Convexity Properties and Comparative Statics for M/M/S Queues with Balking and Reneging

Mor Armony[1] Erica Plambeck[2] Sridhar Seshadri[3]

We use sample path arguments to derive convexity properties of an M/M/S queue with impatient customers that balk and renege. First, assuming that the balking probability and reneging rate are increasing and concave in the total number of customers in the system (head-count), we prove that the expected head-count is convex decreasing in the capacity (service rate). Second, with linear reneging and balking, we show that the expected lost sales rate is convex decreasing in the capacity. Finally, we employ a sample-path sub-modularity approach to comparative statics. That is, we employ sample path arguments to show how the optimal capacity changes as we vary the parameters of customer demand and impatience. We find that the optimal capacity increases in the demand rate and decreases with the balking probability, but is *not* monotone in the reneging rate. This means, surprisingly, that failure to account for customers' reneging may result in *over*-investment in capacity. Finally, we show that a seemingly minor change in system structure, customer commitment during service, produces qualitatively different convexity properties and comparative statics.

1. Introduction and Overview of Results

This paper develops qualitative insights about how the optimal capacity investment for a make-to-order system is influenced by customers' impatience, which may lead them to cancel an order (renege) or not to order at all (balk) when waiting is required. Technically, we prove convexity and comparative statics properties for a M/M/S queue with quite general reneging and balking behavior.

A dominant assumption in the manufacturing operations management literature is that customers will wait for as long as necessary to obtain a product (infinite backordering). In

[1]Stern School of Business, New York University, marmony@stern.nyu.edu
[2]Graduate School of Business, Stanford University, elp@stanford.edu
[3]Stern School of Business, New York University, sseshadr@stern.nyu.edu

reality, only a subset of customers will wait, and only for a limited time. Unfortunately, models that incorporate dynamic balking and reneging are notoriously intractable. There exist many structural results and simple optimal policies for inventory management with infinite backordering, but relatively few for systems with lost sales, and these few require strong assumptions (e.g. at most one order may be outstanding (Johansen and Thorstenson, 1993; 1996; Moinzadeh and Nahmias, 1988)) or approximations (Nahmias, 1979; Cohen, Kleindorfer and Lee, 1988; Johansen and Hill, 2000). The following papers provide analytic results for make-to-order manufacturing systems in which the customer arrival process depends on the static expected waiting cost, but not dynamic state information (Mendelson and Whang, 1990; Van Mieghem 1995, 2000; Armony and Haviv 2000; Lederer and Li, 1997; and Afeche 2004). Duenyas and Hopp (1995) were the first to study a make-to-order system in which the customer arrival process is shaped by dynamically quoting delivery leadtimes. Because dynamic leadtime quotation and scheduling in make-to-order systems is so complex, researchers employ heuristic algorithms, simulation and approximations (Duenyas and Hopp, 1995; Hopp and Sturgis, 2001; Keskinocak, Ravi and Tayur, 2001; Kapusckinski and Tayur, 2002; Plambeck, 2003). All the aforementioned papers model the make-to-order system with a single server queue. In contrast, we provide analytic results for multi-server systems.

Modeling dynamic balking and reneging is difficult but worthwhile, because one obtains qualitatively different managerial insights, and structurally different control policies. For example, Armony and Plambeck (2002) show that failure to account for duplicate ordering and reneging can cause either over- or under-investment in capacity. By incorporating capacity constraints and customer reneging into the well known Bass model, Ho, Savin, and Terwiesch (2002) obtain qualitatively different insights. Kumar and Swaminathan (2003) analyze a related model of new product introduction with balking rather than reneging and find optimal control policies that are structurally different. Plambeck (2004) analyzes an assemble-to-order system in which orders must be filled within a product-specific target leadtime, or they are lost. A simple policy with independent control of each component is near optimal. In contrast, when customers wait for as long as necessary to obtain the product, optimal control becomes more complex: component production and assembly sequencing depend upon the inventory positions for all components (Plambeck and Ward, 2003). In (Li and Lee, 1994) two firms compete by setting prices; customers observe queue lengths and jockey between the firms to minimize delivery-time. In contrast to traditional Bertrand equilibrium with zero prices and profits, because customer orders depend dynamically on

the leadtime, the firms sustain strictly positive profits. In a dynamic Bayesian formulation, Chen and Plambeck (2004) show the value of reducing inventory levels to learn about about customer's willingness to wait.

Most of the existing research assumes the simplest structure for reneging (customers renege after an exponentially distributed amount of time) or balking (balk with probability p if there is any wait, and with probability (1-p) wait until the product is delivered). Two notable exceptions are Ward and Glynn (2004) and Zeltyn and Mandelbaum (2004). Both papers allow general distributions for reneging and balking, and perform asymptotic analysis of these systems under conventional heavy traffic and the many-servers heavy traffic regimes, respectively. Mandelbaum and Shimkin (2000) derive complex dynamic customer behavior from primitives on valuation and waiting costs, for an M/M/S queue with congestion/failure shocks, assuming customers cannot observe the queue length.

Most of the literature on balking and reneging in queues focuses on performance evaluation and estimation (see, for example, Baccelli and Hebuterne (1981), Garnett, Mandelbaum and Reiman (2002), Mandelbaum and Zeltyn (1998) and Mandelbaum, Sakov and Zeltyn (2000) and references therein). One exception, (Kumar and Ward, 2005) proposes an admission control policy for a system with reneging in which revenue from admitting a customer is less than the penalty incurred if that customer later reneges. Recently, researchers have made rapid progress in staffing for call centers (Harrison and Zeevi (2005); Mandelbaum and Zeltyn (2005) and Borst, Mandelbaum, Reiman and Zeltyn (2005)), which involves determining the number of servers of possibly several pools and thus differs from our 1-dimensional model of capacity planning for a make-to-order system.

We derive fundamental properties of an M/M/S queue with state-dependent balking and reneging rates. We adopt the sample path approach of Shaked and Shanthikumar (1988) to verify convexity of stochastic processes and related cost functions. First, in Section 3, assuming that the balking probability and reneging rate are increasing and concave in the head-count, we prove that the expected head-count is convex decreasing in the capacity (service rate). This is complementary to the famous result that in a G/G/1 queue with a convex increasing delay cost and without balking/reneging, a customers' expected cost of delay is a convex decreasing function of capacity (Weber, 1983). Second, in Section 4, we assume linear reneging and constant balking probability, and show that the expected lost sales rate is convex decreasing in the capacity. This is similar to the result by Fridgeirsdottir and Chu (2005) that in a G/G/1 queue with convex nondecreasing delay cost and without

balking/reneging, the expected delay cost rate is convex increasing in the arrival rate, and to the result by Janakiraman and Roundy (2004) that for an inventory system with lost sales and stochastic sequential leadtimes, expected discounted cost is convex in the base stock level. Establishing that the expected cost function is convex in the control parameter (capacity, arrival rate and base stock level in the preceding examples) justifies using a simple search procedure to compute the optimal parameter level and sets the stage for deriving qualitative insights from comparative statics.

Inspired by Shaked and Shantikumar's (1988) concept of sample-path convexity, in Section 4 we employ a sample-path sub-modularity approach to comparative statics. That is, we employ sample path arguments to show how the optimal capacity changes as we vary the parameters of customer demand and impatience. We find that the optimal capacity increases in the demand rate and decreases with the balking probability, but is *not* monotone in the reneging rate. This means, surprisingly, that failure to account for customers' impatience and reneging may result in *over*-investment in capacity.

Finally, in Section 5, we assume commitment during service, i.e., customers cannot balk or renege during service. This seemingly minor change in system structure produces qualitatively different results. The expected rate of reneging from the system in steady-state is convex, but the expected rate of balking and hence expected cost is non-convex in some parameter regions. Furthermore, the optimal capacity is no longer monotone in the balking probability. We conclude that commitment during service strongly impacts the convexity properties and comparative statics of make-to-order systems with impatient customers.

2. Notation and Model Formulation

Consider a make-to-order system modelled by a multi-server, infinite-buffer queue. Customers arrive at the system according to a Poisson process with rate λ. The service time has an exponential distribution with rate μ. We denote the number of customers in the system (head-count) by Y. An arriving customer may decide to balk, namely, to leave upon arrival. The balking probability is a function of the head-count, and is denoted by $\beta(\cdot)$. Finally, customers may decide to cancel their order (renege) at any point during their wait or while being served. The reneging rate is a function of the head-count, and is denoted by $\eta(\cdot)$. All arrivals, service times, balking and reneging are assumed to be independent. Therefore, the head-count process is a continuous time Markov chain (CTMC).

The system manager knows the customers characteristics modelled here by λ, β, and η, wishes to choose the service capacity μ to minimize the cost associated with lost sales and capacity investment:

$$C(\mu) = C(\mu; \lambda, \eta, \beta) = c[\lambda E\beta(Y(\infty)) + E\eta(Y(\infty))] + k\mu,$$

where $Y(\infty)$ is the head-count in steady-state, and, without loss of generality, it is assumed that $c = 1$.

Let θ be an arbitrary parameter, and let $\mu^*(\theta)$ be a value of μ that minimizes a certain function $g(\mu; \theta)$. The meaning of the saying '$\mu^*(\theta)$ is increasing in θ' when $\mu^*(\theta)$ is not necessarily unique, is that if $\theta_L < \theta_H$, and μ_H minimizes $g(\mu; \theta_H)$ then there exists $\mu_L \leq \mu_H$, such that μ_L minimizes $g(\mu; \theta_L)$. Similarly, if μ_L minimizes $g(\mu; \theta_L)$, then there exists $\mu_H \geq \mu_L$, such that μ_H minimizes $g(\mu; \theta_H)$. Throughout the paper we use the term increasing to mean non-decreasing, and the term decreasing to mean non-increasing.

3. Convexity of Cost in Capacity

In this section we address the issue of convexity of the cost function in the capacity variable μ. We start by establishing the convexity of the expected head-count as a function of the capacity. This convexity property is true for very general balking and reneging functions. The only requirements is that both functions are non-decreasing and concave in the head-count.

Proposition 1 *Let $Y(t)$ denote the head-count process for an $M/M/S$ system with reneging and balking. Suppose that the reneging rate $\eta(\cdot)$ and the balking probability $\beta(\cdot)$ are both non-decreasing and concave functions of Y, then the expected head-count in steady state, $EY(\infty)$, is convex in the service rate, μ.*

The proof of Proposition 1 is based on the sample path approach. In particular, we prove that Y satisfies *sample path convexity* (a term that has been introduced by Shaked and Shanthikumar (1988)). More specifically, for any service rates $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$, we show that there exist $Y_1, ..., Y_4$, which are versions of the original head-count processes (Y_i has service rate of μ_i), and which satisfy the following two properties for all $t \geq 0$:

1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$, a.s, and

2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

Then, according to Shaked and Shanthikumar (1988), Y is said to be stochastically decreasing and convex in the sample path sense (SDCX(sp)). From Theorem 3.6, Proposition 2.11 and Remark 2.8 of Shaked and Shanthikumar (1988) it follows that $EY(\infty)$ is decreasing and convex in μ. To prove that 1. and 2. hold, we discretize time, uniformize the transition rates, and finally prove 1. and 2. using path-wise coupling and induction on time. Details are given in the Appendix.

This convexity result is particularly interesting in the case that customers cannot observe the head-count. Note that (with a slight abuse of notation) both simple functions $\eta(y) = \eta y$ for some constant $\eta \geq 0$, and $\beta(y) = \beta$ for some constant $1 \geq \beta \geq 0$ satisfy the assumptions of Proposition 1. Both these functions are likely scenarios when the head-count is unobservable. The reneging function corresponds to the case where every order is cancelled if is not fulfilled by an exponential amount of time with mean $1/\eta$. Similarly, the balking probability function corresponds to the case where the customer is not aware of the head-count and makes her balking decision at random. The cost function that goes along with these reneging and balking functions is $C(\mu; \lambda, \eta, \beta) = \lambda\beta + \eta EY(\infty) + k\mu$. It is easily verified (see Proposition 2) that Proposition 1 implies the convexity of this cost function in μ.

A similar result to proposition 1 appears in Shaked and Shanthikumar (1988) (Theorem 5.5). The similarity is that both our result and theirs assume that the departure rate is increasing and concave in the head-count. The difference is in the conclusions. They show that for a single server queue the head-count is increasing and convex in the *arrival* rate, while we show that for a multi-server queue the head-count is decreasing and convex in the *service* rate.

Convexity of the cost as a function of capacity implies that efficient optimization algorithms can be applied to find the capacity level which minimizes cost. In addition, this convexity allows for comparative statics that evaluate the effect of changes in the model parameters on the optimal capacity level. The latter is pursued in the next section.

4. Optimal Capacity Investment

In this section we investigate the effect of varying fundamental system parameters on the optimal capacity investment. We assume that each customer balks with probability $\beta > 0$ (regardless of the head-count upon arrival) and reneges after an exponential time with rate

$\eta > 0$ (This is the case when the head-count is unobservable). As one would expect, we find that the optimal capacity is increasing in λ and decreasing in β. Surprisingly, we find that the optimal capacity may either increase or decrease in η.

Theorem 1 *Suppose that the balking probability function is constant at $\beta(y) = \beta$, for some $0 \leq \beta < 1$, and that the reneging rate function is $\eta(y) = \eta y$, for some $\eta \geq 0$. Let*

$$C(\mu; \lambda, \eta, \beta) = \lambda\beta + \eta EY(\infty) + k\mu, \tag{1}$$

be the cost function associated with lost sales and capacity investment. Let $\mu^*(\lambda, \eta, \beta)$ be the optimal capacity that minimizes $C(\mu; \lambda, \eta, \beta)$. Then, $\mu^*(\lambda, \eta, \beta)$ is non-decreasing in λ, non-increasing in β, but is not necessarily monotone in η.

The proof of Theorem 1 is based on four propositions. The first shows that $C(\mu; \lambda, \eta, \beta)$ is convex in μ; therefore, for all given values of λ, η, and β, $\mu^*(\lambda, \eta, \beta)$ is well defined (although it may be non-unique), and any local minimum of $C(\mu; \lambda, \eta, \beta)$ is also a global minimum. The second proposition shows that $\mu^*(\lambda, \eta, \beta)$ is increasing in λ. Similarly, the third proposition shows that $\mu^*(\lambda, \eta, \beta)$ is decreasing in the balking probability β. Finally, in the fourth proposition we show that $\mu^*(\lambda, \eta, \beta)$ may either increase or decrease in η. Most arguments are based on the sample path approach. Building on the concept of *sample path convexity,* we define the *sample-path sub-modularity* property the implies monotonicity of the expected-cost-minimizing value of one parameter in a second parameter (Theorem 2).

Proposition 2 *Under the assumptions of Theorem 1, the cost function $C(\mu; \lambda, \eta, \beta)$ is convex in μ for all values of λ, η and β.*

Proof: Fix λ, η and β and let $f(\mu) = \eta EY(\infty)$. Clearly, the convexity of f in μ implies that C is also convex in μ. To establish the convexity of f, note that the reneging rate and the balking probability functions are both non-decreasing and concave, and therefore, by Proposition 1 it follows that $f(\mu)$ is convex in μ. $\square$

In order to show that the optimal capacity is increasing in λ and decreasing in β we introduce and utilize the following concept of sample-path sub-modularity.

Definition: Let $X = X_{\gamma,\delta}$ be a stochastic process which depends on the two parameters γ and δ. We say that X is "*path-wise sub-modular*" with respect to γ and δ if for all $\gamma_L < \gamma_H$ and $\delta_L < \delta_H$ we have four processes $\hat{X}_{\gamma,\delta}, \ \ \gamma = \gamma_L, \gamma_H, \ \delta = \delta_L, \delta_H$ which are defined on the same probability space, such that

1. $\hat{X}_{\gamma,\delta}$ is a version of $X_{\gamma,\delta}$ for every fixed pair (γ,δ) (that is, $\hat{X}_{\gamma,\delta} \stackrel{st}{=} X_{\gamma,\delta}$), and

2. $\hat{X}_{\gamma_H,\delta_H} - \hat{X}_{\gamma_H,\delta_L} \leq \hat{X}_{\gamma_L,\delta_H} - \hat{X}_{\gamma_L,\delta_L}$, a.s.

The next theorem establishes the connection between the sample-path sub-modularity property and monotonicity.

Theorem 2 *Let $X = X_{\gamma,\delta}$ be a stochastic process, and let $g(\gamma,\delta) = EX_{\gamma,\delta}$ be its expected value in steady-state. Suppose that $g(\cdot)$ is convex in γ for every fixed δ and that it is path-wise sub-modular with respect to these two variables. Let $\gamma^*(\delta)$ be the (possibly non-unique) value of γ that minimizes $g(\gamma,\delta)$ for every fixed value of δ, then $\gamma^*(\delta)$ is increasing in δ.*

Proof: Let δ_L, δ_H be two values of δ such that $\delta_L < \delta_H$. Let $\gamma^*(\delta_L)$ be a value of γ that minimizes $g(\gamma,\delta_L)$. We need to show that there is $\hat{\gamma} \geq \gamma^*(\delta_L)$ such that $\hat{\gamma}$ minimizes $g(\gamma,\delta_H)$. By contradiction, assume that for all optimal solutions $\gamma^*(\delta_H)$ of $g(\gamma,\delta_H)$, we have $\gamma^*(\delta_H) < \gamma^*(\delta_L)$. In particular,

$$0 \leq g\left(\gamma^*(\delta_H),\delta_L\right) - g\left(\gamma^*(\delta_L),\delta_L\right) \leq g\left(\gamma^*(\delta_H),\delta_H\right) - g\left(\gamma^*(\delta_L),\delta_H\right) \leq 0, \tag{2}$$

where the first inequality follows from the optimality of $\gamma^*(\delta_L)$, the second on follows from the sample-path sub-modularity and the assumption that $\gamma^*(\delta_H) < \gamma^*(\delta_L)$, and the third one follows from the optimality of $\gamma^*(\delta_H)$. In particular, (2) implies that $g\left(\gamma^*(\delta_H),\delta_H\right) = g\left(\gamma^*(\delta_L),\delta_H\right)$, which in turn implies that $\gamma^*(\delta_L)$ is minimizes $g(\gamma,\delta_H)$. This leads to a contradiction. □

The next proposition establishes that the head-count process is path-wise sub-modular in μ and λ (β and η will be omitted from the current expressions for expository purposes). From Theorem 2 it then follows that $\mu^*(\lambda)$ is non-decreasing in λ.

Proposition 3 *For any values of λ and μ, let $Y_{\lambda,\mu}$ represent the head-count process when the arrival rate is λ and the service capacity is μ. Then $Y_{\lambda,\mu}$ is path-wise sub-modular in λ and μ.*

Note that the proposition only establishes the path-wise sub-modularity of $Y_{\lambda,\mu}$. However, it is straightforward to verify that this implies the sample-path sub-modularity of the entire cost function in these two parameters. The proof of Proposition 3 follows the sample path

approach. More specifically, we show that for all $\lambda_L < \lambda_H$ and $\mu_L < \mu_H$ there exist versions of $Y_{\lambda,\mu}$, for $\lambda \in \{\lambda_L, \lambda_H\}$ and $\mu \in \{\mu_L, \mu_H\}$, such that the following three properties hold at all times $t \geq 0$:

I. $Y_{\lambda_H,\mu_L}(t) = \max[Y_{\lambda,\mu}(t) : \ \lambda \in \{\lambda_L, \lambda_H\}, \ \mu \in \{\mu_L, \mu_H\}]$, a.s.,

II. $Y_{\lambda_L,\mu_H}(t) = \min[Y_{\lambda,\mu}(t) : \ \lambda \in \{\lambda_L, \lambda_H\}, \ \mu \in \{\mu_L, \mu_H\}]$, a.s., and

III. $Y_{\lambda_H,\mu_H}(t) - Y_{\lambda_H,\mu_L}(t) \leq Y_{\lambda_L,\mu_H}(t) - Y_{\lambda_L,\mu_L}(t)$, a.s.

Similarly to the proof of Proposition 1, we show I. II. and III. using time discretization and uniformization, and then establishing these properties using sample-path coupling and induction on time. Details are given in the appendix.

The next proposition establishes the monotonicity of the optimal capacity in the balking probability. Specifically, we show that if the balking probability function is constant then the optimal capacity is decreasing (in fact, non-increasing) in this constant balking probability.

Proposition 4 *Let λ and η be fixed. Then the optimal capacity $\mu^*(\beta)$ which minimizes the cost $C(\mu; \lambda, \eta, \beta)$ is non-increasing in β.*

Proof: Consider another system with arrival rate equal to $\lambda(1-\beta)$, no balking (balking probability $= 0$), and reneging rate η. It is easy to see that the head-count process for the new system evolves the same as the head-count process for the original system. According to Proposition 3 the optimal capacity that minimizes $C(\mu; \lambda(1-\beta), \eta, 0)$ is non-decreasing in $\lambda(1-\beta)$ and is, therefore, non-increasing in β. But the actual cost we seek to minimize is $C(\mu; \lambda, \eta, \beta) = C(\mu; \lambda(1-\beta), \eta, 0) + \lambda\beta$. Since this additional term is not a function of μ, it follows that $\mu^*(\lambda, \eta, \beta) = \mu^*(\lambda(1-\beta), \eta, 0)$, and hence is also non-increasing in β. $\square$

Corollary 1 *Let η be fixed. Then the optimal capacity $\mu^*(\lambda, \beta)$ which minimizes the cost $C(\mu; \lambda, \eta, \beta)$ is non-increasing in the balking rate ($\lambda\beta$).*

Proof: The proof follows immediately from the proof of Proposition 4. $\square$

The final proposition establishes that $\mu^*(\lambda, \eta, \beta)$ may be either increasing or decreasing in η. This counterintuitive result will be contrasted in the discussion with the traditional model of infinite backordering, in which such phenomenon does not occur. This underlines the importance of modelling order cancellation explicitly.

Proposition 5 *Fix the values of λ and β, and let $\mu^*(\eta)$ be the optimal capacity which minimizes the cost function $C(\mu;\lambda,\eta,\beta)$, then $\mu^*(\eta)$ can either increase or decrease in η.*

Proof: Recall the cost function $C(\mu;\eta) := C(\mu;\lambda,\eta,\beta) = \beta\lambda + \eta EY(\infty) + k\mu$. Suppose that $S = 1$ and $\beta = 0$. In order to prove the proposition we first show that for arbitrary values of λ and k with $0 < k < 1$, $\mu^*(\eta)$ may decrease in η. To show that, we note that the definitions $C(\mu;\eta = 0) = (\lambda - \mu)1_{\{\mu \le \lambda\}} + k\mu$, and $C(\mu;\eta = \infty) = \lambda + k\mu$ are continuous extensions of the cost function $C(\cdot)$ for all values η in the closed interval $[0,\infty]$. However, notice that $\mu^*(\eta = 0) = \lambda$, whereas, $\mu^*(\eta = \infty) = 0$, that is, $\mu^*(\eta)$ may decrease with η.[4]

To show that $\mu^*(\eta)$ may also increase in η, all we have to show is that there exist $0 < k < 1$ and $\eta_k > 0$ such that $\mu_k^*(\eta_k) > \lambda$ (recall that $\mu^*(\eta = 0) = \lambda$). We show that, in fact, a stronger result applies; namely, that for every fixed value of $\eta > 0$, there exists a value $k = k(\eta)$, $0 < k < 1$, such that $\mu_k^*(\eta) > \lambda$, where $\mu_k^*(\eta)$ stands for the optimal capacity that minimizes the cost function $C(\mu;\lambda,\eta,\beta) = \eta EY(\infty) + k\mu$. To show that, fix the value of $\eta > 0$, and note that the function $f(\mu) = \eta EY(\infty)$ is decreasing and convex in μ (Proposition 1). We claim that it is sufficient to show that:

$$\text{There exists } \mu_0 \text{ such that: } \mu_0 > \lambda, \quad f(\mu_0) < f(\lambda) \quad \text{and} \quad f'(\mu_0^-) > -1, \tag{3}$$

where $f'(\mu_0^-)$ is the directional derivative of f at $\mu = \mu_0$ from below (exists due to Lemma 3.1.5 of Bazaraa, Sherali and Shetty (1993)). If (3) is true then the convexity of $f(\mu)$ implies that $f'(\mu_0^-) \le f'(\mu_0^+)$ (here, $f'(\mu_0^+)$ is the directional derivative of f at $\mu = \mu_0$ from above). Let k be such that $f'(\mu_0^-) \le -k \le f'(\mu_0^+)$, then $C'(\mu_0^-) = f'(\mu_0^-) + k \le 0$, and $C'(\mu_0^+) = f'(\mu_0^+) + k \ge 0$. In particular, $\mu_k^*(\eta) = \mu_0$ is a local minimum for $C(\cdot)$, and from convexity, it is also a global minimum.

To establish (3), note that from flow conservation $f(\mu) = \eta EY(\infty) = \lambda - \mu P(Y(\infty) > 0)$. In particular, $f(\mu = \lambda) > 0$, and $\lim_{\mu\to\infty} f(\mu) = 0$. Since $f(\mu)$ is a non-increasing function of μ, this implies that there exists $\mu_1 > \lambda$ such that $f(\mu) < f(\lambda)$, for all $\mu \ge \mu_1$. Now note that if $f'(\mu^-) \le -1$ for all $\mu \ge \mu_1$, then $f(\mu) < 0$ for μ large enough, which is a contradiction. This shows that μ_0 is well defined. □

[4]The continuity of $\mu^*(\eta)$ (which follows from Theorem 3.1.3 of Bazaraa, Sherali and Shetty (1993) and the implicit functions theorem) may be used to show that $\mu^*(\eta)$ indeed decreases for some points on the interval $(0,\infty)$.

5. Customer Commitment During Service

For many service systems and some make-to-order manufacturing systems, it is reasonable to assume that customers will not balk or renege during service. According to Farlie (2004), small manufacturers of customized computers charge a customer's credit card before initiating assembly, to prevent cancellations during the assembly process. The assumption that customers cannot balk or renege during service (which we call 'customer commitment') makes derivation of convexity and comparative statics results much more difficult. In fact, some of our previous results are no longer true under this assumption.

To illustrate the effect of customer commitment during service on convexity we use the simplest form of balking and reneging that falls within this framework. Specifically, throughout this section, we assume that the balking probability function is of the form $\beta(Y) = \beta 1_{\{Y \geq S\}}$ and the reneging rate function is of the form $\eta(Y) = \eta(Y - S)^+$, for some positive constants $\beta \leq 1$ and η. These balking and reneging functions are likely scenarios when customers cannot observe the head-count but are aware of whether their service is in progress, is about to begin, or is going to be delayed. Consequently, each customer will balk with probability β if and only if no server is available when she arrives. Similarly, she will renege after an exponential time with rate η as long as she is waiting in line. The expected balking rate and reneging rate in steady-state associated with the above functions are $b(\mu; \lambda, \eta, \beta) = \lambda\beta P(Y(\infty) \geq S)$ and $r(\mu; \lambda, \eta, \beta) = \eta E[Y(\infty) - S]^+$, respectively.

In this section we also allow for the cost associated with a customer balking (c_b) to differ from the cost associated with a customer reneging (c_r). Let

$$C(\mu; \lambda, \eta, \beta) = c_b \lambda \beta P(Y(\infty) \geq S) + c_r \eta E[Y(\infty) - S]^+ + k\mu, \tag{4}$$

denote the cost function associated with lost sales and capacity investment. It is straightforward to see that all our results in the previous sections hold when the cost of balking differs from the cost of reneging. Surprisingly, with customer commitment during service, important system properties (convexity of the cost (4) as a function capacity μ and monotonicity of the optimal capacity μ^* in the balking probability β) depend upon the relative costs of balking and reneging.

We start by establishing that the expected reneging rate from the system in steady-state is a convex function of μ for $\mu \geq \eta$.

Proposition 6 *Suppose that $\eta > 0$ and let*

$$r(\mu; \lambda, \eta, \beta) = \eta E[Y(\infty) - S]^+, \tag{5}$$

denote the expected reneging rate in steady-state. Then if either $S = 1$ or $\beta = 0$ then $r(\cdot)$ is convex in μ for $\mu \geq \eta$.

Before proving this proposition we introduce the following Lemma:

Lemma 1 *Under the assumptions of Proposition 6, the head-count process Y is stochastically decreasing and convex in μ for $\mu \geq \eta$.*

The proof of Lemma 1 appears in the Appendix. It is similar to the proof of Proposition 1, but it does not follow for this proposition because the assumptions of the concavity of the balking probability and reneging rate in the head-count do not hold.

Proof of Proposition 6: Suppose that $\mu \geq \eta$. By Lemma 1 the head-count process Y is stochastically decreasing and convex in μ. Now, since the function $\eta(y) = \eta(y - S)^+$ is increasing and convex in y, it follows that the process $\eta(Y - S)^+$ is also stochastically decreasing and convex in μ. Finally, it follows that $r(\mu; \lambda, \eta, \beta) = \eta E[Y(\infty) - S]^+$ is decreasing and convex in μ. $\square$

In contrast, the expected balking rate from the system in steady-state is *non-convex* in μ, when the reneging rate η is small. This result seems counter-intuitive, especially in light of Lemma 1. However, note that in the customer commitment case, the balking probability $\beta 1_{\{Y \geq S\}}$ is not convex in the head-count, and therefore the convexity of this rate in μ does not follow from Lemma 1. The direct sample-path argument for convexity also fails. To see this, note that the direct approach requires establishing sample path convexity for $1_{\{Y \geq S\}}$, analogously to the proof of Proposition 1. However, the quantity $1_{\{Y \geq S\}}$ does not carry enough information for such arguments to work. For example, consider four systems with service rates $\mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4$, and $\mu_1 + \mu_4 = \mu_2 + \mu_3$. Pathwise convexity requires that

$$1_{\{Y_1(t) \geq S\}} + 1_{\{Y_4(t) \geq S\}} \geq 1_{\{Y_2 \geq S\}} + 1_{\{Y_3 \geq S\}}, \quad \forall t \geq 0. \tag{6}$$

But (6) could work at time t_0, with $Y_1(t_0) = Y_4(t_0) = S$ and $Y_2(t_0) = Y_3(t_0) = S + 1$. In this case, the next departure from all systems will result in $Y_1(t_0) = Y_4(t_0) = S - 1$, and $Y_2(t_0) = Y_3(t_0) = S$, which violates (6). The next proposition states the non-convexity of the balking rate in μ.

Proposition 7 *Let*

$$b(\mu;\lambda,\eta,\beta) = \lambda\beta P(Y(\infty) \geq S), \tag{7}$$

denote the expected balking rate in steady-state. Then $b(\cdot)$ is not necessarily convex in μ. In particular, if $S = 1$, then for sufficiently small values of η, $b(\cdot)$ is not convex in μ.

Proof: Let $S = 1$, and fix λ and β. First we examine the limit of the balking rate function as the reneging rate η approaches zero. Note that, from the monotone convergence theorem, we have that

$$b(\mu) := \lim_{\eta \downarrow 0} b(\mu;\lambda,\eta,\beta) = \begin{cases} \lambda\beta, & \mu \leq \lambda(1-\beta), \\ \frac{\lambda^2\beta}{\mu+\lambda\beta}, & \mu > \lambda(1-\beta). \end{cases}$$

In particular, $b(\mu)$ is not convex at the point $\mu_0 = \lambda(1-\beta)$. Moreover, if we let $\mu_1 = \mu_0/2$ and $\mu_2 = 3\mu_0/2$, then $\mu_0 = (\mu_1+\mu_2)/2$, but $b(\mu_0) > (b(\mu_1)+b(\mu_2))/2$. We next show that for sufficiently small values of $\eta > 0$, $b(\mu;\eta) := b(\mu;\lambda,\eta,\beta)$ is not convex in μ. Let $\epsilon = b(\mu_0) - (b(\mu_1)+b(\mu_2))/2$. By the definition of $b(\mu)$, there exists $\eta(\epsilon)$ such that $|\, b(\mu;\eta) - b(\mu)| < \epsilon/3$, for all $\eta \leq \eta(\epsilon)$ and $\mu = \mu_0, \mu_1, \mu_2$. This implies, that for all $\eta \leq \eta(\epsilon)$

$$b(\mu_0;\eta) - \frac{b(\mu_1;\eta)+b(\mu_2;\eta)}{2} \geq b(\mu_0) - \frac{\epsilon}{3} - \frac{b(\mu_1)+\frac{\epsilon}{3}+b(\mu_2)+\frac{\epsilon}{3}}{2} > 0.$$

□

Figure 1 illustrates the non-convexity of the balking rate as a function of μ for the special case where $S = 1$, $\lambda = 50$, $\eta = 0.5$ and $\beta = 0.2$.

In light of the proof of Proposition 7, one might think that the non-convexity in μ may only occur if we allow for traffic intensity ($\rho := \lambda(1-\beta)/\mu$) values which are greater than or equal to 1. An exhaustive numerical search over the parameter values reveals that this is not the case. Specifically, for high traffic intensity which is close to 1 (but still less than 1) and low reneging rate, the balking rate is not convex. This numerical result is illustrated in Figure 2. The surface in the figure displays, for each pair of values of β and (λ/μ), the highest value of η for which the steady-state expected rate of balking is non-convex in the capacity μ. That is, nonconvexity occurs below the surface in Figure 2 and convexity above. Note that for an arrival rate $\lambda \neq 1$, the value of η in the figure would be scaled by λ, but the region of non-convexity in the $(\beta, \lambda/\mu)$ space will not change.

Finally, we prove that the convexity properties of the cost function (4) depend upon the relative costs of balking and reneging. Proposition 8 focuses on the single-server case; we conjecture that similar convexity properties hold for the multi-server case ($S > 1$) but have not been able to prove this.

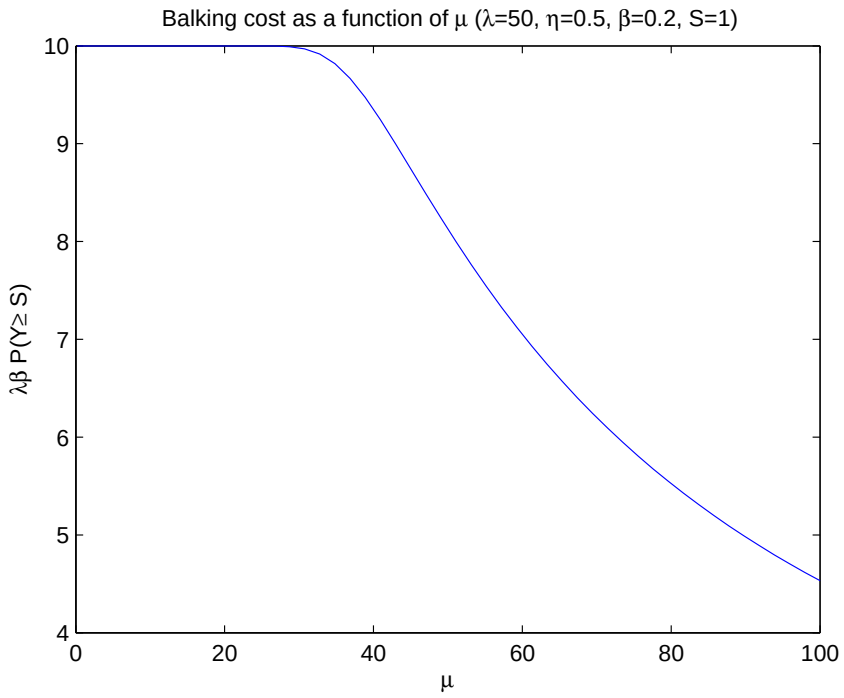

Figure 1: Steady-state expected balking rate is non-convex in capacity μ

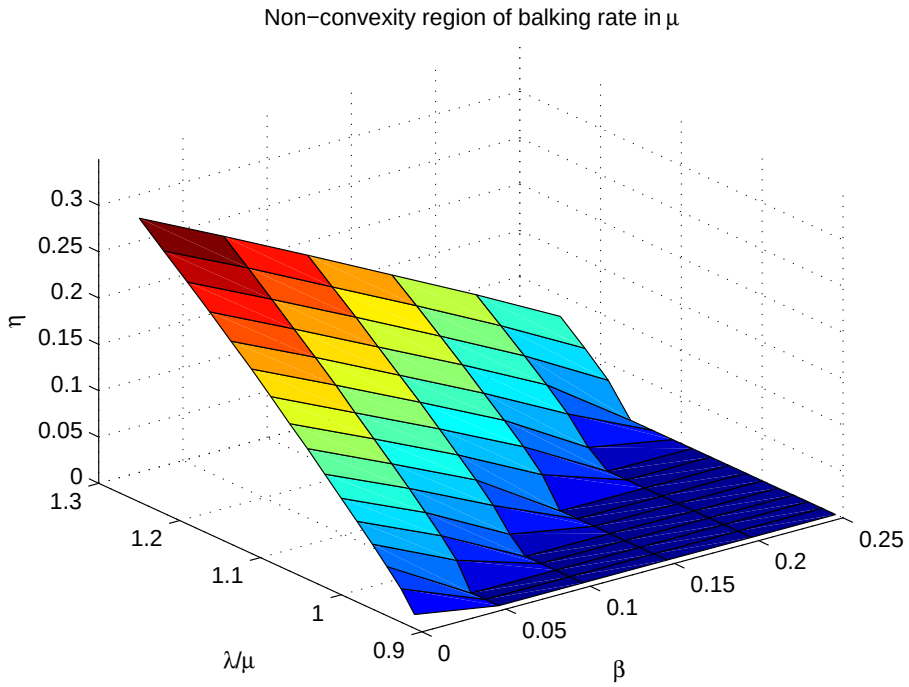

Figure 2: M/M/1 systems non-convexity region for steady-state expected balking rate as function of capacity μ (for arrival rate $\lambda = 1$).

Proposition 8 *Suppose that $S = 1$. If $c_r \geq c_b$, then the cost function $C(\cdot)$ defined in (4) is convex in μ for all $\mu \geq \eta$. Also, if $c_b\beta \leq c_r < c_b$ then $C(\cdot)$ is convex in μ for $\mu \geq \max\{\eta, \lambda(1-\beta)\}$. On the other hand, if $c_r < c_b\beta$, then, for sufficiently small $\eta > 0$, $C(\cdot)$ is non-convex in μ.*

Proof: Fix λ, β, and $\eta > 0$ (to be omitted as parameters of $C(\cdot)$ for brevity). Suppose that $\mu \geq \eta$ and that $c_r = c_b = 1$. We will prove that $C''(\mu) \geq 0$. From this, convexity of $C(\cdot)$ in μ for any c_r and c_b satisfying $c_r \geq c_b$ follows immediately from Proposition 6. From flow conservation we have that

$$C(\mu) = \lambda - \mu P(Y \geq 1) + k\mu.$$

Let $P(\mu) = P(Y \geq 1)$. Then,

$$C'(\mu) = -\mu P'(\mu) - P(\mu) + k,$$

and

$$C''(\mu) = -2P'(\mu) - \mu P''(\mu).$$

Therefore, if

$$-2P'(\mu) \geq \mu P''(\mu), \tag{8}$$

then $C(\cdot)$ is convex. Notice that $P(\cdot)$ is decreasing in μ. Therefore, the right-hand-side of (8) is non-negative. Hence, if the left-hand-side of (8) is negative the proof is complete. Otherwise, if $P''(\mu) \geq 0$, then the balking rate $\lambda\beta P(\mu)$ is convex in μ. In this case, by (4), we only need to establish the convexity of the reneging rate in μ. But this has been shown in Proposition 6.

Suppose now that $c_b\beta \leq c_r < c_b$ and that $\mu \geq \max\{\eta, \lambda(1-\beta)\}$. We show that if $c_r = 1$ and $c_\beta = 1/\beta$, then $C(\cdot)$ is convex in μ. Convexity for the general $c_b\beta \leq c_r < c_b$ case will immediately follow from Proposition 6. If indeed $c_r = 1$ and $c_\beta = 1/\beta$, then, from flow conservation,

$$C(\mu) = \lambda + (\lambda(1-\beta) - \mu)P(Y \geq 1) + k\mu.$$

Recall that $P(\mu) = P(Y \geq 1)$. Then,

$$C'(\mu) = (\lambda(1-\beta) - \mu)P'(\mu) - P(\mu) + k,$$

and

$$C''(\mu) = -2P'(\mu) + (\lambda(1-\beta) - \mu)P''(\mu).$$

Therefore, if

$$-2P'(\mu) \geq (\mu - \lambda(1-\beta))P''(\mu), \tag{9}$$

then $C(\cdot)$ is convex. Following considerations similar to those in the previous case, and noting that $\mu - \lambda(1-\beta) \geq 0$, we conclude the convexity of $C(\cdot)$ in μ for this region.

Finally, suppose that $c_r < c_b\beta$. We show that the limit of the cost of lost sales as $\eta > 0$ approaches zero is not convex. The rest will follow analogously to the proof of Proposition 7. Note that, from flow conservation and from the monotone convergence theorem, we have that

$$\begin{aligned}\tilde{C}(\mu): &= \lim_{\eta\downarrow 0}\{c_b\lambda\beta P(Y(\infty)\geq 1) + c_r\eta E[Y(\infty)-1]^+\} \\ &= \begin{cases} c_b\lambda\beta + c_r(\lambda(1-\beta)-\mu), & \mu \leq \lambda(1-\beta), \\ c_b\frac{\lambda^2\beta}{\mu+\lambda\beta}, & \mu > \lambda(1-\beta). \end{cases}\end{aligned}$$

Let $\mu_0 = \lambda(1-\beta)$. Then the left derivative of $\tilde{C}(\cdot)$ at $\mu = \mu_0$ is

$$\tilde{C}'(\mu = \mu_0^-) = -c_r.$$

Also, its right derivative at $\mu = \mu_0$ is

$$\tilde{C}'(\mu = \mu_0^+) = -c_b\beta.$$

Clearly, if $c_r < c_b\beta$, then $\tilde{C}(\cdot)$ is not convex at μ_0. $\square$

In light of the above, one might think that convexity properties change but, fundamentally, comparative statics do not. Surprisingly, customer commitment destroys one of the monotonicity results obtained in the previous section. Recall that according to Theorem 1 the optimal capacity is increasing in the arrival rate and decreasing in the balking probability for the non-commitment case. For the case of customer commitment during service, while we believe that the monotonicity in the arrival rate still holds, we prove that the optimal capacity is not necessarily monotone in the balking probability.

Proposition 9 *Let $C(\mu;\lambda,\eta,\beta)$ be the cost function defined in (4) and let $\mu^*(\lambda,\eta,\beta)$ be the optimal capacity that minimizes $C(\cdot)$. Then, $\mu^*(\cdot)$ is not necessarily monotone in the balking probability β. In particular, with a single server ($S = 1$), for all sufficiently small η, the optimal capacity $\mu^*(\lambda,\eta,\cdot)$ exhibits non-monotonicity in β.*

Proof: Suppose that $S = 1$, $c_b = c_r = 1$ and fix $\lambda > 0$. The limit of the cost function as $\eta \downarrow 0$ satisfies:

$$C(\mu;\beta) := \lim_{\eta\downarrow 0} C(\mu;\lambda,\eta,\beta) = \begin{cases} \lambda - (1-k)\mu, & \mu \leq \lambda(1-\beta), \\ \frac{\lambda^2\beta}{\mu+\lambda\beta} + k\mu, & \mu > \lambda(1-\beta). \end{cases}$$

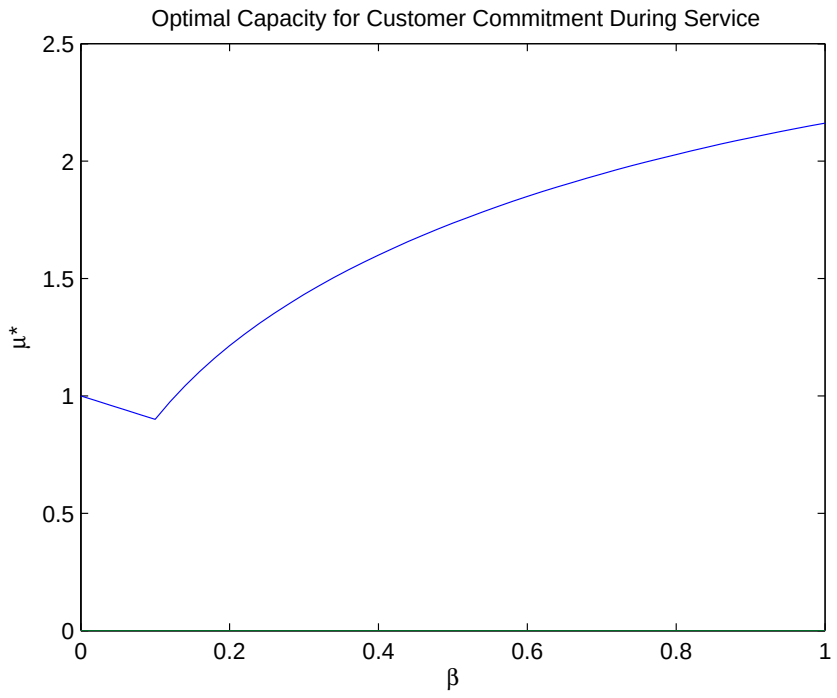

Figure 3: The optimal capacity is non-monotone in the balking probability ($S = 1, \lambda = 1, k = 1$ and $\eta \downarrow 0$).

It is easy to see that the optimal value of μ that minimizes $C(\mu; \beta)$ satisfies:

$$\mu^*(\beta) = \begin{cases} \lambda(1-\beta), & k \geq \beta, \\ \lambda\left(\sqrt{\beta/k} - \beta\right), & k < \beta. \end{cases}$$

Suppose that $\lambda = 1$ and $k = 0.1$. In this case, as shown in Figure 3, $\mu^*(\beta)$ is first decreasing in β and then it is increasing. In particular, for $\beta_1 = 0 < \beta_2 = 0.1 < \beta_3 = 0.2$, we have $mu^*(\beta_1) > \mu^*(\beta_2) < \mu^*(\beta_3)$ and $\mu^*(\beta)$ is non-monotone. Arguments analogous to the proof of Proposition 7 show that for sufficiently small values of η, $\mu^*(\beta, \eta)$ is not necessarily monotone in β. $\square$

Intuitively, if $\eta \downarrow 0$, then for relatively small values of β, the dominant cost is the cost of capacity. In this case, the optimal capacity is the minimum that guarantees stability ($\mu = \lambda(1-\beta)$), which is decreasing in β. For higher values of β, if k is sufficiently small, then the dominant cost becomes the balking cost. In this case, to counteract the increasing balking rate, the optimal μ is increasing in β.

We conclude that customer commitment during service strongly influences the convexity properties and comparative statics of make-to-order systems with impatient customers.

6. Discussion

We have derived convexity properties for the cost of capacity and lost sales, as a function of capacity, and evaluated how balking and reneging influence the optimal capacity investment. Some of our results (particularly Proposition 5) are counterintuitive. Furthermore, we show that a seemingly minor change in system structure, customer commitment during service, leads to qualitatively different results. These results underline the importance of painstakingly accounting for balking and reneging in the design of make-to-order or service systems.

Proposition 5 demonstrates an important difference between systems with backordering costs and systems with reneging. It suggests that, before investing in capacity, managers need to carefully model and estimate customers' willingness to wait for their orders to be fulfilled. Surprisingly, a larger reneging rate does not necessarily imply that greater capacity is needed. To contrast this result with the more traditional models of inventory theory, suppose that one assumes that customers will wait indefinitely for their order, but the manufacturer will incur a backordering cost in addition to the cost of capacity. In this paper's notation, one can write down the cost function as $\widetilde{C}(\mu;\lambda) = cEY(\infty) + k\mu$, where c is the cost per backlogged order per time unit, and no balking or reneging occurs. Note that $\widetilde{C}(\mu;\lambda)$ and $C(\mu;\lambda,\eta,\beta=0)$ are almost identical in form. The one crucial difference is that $EY(\infty)$ in the infinite-backordering model is independent of its coefficient c, whereas $EY(\infty)$ in our model depends on its coefficient η in a non-trivial manner. In particular, in the infinite-backordering model, the optimal capacity is always increasing in the backordering cost c. In contrast, the optimal capacity in a system with reneging may *decrease* with the reneging rate. The operations management literature widely assumes infinite backordering (rather than lost sales) for analytic tractability. Customer impatience is represented by a high backordering cost c, which is said to account for the "loss of good will" from forcing customers to wait. The striking qualitative difference in results (that optimal capacity always increases with c in infinite backordering model but may decrease with η in model with explicit reneging) shows that, in deriving qualitative or structural insights, one cannot rely on a backorder penalty to represent customers' impatience. More specifically, in making decisions about capacity investment for a make-to-order system, failure to explicitly account for reneging may result in *over*-investment in capacity. Further research is needed to understand the implications of balking and reneging for more general production-inventory systems.

A. Appendix: Proofs

Proof of Proposition 1: We first prove the proposition for the single-server case ($S = 1$). The general multi-server case is dealt with at the end of this proof. The proof is based on the sample path approach. Specifically, we prove that Y (viewed as a function of μ) satisfies *sample path convexity* (a term that has been introduced by Shaked and Shanthikumar (1988)). Specifically, let $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ be four service rates such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$, and fix λ, $\beta(\cdot)$ and $\eta(\cdot)$. Suppose that there exist $Y_1, ..., Y_4$, which are versions of the original head-count processes (Y_i has service rate of μ_i) that satisfy the following two properties for all $t \geq 0$:

1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$, a.s.

2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

Then, according to Shaked and Shanthikumar (1988), Y is said to be stochastically decreasing and convex in the sample path sense (SDCX(sp)). From Theorem 3.6, Proposition 2.11 and Remark 2.8 of Shaked and Shanthikumar (1988) it follows that $EY(\infty)$ is decreasing and convex in μ.

To construct the coupled versions $Y_1, ..., Y_4$ we wish to come up with appropriate uniformized discrete versions of the original processes. However, for uniformization to work one needs bounded transition rates of the original Markov chain, which is not the case in this paper (we do *not* assume boundedness of the reneging rates $\eta(y)$). To resolve this problem we define for all $M > 0$ a truncated reneging function $\eta_M(y) = \min\{\eta(y), M\}$. Clearly, since $\eta(\cdot)$ is concave, and $\min\{\cdot, M\}$, is non-decreasing and concave, $\eta_M(\cdot)$ is also concave. Moreover, for any fixed $M > 0$, $\eta_M(\cdot)$ is bounded. Let $Y_1^M, ..., Y_4^M$ be uniformized discrete versions of the head-count processes with arrival rate λ, balking probability function $\beta(\cdot)$, service capacity μ_i, $i = 1, ..., 4$, and reneging rate function $\eta_M(\cdot)$. We will show that for each $M > 0$ and for every $n \in \mathbb{Z}_+$ properties 1. and 2. hold at time n, with respect to $Y_1^M, ..., Y_4^M$. It will then follow that $EY^M(\infty)$ is decreasing and convex in μ. But since $Y^M(\infty)$ weakly converges to $Y(\infty)$[5] it follows from Proposition 2.11 of Shaked and Shanthikumar (1988) that $EY(\infty)$ is a decreasing and convex function of μ.

[5]This can be shown by writing down the stationary distributions of the corresponding birth and death processes explicitly, and show that those distributions converge to the limiting one, with unbounded reneging rates.

We now fix $M > 0$, and establish, by induction, that if 1. and 2. hold at time $n = 0$ for Suppose that 1. and 2. hold for $Y_1^M, ..., Y_4^M$, then they hold for all $n = 1, 2, ...$ For brevity, we omit the superscript M from the subsequent terms. In addition to 1. and 2. we define a third property as follows:

$\widetilde{1}$. $Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n)$,

that is, property $\widetilde{1}$. is property 1. with an equality replacing the inequality. We first establish that if properties $\widetilde{1}$. and 2. are satisfied at time n, then properties 1. and 2. hold at time $n + 1$. Let $v = \lambda + \mu_4 + M$. be an upper bound on the total transition rate of the processes $Y_1, ..., Y_4$. For n, such that $\widetilde{1}$. and 2. hold, we define the following possible uniformized and coupled transitions:

Arrival + balking: With probability $\frac{\lambda}{v}$ we have a new order arriving into all four systems. When a new order arrives, it balks system i with probability $\beta(Y_i(n))$. This is done as follows: Let $Y_{(1)}(n) \geq Y_{(2)}(n) \geq Y_{(3)}(n) \geq Y_{(4)}(n)$ be the order statistics for $Y_i(n)$, $i = 1, .., 4$. Respectively, refer to system (i) as the systems whose head-count is $Y_{(i)}(n)$. Note that from properties $\widetilde{1}$ and 2, it follows that $Y_{(1)}(n) = Y_1(n)$ and $Y_{(4)}(n) = Y_4(n)$. Now let $\beta_i = \beta(Y_{(i)}(n))$. From the monotonicity and concavity of $\beta(\cdot)$ is follows that:

a. $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$,

b. $\beta_1 + \beta_4 \leq \beta_2 + \beta_3$.

Now, let $U \sim Uniform(0, 1)$. U will determine in which systems the order just arrived will immediately balk according to the following rules:

i. If $U \leq \beta_4$, then balk in all four systems.

ii. Else, if $U \leq \beta_2 + \beta_3 - 1$, then balk from queues $1, (2)$ and (3).

iii. Else, if $U \leq \beta_3$, then balk in queues (3) and 1 only.

iv. Else, if $U \leq \beta_1$, then balk in queues (2) and 1 only.

v. Else, if $U \leq \beta_2 + \beta_3 - \beta_4$, balk in queue (2) only.

To verify that the balking occur according to the right probabilities, note that in systems $1, (3)$ and 4 the balking probabilities are trivially equal to the required probabilities. In queue (2), if $\beta_2 + \beta_3 - \beta_4 < 1$ balking will occur with probability: $\beta_4 + (\beta_2 + \beta_3 - \beta_4 - \beta_3) = \beta_2$. Similarly, if $\beta_2 + \beta_3 - \beta_4 \geq 1$, balking in this queue will occur with probability: $(\beta_2 + \beta_3 - 1) + (1 - \beta_3) = \beta_2$.

Service Completion: With probability $\frac{\mu_4}{v}$ we have a service completion event. To determine which systems are going to indeed have service completions (as opposed to a transition from a state to itself), let $U \backsim Uniform(0, 1)$.

a. If $U < \frac{\mu_1}{\mu_4}$ we have service completions from all systems for which $Y_i(n) > 0$.

b. If $\frac{\mu_1}{\mu_4} \leq U < \frac{\mu_2}{\mu_4}$, we have departures in systems 2 and 4 only, whenever the corresponding queues are non-empty.

c. If $\frac{\mu_2}{\mu_4} \leq U < 1$, we have departures in systems 3 and 4 only, whenever the corresponding queue are non-empty.

It is easy to see, that system i has a service completion with probability $\frac{\mu_i}{v}$ as long as $Y_i(n) > 0$ (recall that $\mu_1 + \mu_4 = \mu_2 + \mu_3$). Note that the reason why we do not simply have a service completion from system i whenever $U < \frac{\mu_i}{\mu_4}$, is that in this case we may have a service completion from system 4 only, which may violate property 1.

A Reneging Job (order cancellation) : Finally, with probability $[\eta(Y_i(n)) \wedge M]/v$ we have an order cancellation from system i. The coupling works as follows: let $Y_{(1)}(n) \geq Y_{(2)}(n) \geq Y_{(3)}(n) \geq Y_{(4)}(n)$ be the ordered statistics of $Y_1(n), ..., Y_4(n)$, and let $\xi_{(i)} = \eta_M(Y_{(i)}(n)) = \min\{\eta(Y_{(i)}(n)), M\}$. Note that property $\widetilde{1}$. and the convexity of $\eta_M(\cdot)$ imply that $\xi_{(1)} + \xi_{(4)} \leq \xi_{(2)} + \xi_{(3)}$ (that is, the inequality with respect to the ξ_i's is the opposite of property 1.) Let $U \backsim Uniform(0, 1)$ be the random variable that determines the reneging from all systems. Let $m = \max\{M, \xi_{(3)} + \xi_{(2)} - \xi_{(4)}\}$.

a. If $U < \frac{\xi_{(4)}}{m}$, we will have one order cancellation from all the systems such that $Y_i(n) > 0$.

b. If $\frac{\xi_{(4)}}{m} \leq U < \frac{\xi_{(3)}}{m}$, we have one order cancellation from each of the systems (3) and (1) (provided that $Y_{(i)}(n) > 0$, for $i = 1, 3$).

c. If $\frac{\xi_{(3)}}{m} \leq U < \frac{\xi_{(1)}}{m}$, we have one order cancellation from each of the systems (2) and (1) (provided that $Y_{(i)}(n) > 0$, for $i = 1, 2$).

d. If $\frac{\xi_{(1)}}{m} \leq U < \frac{\xi_{(3)}+\xi_{(2)}-\xi_{(4)}}{m}$, we have one order cancellation from system (2), provided that $Y_{(2)}(n) > 0$.

Note that given this setup, an order cancellation occurs in system (i) with probability $[\eta(Y_{(i)}(n)) \wedge M]/v$.

We will now show that if properties $\widetilde{1}$. and 2. hold at time n, 1.-2. are satisfied at time $n+1$. We will go over the different types of events, to show that 1.-2. still hold at time $n+1$:

Arrival + balking: Since we have arrivals coming into all systems at the same time, properties 1.-2. will still hold at time $n+1$, if no balking occurs. To verify that properties 1. and 2. hold at time $n+1$ in case of balking note that these can happen only if from time n to $n+1$ one of the following occurs:

I. The LHS of 1. stays the same, while the RHS of 1. increases by 1 or 2: This will only occur when there is balking in both queues 1 and 4, which implies balking in queues (2) and (3) as well.

II. The LHS of 1. increases by 1, while the RHS of 1. increases by 2: This change in the LHS of 1. can only occur when the arrival to queue 4 does not balk, while the arrival to queue 1 balks. However, in this case, at least one of the arrival to queue (2) or (3) will balk.

III. $Y_i(n) = Y_1(n)$ for some $i \neq 1$, and Y_1 stays the same, while Y_i increases by 1 (this will violate 2.): This would occur only if $Y_{(2)}(n) = Y_1(n)$ and there will be balking in queue 1 and not in queue (2). However, if $Y_{(2)}(n) = Y_1(n)$, then $Y_{(3)}(n) = Y_4(n)$, and in particular $\beta_3 = \beta_4$. In this case, it is easily verified that balking in queue 1 implies balking in queue (2) as well.

Service Completion: Here we have to make sure we are avoiding the following:

I. The LHS of 1. decreases by 1, while the RHS does not change:

II. The LHS of 1. decreases by 2, while the RHS decreases by 1 or does not change.

III. $Y_i = Y_1$ for some $i \neq 1$, and Y_1 decreases by 1, while Y_i does not change (hence property 2. is violated).

Observe that none of these can happen because whenever $Y_i = 0$ for either $i = 2$ or 3, we have $Y_4 = 0$. Moreover, if $Y_i = Y_1$, then if Y_1 decreases, Y_i will also decrease.

Order Cancellation: In this case, properties 1. - 2. will be violated if any of the above I.-III. occur. We show that this cannot happen by going over the different values of the uniform variable U. First note that here $Y_{(1)} = Y_1$ and $Y_{(4)} = Y_4$. Without loss of generality, assume that $Y_{(2)} = Y_2$, and $Y_{(3)} = Y_3$, and omit the $(\cdot)$ from the subscript. Also, recall that $\xi_1 + \xi_4 \leq \xi_2 + \xi_3$.

a. If $U < \frac{\xi_4}{m}$, then $Y_4(n) > 0$, which implies that $Y_i(n) > 0$ for all i, which means that all values of $Y_i(n)$ will be reduced by 1.

b. If $\frac{\xi_4}{m} \leq U < \frac{\xi_3}{m}$, then $Y_3(n) > Y_4(n)$. This implies that $Y_1(n) > Y_2(n)$ (from property $\widetilde{1}$.), and therefore the fact that $Y_1(n)$ and $Y_3(n)$ are the only processes reduced by 1, will not violate 1.-2.

c. If $\frac{\xi_3}{m} \leq U < \frac{\xi_1}{m}$, then $Y_1(n) > Y_3(n)$. This implies that $Y_2(n) > Y_4(n) \geq 0$ (see property $\widetilde{1}$.), and therefore the fact that $Y_1(n)$ and $Y_2(n)$ are the only processes reduced by 1, will not violate1.-2.

d. If $\frac{\xi_1}{m} \leq U < \frac{\xi_2+\xi_3-\xi_4}{m}$, then $1. - 2.$ will clearly not be violated.

So far we have shown that if at time n properties $\widetilde{1}$. and 2. hold, then at time $n+1$ both properties 1 and 2 will hold. Suppose that at time n property 1. holds with a strict inequality, that is:

$$Y_1(n) + Y_4(n) > Y_2(n) + Y_3(n).$$

In order to describe the transitions in this case, we first define the following transformation of $Y_1(n)$ and $Y_4(n)$: $\widetilde{Y}_1(n) = \max\{0, Y_2(n) + Y_3(n) - Y_4(n)\}$ and $\widetilde{Y}_4(n) = \min\{Y_2(n) + Y_3(n), Y_4(n)\}$. It is easy to see that $\widetilde{Y}_i(n) \leq Y_i(n)$ for $i = 1, 4.$ and that $\widetilde{Y}_1(n) + \widetilde{Y}_4(n) = Y_2(n) + Y_3(n)$. That is, property $\widetilde{1}$. holds for the modified values of $Y_i(n)$. Let $\widetilde{Y}_i(n+1), i = 1, 2, 3, 4$, be the values of these processes after one transition, that occurred according to the above rules. In particular, we know that properties 1. and 2. hold for $\widetilde{Y}_i(n+1), i = 1, 2, 3, 4$. Let $\overline{F}_{i,x}(y) = P_{\mu=\mu_i}\{Y_i(n+1) > y \mid Y_i(n) = x\}$, then it is easy to verify that $\overline{F}_{i,x}(y)$ is

non-decreasing in x. In particular, for $i = 1, 4$, let $Y_i(n+1) = \overline{F}^{-1}_{i,Y_i(n)}(\overline{F}_{i,\widetilde{Y}_i(n)}(\widetilde{Y}_i(n+1)))$, and for $i = 2, 3$, simply let $Y_i(n+1) = \widetilde{Y}_i(n+1)$. One can now easily verify that for all i, $Y_i(n+1) \geq \widetilde{Y}_i(n+1)$, that properties 1. and 2. hold for $Y_i(n+1), i = 1, .., 4$, and that $Y_i(n+1)$ has the right distribution (i.e. for all y, $P_{\mu=\mu_i}\{Y_i(n+1) > y | Y_i(n)\} = \overline{F}_{i,Y_i(n)}(y)$. This completes the proof of the Proposition for the single server case.

It is left to prove the proposition for the general multiserver ($S > 1$) case. To extend the above proof to the M/M/S system, the only case that needs to be treated is service completions. We first establish the following relation:

If $\mu_1 + \mu_4 = \mu_2 + \mu_3$, $Y_1 + Y_4 = Y_2 + Y_3$ and Y_1 is greater than $\max\{Y_2, Y_3\}$ then $\mu_1 Y_1 + \mu_4 Y_4 \leq \mu_2 Y_2 + \mu_3 Y_3$.

However, $\mu_4 Y_1 + \mu_4 Y_4 = \mu_4 Y_2 + \mu_4 Y_3$. Therefore, in order to prove the relation it is sufficient to show: $(\mu_1 - \mu_4)Y_1 \leq (\mu_2 - \mu_4)Y_2 + (\mu_3 - \mu_4)Y_3$. The last is true because $\mu_i - \mu_4 \leq 0$, $\mu_1 - \mu_4$ is equal to $(\mu_2 - \mu_4) + (\mu_3 - \mu_4)$, and Y_1 is greater than $\max\{Y_2, Y_3\}$.

Once we have this relation in hand it follows that: $\mu_1 \min\{Y_1, S\} + \mu_4 \min\{Y_4, S\} \leq \mu_2 \min\{Y_2, S\} + \mu_3 \min\{Y_3, S\}$.

To see this, notice that because Y_1 and Y_4 are more spread out than Y_2 and Y_3 and because min is a concave function, $\min\{Y_1, S\} + \min\{Y_4, S\} \leq \min\{Y_2, S\} + \min\{Y_3, S\}$. (For a proof assume that there are two random variables, the first of which takes values Y_1 and Y_4 with probability 0.5 each, whereas the second takes the other two values with equal probability. The random variables have the same expected value but one dominates the other in the convex order.)

This observation and the earlier proved relation complete the proof that $\mu_1 \min\{Y_1, S\} + \mu_4 \min\{Y_4, S\} \leq \mu_2 \min\{Y_2, S\} + \mu_3 \min\{Y_3, S\}$.

Finally, this shows that we can couple the four systems such that the second and third have more service completions on each sample path and that $\widetilde{1}$. and 2. hold at each service completion.

□

Proof of Proposition 3: The proof follows the sample path approach. In particular, we discretize time, and uniformize the transition rates in an analogous way to what was done in the proof of Proposition 1. Specifically, we bound the reneging rate from above by M, and after we prove the result for any M, we let $M \to \infty$, to get the desired result. Given a value of M, we show that we have sample-path sub-modularity for all n. More specifically, suppose

that the following three properties hold at time $n = 0$, for all $\lambda_L < \lambda_H$ and $\mu_L < \mu_H$:

I. $Y_{\lambda_H,\mu_L}(n) = \max\{Y_{\lambda,\mu}(n)\ ;\ \lambda = \lambda_L, \lambda_H,\ \mu = \mu_L, \mu_H\}$,

II. $Y_{\lambda_L,\mu_H}(n) = \min\{Y_{\lambda,\mu}(n)\ ;\ \lambda = \lambda_L, \lambda_H,\ \mu = \mu_L, \mu_H\}$,

III. $Y_{\lambda_H,\mu_H}(n) - Y_{\lambda_H,\mu_L}(n) \leq Y_{\lambda_L,\mu_H}(n) - Y_{\lambda_L,\mu_L}(n)$,

then we show by induction that they hold for all $n \geq 0$.

Suppose that $S = 1$, and let $v = \lambda_H + \mu_H + M$. That is, v is the maximal transition rate in all four systems given any state. Now suppose that I.-III. hold at time n, where III. holds with an equality (we will call this property $\widetilde{\text{III}}$). In this case we have three types of transitions:

Arrival + balking: With probability $\frac{\lambda_H}{v}$ we have an arrival event. The coupling works as follows: let $U \backsim Uniform(0,1)$.

1. If $U < \frac{\lambda_L}{\lambda_H}$ we have one arrival into each of the four systems.
2. If $U \geq \frac{\lambda_L}{\lambda_H}$ we have arrivals into the systems with $\lambda = \lambda_H$ only.

Once it has been determined which systems will have new arrivals, these new arrivals all balk together with probability β, and otherwise they join the queue.

Service Completion: With probability $\frac{\mu_H}{v}$ we have a service completion event. To determine which systems have a departure, let $U \backsim Uniform(0,1)$.

1. If $U < \frac{\mu_L}{\mu_H}$ we have a service completion for each one of the systems for which $Y_{\lambda,\mu}(n) > 0$.
2. If $U \geq \frac{\mu_L}{\mu_H}$ we have a service completion for those systems with $\mu = \mu_H$ only, whenever $Y_{\lambda,\mu_H}(n) > 0$.

Order Cancellation: With probability $\frac{M}{v}$ we have an order cancellation event. Let $\eta_M(y) = \min\{\eta y, M\}$ be the reneging rate function. Let $Y_{(i)}, i = 1,2,3,4$ be a permutation of $\{Y_{\lambda,\mu}(n); \lambda = \lambda_L, \lambda_H, \mu = \mu_L, \mu_H\}$ such that $Y_{(1)} \geq Y_{(2)} \geq Y_{(3)} \geq Y_{(4)}$. Let $\xi_{(i)} = \eta_M(Y_{(i)})$. Note that I.,II, and $\widetilde{\text{III}}$. and the concavity of $\eta_M(\cdot)$ imply that $\xi_{(1)} + \xi_{(4)} \leq \xi_{(2)} + \xi_{(3)}$. Finally, let $m = \max\{M, \xi_{(2)} + \xi_{(3)} - \xi_{(4)}\}$. To determine which systems have a service cancellation, let $U \backsim Uniform(0,1)$.

1. If $U < \frac{\xi_{(4)}}{m}$, we have a service cancellation from each one of the systems, provided that the corresponding head-count is positive.
2. If $\frac{\xi_{(4)}}{m} \leq U < \frac{\xi_{(3)}}{m}$, we have a service cancellation in systems (1) and (3), provided that $Y_{(i)} > 0, i = 1, 3$.
3. If $\frac{\xi_{(3)}}{m} \leq U < \frac{\xi_{(1)}}{m}$, we have a service cancellation in systems (1) and (2), provided that $Y_{(i)} > 0, i = 1, 2$.
4. If $\frac{\xi_{(1)}}{m} \leq U < \frac{\xi_{(2)}+\xi_{(3)}-\xi_{(4)}}{m}$, we have a service cancellation is system (2), provided that $Y_{(2)} > 0$.

Verifying that if $I., II.$, and $\widetilde{III}$ hold at time n, then $I., II.$ and $III.$ hold at time $n+1$ is straightforward, and is analogous to proving Proposition 1. We omit the details. If instead of $\widetilde{III}$, we have III at time n, proceed similarly to the proof of the same proposition to validate the induction step. If $S > 1$ proceed similarly to the general proof of Proposition 1, realizing that the only case to be concerned about is the service completion. However, the service rates of the four systems being compared can be ordered as $\mu_1, ..., \mu_4$ as in the proof of Proposition 1. Therefore, this part of the proof extends without modifications.

This completes the proof of the proposition. □

Proof of Lemma 1: Following the notation of the proof of Proposition 1, let $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ be four service rates such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$. Assume that the reneging rate η is bounded above by μ_4. This is a weaker condition than the one stated in the Lemma, but it turns out to be sufficient in establishing the its results.

Analogously to the proof of Proposition 1, let $Y_1, ..., Y_4$ be discretized and uniformized versions of the head-count with service rates $\mu_1, ..., \mu_4$, respectively, that satisfy properties:

1. $Y_1(n) + Y_4(n) \geq Y_2(n) + Y_3(n)$, a.s.
2. $Y_1(n) \geq \max\{Y_2(n), Y_3(n), Y_4(n)\}$, a.s.

at time $n = 0$. By induction, we wish to show that properties 1. and 2. hold for all $n \geq 0$.

The induction proof of 1. and 2. goes through by the simple construction explained next. Note that arrivals, balking and service completion do not introduce a problem. For reneging,

one can transfer customers from system 1 to system 4 until one of two events happens: either Y_4 equals the minimum of Y_2 and Y_3, or Y_4 equals S. In the first case, after the transfer Y_1 will equal the maximum of Y_2 and Y_3. In the second case all systems will have S or more customers. The transfer will not decrease the rate at which queues deplete in systems 1 and 4 due to the assumption on the reneging rate. Moreover, $\widetilde{1}.$ (or 1.) and 2. will continue to hold. It thus follows that the induction proof goes through after this modification. In detail, in the first case the two sets of systems will have equal reneging rate. In the second case, $(Y_1 - S) + (Y_4 - S) = (Y_2 - S) + (Y_3 - S)$. The reneging rates depend on these four quantities and the earlier proof for Proposition 1 goes through.

Notice that if the condition $\eta \leq \mu_4$ does not hold then the induction step will not work. For example, if $S = 1$, then when $Y_1 = 2,\ Y_2 = Y_3 = 1,\ Y_4 = 0$, reneging can take place only in the first system. Thus, 1. will get violated when there is a reneging. Similarly, if $\beta > 0$ and $S > 1$, then the induction will not work either. For example, if $S = 2$, then when $Y_1 = 2,\ Y_2 = Y_3 = 1,\ Y_4 = 0$, balking may only occur in system 1, and if it does occur condition 1. will again be violated. Therefore, it appears that the conditions of the lemma are not only sufficient but also necessary. □

Acknowledgments

This research was supported by the National Science Foundation under grant DMI-0239840.

References

Afeche, P. 2004, Incentive-compatible revenue management in queueing systems: optimal strategic idleness and other delaying tactics, working paper. Kellogg School of Management.

Armony, M. and M. Haviv. 2000. Price and delay competition between two service providers. European Journal of Operational Research 147(1) 32-50.

Armony, M. and E. L. Plambeck. 2002. The impact of duplicate orders on demand estimation and capacity investment. forthcoming in Management Science.

Baccelli, F. and Hebuterne G. 1981. On queues with impatient customers. In: F.J. Kylatra (Ed.), Performance '81. North-Holland Publishing Company, 159-179.

M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. 1993. Nonlinear Programming: Theory and Algorithms. 2nd edition. John Wiley & Sons, Inc.

Borst, S., Mandelbaum, A., Reiman M. and Zeltyn, S. 2005. Dimensioning call centers with abandonment. In preparation.

Chen, L. and E.L. Plambeck. 2004. Dynamic inventory management with learning about the demand distribution and substitution probability. Working Paper, Stanford Graduate School of Business, Stanford, CA.

Cohen, M., P. Kleindorfer, and H. Lee. 1988. Service constrained (s,S) inventory systems with priority demand classes and lost sales. Management Science 34 (4) 482-499.

Fridgeirsdottir, K. and S. Chiu. 2005 A note on convexity of the expected delay cost in single server queues, forthcoming in Operations Research 53 (3).

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

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

Fairlie, R. 2004. How a custom PC can come with some very alien payment policies. Computer Shopper, March 10.

Garnett O., Mandelbaum A. and Reiman M. 2002. Designing a Call Center with Impatient Customers. Manufacturing and Service Operations Management, 4(3), 208-227.

Harrison and Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. Manufacturing and Service Operations Management 7 (1) 20-36.

Ho, T.-H., S. Savin and C. Terwiesch. 2002. Managing demand and sales dynamics in new product diffusion. Management Science, 48 (2) 187-206.

Hopp, W.J. and M.R. Sturgis. 2001. A simple and robust leadtime-quoting policy. Manufacturing and Service Operations Management 3 (4) 331-336.

Janakiraman, G. and R.O. Roundy. 2004. Lost-sales problems with stochastic leadtimes: convexity results for base-stock policies. Operations Research. 52 (5) 795-803.

Johansen, S.G. and R.M. Hill. 2000. The (r,Q) control of a periodic-review inventory system with continuous demand and lost sales. International Journal of Production Economics 68 279-286.

Johansen, S.G. and A. Thorstenson. 1993. Optimal and approximate (Q,r) inventory policies with lost sales and gamma-distributed leadtimes. International Journal of Production Economics 30-31 179-194.

Johansen, S.G. and A. Thorstenson. 1996. Optimal (r,Q) inventory policies with Poisson demands and lost sales: discounted and undiscounted cases. International Journal of Production Economics 46-47 359-371.

Kapuscinski, R. and S. Tayur. 2002. Reliable due date setting in a capacitated MTO system with two customer classes. Michigan Business School Working Paper.

P. Keskinocak, R. Ravi, S. Tayur. 2001. Scheduling and reliable lead time quotation for orders with availability intervals and lead time sensitive revenues. Management Science 47 (2) 264-279.

Kumar, S. and J. Swaminathan. 2003. Diffusion of innovations under supply constraints. Operations Research 51 (6) 866-879.

Lederer, P.J. and L. Li. 1997. Pricing, production, scheduling and delivery-time competition. Operations Research 45 (3) 407-420

Li, L. and Y.S. Lee. 1997. Pricing and delivery-time performance in a competitive environment. Management Science 40 (5) 633-646.

Mandelbaum A., Sakov A. and Zeltyn S. 2000. Empirical Analysis of a Call Center. Technical Report.

Mandelbaum, A. and N. Shimkin. 2000. A model for rational abandonments from invisible queues. Queueing Systems 36 (1-3) 1084-1134.

Mandelbaum A. and Zeltyn S. 1998. Estimating Characteristics of Queueing Networks Using Transactional Data. Queueing Systems 29, 75-127.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. Operations Research 38 (5) 870-883.

Moinzadeh, K. and S. Nahmias 1988. A continuous review model for an inventory system with two supply modes. Management Science 6 761-773.

Nahmias, S. 1979. Simple approximations for a variety of dynamic leadtime lost-sales inventory models. Operations Research 27 (5) 904-924.

Plambeck, E.L. 2004. Optimal leadtime differentiation via diffusion approximations. Operations Research 52 (2) 213-228.

Plambeck, E.L. 2004. Asymptotically optimal control for an assemble-to-order system with capacitated component production and fixed transport cost. Working Paper, Stanford Graduate School of Business, Stanford, CA.

Plambeck, E.L. and A.R. Ward. 2003. Optimal control of high-volume assemble-to-order systems, Working Paper, Stanford Graduate School of Business, Stanford, CA.

Van Mieghem, J.A. 1995. Dynamic scheduling with convex delay costs: the generalized cμ rule. Annals of Applied Probability 5 (3) 809-833.

Van Mieghem, J. 2000. Price and service discrimination in queueing systems: incentive compatibility of Gcμ scheduling. Management Science 46 (9) 1249-1267.

Ward, A.R. and P. Glynn. 2004. A diffusion approximation for a GI/GI/1 queue with balking or reneging. Working Paper, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Ward and Kumar. 2005. Asymptotically optimal admission control of a queue with impatient customers. Working paper. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Weber, R.R. 1983. A note on waiting times in single server queues. Operations Research 31 (5) 950-951.

Wein, L.M. 1991. Due date setting and priority sequencing in a multiclass M/G/1 queue.

Management Science 37 (7) 834-80.

Wein, L.M. and P. Chevalier. 1992. A broader view of the job-shop scheduling problem. Management Science 38 (7) 1018-1033.

Zeltyn S. and Mandelbaum A. 2004. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. Working paper.