume that the factory processes first-in–first-out. This is consistent with many studies of manufacturing delays using first-come–first-served sequencing. See, for example, Palaka et al., 1998; Weng, 1999; Chatterjee et al., 2002; Boyaci and Ray, 2003, and the literature reviews therein. If customer k leaves, the backlog remains at B_k. By interpreting the revenue and tardiness penalty as expected present values, this notation encompasses various assumptions concerning the actual timing of revenues and penalties. We give examples of the general notation after the model is completely specified. Let $\alpha > 0$ be the continuous-time discount factor.

Generally, we assume for each $s \geqslant 0$ that $a(s,\cdot)$ is continuous on $[0,\infty)$ and $a(s,L) \to 0$ as $L \to \infty$. So long lead-time quotations make it unlikely that a customer will stay. Only in Proposition 4 do we also assume that customers do not balk if their lead-time quotation is zero, i.e., $a(s,0) = 1$ (for all s). We assume that r, τ, and z are nonnegative and that $r(\cdot,\cdot)$ and $\tau(\cdot,\cdot)$ are continuous on their domain, $z(\cdot)$ is continuous on $(0,\infty)$ and $z(x) = 0$ if $x < 0$, $r(s,0) \geqslant r(s,b)$ for all $s \geqslant 0$ and $b \geqslant 0$, and $r(0,b) = 0$ for all b.

Examples

If the balking probability is $1 - e^{-\xi L}$, so $a(s,L) = e^{-\xi L}$, then ξ parameterizes a customer's sensitivity to due-date quotations. For fixed L, a customer is more likely to leave if ξ is higher. In Section 3.3, we analyze the following combination of exponential balking, linear revenue, and proportional tardiness penalty:

$$a(s,L) = e^{-\xi L}, \quad r(s,b) = \pi s, \quad \tau(\cdot,\cdot) \equiv 1. \tag{1}$$

Thus, π represents the ratio of the unit net profit to the unit tardiness cost.

Another example of the notation $a(s,L)$ arises when a customer stays with probability $e^{-\xi L}$ but ξ varies among customers and is not known when a customer arrives. Suppose that the ξ's of successive customers are independent random variables with the distribution function $H_s(\cdot)$ if their processing time is s. Then the general model is appropriate with

$$a(s,L) = \int_0^\infty e^{-cL}\, dH_s(c).$$

If customers who bring bigger jobs are more impatient, i.e. if $s \leqslant s'$ implies $H_s(g) \leqslant H_{s'}(g)$ for all g, then this manifestation of $a(s,L)$ satisfies all the assumptions made in Section 3.2. Although it is standard practice for purchasing agents to insist on more favorable terms for larger jobs (Carter, 1993), only Proposition 5 depends on the assumption that customers with bigger jobs are more impatient.

An example where $r(s,b)$ depends on b is $r(s,b) = \pi s e^{-\alpha(s+b)}$ which is the present value of a revenue πs that is received upon completion of processing. Similarly, an example where $\tau(s,b)$ depends on b is $\tau(s,b) = e^{-\alpha(s+b)}$ which is the present value of the unit cost of tardiness that is rebated when the job is finished. An example where $a(s,L)$ depends on s is $a(s,L) = e^{-\xi_s L}$, where customer impatience grows with processing time if ξ_s increases with s.

The model corresponds to a semi-Markov decision process in which the state is (s,b) and the action is L. Let $f(s,b)$ be the maximal expected present value of the infinite-horizon time stream of revenues offset by penalties when the initial state is (s,b). We say that a policy, i.e. a decision rule for choosing lead times, is *optimal* if its expected present value is $f(s,b)$ when (s,b) is the initial state. Let $f_n(s,b)$ be the maximal expected present value if the current customer has processing time s, the backlog is b, and the process ends after $n-1$ further customers arrive. We use $\{f_n(\cdot,\cdot)\}$ to approximate $f(\cdot,\cdot)$ because the latter inherits essential properties of the former. The $\{f_n(\cdot,\cdot)\}$ satisfy the following recursion with $f_0(\cdot,\cdot) \equiv 0$:

$$f_n(s,b) = \sup_{L \geqslant 0} \{[1 - a(s,L)] f_{n-1}(0,b) + a(s,L)[r(s,b) - \tau(s,b) z(b-L) + f_{n-1}(0, s+b)]\}, \tag{2}$$

$$f_n(0,b) = \int_0^b \lambda e^{-\lambda u} e^{-\alpha u} E[f_{n-1}(S, b-u)]\, du + \int_b^\infty \lambda e^{-\lambda u} e^{-\alpha u} E[f_{n-1}(S,0)]\, du. \tag{3}$$

The maximization in (2) refers to a job that has just arrived and on which a lead time must be quoted. The first term in the maximand of (2), $[1 - a(s,L)] f_{n-1}(0,b)$, refers to the outcome when the customer balks. It is the product of the probability of balking and the value of the process at the resulting state. That state would be $(0,b)$ because the backlog would remain at b, while no job that invites a due-date quotation would be at hand. The second term in the maximand refers to the outcome when the customer stays. Then the firm receives revenue minus tardiness penalty and the backlog rises to $s+b$.

The expected present value is given by (3) when a customer has just balked and the state is $(0,b)$. The first term (second term) refers to the event that the next customer arrives before (after) the backlog is eliminated.

3.2. The value of an optimal due-date quotation

The infinite-horizon dynamic program that corresponds to (2) and (3) is (a) a vehicle for characterizing an optimal policy for quoting due dates, and (b) a means of calculating an optimal policy. In this subsection we elicit properties of $f(\cdot,\cdot)$ and $f_n(\cdot,\cdot)$ that are used later to characterize optimal due-date quotations and to accelerate algorithms. In spite of the general structure of the revenue function and tardiness penalty, the problem has an intuitive tradeoff based on a limited range of optimization (Proposition 1), the infinite-horizon decision problem has a well-defined value function linked to an optimal policy (Proposition 2), f is a nonincreasing function of the backlog, b (Proposition 3), and there are bounds on f and its gradients (Propositions 4 and 5).

The unbounded range of due dates in (2) is an obstacle to achieving both (a) and (b); the first result is intuitive and justifies restricting L to $[0,b]$ rather than $[0,\infty)$ as in (2). So either a job should be rejected (i.e., $L = \infty$) or the lead time should be restricted to $[0,b]$. There is no reason to assign a lead time that is longer than the current backlog, unless it is unprofitable to process the job at all (in which case a very long due-date quotation is used to "reject" the customer). Let Y be an exponential random variable with mean $(\lambda + \alpha)^{-1}$. Also, let

$$K_n(b) = E[f_n(S,b)] \quad (b \geqslant 0). \tag{4}$$

Proposition 1

$$f_n(s,b) = f_{n-1}(0,b) + \left[\max_{0\leqslant L\leqslant b}\{a(s,L)[r(s,b) - \tau(s,b)z(b-L) + f_{n-1}(0,s+b) - f_{n-1}(0,b)]\}\right]^+, \tag{5}$$

$$(\lambda+\alpha)f_n(0,b) = \lambda E(f_{n-1}[S,(b-Y)^+]) = \lambda E(K_{n-1}[(b-Y)^+]). \tag{6}$$

In particular,

$$(\lambda+\alpha)f_n(0,0) = \lambda E f_{n-1}[(S,0)]. \tag{7}$$

All proofs can be found in Appendix A.

The brackets in (5) enclose the basic tradeoff in due-date quotation. If the job moves to the factory, the immediate increment to net profit is $r(s,b) - \tau(s,b)z(b-L)$. However, raising the backlog from b to $s+b$ will reduce the subsequent expected present value from $f_{n-1}(0,b)$ to $f_{n-1}(0,s+b)$. The reduction is due to higher tardiness penalties caused by the higher backlog and the possible rejection of future customers who would otherwise have been worthwhile.

The next result justifies the attention paid to f_n. It asserts that $f_n(s,b)$ has a limit f as $n\to\infty$ which is the value function of an optimal due date policy, $f(\cdot,\cdot)$ satisfies a functional equation, and the value function identifies an optimal due-date policy.

Let $L = \mathscr{A}$ denote the rejection of a customer (i.e., $L=\infty$),

$$J(s,b,L) = r(s,b) - \tau(s,b)z(b-L) + f(0,s+b) - f(0,b)$$

and let $\Phi(s,b) = \{\mathscr{A}\}$ if $J(s,b,L) < 0$ for all $L\in[0,b]$; otherwise, $\Phi(s,b) = \{L\in[0,b] : f(s,b) = f(0,b) + a(s,L)J(s,b,L)\}$.

Proposition 2. *For each $s\geqslant 0$ and $b\geqslant 0$,*

$$f(s,b) = \lim_{n\to\infty} f_n(s,b), \tag{8}$$

$$f(s,b) = f(0,b) + \left[\max_{0\leqslant L\leqslant b}\{a(s,L)J(s,b,L)\}\right]^+, \tag{9}$$

$$(\lambda+\alpha)f(0,b) = \lambda E(f[S,(b-Y)^+]) = \lambda E(K[(b-Y)^+]). \tag{10}$$

In particular,

$$(\lambda+\alpha)f(0,0) = \lambda E[f(S,0)]. \tag{11}$$

Let $\delta(s,b)\in\Phi(s,b)$ for each (s,b). Then δ is an optimal lead-time quotation policy.

Lemma 1. *For each n and (s,b),*

$$f_n(s,b) \leqslant f_{n+1}(s,b) \leqslant r(s,b) + \frac{\lambda E[r(S,0)]}{\alpha}. \tag{12}$$

We note that (12) implies the existence of the limit in (8), but without an additional argument such as in Schäl (1975), it does not follow that the limit is the value of an optimal policy or that the limit satisfies (9).

The following result specifies that the value function (of an optimal due-date policy) diminishes as the shop backlog increases. This property is intuitive because the revenue and cost of an arriving job are posted

at the time of its arrival. Thereafter, its presence increases the backlog which can only force the payment of a tardiness penalty that might have been avoided or the rejection of a job that would otherwise have been worth accepting. The result is based on the assumption that as the backlog increases, the revenue brought by a job is more likely to be offset by the penalty it may incur (if revenue does not rise with service time, but penalty does). So profit decreases as backlog increases.

Proposition 3. *If $z(\cdot)$ is nondecreasing and, for all s, $r(s,\cdot)$ is nonincreasing and $\tau(s,\cdot)$ is nondecreasing, then for each n, $f_n(s,\cdot)$ is nonincreasing. So $f(s,\cdot)$ is nonincreasing for each s.*

Bounds

Bounds on value functions and their gradients accelerate computations and characterize optimal policies (Veinott, 1966; MacQueen, 1967; Sobel, 1971; Lovejoy, 1986). The next result provides general bounds on the gradients of the value function. Later bounds depend on the form of the revenue, penalty and balking probability functions. We use the following notation for partial derivatives: $r^{(1)}(s,b) = \partial r(s,b)/\partial s$, $f^{(1)}(s,b) = \partial f(s,b)/\partial s$, $f^{(2)}(s,b) = \partial f(s,b)/\partial b$, etc. The assumption added to the following result is that customers do not balk if their lead-time quotation is zero.

Proposition 4. *Under the assumptions of Proposition 3, if $a(s,0) = 1$ (for all s) then*

$$e^{-\alpha(s+x)} f(0,b-x) \leqslant f(s,b) \quad (0 \leqslant x \leqslant b), \tag{13}$$

$$\frac{f(0,b) - f(0,s+b)}{1 - e^{-\alpha s}} \leqslant f(0,b), \tag{14}$$

$$-\lambda E[r(S,0)] \leqslant -\alpha f(0,b) \leqslant f^{(2)}(0,b) \leqslant 0. \tag{15}$$

As an example of the usefulness of Proposition 4, one can accelerate a successive approximations algorithm by exploiting (13) and (14) as follows. Begin with $b = 0$ and use $f(0,b+1) \geqslant e^{-\alpha} f(0,b)$ and $f(s,b) \geqslant e^{-\alpha s} f(0,b)$ for each s.

The next result suggests an interdependence between the rates at which the value function changes as backlog grows and as processing time grows. The additional restriction is a proportional tardiness penalty and the added assumptions are that customers who bring jobs with larger processing times are more impatient and insist on higher performance penalty rates but their jobs contribute more net revenue. The consequence is that $f(s,b)$ grows with s no faster than the rate at which revenue grows with s. A sharper bound requires more computation.

Proposition 5. *If $z(u) = (u)^+$, $\tau(\cdot,b)$ and $r(\cdot,b)$ are nondecreasing (for all b), and $a(\cdot,L)$ is nonincreasing (for all L), then*

$$f^{(1)}(s,b) \leqslant [r^{(1)}(s,b) + f^{(2)}(0,s+b)]^+ \leqslant r^{(1)}(s,b). \tag{16}$$

3.3. Linear revenue, exponential balking and proportional tardiness penalty

Restrictions of the model lead to further results. For example, one might expect that unrestrictive assumptions would imply that an optimal due-date quotation is a nondecreasing function of the backlog and

a nonincreasing function of job processing time. The benefits of making the following structural assumptions includes these results, other characterizations of an optimal due-date quotation, and properties that accelerate calculating an optimal quotation. Throughout this subsection, structure (1) is valid. So the penalty is proportional to tardiness (a common practice in manufacturing; for example, Shapiro et al., 1992; Stansbury, 2000; Pinedo, 1995), revenue is linear (the prevalent pricing policy in industry; Johansen, 1991), the balking probability is exponential, and the unit tardiness penalty is constant: $z(b-L)=(b-L)^+$, $r(s,b)=\pi s,\ a(s,L)=\mathrm{e}^{-\xi L}$, and $\tau(s,b)$ is the same constant for all (s,b). Without loss of generality let that constant be unity. These assumptions yield a further characterization of an optimal policy and additional bounds on the value function and its gradients. So henceforth we let $\tau\equiv 1$ and interpret π as a ratio of revenue to unit tardiness penalty.

It is apparent from Proposition 2 and (9) that an optimal policy corresponds to partitioning the set of $\{(s,b)\}$ into four regions: the set where jobs are rejected ($L=\mathscr{A}$); the set where jobs are quoted zero lead times ($L=0$); the set where jobs are quoted maximal lead times ($L=b$); and the set where $0<L<b$. The next result characterizes these regions using this notation:

$$J(s,b,L)=\pi s-b+L-f(0,b)+f(0,s+b), \tag{17}$$

$$\mathscr{L}(s,b)=\frac{1}{\xi}-J(s,b,0)=\frac{1}{\xi}-\pi s+b+f(0,b)-f(0,b+s). \tag{18}$$

So (9) corresponds to

$$f(s,b)=f(0,b)+\left[\max_{0\leqslant L\leqslant b}\{\mathrm{e}^{-\xi L}[J(s,b,0)+L]\}\right]^+. \tag{19}$$

Proposition 6. *$L=\mathscr{A}$ is optimal if $J(s,b,0)+\min\{[\mathscr{L}(s,b)]^+,b\}<0$. Otherwise, an optimal selection of L is the following, where $\mathscr{L}(s,b)\geqslant 1/\xi-\pi s+b$:*

$$\left.\begin{array}{ll} L=0 & \text{if } \frac{1}{\xi}<J(s,b,0) \\ L=\mathscr{L}(s,b) & \text{if } \frac{1}{\xi}-b\leqslant J(s,b,0)\leqslant\frac{1}{\xi} \\ L=b & \text{if } J(s,b,0)<\frac{1}{\xi}-b \end{array}\right\}. \tag{20}$$

The lower bound $\mathscr{L}(s,b)\geqslant 1/\xi-\pi s+b$ has the same flavor as bounds in Duenyas and Hopp (1995, Theorem 4) and Duenyas (1995, Theorem 1) but entails much less computation (and is affected by the current backlog and the processing time of the prospective customer).

It follows from (20) that $L=\mathscr{L}(s,b)$ is optimal if $0<L<b$ is optimal. Then the gradients of L are the gradients of an optimal lead-time quotation. The next result bounds these gradients. Let $\mu^{-1}=E(S)$ be the population mean processing time and let $\rho=\lambda/\mu$. Notice that $\rho>1$ is possible because the input to the shop is a filtration of the jobs seen by sales.

Lemma 2

$$-\pi\rho\leqslant f^{(2)}(0,b)\leqslant\mathscr{L}^{(2)}(s,b)-1\leqslant -f^{(2)}(0,s+b)\leqslant\pi\rho, \tag{21}$$

$$0\leqslant\mathscr{L}^{(1)}(s,b)+\pi\leqslant\pi\rho. \tag{22}$$

The following result asserts that if ρ is not too high then: bigger backlogs generate longer lead-time quotations, larger processing times result in shorter quotations, and customers are never rejected (which simplifies (9)).

Proposition 7. *If $\rho \leqslant 1/\pi$ then $\mathscr{L}(s,\cdot)$ is nondecreasing (for all s). If $\rho \leqslant 1$ then $\mathscr{L}(\cdot,b)$ is nonincreasing (for all b) and customers are never rejected, i.e.,*

$$f(s,b) = f(0,b) + \max_{0 \leqslant L \leqslant b} \{e^{-\xi L}[J(s,b,0) + L]\}.$$

Further Bounds

Linear revenue and exponential balking lead to further bounds on the value function and its gradients. As usual, looser bounds are easier to calculate.

Proposition 8

$$\pi(1-\rho)e^{-\xi b} \leqslant e^{-\xi b}[\pi - \alpha f(0,b)]^+ \leqslant f^{(1)}(0,b) = e^{-\xi b}[\pi + f^{(2)}(0,b)]^+ \leqslant \pi e^{-\xi b}, \tag{23}$$

$$f^{(1)}(s,b) \leqslant \pi, \tag{24}$$

$$f^{(2)}(0,b) \leqslant -[\pi - e^{\xi b} f^{(1)}(0,b)]^+. \tag{25}$$

The following bounds are explicit, and therefore easy to calculate, i.e. the bounds do not depend on values of the value function or its gradients. Let $\phi(\alpha) = E(e^{-\alpha S})$ be the characteristic function of the processing time and let $D = (\lambda[1-\phi(\alpha)] + \alpha)^{-1}$.

Proposition 9

$$\pi s - b + \pi\rho D e^{-\alpha(s+b)} \leqslant f(s,b) \leqslant \pi\left(s + \frac{\rho}{\alpha}\right), \tag{26}$$

$$\pi\{s + \rho D e^{-\alpha s} + [\rho D(1 - e^{-\alpha s}) - s]^+\} \leqslant f(s,0). \tag{27}$$

The next result is an upper bound on $f(s,b)$ (hence on $f_n(s,b)$ due to Lemma 1) that accelerates the solution of recursive equations (5) and (6) because it is not apparent that either (26) or the following upper bound is uniformly sharper than the other.

Proposition 10

$$f(s,b) \leqslant f(0,b) + [\pi s + f(0,s+b) - f(0,b)]^+. \tag{28}$$

We already know that $f(s,\cdot)$ is nonincreasing for each s, $\mathscr{L}(s,\cdot)$ is nondecreasing (for each s) if $\rho \leqslant 1$, and $\mathscr{L}(\cdot,b)$ is nonincreasing (for each b) if $\rho \leqslant 1$. Under the assumptions in this subsection, we would find the following additional properties intuitive: (i) $f(\cdot,b)$ is convex (for each b); (ii) $\mathscr{L}(\cdot,b)$ is nonincreasing (for each b) regardless of the value of ρ; and (iii) $f(\cdot,b)$ is nondecreasing (for each b) regardless of the value of ρ. However, Figs. 1 and 2 (taken from the computational study described in Section 5) are counterexamples to the convexity of $f(\cdot,\cdot)$ in either argument. We shall see that the absence of convexity casts doubt on the general validity of (ii) and (iii).

Proposition 11. *On the set of $\{(s,b)\}$ where $L = \mathscr{L}(s,b)$ is optimal,* (i) *implies* (iii), *and* (ii) *and* (iii) *imply each other.*

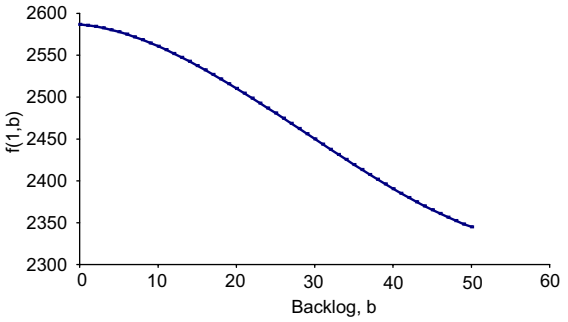

Fig. 1. $f(1,b)$.

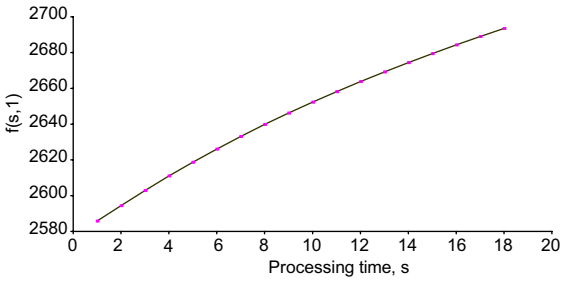

Fig. 2. $f(s,1)$.

4. Extensions

4.1. Heterogeneous customer classes

Large customers, regardless of the processing times of their jobs, tend to insist on better contract terms than their smaller counterparts (Carter, 1993). For example, a firm that does custom metal fabrication may take orders from small businesses as well as cater to a major automobile manufacturer. So two customers with the same processing time may have characteristics that are known by the sales department and which yield different profitability and balking probability functions. Even if the backlog is the same when quotes are made to both of them, it may be appropriate to give them different due-date quotations.

Suppose that the model in Section 3.1 is augmented with multiple *classes* of customers. For each class k of customers, $k = 1, \ldots, K$, let $r_k(s,b)$ be the expected value of the net revenue, not including tardiness penalty, from a job with processing time s brought by a type k customer at a moment when the backlog is b. As in Duenyas (1995), we assume that sales knows a customer's class at the moment of arrival, and the classes of customers arrive according to independent Poisson processes with intensities $\lambda_1, \ldots, \lambda_K$. However, we assume that sales also knows the processing time of the current prospective job and the current factory backlog. It is convenient to choose the unit of time so that $\sum_{k=1}^{K} \lambda_k = 1$. We assume for each k that the processing times of type k customers are independent and identically distributed with the same distribution as the generic random variable S_k.

Straightforward generalizations of the results in Sections 3.2 and 3.3 remain valid when the state (s,b) is replaced with (s,b,k) to include the class of customer who is about to receive a due-date quotation. We shall not list those results, but the obvious variants of the assumptions in Section 3 lead to extensions of Propositions 1–11.

A comparison of due-date quotations given to customers from different classes has no parallel in Section 3.3. The following result, similar in spirit to Theorem 1 in Duenyas (1995), gives conditions under which shorter lead-time quotations should be given to customers who are more profitable and more impatient. The model with linear revenue, exponential balking and proportional tardiness penalties has $r_k(s,b) = \pi_k s$ and $a_k(s,L) = 1 - e^{\xi_k L}$ for each k, s, b and L. The analogue of (18) plays the same role as in Proposition 6: for each s, b and k, let $\mathscr{L}_k(s,b) = 1/\xi_k - J(s,b,k,0) \geqslant 1/\xi_k - \pi_k s + b$.

Proposition 12. *If $\xi_k \geqslant \xi_i$ and $\pi_k \geqslant \pi_i$ then*

$$\mathscr{L}_k(s,b) \leqslant \mathscr{L}_i(s,b) \quad (s \geqslant 0, b \geqslant 0).$$

4.2. Average profit per unit time

The long-run average net profit per unit time is an important criterion in practice and in research, whereas Section 3 utilizes the discounted criterion. So this subsection sketches (a) why the model with the former criterion inherits some important properties of the model with the latter, and (b) obstacles to complete inheritance. The extant theory for a discrete-time MDP (Markov decision process) is stronger with the discounted criterion than with the average-profit criterion. To confirm the strong connection between the two criteria as the single-period discount factor tends to unity, it is generally necessary to make assumptions beyond those needed to establish that the discounted model is well behaved.

This difficulty is compounded with a semi-Markov decision process (SMDP for short) which is a generalization of an MDP where the transition time between successive states may be a random variable whose distribution depends on the current state, the action taken, and the next state. Although there are strong connections between MDP's and SMDP's (cf. Heyman and Sobel, 1984, Sections 5.1 and 5.2), the analysis of the connection between the two criteria for SMDP's encounters the same obstacles as for MDP's plus the complexity that transition times can depend on the actions taken. That is, letting the continuous-time discount factor (α) tend to zero in a discounted SMDP does not necessarily yield the correct average-reward SMDP. However, the SMDP in the previous sections has an enviable property: the probability distribution of transition times is the same for all pairs of states and actions, namely an exponential distribution with mean λ^{-1}. This property permits the use of MDP methods in Sections 3.2, 3.3, 4.1 and in the current subsection.

We assume that there is a large finite number b^*, such that the firm rejects any job whose processing time would raise the backlog above b^*. If $s + b > b^*$ the only feasible decision in state (s,b) is $L = \mathscr{A}$ (i.e., $L = \infty$). As an alternative to this assumption, one could use more fundamental assumptions to infer the existence of $b^* < \infty$ such that $L = \mathscr{A}$ is optimal if $s + b > b^*$. In either case, the original discounted model can be confined to the set of states $\Upsilon = \{(s,b) : 0 \leqslant s \leqslant b^*, 0 \leqslant b \leqslant b^*\}$ because a backlog larger than b^* would result from the acceptance of a job with processing time $s \geqslant b^*$. It follows that state $(0,0)$ is recurrent under every stationary policy and, therefore, the model satisfies a *unichain* assumption.

The single-transition net profit function is continuous, Υ is compact, and the transition times are exponential random variables with mean λ^{-1} (i.e., $\alpha = 0$); so the absolute value of the expected net profits during a transition has a finite upper bound. It follows that there is a scalar g (which is unique) and a real-valued function $\{w(s,b) : (s,b) \in \Upsilon\}$ that is a solution to a functional equation that is the average-profit counterpart to (9) for the discounted criterion:

$$g + w(s,b) = w(0,b) + \left[\max_{0 \leqslant L \leqslant b} \{e^{-\xi L}[\pi s - b + L + w(0, s+b) - w(0,b)]\} \right]^+, \tag{29}$$

$$g + w(0,b) = \int_0^b \lambda e^{-\lambda u} E[w(S, b-u)]\,du + E[w(S,0)]e^{-\lambda b}. \tag{30}$$

Here, g is the *gain rate* (maximal long-run net profit per unit time), and $w(s,b)$ is a *relative-value* term. Equations (29) and (30) are useful for characterizing and computing optimal policies. The parallels between [(9), (10)] and [(29), (30)] yield variants of Propositions 2, 3, 6 and 12. For example, Υ can be partitioned into four regions where three are analogous to (20) in which $\mathscr{L}(s,b)$ is replaced by

$$\mathscr{M}(s,b) = \frac{1}{\xi} - \pi s + b + w(0,b) - w(0,s+b).$$

It seems difficult to obtain bounds on the gradients of $w(s,b)$ that correspond to the gradient bounds in Sections 3.2 and 3.3. Nevertheless, there are immediate bounds on the maximal gain rate.

Proposition 13

$$\frac{\pi\rho}{1+\rho} \leqslant g \leqslant \pi\rho.$$

5. Computational study

5.1. Comparison of two methods

We designed a computational study to investigate the conditions under which it is advantageous for a firm to improve communications between sales and operations, i.e. to provide current manufacturing backlog information to sales. In other words, how much will accurate information enhance profit, when considered in the light of what it costs to implement such a change? Relevant costs might include the considerable expenses of implementation and maintenance of an ERP system, which tracks production status and provides information to various functional areas.

The computational study compares the impact of full information, using the dynamic programming results in Section 3, with the consequences of optimal behavior under partial information, i.e. the log-linear due-date assignment rule. The log-linear due-date assignment rule (mentioned in Section 1.2 and specified below) is optimal when sales knows the processing time of the job on which it issues a quotation, but does not have current backlog information, and assumes that the probability distribution of shop delay is that of first-come–first-served M/M/1 queueing time. So sales is assumed to know the following historic attributes: the mean processing time in manufacturing, v, and the arrival rate of jobs to manufacturing, Λ. In the study, S, the generic processing time seen by sales, is an exponential random variable and the tardiness penalty is proportional, as in (1), i.e. $z(b-L) = (b-L)^+$.

The log-linear rule quotes a due date as a function of a job's processing time. Thus the flow of jobs to manufacturing is a Poisson process with intensity $\Lambda < \lambda$. Let $R = \Lambda v$ which is the shop utilization, $\gamma = (1-R)/v$, and $r(s,b) = \pi s$ defined as in Section 3.3. Then the log-linear rule is $L = (L^*)^+$ where $L^* = x - y \cdot \ln s$ and

$$x = y \ln\left\{ \frac{R\gamma(\alpha+\gamma+\xi)}{[\alpha+\gamma]\xi\pi[\alpha(1-R)+\gamma]} \right\} \quad \text{and} \quad y = (\alpha+\gamma)^{-1}. \tag{31}$$

Let W_Q be the equilibrium queueing time in a queueing model. The derivation of (31) in Chatterjee et al. (2002) uses the following property of first-come first-served M/M/1. There are constants a and b such that

$P\{W_Q > x\} = ae^{-bx}$ for all $x \geqslant 0$. However, Abate et al. (1995) cite a substantial literature that proves that, in a broad array of queueing models, there are constants a and b such that $P\{W_Q > x\} \approx ae^{-bx}$ for suitably large x. This suggests that the log-linear rule would perform well with many other specific assumptions concerning the queueing time distribution (besides M/M/1).

In summary, the difference between the two methods in this numerical study is that the log-linear rule is used when actual shop status is not known, and delay time is inferred from characteristics of the shop, while the sequential decision model described in Section 3 uses backlog information directly to calculate an optimal policy.

5.2. Description of the study

We compare the two methods by running each on the undiscounted ($\alpha = 0$) discretized model with the spectrum of parameters (π, γ, ξ) specified in Table 1. The discretization, discussed further below and in Appendix A, consists of using the geometric distribution $P\{Y = k\} = \gamma(1-\gamma)^{k-1}$ $(k = 1, 2, \ldots)$ instead of an exponential distribution for interarrival times, and using the geometric distribution $P\{S = k\} = 0.15(0.85)^{k-1}$ $(k = 1, 2, \ldots)$ for processing times. That is, values are varied for the profit ratio (π, which corresponds to $[r(s,b)/s]/\tau(s,b)$), external arrival rate (γ), and customer impatience (ξ) for a total of 315 tests. A finite state space is obtained with the truncation $s \leqslant 18$ and $b \leqslant 50$. We note that $P\{S \leqslant 18\} = 0.9464$ and $P\{Y \leqslant 50\} \geqslant 0.9948$ with the parameters in Table 1. A discussion of convergence properties is included in Appendix A.

Each routine in the numerical study was coded in FORTRAN 77 and run on a Sun Ultra workstation under the Solaris 2.6 operating system. The fractional error, or normalized difference between profit obtained by the optimal algorithm and log-linear rule $(V^{\mathrm{OPT}} - V^{\mathrm{LL}})/V^{\mathrm{OPT}}$ (see Appendix A for details) was computed for each vector of input parameters. In addition to comparing the net profits generated by the two procedures, we used two measures to assess the difference in lead times. ABS is the sum of the weighted absolute value of the difference of the lead times:

$$\mathrm{ABS} = \sum_{s,b} |L(s,b) - LL(s)| \cdot p(s,b),$$

where $p(s,b)$ is the stationary probability of being in state (s,b). So $p(s,b) = p_b(0.15)(0.85)^{s-1}$ if $s \leqslant 17$ and $p(18,b) = p_b(0.85)^{17}$ (where p_b is defined in (A.3)). DIFF is the weighted sum (no absolute value), which measures whether log-linear lead times were too long or too short on average:

$$\mathrm{DIFF} = \sum_{s,b} (L(s,b) - LL(s)) \cdot p(s,b).$$

So negative (positive) values of DIFF indicate where the log-linear rule is generating lead times that are higher (lower) than optimal.

5.3. Discussion of results

A comparison of the value functions of the two procedures at different values of parameters π, γ and ξ yields patterns that characterize conditions under which the log-linear rule does well, and when it does not. We also examined the differences between lead-time quotations generated by the two procedures, in order to compare them in decision space as well as payoff space, and to help explain performance variations under different conditions. Results are illustrated by sets of graphs, where parameters range over the values used for the study. A complete set of diagrams (spanning all of the data) is available from the authors.

Table 1
Input values for numerical example

γ	0.1	0.15	0.2												
π	5	7.5	10	12.5	15	17.5	20								
ξ	0.001	0.006	0.011	0.016	0.021	0.026	0.031	0.036	0.041	0.046	0.051	0.056	0.061	0.066	0.071

5.3.1. Monotonicity

If $\rho \leqslant 1/\pi$ (where $\rho = \lambda/\mu$), Proposition 7 asserts that an optimal lead-time quotation is a nondecreasing function of the backlog. At all parameter combinations in the numerical study, $\rho > 1/\pi$ because $\rho = \gamma/0.15$, $\pi \geqslant 5$ and $\gamma \geqslant 0.1$. Nevertheless, *the optimal lead-time quotation was monotone and nondecreasing with respect to backlog at all 289,170 states in the study* (18 processing times × 51 backlogs × 315 parameter vectors). Similarly, if $\rho \leqslant 1$, Proposition 7 asserts that an optimal lead-time quotation is a nonincreasing function of the processing time and customers are not rejected. Since $\gamma \in \{0.1, 0.15, 0.2\}$, $\rho = \gamma/0.15 > 1$ for one-third of the numerical study. Nevertheless, *at all states in the study the customer was not rejected and the optimal lead-time quotation was monotone nonincreasing with respect to processing time*. Indeed, there are typical situations in congested shops where customers are quoted attractive lead times in order to keep their business (cf. Dwyer, 2000; APICS, 1991).

5.3.2. Higher profit margin

The log-linear rule is closer to optimal in terms of profit and lead-time performance when profit margin (π) is relatively higher. At all values of γ, higher values of π result in lower fractional error (Fig. 3(a)–(c)), and in lead-time estimates that are closer to optimal (Figs. 4(a) and (b), 5(a) and (b)) except when $\pi = 5$ when $\gamma \geqslant 0.15$ (Figs. 4(c) and 5(c)), and at $\gamma = 0.2$ for low values of ξ (Fig. 4(c)). For $\pi = 20$, the explanation for the near-optimal log-linear lead-time quotations in Fig. 5(a)–(c) is that here the optimal lead-time quotation is 0 (since admitting additional jobs results in more revenue than is lost by tardiness penalties resulting from congestion), and so the log-linear lead times are exactly right.

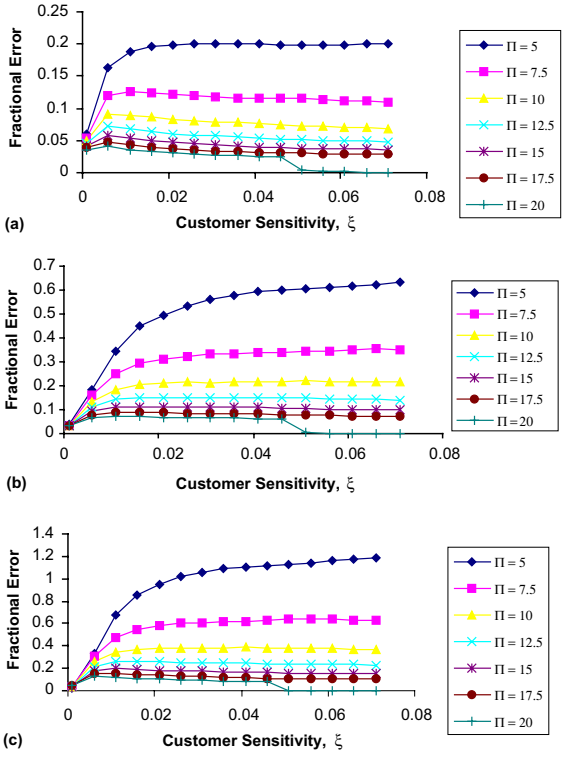

Fig. 3. Fractional error when (a) $\gamma = 0.1$, (b) $\gamma = 0.15$, (c) $\gamma = 0.2$.

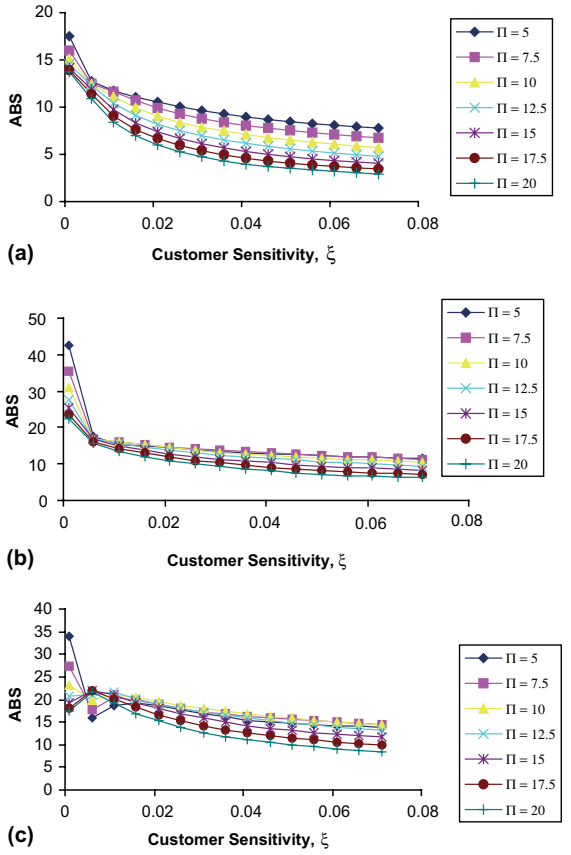

Fig. 4. ABS when (a) $\gamma = 0.1$, (b) $\gamma = 0.15$, (c) $\gamma = 0.2$.

5.3.3. Lower arrival rate

The log-linear rule is closer to optimal in terms of profit and lead-time performance when the "external" arrival rate to sales (γ) is lower. For all values of π, lower values of γ result in lower fractional error, except when $\xi = 0.001$ (Fig. 6(a)–(c)). When $\xi \geqslant 0.006$, lower values of γ also result in lead-time estimates that are closer to optimal (Figs. 7(a)–(c) and 8(a)–(c)). This reflects the fact that the log-linear rule generally is closer to optimal when the shop is less congested, since it does not consider actual backlog.

5.3.4. More patient customers

The log-linear rule is generally closer to optimal in terms of profit when customers are more patient (ξ is lower); however, the effect of varying customer sensitivity to lead-time quotations (ξ) is complicated, and must be considered in combination with other parameters. As ξ increases, a longer lead time is more likely to result in balking. Fig. 3(a)–(c) show that fractional error rises and then levels off as ξ increases, for small values of π $(5, 7.5, 10)$. For larger π $(12.5, 15, 17.5, 20)$, fractional error is fairly level or slightly *decreasing* as ξ increases. Fig. 9 also demonstrates this relationship of π and ξ when $\gamma = 0.15$.

In contrast, the lead-time error generally decreases as customers become more impatient, i.e., as ξ gets larger (Figs. 4(a)–(c) and 5(a)–(c)). For very small values of ξ (smaller than 0.006), lead-time error is highest. Although the lead times generated by the log-linear rule are generally too low (DIFF > 0; see Fig. 5(a)–(c)), when customers are very patient (small ξ), the log-linear rule quotes lead times that are too high (DIFF < 0),

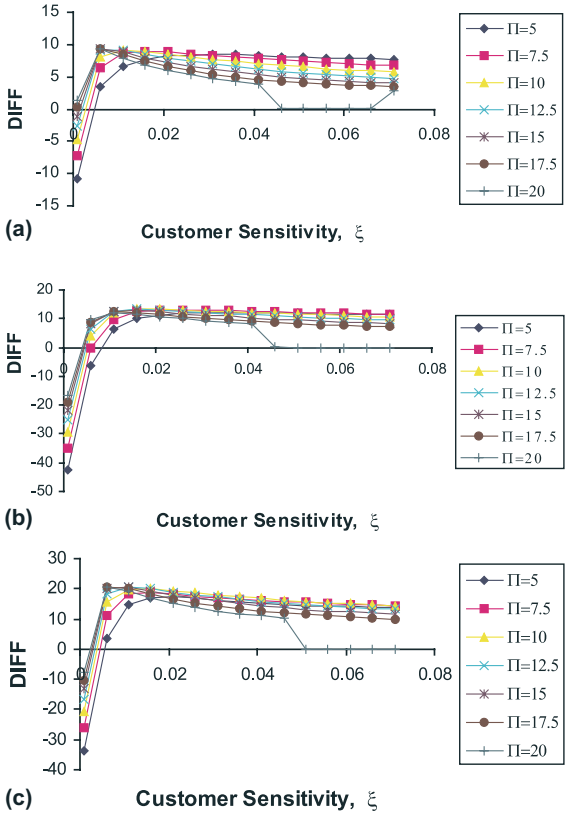

Fig. 5. DIFF when (a) $\gamma = 0.1$, (b) $\gamma = 0.15$, (c) $\gamma = 0.2$.

and could result in unwarranted loss of revenue by turning away customers. However, the probability of balking at these values of ξ is very low (less than 0.05), and so these excessive lead times do not result in much loss of revenue; they do prevent loss due to tardiness penalties. This explains why the log-linear rule does well in terms of *fractional error* at low values of ξ, despite the higher *lead-time error* (compare Figs. 9 and 10).

Looking further at combinations of parameters, we note that the highest fractional errors occur at low values of π combined with high values of γ (Fig. 11). The overall worst-case error was 1.18912, i.e., the log-linear rule resulted in tardiness penalties that exceeded revenues. This occurred when $\pi = 5$, $\gamma = 0.2$ and $\xi = 0.071$. So the log-linear rule does worst in terms of profit when profit margins are relatively low, arrival rates are high, and customers are impatient. Lead times are also furthest from optimal when profit margins are low and arrival rates are high (Fig. 12). Fractional error declines at a decreasing rate, i.e. is convex in π for most of its range (Fig. 11), while it is mostly concave for absolute lead-time error (Fig. 12).

In order to understand these results it is essential to remember that the log-linear rule does not consider shop backlog, and so is likely to quote inappropriate lead times when backlog is relatively high.

6. Conclusions and future work

We analyze the sequential decision process for lead-time quotation at a firm that manufactures products to order, faces an increased likelihood of balking as lead times grow, pays tardiness penalties for late

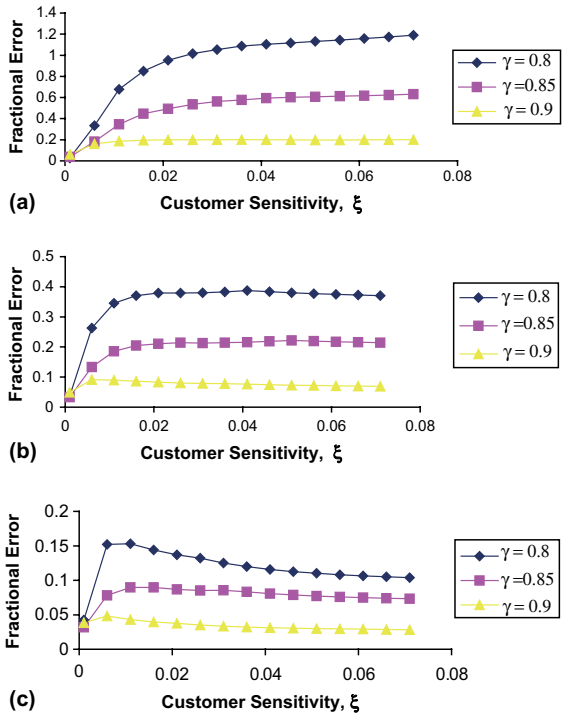

Fig. 6. Fractional error when (a) $\Pi = 5$, (b) $\Pi = 10$, (c) $\Pi = 17.5$.

deliveries, and books revenue when a customer enters the system. The firm has accurate information about the status of the shop, that is, it knows the manufacturing backlog. We characterize the optimal policy and its value function. When net revenue is proportional to processing time, tardiness penalty is proportional to lateness, and balking is an exponential function of lead time: bottom-line profit decreases as shop backlog increases, and increases as processing time increases, lead-time quotations increase as backlog increases and decrease as processing time increases. When the shop is not too busy, it does not pay to reject customers. These results are extended to include cases with multiple customer classes.

The computational study compares the performance of the optimal policy when the sales department has complete information about shop status with a previously developed log-linear rule which is optimal when sales has only historical information on backlog. We found conditions under which the log-linear rule is likely to do well, and when it is likely to do badly. Except when customers are relatively patient, the log-linear rule generates lead-time quotations that are lower than optimal. When profit margins are high, all else equal, optimal lead times become lower (approach zero), and so the log-linear rule does relatively well. It does worse as the arrival rate increases, all else equal, since this is apt to result in backlogs which it does not take into account, and so more tardiness penalties are incurred than for the optimal decision rule. With regard to customer sensitivity, the log-linear rule generally is closer to optimal when customers are relatively patient. Good performance in lead time tends to mirror good performance in profit, except when lead times are too high and customers are patient; here the lead-time error does not result in much diminution of profit since these patient customers are most likely to stay.

What does this mean for the value of accurate shop information? In a highly competitive industry that operates with relatively low profit margins, shop status information that permits more accurate lead-time estimates will enhance profits. Intuitively, there is less margin for error (i.e. tardiness costs). Similarly, firms

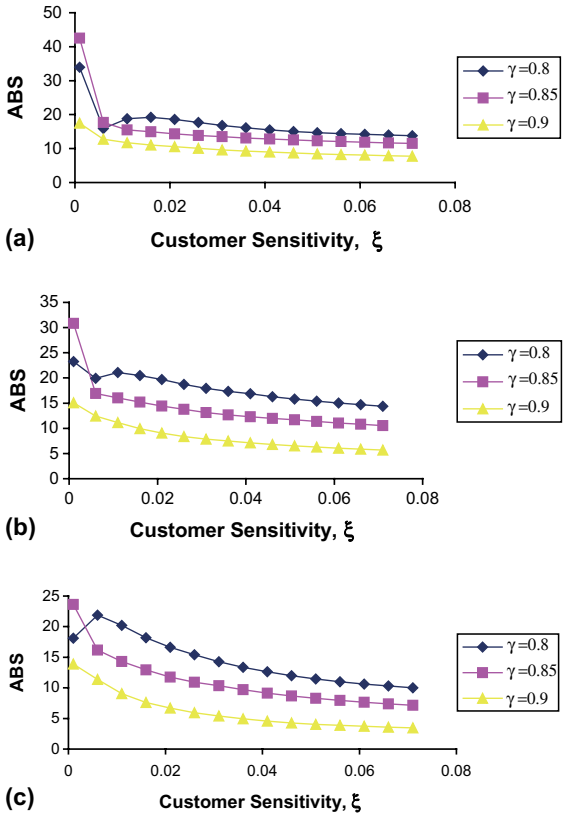

Fig. 7. ABS when (a) $\Pi = 5$, (b) $\Pi = 10$, (c) $\Pi = 17.5$.

that experience relatively high volumes of orders will benefit more from providing sales departments with current information that enables accurate lead-time quotations. The higher the order volume, the more likely the shop is to be congested, resulting in higher backlogs which should be taken into account when promising delivery dates to the customer. Customer characteristics also influence the value of accurate shop-status information. In particular, when customers are more sensitive to lead times (i.e. less "patient"), there is more benefit from information systems that allow accurate quotation of delivery dates based on the status of the production system. In this case, the monetary benefit realized is from keeping customers and the revenues that they bring (rather than saving tardiness costs).

In summary, there are important interaction effects among these three factors. When profit margins are relatively tight, there is a relatively high flow of customers to sales, and customers are relatively impatient, it is worth spending resources to track manufacturing backlogs accurately in order to improve the lead-time decision process. The sales department's due-date quotations could be much improved if it used accurate and current information about backlogs in the shop, and this would significantly enhance profitability under those conditions. On the other hand, when the profit margin is high, customers arrive at a relatively low rate to sales, and they are patient, the lead-time estimates with statistical information are apt to be close to the optimal ones.

Future research might include extensions of this model to sequencing, pricing and quality considerations. For a firm that receives multiple orders at once, and has the opportunity to decide upon the sequence in which they will be processed, how can the scheduling and lead-time decisions be coordinated?

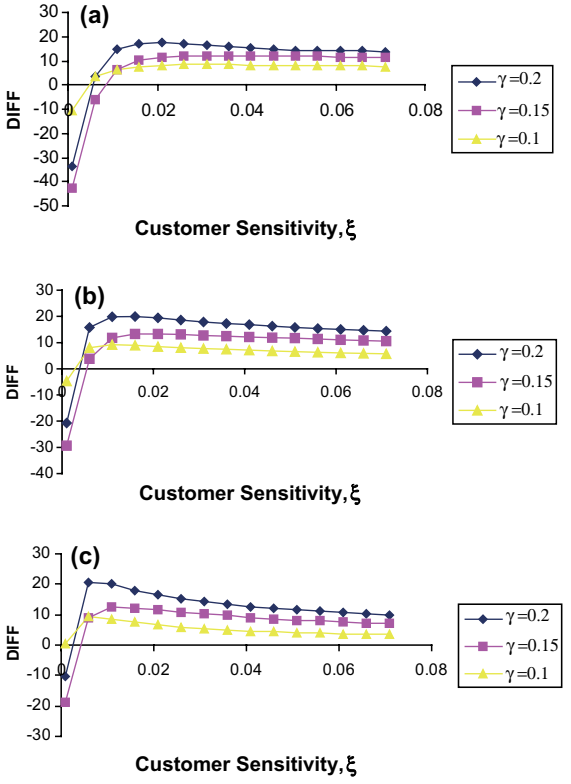

Fig. 8. DIFF when (a) $\Pi = 5$, (b) $\Pi = 10$, (c) $\Pi = 17.5$.

Another variation is the option of offering customers expedited service (i.e. shorter or negligible processing times) for a premium price. Such expedited services might include fewer features (lower design or performance quality) or lower yield (higher defect rates, or conformance quality).

Appendix A

A.1. Proof of Proposition 1

From (2),

$$f_n(s,b) = f_{n-1}(0,b) + \sup_{0 \leqslant L} \{a(s,L) J_n(s,b,L)\},$$

where

$$J_n(s,b,L) = r(s,b) - \tau(s,b) z(b-L) + f_{n-1}(0,s+b) - f_{n-1}(0,b)$$

is constant with respect to L on $[b,\infty)$. So

$$f_n(s,b) = f_{n-1}(0,b) + \max\left\{ \sup_{0 \leqslant L \leqslant b} \{a(s,L) J_n(s,b,L), 0 \right\},$$

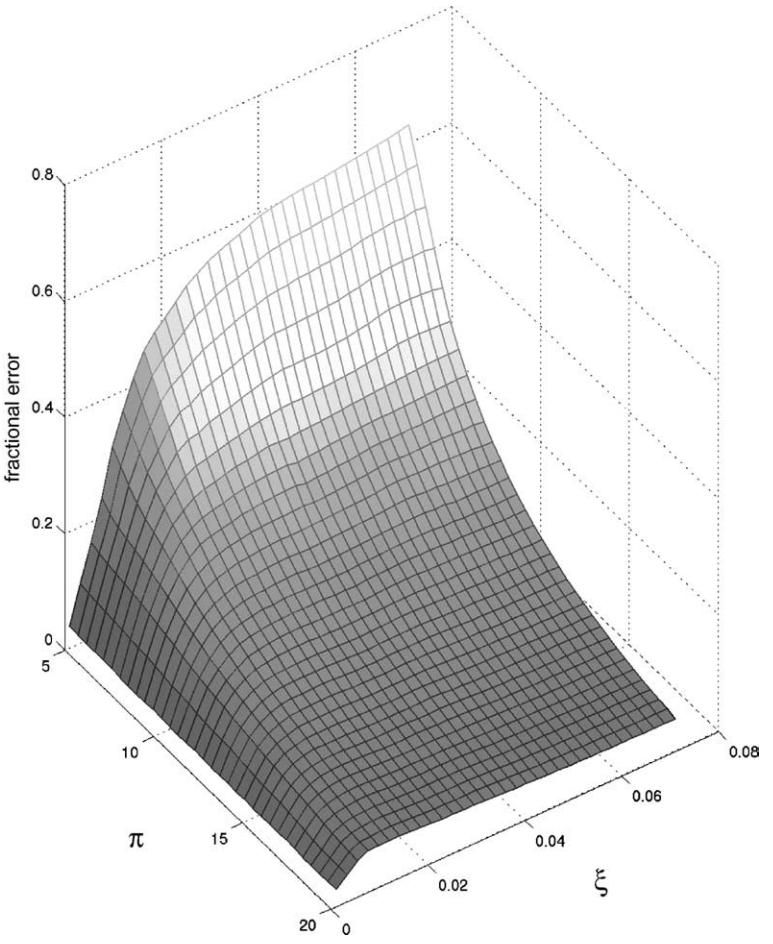

Fig. 9. Fractional error when $\gamma = 0.15$.

because if $J_n(s,b,b) \geqslant 0$ then

$$\sup_{b \leqslant L} \{a(s,L)J_n(s,b,L)\} = J_n(s,b,b) \sup_{b \leqslant \mathrm{L}} \{a(s,L)\} = J_n(s,b,b)a(s,b)$$

and if $J_n(s,b,b) < 0$, then

$$\sup_{b \leqslant L} \left\{ a(s,L)J_n(s,b,L)\} = J_n(s,b,b) \inf_{b \leqslant L} \{a(s,L) \right\} = 0.$$

So

$$f_n(s,b) = f_{n-1}(0,b) + \sup_{0 \leqslant L \leqslant b} \{a(s,L)[J_n(s,b,L)]^+\} = f_{n-1}(0,b) + \left[\sup_{0 \leqslant L \leqslant b} \{a(s,L)J_n(s,b,L)\} \right]^+.$$

The continuity of $a(s,\cdot)$ on $[0,\infty)$ and of $z(\cdot)$ on $(0,\infty)$ implies that the supremum is achieved for each (s,b) and n. Expanding (3) and inserting (4) yields

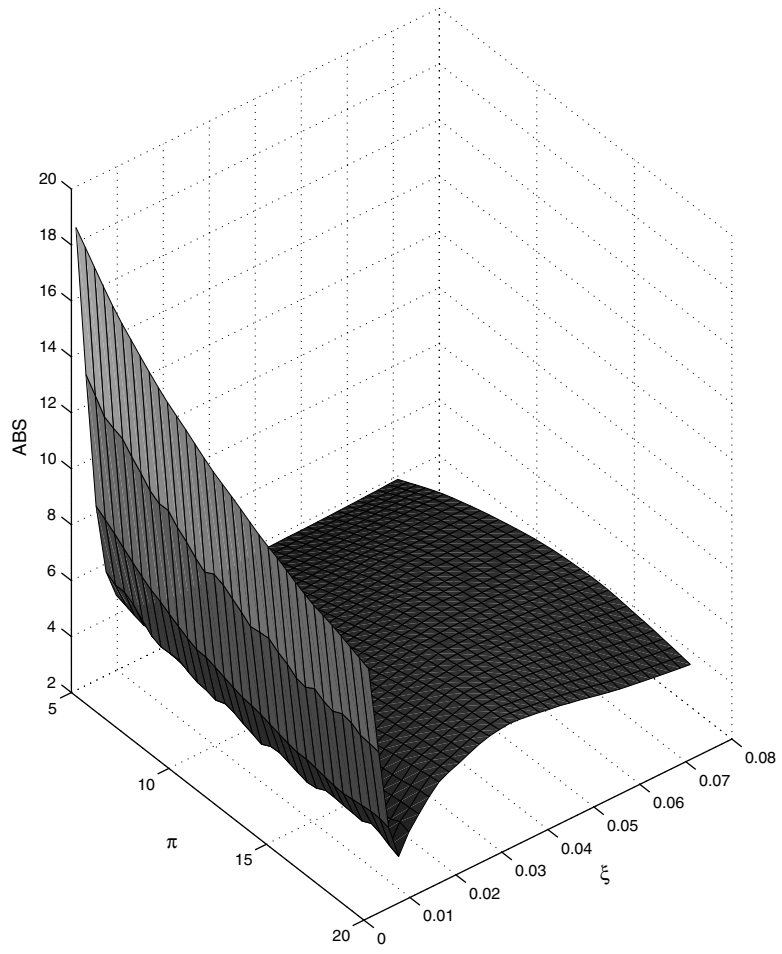

Fig. 10. ABS when $\gamma = 0.15$.

$$(\lambda+\alpha)f_n(0,b) = \lambda\left\{\int_0^b (\lambda+\alpha)\mathrm{e}^{-(\lambda+\alpha)u}E[f_{n-1}(S,b-u]\,\mathrm{d}u + E[f_{n-1}(S,0]\int_b^\infty (\lambda+\alpha)\mathrm{e}^{-(\lambda+\alpha)u}\,\mathrm{d}u\right\}$$
$$= \lambda E(f_{n-1}[S,(b-Y)^+]) = \lambda E(K_{n-1}[(b-Y)^+]).$$

When $b = 0$, (6) becomes (7). □

A.2. Proof of Proposition 2

The model is a semi-Markov programming problem with a discounted infinite-horizon criterion which corresponds to a discrete-time Markov decision process with a discounted infinite-horizon criterion whose discount factor depends on the state and action. It follows from Proposition 1 that the set of feasible actions is compact at each state. So (8) and (9) follow from Lemma 1 which enables us to invoke Schäl (1975). Then (10) results from (6) and (8), and (11) is a special case of (10). The final assertion follows from Lemma 1 and Schäl (1975). □

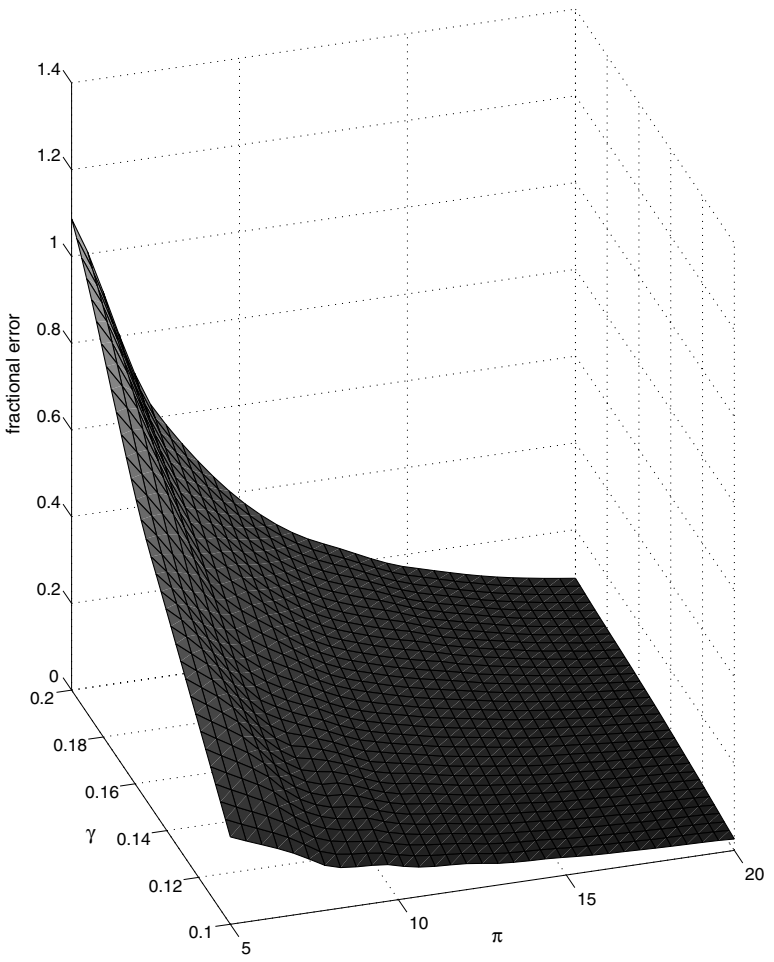

Fig. 11. Fractional error when $\xi = 0.036$.

A.3. Proof of Lemma 1

An inductive proof using (2) and (5) and starting with $f_0 \equiv 0$ establishes that $\{f_n(s,b)\}$ is a monotone sequence for each (s,b). An upper bound on $f_n(s,b)$ is obtained from the expected present value of the denumerable sequence of revenues, i.e.

$$f_n(s,b) \leqslant r(s,b) + E\left[r(S,0)\sum_{n=1}^{\infty}\exp\left(-\alpha\sum_{k=1}^{n}Y_k'\right)\right] = r(s,b) + E[r(S,0)]\frac{\lambda}{\alpha},$$

where $\{Y_k'\}$ are independent and identically distributed exponential random variables with mean λ^{-1}. □

A.4. Proof of Proposition 3

We use (2) and initiate an inductive proof with $f_0(s,\cdot) \equiv 0$ trivially nonincreasing for each s. If $f_{n-1}(s,\cdot)$ is nonincreasing for each s then $[1 - a(s,L)]f_{n-1}(0,b)$ and $f_{n-1}(0,s+b)a(s,L)$ are nonincreasing in b (for each s and L). By assumption, $a(s,L)r(s,b)$ and $-a(s,L)\tau(s,b)z(b-L)$ are nonincreasing in b (for each s and L).

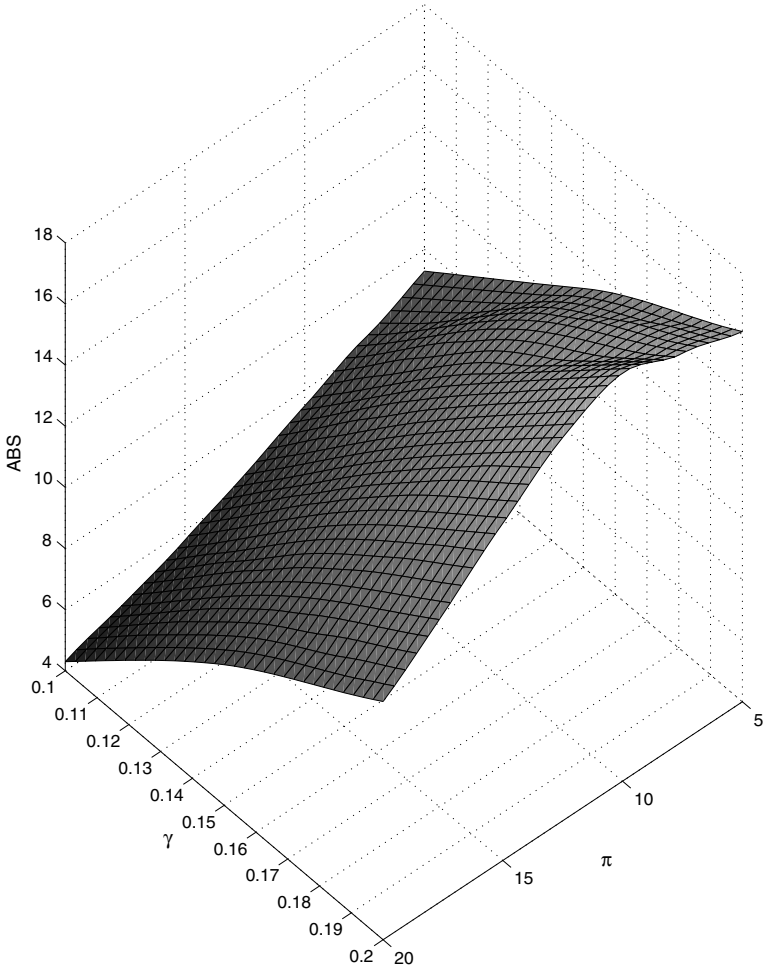

Fig. 12. ABS when $\xi = 0.036$.

So $f_n(s,\cdot)$ is nonincreasing (for each n and s). Therefore, (8) and Lemma 1 imply that $f(s,\cdot)$ is nonincreasing because the finite point-wise limits of monotone functions are monotone. □

A.5. Proof of Proposition 4

Quoting $L = 0$ in state (s,b) is no better than optimal, rejecting all arriving customers for x units of time is no better than optimal in state $(0,s+b)$, and rejecting all arrivals for s units of time is no better than optimal in state $(0,s+b-x)$. So (13) results from

$$f(s,b) \geqslant f(0,s+b) \geqslant e^{-\alpha x} f(0,s+b-x) \geqslant e^{-\alpha x} e^{-\alpha s} f(0,b-x).$$

For (14), we note that using an optimal policy in state $(0,s+b)$ is at least as good as rejecting all arriving customers for the next s units of time and thereafter behaving optimally:

$$f(0,s+b) \geqslant e^{-\alpha s} f(0,b).$$

Multiply both sides by -1 and add $f(0,b)$.

Proposition 3 implies $f^{(2)}(0,b) \leqslant 0$ in (15). For $-\alpha f(0,b) \leqslant f^{(2)}(0,b)$, substitute $e^{-\alpha s} = 1 - \alpha s + o(s)$ (where $o(s)/s \to 0$ as $s \to 0$) in the denominator of (14) and let $s \downarrow 0$. Lemma 1 implies the leftmost inequality in (15). □

A.6. Proof of Proposition 5

Let $\Delta(s,b) = f(0,s+b) - f(0,b)$; here,

$$J(s,b,L) = r(s,b) - \tau(s,b)(b-L)^+ + \Delta(s,b).$$

Since $\tau(\cdot,b)$ is nondecreasing, if $L \leqslant b$,

$$\begin{aligned} J(s+\epsilon,b,L) &= J(s,b,L) + r(s+\epsilon,b) - r(s,b) - (b-L)[\tau(s+\epsilon,b) - \tau(s,b)] + \Delta(\epsilon,s+b) \\ &\leqslant J(s,b,L) + r(s+\epsilon,b) - r(s,b) + \Delta(\epsilon,s+b). \end{aligned}$$

Since $a(\cdot,L)$ is nonincreasing, this inequality implies

$$\begin{aligned} f(s+\epsilon,b) &= f(0,b) + \left[\max_{0\leqslant L\leqslant b}\{a(s+\epsilon,L)J(s+\epsilon,b,L)\}\right]^+ \\ &\leqslant f(0,b) + \left[\max_{0\leqslant L\leqslant b}\{a(s+\epsilon,L)[J(s,b,L) + r(s+\epsilon,b) - r(s,b) + \Delta(\epsilon,s+b)]\}\right]^+ \\ &\leqslant f(0,b) + \left[\max_{0\leqslant L\leqslant b}\{a(s,L)J(s,b,L)\}\right]^+ + [r(s+\epsilon,b) - r(s,b) + \Delta(\epsilon,s+b)]^+ \\ &= f(s,b) + [r(s+\epsilon,b) - r(s,b) + \Delta(\epsilon,s+b)]^+. \end{aligned}$$

This inequality, $\Delta(\epsilon,s+b)/\epsilon \to f^{(2)}(0,s+b)$ as $\epsilon \to 0$, and Proposition 3 yield (16). □

A.7. Proof of Proposition 6

Let $\theta(L)$ denote the maximand in (19). Then

$$\frac{d\theta(L)}{dL} = e^{-\xi L}\{1 - \xi[J(s,b,0) + L]\},$$

$$\frac{d^2\theta(L)}{dL^2} = -\xi e^{-\xi L}\{2 - \xi[J(s,b,0) + L]\}.$$

So $\theta(\cdot)$ is maximized on $(-\infty,\infty)$ at $L = \mathscr{L}(s,b)$, it is concave on $(-\infty, \mathscr{L}(s,b) + 1/\xi]$, and it is convex on $[\mathscr{L}(s,b) + 1/\xi, \infty)$. Therefore, $L = \min\{[\mathscr{L}(s,b)]^+, b\}$ maximizes $\theta(L)$ subject to the constraint $0 \leqslant L \leqslant b$.

From (18), the regions in (20) correspond to $\mathscr{L}(s,b) < 0, 0 \leqslant \mathscr{L}(s,b) \leqslant b$, and $b < \mathscr{L}(s,b)$. Since $f(0,b) - f(0,b+s) \geqslant 0$ from Proposition 3, (18) implies $\mathscr{L}(s,b) \geqslant 1/\xi - \pi s + b$. □

A.8. Proof of Lemma 2

From (18), $\mathscr{L}^{(2)}(s,b) = 1 - f^{(2)}(0,s+b) + f^{(2)}(0,b)$ with the inner inequalities of (21) implied by $f^{(2)}(0,b) \leqslant 0$ and $f^{(2)}(0,s+b) \leqslant 0$. The outer inequalities are implied by (15) because $\lambda E[r(S,0)] = \pi\rho$. For (22), use (21) and $f^{(2)}(0,s+b) \leqslant 0$ in $\mathscr{L}^{(1)}(s,b) = -\pi - f^{(2)}(0,s+b)$. □

A.9. Proof of Proposition 7

From (21), $\mathscr{L}^{(2)}(s,b) \geqslant 1-\pi\rho \geqslant 0$ if $\rho \leqslant 1/\pi$. From (22), $\mathscr{L}^{(1)}(s,b) \leqslant \pi(\rho-1) \leqslant 0$ if $\rho \leqslant 1$.

Since $0 \leqslant L \leqslant b$ in (9), the value of the maximization is negative if, and only if, $\pi s + f(0,s+b) - f(0,b) < 0$. However, (21) implies $\pi s + f(0,s+b) - f(0,b) \geqslant \pi s - s\pi\rho = \pi s(1-\rho) \geqslant 0$ if $\rho \leqslant 1$. □

A.10. Proof of Proposition 8

Proof. For (23), inserting (1) in (9) yields

$$\frac{f(s,b)-f(0,b)}{s} = \left[\max_{0\leqslant L\leqslant b}\left\{e^{-\xi L}\left(\pi - \frac{(b-L)}{s} + \frac{[f(0,s+b)-f(0,b)]}{s}\right)\right\}\right]^{+}. \tag{A.1}$$

If $L < b$ then $-(b-L)/s \to -\infty$ as $s \downarrow 0$. So maximization implies $L = b - o(s)$ as $s \downarrow 0$. Therefore, letting $s \downarrow 0$ in (A.1) yields the equality in (23) whose left inequalities follow from (15). The right inequality is implied by Proposition 3. Proposition 5 and $r^{(1)}(s,b) = \pi$ imply (24).

For (25), Proposition 3 asserts $-f^{(2)}(0,b) \geqslant 0$ and (23) implies $-f^{(2)}(0,b) \geqslant \pi - e^{\xi b} f^{(1)}(0,b)$. So $-f^{(2)}(0,b) \geqslant \max\{0, \pi - e^{\xi b} f^{(1)}(0,b)\}$. □

A.11. Proof of Proposition 9

First we prove (27) at $s = 0$, namely,

$$f(0,0) \geqslant \pi\rho D. \tag{A.2}$$

Quoting $L = 0$ at $(s,0)$ is no better than optimal and rejecting all customers for s units of time at $(0,s)$ is no better than optimal. So (9) yields

$$f(s,0) \geqslant f(0,0) + [\pi s + f(0,s) - f(0,0)]^{+} \geqslant \pi s + f(0,s) \geqslant \pi s + e^{-\alpha s} f(0,0).$$

Therefore,

$$\lambda E[f(S,0)] \geqslant \pi\rho + \lambda E[f(0,S)] \geqslant \pi\rho + \lambda\phi(\alpha) f(0,0).$$

Hence, (7) yields

$$f(0,0) = \frac{\lambda E[f(S,0)]}{\lambda+\alpha} \geqslant \frac{\pi\rho + \lambda\phi(\alpha) f(0,0)}{\lambda+\alpha},$$

which implies (A.2) because $\alpha > 0$ implies $D > 0$. Now (A.2) and Proposition 4 imply the lower bound in (26) because quoting $L = 0$ at (s,b) cannot be better than optimal:

$$f(s,b) \geqslant \pi s - b + f(0,s+b) \geqslant \pi s - b + e^{-\alpha(s+b)} f(0,0) \geqslant \pi s - b + \pi\rho D e^{-\alpha(s+b)}.$$

The upper bound in (26) follows from (8) and (12).

In order to establish (27), (19) implies

$$f(s,0) = f(0,0) + [\pi s + f(0,s) - f(0,0)]^{+} = \max\{f(0,0), \pi s + f(0,s)\}.$$

So (13) and (A.2) imply

$$\begin{aligned} f(s,0) &\geqslant \max\{f(0,0), \pi s + \mathrm{e}^{-\alpha s} f(0,0)\} \\ &= \pi s + \mathrm{e}^{-\alpha s} f(0,0) + [(1-\mathrm{e}^{-\alpha s}) f(0,0) - \pi s]^+ \\ &\geqslant \pi s + \pi\rho D \mathrm{e}^{-\alpha s} + [(1-\mathrm{e}^{-\alpha s})\pi\rho D - \pi s]^+. \qquad \square \end{aligned}$$

A.12. Proof of Proposition 10

Dropping the penalty function in the infinite-horizon counterpart of (2),

$$f(s,b) \leqslant f(0,b) + \max_{L \geqslant 0} \{\mathrm{e}^{-\xi L}[\pi s + f(0, s+b) - f(0,b)]\} = f(0,b) + [\pi s + f(0,s+b) - f(0,b)]^+. \qquad \square$$

A.13. Proof of Proposition 11

On the set of $\{(s,b)\}$ where $L = \mathscr{L}(s,b)$ is optimal, it follows from (18) and (19) that

$$f(s,b) = f(0,b) + \frac{\mathrm{e}^{-\xi \mathscr{L}(s,b)}}{\xi} = f(0,b) + \frac{\exp[-1 + \xi J(s,b,0)]}{\xi}.$$

So

$$f^{(1)}(s,b) = -\mathscr{L}^{(1)}(s,b)\mathrm{e}^{-\xi \mathscr{L}(s,b)},$$

which implies that $f^{(1)}(s,b)$ and $\mathscr{L}^{(1)}(s,b)$ have opposite signs; so (ii) and (iii) are equivalent.

Proposition 8 asserts that $f^{(1)}(0,b) \geqslant 0$. Therefore, if $f(\cdot,b)$ were convex, $f^{(1)}(s,b) \geqslant f^{(1)}(0,b) \geqslant 0$ for all $s \geqslant 0$. That is, (i) implies (iii). $\square$

A.14. Proof of Proposition 12

$$\mathscr{L}_i(s,b) - \mathscr{L}_k(s,b) = \frac{1}{\xi_i} - \frac{1}{\xi_k} + (\pi_k - \pi_i)s \geqslant 0. \qquad \square$$

A similar result is valid without the counterpart of the additional assumptions in Section 3.3, if $r_k(s,b) \geqslant r_i(s,b)$ (for all s and b), $a_k(s,L) \leqslant a_i(s,L)$ (for all s and L), and

$$\partial[a_j^{(2)}(s,L)]^2 + a_j(s,L)a_j^{(22)}(s,L) \geqslant 0$$

for $j = i,k$ (and all s and L).

A.15. Proof of Proposition 13

An optimal due-date quotation rule cannot have a net profit rate that is higher than the average *revenue* per unit time if all customers are accepted, i.e., $L = 0$ so none balk. Since no tardiness penalties are paid, the expected net profit during a transition is $\pi E(S) = \pi/\mu$ and the expected duration of the transition is λ^{-1}. So $g \leqslant \pi\lambda/\mu = \pi\rho$. On the other hand, the policy of admitting a customer only if the backlog is 0 cannot be better than optimal. This customer incurs no tardiness penalty and brings expected revenue $\pi E(S) = \pi/\mu$. The expected duration of the cycle between processing successive customers is $E(S) + 1/\lambda = 1/\mu + 1/\lambda$ with

$1/\lambda$ due to the memorylessness of the residual of the exponential time until the next customer arrives. So $g \geqslant [\pi/\mu]/[1/\mu + 1/\lambda] = \pi\rho/(1+\rho)$. □

A.16. Convergence properties for the computational study

Let $LL(s) = (x - y \cdot \ln s)^+$ where x and y are specified in (31) i.e., $LL(s)$ is the lead-time quotation with the log-linear rule. Similarly, let $L(s,b)$ denote a value of L that achieves the maximum in (5) and (6) when $n = 50$. So $L(\cdot,\cdot)$ is an optimal due-date quotation in state (s,b) with 50 iterations remaining and is associated with the value function $f_{50}(\cdot,\cdot)$. We use two consequences of the *unichain property*, i.e., $L(\cdot,\cdot)$ induces a Markov chain with exactly one ergodic class (and perhaps transient states). Let p_b be the stationary probability of state $(0,b)$ induced by $L(\cdot,\cdot)$ over an infinite horizon. The first consequence is that $\{p_b\}$ is the unique solution to the usual Markov chain balance equations where $\eta(b) = \min\{18, 50-b\}$:

$$p_b = (1-\gamma)p_{b+1} + \gamma p_{b+1}\sum_{s=1}^{\eta(b)-1}(0.15)(0.85)^{s-1}(1-\mathrm{e}^{-\xi L(s,b)}) + \gamma p_{b+1}(0.85)^{\eta(b)-1}(1-\mathrm{e}^{-\xi L(\eta(b),b)}) + \gamma\sum_{j=1}^{b-1}p_{j+1}(0.15)(0.85)^{b-j}\mathrm{e}^{-\xi L(b-j,b)} \quad (b=0,1,\ldots,49), \tag{A.3}$$

$$p_{50} = \gamma\sum_{j=1}^{49}p_{j+1}(0.15)(0.85)^{50-j}\mathrm{e}^{-\xi L(50-j,50)},$$

$$\sum_{b=0}^{50}p_b = 1.$$

In the right-hand side of (A.3), the first term arises from the absence of an arriving customer, the second and third terms correspond to a customer arriving and balking, and the fourth term refers to a customer arriving and not balking.

The second consequence of the unichain property links $f_n(\cdot,\cdot)$ with the solution to (29) and (30): as $n \to \infty$, $f_n(s,b) \approx ng + w(s,b)$. So $\Delta_n(s,b) \to 0$ as $n \to \infty$ where $\Delta_n(s,b) = |f_n(s,b) - f_{n-1}| - |f_{n-1}(s,b) - f_{n-2}(s,b)|$. As an indication that $n = 50$ is sufficient for asymptote-like behavior, each of the 315 vectors of input parameters yielded $\sum_{b=0}^{50} p_b \Delta_n(0,b) \leqslant 4.73 \times 10^{-6}$. Accordingly, let $V^{\mathrm{OPT}} = \sum_{b=0}^{50} p_b f_{50}(0,b)$ denote the weighted average value of the value function with 50 iterations remaining. States (s,b) with $s > 0$ are not included in the weighted average because they are occupied only momentarily; the period is actually spent either in state $(0,b)$ with probability $1 - \mathrm{e}^{-\xi L(s,b)}$ or in state $(0, s+b)$ with probability $\mathrm{e}^{-\xi L(s,b)}$.

Let $V^{\mathrm{LL}} = \sum_{b=0}^{50} p_b f_{50}^{\mathrm{LL}}(0,b)$ be the corresponding weighted average for the log-linear rule. Here, $f_{50}^{\mathrm{LL}}(0,b)$, the value function of the log-linear due-date quotation policy in (31) with 50 iterations remaining, satisfies the following recursion with $f_0^{\mathrm{LL}}(\cdot,\cdot) \equiv 0$, $f_{n-1}^{\mathrm{LL}}(0,s+b) = f_{n-1}^{\mathrm{LL}}(0,50)$ if $s+b > 50$, $\mu(b) = \min\{17, b-1\}$ and $\delta(b-18) = 1(0)$ if $b \geqslant 19$ $(b \leqslant 18)$:

$$f_n^{\mathrm{LL}}(s,b) = f_{n-1}^{\mathrm{LL}}(0,b) + \mathrm{e}^{-\xi L(s,b)}[\pi s - (b - L(s))^+ + f_{n-1}^{\mathrm{LL}}(0,s+b) - f_{n-1}^{\mathrm{LL}}(0,b)],$$

$$f_{n-1}^{\mathrm{LL}}(0,b) = \gamma\sum_{s=1}^{\mu(b)}(0.15)(0.85)^{s-1}\mathrm{e}^{-\xi L(s,b-s)}f_{n-1}^{\mathrm{LL}}(s,b-s) + \gamma(0.85)^{17}f_{n-1}^{\mathrm{LL}}(18,b-18)\delta(b-18) + (1-\gamma)f_{n-1}^{\mathrm{LL}}(0,(b-1)^+).$$

Let $\Delta_n^{\mathrm{LL}}(s,b) = |f_n^{\mathrm{LL}}(s,b) - f_{n-1}^{\mathrm{LL}}| - |f_{n-1}^{\mathrm{LL}}(s,b) - f_{n-2}^{\mathrm{LL}}(s,b)|$. As an indication that $n = 50$ is sufficient for asymptote-like behavior with the log-linear rule, each of the 315 vectors of input parameters yielded $\sum_{b=0}^{50} p_b \Delta_n^{\mathrm{LL}}(0,b) \leqslant 1.4 \times 10^{-4}$.

References

Abate, J., Choudhury, G., Whitt, W., 1995. Exponential approximations for tail probabilities in queues, I: Waiting times. Operations Research 43, 885–901.

Aerospace Industries Association, 2002. Net new orders, shipments and backlog for large civil jet transport aircraft. Aerospace Statistics. Downloadable from <http://www.aia-aerospace.org/stats/aero_stats/jets/q01.pdf>.

Andel, T., 2002. From common to custom: The case for make to order. Material Handling Management 57, 24–31.

Ansberry, C., 2002. A new hazard for recovery: Last-minute pace of orders. Wall Street Journal. A1, A12.

APICS, 1991. Processing products into profits. Production and Inventory Management Review and APICS News 11, 40–41.

Baker, K., 1984. Sequencing rules and due-date assignments in a job shop. Management Science 30, 1093–1104.

Baker, K., Bertrand, J., 1981. A comparison of due-date selection rules. AIIE Transactions, 123–131.

Bartholomew, D., 1996. Boost to response time. Informationweek, 73.

Boots, N., Tijms, H., 1999. A multiserver queueing system with impatient customers. Management Science 45, 444–448.

Boyaci, T., Ray, S., 2003. Product differentiation and capacity cost interaction in time and price sensitive markets. Manufacturing and Service Operations Management 5, 18–36.

Bureau of Census, 2000. New orders, shipments and backlog of orders for selected industrial air pollution control equipment: 1998 and 1997. Current Industrial Reports, 9–4.

Bureau of Census, 2002. Value of manufacturers shipments. Manufacturing, Mining and Construction Statistics. Downloadable from <http://www.census.gov/indicator/www/m3/prel/table1p.xls>.

Carter, J., 1993. Purchasing: Continued Improvement Through Integration. Business One Irwin, Homewood, IL.

Chatterjee, S., Slotnick, S.A., Sobel, M.J., 2002. Delivery guarantees and the interdependence of marketing and operations. Production and Operations Management 11, 393–410.

Cheng, T., Gupta, M., 1989. Survey of scheduling research involving due date determination decisions. European Journal of Operational Research 38, 156–166.

Cheung, K.L., 1998. A continuous review inventory model with time discount. IIE Transactions 30, 747–757.

Czech Statistical Office, 2002. Summary for legal and natural persons; basic indicators for industrial enterprises. Downloadable from <http:www.czso.cz/eng/figures/8/80/800602/data/8006rr03.xls>.

Duenyas, I., 1995. Single facility due date setting with multiple customer classes. Management Science 41, 608–619.

Duenyas, I., Hopp, W., 1995. Quoting customer lead times. Management Science 41, 43–57.

Dwyer, J., 2000. Cycle time revolution. Works Management 53, 54–57.

ElHafsi, M., 2000. An operational decision model for lead-time and price quotation in congested manufacturing systems. European Journal of Operational Research 126, 355–370.

Elimam, A.A., Dodin, B.M., 2001. Incentives and yield management in improving productivity of manufacturing facilities. IIE Transactions 33, 449–462.

Freeman, L., 1996. Beating a strategic retreat. Business Marketing 81, 44.

Heyman, D.P., Sobel, M.J., 1984. Stochastic Models in Operations Research, Volume II: Stochastic Optimization. McGraw-Hill, New York.

Johansen, S.G., 1991. Optimal prices of a jobshop with a single work station: A discrete time model. International Journal of Production Economics 23, 129–138.

Johansen, S.G., 1994. Optimal prices of an M/G/1 jobshop. Operations Research 42, 765–774.

Keskinocak, P., Ravi, R., Tayur, S., 2001. Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. Management Science 47, 264–279.

Krajewski, L., Ritzman, L., 2002. Operations Management: Strategy and Analysis. Prentice Hall, Englewood Cliffs, NJ.

Lawrence, S.R., 1995. Estimating flowtimes and setting due-dates in complex production systems. IIE Transactions 27, 657–688.

Leachman, R., Benson, R., Liu, C., Raar, D., 1996. An automated production-planning and delivery-quotation system at Harris Corporation-Semiconductor sector. Interfaces 26, 6–37.

Lederer, P.J., Li, L., 1997. Pricing, production scheduling and delivery-time competition. Operations Research 45, 407–420.

Li, L., 1992. The role of inventory in delivery-time competition. Management Science 38, 182–197.

Lovejoy, W., 1986. Policy bounds for Markov decision processes. Operations Research 34, 630–637.

MacQueen, J., 1967. A test for suboptimal actions in Markov decision problems. Operations Research 15, 559–561.

Migliorelli, M., Swan, R.J., 1988. MRP and aggregate planning—A problem solution. Production and Inventory Management Journal 29, 42–45.
Moodie, D., 1999. Demand management: The evaluation of price and due date negotiation strategies using simulation. Production and Operations Management 8, 151–162.
Moodie, D., Bobrowski, P., 1999. Due-date demand management: Negotiating the trade-off between price and delivery. International Journal of Production Research 16, 305–318.
Ogden, H.J., Turner, R.E., 1996. Customer satisfaction with delivery scheduling. Journal of Marketing Theory and Practice 4, 79–94.
Palaka, K., Erlebacher, S., Kropp, D.H., 1998. Lead-time setting, capacity utilization and pricing decisions under lead-time dependent demand. IIE Transactions 30, 151–163.
Palmatier, G., Shull, J., 1989. The Marketing Edge: The New Leadership Role of Sales & Marketing in Manufacturing. Oliver Wight Limited Publications, Essex Junction, VT.
Pinedo, M., 1995. Scheduling: Theory, Algorithms, and Systems. Prentice Hall, Englewood Cliffs, NJ.
Ragatz, G., Mabert, V., 1984. A framework for the study of due date management in job shops. International Journal for Production Research 22, 685–695.
Rajagopalan, S., 2002. Make to order or make to stock: Model and application. Management Science 48, 241–256.
Schäl, M., 1975. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 32, 179–196.
Shapiro, R.D., 1988. Donner. Harvard Business School Case Study No. 9-689-030. Boston, MA.
Shapiro, B., Moriarty, R., Cline, C., 1992. Fabtek(A). Harvard Business School Case Study No. 9-592-095. Boston, MA.
So, K., Song, J., 1998. Price, delivery time guarantees and capacity selection. European Journal of Operational Research 111, 28–49.
Sobel, M.J., 1971. Production smoothing with stochastic demand, II: Infinite horizon case. Management Science 17, 724–735.
Stansbury, T., 2000. Global sourcing, offset and risk/revenue sharing partnerships. Lecture delivered at Arizona State University West, April 4. Mr. Stansbury was Director of the International Programs Group at Honeywell Engines and Systems.
Statisches Bundesamt Deutschland, 2002. Index of orders received—Germany—manufacturing. Downloadable from <http://www.destatis.de/indicators/e/tkae211.htm>.
Statistics Bureau and Statistics Center (Japan), 2002. Values of orders received for machinery. Downloadable from <http://www.stat.go.jp/english/data/geppou/zuhyou/e08.xls>.
Statistics Canada, 2002. Business conditions survey: Manufacturing industries. Downloadable from <http://www.statcan.ca/english/020201/d020201a.htm>.
Stidham, S., 1985. Optimal control of admission to a queueing system. IEEE Transactions on Automatic Control AC-30, 705–713.
Veinott, A.F., 1966. On the optimality of (s,S) inventory policies: New conditions and a new proof. SIAM Journal 14, 1067–1083.
Weeks, J., 1979. A simulation study of predictable due-dates. Management Science 25, 363–373.
Wein, L., 1991. Due-date setting and priority sequencing in a multiclass M/G/1 queue. Management Science 37, 834–850.
Weng, Z.K., 1999. Strategies for integrating lead time and customer-order decisions. IIE Transactions 31, 161–171.
Whitt, W., 1999. Improving service by informing customers about anticipated delays. Management Science 45, 192–207.
Yeager, B., 1997. Keep it on-hand. Manufacturing Systems 15, 47–52.