Manufacturing & Service Operations Management

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

Speed-Quality Trade-Offs in a Dynamic Model

Vasiliki Kostami, Sampath Rajagopalan

To cite this article:
Vasiliki Kostami, Sampath Rajagopalan (2014) Speed-Quality Trade-Offs in a Dynamic Model. Manufacturing & Service Operations Management 16(1):104-118. http://dx.doi.org/10.1287/msom.2013.0458

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 1, Winter 2014, pp. 104–118
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2013.0458

Speed–Quality Trade-Offs in a Dynamic Model

Vasiliki Kostami
London Business School, London NW1 4SA, United Kingdom, vkostami@london.edu

Sampath Rajagopalan
Marshall School of Business, University of Southern California, Los Angeles, California 90089, srajagop@marshall.usc.edu

An important trade-off organizations face in many environments is one between quality and speed. Working faster may result in greater output and less delay, but may result in lower quality and dissatisfied customers. In this work, we consider dynamic models in a monopoly setting to explore the optimal balance among the multiple dimensions of speed, price, and wait time. The impact of quality is captured via the market demand potential, which is a function of the speed (quality) in the previous period. We obtain several results and insights. First, in scenarios where speed may be difficult to change over time (e.g., some automated production lines) but price can be changed, we show that the optimal price charged is such that the demand rate remains constant over time, even though the price and market potential are changing. Furthermore, we identify conditions when the firm will work at a speed that is higher or lower than a benchmark speed and characterize the behavior of prices over time. Second, in scenarios where a firm may not be able to change prices but can adjust the speed each period, the firm starts at a speed that may be faster or slower than a benchmark speed but converges to it over time. In this constant price case, as the benchmark speed increases, the initial speed adopted by the firm is actually lower but increases more quickly thereafter. We also characterize the behavior of price and speed in settings where both can be changed over time. Interestingly, a firm typically starts at a slow speed and increases the speed, price, and demand over time. Although our main model assumes that the firm internalizes the congestion cost, several of our results extend to a scenario where the demand rate is impacted by the congestion level.

Key words: service operations; quality management; dynamic programming; queuing theory
History: Received: February 14, 2012; accepted: June 8, 2013. Published online in *Articles in Advance* October 23, 2013.

1. Introduction

"I want a caffeine rush, not to be in a rush," says a Starbucks customer. Managing a customer-oriented business such as Starbucks is challenging and usually requires a careful trade-off of speed and quality. Consider a high-end restaurant that attempts to provide high-quality service to customers by paying careful attention to their needs. This results in highly satisfied customers who keep returning to the restaurant even if it charges high prices. At the same time, this may result in slower service and a longer undesirable wait. On the other hand, rushing the customers to provide quick service may result in lower satisfaction and unhappy customers who may never return. This problem becomes more interesting in a dynamic setting. As the restaurant provides great service, the happy customers not only return, but also tell their friends, and the restaurant will get more customers in the future. This will increase wait times in the future, and the restaurant may be forced to either raise the price or speed up the process. A restaurant has to carefully balance the multiple dimensions of speed, price, and wait time in effectively managing its operations. This balance gets challenging in a dynamic environment as it has to trade off the current and future effects of its decisions.

Many make-to-order manufacturing firms that promise a certain delivery lead time also have to balance similar trade-offs. Consider Coverking (http://www.coverking.com), a privately owned firm that designs and manufactures custom car accessories. Their value proposition is in producing custom, make-to-order products (e.g., car covers) designed to fit a particular vehicle with a short turnaround time. They are able to charge a higher price due to the customization and high quality. If prices are lowered, demand increases, but congestion levels and lead times increase and they cannot meet promised lead times. If they speed up the production process, quality often suffers Gupta and Rajagopalan (2012). In such manufacturing contexts, a customer may experience a lower-quality product that results from a higher speed, and this will, in turn, impact customer satisfaction and possibly future customer demand. Such products whose quality is known only after the service or product is purchased are referred to as *experience goods* (Nelson 1970).

The speed–quality trade-off has been discussed both in the academic and trade literature. Oliva and Sterman (2001) and Anand et al. (2011) provide several examples of such trade-offs. The coffee retailer Starbucks, whose main goal is to sell high-quality coffee, has asked its baristas to focus on one drink at a time and stop multitasking to ensure high-quality levels across its shops and reduce errors, recognizing the consequent longer wait times (Jargon 2010). In a study of endoscopy procedures, Cohen (2008) points out that increased emphasis on doing more procedures can result in more errors due to factors such as not following standard processes, working faster to finish the procedure, and fatigue. A study using over 20 years of data found that hospital pharmacists committed more prescription errors during periods of increased workload, which resulted in significantly higher fatalities (Croasmun 2005), and "the risk of error increases when the pharmacist fills more than 10–12 prescriptions per half-hour" (Flynn and Barker 2000, p. 34).

In this work, we consider the trade-offs discussed above between speed of processing and quality of the service (product) delivered and the role of price in balancing these trade-offs. We propose a multiperiod model that captures these trade-offs in a dynamic setting wherein current decisions about speed and price impact not only current period revenues and congestion costs, but also future profits. In particular, if the firm works at a speed faster than a benchmark speed, quality may suffer, and customers may be unhappy; conversely, working at a slower speed improves quality, but may increase congestion costs. The benchmark speed is the service rate or process speed that achieves a level of quality expected by customers, i.e., a benchmark quality standard; we discuss this aspect of the model in greater detail in §§2 and 3. In any period, the speed of the process relative to the "benchmark speed" determines the quality of the output and customer satisfaction, which impacts the demand potential in the next period. The actual demand realized in a period is a function of this demand potential and the price charged. Thus, the two levers the firm can use to optimize profits in each period are the price charged and the process speed.

We highlight a few interesting managerial insights from our analysis. Coverking has a process design and some equipment characteristics such that its production rate, once set, cannot be changed easily. Although this constant production speed may be faster or slower than the benchmark speed, we find that it should set its optimal price such that the demand rate remains constant over time. Furthermore, if the firm had a higher benchmark speed due to a different process or faster equipment, it would satisfy additional demand to the point where, counterintuitively, utilization and congestion are higher. Starbucks, on the contrary, rarely changes its prices over the course of a year, but constantly adjusts the pace of its employees to achieve the *Starbucks experience*. It may start at a slower or faster speed than the benchmark speed depending on its market potential and customer sensitivity to quality and wait times, but it should follow a policy wherein the speed gradually increases or decreases, respectively, to the benchmark speed over time and remains stable thereafter. We also find that a dynamic policy wherein either price or speed, or both, can be changed is far superior to a static policy, and furthermore, pricing flexibility is more valuable than speed flexibility in exploiting the speed–quality trade-off.

The rest of this paper is organized as follows. In §2, we discuss and highlight our contribution relative to the existing literature. In §3, we introduce our main model formulation, discuss our assumptions, and present the main insights from a single-period model. In §§4–6, we consider a multiperiod setting with three different scenarios wherein either speed or price is constant over time or both can vary over time. Unlike in §§3–6, where we assume that the congestion delay is a cost absorbed by the firm, in §7 we consider a scenario where congestion delay impacts the demand rate. Finally, in §8, we highlight the important takeaways, summarize the main results of two extensions, and propose further research directions. Henceforth, we use the terms *service* and *product* interchangeably because their role in the model is the same. We also refer to the *speed of the service* and the *service rate* interchangeably.

2. Literature Review

There is limited literature in operations management that studies the speed–quality interaction. Lovejoy and Sethuraman (2000) is among the earliest papers to consider the trade-off between quality and speed using an M/M/1 queuing model in a specific manufacturing context—rushing jobs to meet scheduled deadlines, which can reduce production yield and increase defects. They consider many details such as inspection, labor and material costs, etc., that we do not consider, but, unlike us, they do not consider price decisions and dynamic trade-offs. They use the concept of a "base" time for processing a unit, similar to the concept of benchmark speed in our work. In the same vein, Lu et al. (2009) consider the fact that the time spent with a product determines the probability of a good quality product, but our focus is very different. Hasija et al. (2010) discuss models of speed up behavior and explain why the staffing level cannot determine the actual service speed that can vary.

There exists some additional literature that studies the trade-off between quality and congestion, but these papers consider waiting time or service level as the dimension of quality, whereas we focus on a different dimension of quality, which is the customer experience with the service or product. Some examples of this literature are the paper by Allon and Federgruen (2007), which considers competition between firms in a service setting where the demand rate depends on the price and the service level, as well as the papers by Gans (2002) and Png and Reitman (1994), which measure quality in terms of the delay but use the idea of a threshold in the service speed that determines the quality level. Hall and Porteus (2000), also in a service competition setting, discuss the optimal dynamic capacity decision, where the customers can seek out another service provider when faced with service failures associated with low capacity. Hall and Porteus (2000) also use the concept of a benchmark speed, and in the paper by Gans (2002), distance from a benchmark speed determines the customer's satisfaction. Aflaki and Popescu (2010) model the evolution of service quality expectations and their impact on customer retention and profitability in a novel model that incorporates behavioral theories and bounded rationality. They show that the optimal service policy converges to a steady-state service level, a result similar to the one we obtain in the constant price case. Although we do not consider some of the behavioral aspects, unlike them, we explicitly capture the speed–quality trade-off.

In a novel model, Hopp et al. (2007) allow workers to adjust time spent with the customer depending on current workload and the value or revenue from the customer is an increasing function of the time spent. We do not assume this to be the case, and our approach is more appropriate for services where the price charged is based on a specific service provided and not a function of the time spent. Veeraraghavan and Debo (2009) present counterintuitive findings to explain why longer queues may imply better service in some environments.

The work closest to ours is by Anand et al. (2011). They model explicitly the dependence of service quality on service duration and quality–speed trade-offs in both single- and multiple-server queuing settings to determine equilibrium prices and speeds. Also, they use the concept of a benchmark service speed. Although our results are similar to theirs in the single-period setting, our focus is on a dynamic setting wherein the demand potential is a function of the speed in the previous period and so the firm has to consider the impact of its speed–quality trade-offs and pricing decisions on future profits too. Our main contribution lies in characterizing the optimal dynamic policy with respect to the service speed and price in three different scenarios, where either speed or price, or both, are changing over time. To our knowledge, there is no previous literature that has addressed such a dynamic trade-off between quality and speed.

The productivity, congestion, and quality trade-off often arises in the call center literature, although quality in this literature is often a function of wait time. Hasija et al. (2008) discuss call center outsourcing contracts where penalties are imposed when agreements on the wait times or service levels are not satisfied. In Ren and Zhou (2008), service quality depends on the staffing level; longer time with the customer often leads to better call resolution and thus to high service quality, an assumption similar to ours. De Véricourt and Zhou (2005) and Armony and Gurvich (2009) study the trade-offs between server speed and quality (that is associated with service failures in de Véricourt and Zhou (2005) and waiting times in Armony and Gurvich 2009) in making optimal call-routing decisions in a call center environment. There is no pricing decision in these works. However, Gurvich et al. (2009) include a pricing decision in a call center with cross-selling capabilities but do not consider speed–quality trade-offs.

Time-varying pricing strategies have been studied in the marketing literature in the product diffusion context, and Mahajan et al. (1990) provide a thorough review of the relevant work. Although speed and quality are not decision variables in these works, quality can be signaled through different pricing strategies, such as skimming and penetration.

There are a few recent empirical papers in the operations literature that explore the speed–quality trade-off. Kc and Terwiesch (2009) verify that overworking is associated with quality reduction in hospital operations, a situation also discussed by Needleman et al. (2002). Ren and Wang (2009) discuss an empirical study of the relationship between service quality and patient volume in hospitals. Oliva and Sterman (2001) provide some anecdotal evidence and show, using simulation, that working overtime may lead to quality issues. Also, they model customers' expectations as a function of the gap between the time allocated to the customer and their expectations (equivalent to speed being higher than benchmark speed in our approach); if the time falls below expectation, quality drops. Several experimental studies in the marketing literature (Gronroos 1983, Parasuraman et al. 1985, Zeithaml et al. 1993) show how the service quality is measured by comparing the service level with customers' expectations. Customers in anticipation of the service have some expectations based on prior experience and reputation of the provider and they "confirm" or "disconfirm" their expectations (Smith and Houston 1983, Churchill and Surprenant 1982) after the actual service occurs.

3. Model Setup

We consider a monopolist firm selling a service or a make-to-order product to homogenous consumers in a multiperiod setting. A high-end restaurant or a make-to-order firm that produces custom products (such as Coverking) has some monopolistic power with respect to pricing and so a monopoly model may be reasonable in such settings. The model and results discussed in this work are not applicable to firms selling make-to-stock products, because the ability to carry inventory will strongly influence both production rate and pricing decisions, as is clear from the inventory literature. There is a demand potential Λ_j at the beginning of each period j $(=1,\ldots,N+1)$ in the market, where $\Lambda_1 > 0$ is exogenously specified.

An important component of our work is the modeling of service speed and its impact on customer satisfaction. Consistent with the empirical findings and related prior works mentioned earlier, we assume that the quality of service increases with service time. We assume that customers are homogenous in their perception of the quality of the product. We define a *benchmark speed*, denoted as $\hat{\mu}$, to be the service speed at which the *expected quality level or a benchmark quality standard* is achieved. We assume that if the provider works faster (slower) than the benchmark speed, then the service has lower (higher) quality, and the company hurts (enhances) its reputation. Depending on the context, the benchmark quality standard may be based on some internal standards or norms at the firm or it could be based on customer expectations or a combination of both. This quality can be seen as a systemic quality, and the actual quality may be higher or lower depending on the actual speed. For example, in a call center, the benchmark speed may depend on the nature of the scripts used for greetings and other routine exchanges with the customer. Customer expectations about these exchanges may be different in different regions and so will be the benchmark speed. But a call center employee may work faster or slower than the benchmark speed depending on how he strays from the script. Even if customer expectations are the same, benchmark speeds could be different across two units within a firm because they are endowed with different processes or technologies. The benchmark speed is assumed to be exogenous and cannot be changed.

The impact of the quality experienced in a period is reflected in the market potential in the following period. In particular, if the speed is slower than the benchmark speed, quality will be higher than what customers were expecting and they will be happier, the firm's goodwill increases, and the market potential increases and vice versa. Consistent with the experience goods literature in economics (Nelson 1970), quality level is observed or experienced after purchase and can only impact future customer purchase behavior. Therefore, we define the demand potential recursively as

$$\Lambda_{j+1} = \Lambda_j - \delta\lambda_j(\mu_j - \hat{\mu}), \quad j = 1,\ldots,N, \tag{1}$$

where λ_j is the demand rate, and μ_j is the service speed (rate) in period j. The parameter $\delta > 0$ reflects the sensitivity to quality. Depending on whether the customer is part of the process or not, δ may reflect customer perceptions about the impact of speed on quality or the actual impact of speed on quality (e.g., defects) in the production process.

We include the term λ_j in the expression to capture the fact that the change in market potential is proportional to the current set of customers. When $\mu_j > \hat{\mu}$, a fraction of the current set of customers is not satisfied and will not return for service any longer. When $\mu_j < \hat{\mu}$, the additional market potential may come from more repeat purchases by existing customers or from new customers that are referred to by the highly satisfied customers in period j. This formulation also captures what has been shown in the customer loyalty literature (Hallowell 1996), that customer satisfaction accounts for 40% of the customer loyalty, which, in turn, translates to profitability for the firm (other factors include activities of competitors, the actual product, demographics, etc.). For simplicity we assume the loss (gain) of customers to be a linear symmetric function of the difference between actual service speed and benchmark speed. In reality, the impact of the difference between actual and benchmark speed on market potential may not be symmetric. We could extend the model using two different δ values to capture the asymmetry, but the model becomes more complex, and the results become dependent on more parameters and yield shallower insights. Also, we could replace $(\mu_j - \hat{\mu})$ by $(\mu_j - \hat{\mu})/\hat{\mu}$ to represent the percentage impact of service speed on market Potential, but this will not change our results; we effectively assume that $1/\hat{\mu}$ is incorporated in δ. The parameter δ controls the rate of change in market potential and plays a key role. We have assumed that the change in market potential is deterministic. In reality, the change in market potential may be stochastic, and this could be captured by making δ random.

A period in the model is not long, typically a month, and correspondingly, the time horizon N is probably not longer than a few years. In this time horizon, as stated earlier, we assume that the *benchmark* speed cannot be changed. On the other hand, the actual speed can be changed (constrained by the process design) by working faster or slower, impacting quality. For example, in a lab that does chemical tests on soil and water samples, the average time

required to finish a certain test depends on the equipment and process flow used by the lab (see Chase and Rajagopalan 1993 for a description of such a process) and test quality standards or specifications, so the benchmark speed is determined by these factors and cannot be changed easily. But the actual speed can be higher or lower depending on the care with which equipment is calibrated and the speed at which some steps such as data analysis and report preparation are done by the analyst.

The firm charges a price p_j in period j for its product. We consider a typical downward sloping demand function λ_j in period j, given by

$$\lambda_j = \Lambda_j - \alpha p_j, \quad j = 1, \ldots, N, \tag{2}$$

where α (>0) is the customer's price sensitivity. So, the effective demand rate λ_j is a function of the demand potential Λ_j and the price p_j. It is easy to show that $\lambda_j < \Lambda_j/2$ due to the presence of congestion costs ($\lambda_j = \Lambda_j/2$ maximizes revenue in the absence of congestion costs), and so the condition $\mu_j < 2/\delta$ ensures that $\Lambda_j > \delta\lambda_j(\mu_j - \hat{\mu})$, i.e., the demand potential is nonnegative. We assume the parameters are such that this condition will be satisfied; in particular, we need $\hat{\mu} < 2/\delta$.

We model our system as an M/M/1 queue with a first-come, first-served queue discipline similar to the approach in Anand et al. (2011) and Ren and Zhou (2008). This assumption should not restrict the spectrum of potential applications of the model to ones with a single server, as our objective here is not to model the physical queue, but to focus on capturing congestion and responsiveness of the system. Customers arrive to the system in every period j according to a Poisson process with demand rate λ_j. Other than price, which controls the demand rate, the firm also determines the speed of service (rate μ_j); service times are assumed to be independent exponential.

The firm's decision on speed and price also impacts congestion levels. Two approaches are common in the queuing literature to capture congestion effects. One approach is to assume that the firm incurs a congestion cost that is linear in the wait time or delay (see Ren and Zhou 2008, Whitt 2006, Hasija et al. 2008, Urban 2009). There is also a long tradition of modeling delay cost as tardiness or backorder cost that is a linear function of delay in the production control and scheduling literature. In the second approach, common in the congestion pricing literature (Allon and Federgruen 2007, Anand et al. 2011), the congestion delay is a cost that the customer pays in addition to the nominal price. We focus on the first approach, but we also consider the second approach in §7. If $\gamma > 0$ denotes the waiting or congestion cost per customer incurred by the firm, then the firm incurs a total cost that depends on the demand rate and is given by

$$\gamma\lambda_j \frac{1}{\mu_j - \lambda_j}, \quad j = 1, \ldots, N. \tag{3}$$

Note here that we consider the total sojourn time in the system, consistent with the existing literature. An identical linear waiting cost function is also used in Whitt (2006), Ren and Zhou (2008), and other works mentioned earlier. In some scenarios, the firm may actually pay a penalty that is linear in the delay (e.g., see Hasija et al. 2008, Urban 2009). In other scenarios, the firm may commit to a lead time, and, as pointed out in So and Song (1998), the service level constraint associated with the lead time can be viewed as equivalent to a linear delay cost structure using a Lagrangian approach. In this case, our approach is a good approximation.

The firm's total expected profit consists of the revenue from customers paying price p_j for the service in period j less the congestion cost. We denote the firm's expected profit as $\mathcal{R}$, and so

$$\mathcal{R} = \sum_{i=1}^{N} \left[p_i\lambda_i - \gamma \frac{\lambda_i}{\mu_i - \lambda_i} \right] + \theta\Lambda_{N+1}. \tag{4}$$

The service provider's objective is to maximize her expected profit by choosing the price p_j and the service speed μ_j in every period. The term $\theta\Lambda_{N+1}$ has been included to incorporate an appropriate ending condition, and θ (>0) represents the salvage value of the market potential at the end of the horizon. If this term is not included, note that μ_N can be increased with no penalty, but this will result in a negative Λ_{N+1}. We ignore discounting of profits, but the results for a discounted formulation are briefly summarized in §8, and details are available from the authors. Note that we have not included an explicit cost for increasing the speed so as to focus on the speed–quality trade-off and not get distracted by other trade-offs. However, including a linear cost for increasing the speed will not change the insights derived here.

In the next proposition, we verify that the profit function is concave under certain conditions. Hence, there exists a unique set of μ_j and p_j values that maximizes the profit.

Proposition 1. *When δ is small and $\lambda_j > \alpha\sqrt{\gamma\delta\theta}/4$ ($\forall j$), the profit function in* (4) *is concave.*

The condition on δ (provided in the proof of Proposition 1 in the online appendix, available as supplemental material at http://dx.doi.org/10.1287/msom.2013.0458) will be typically satisfied given the definition of δ and the fact that it is divided by $\hat{\mu}$ as discussed earlier. We further assume that the parameters are such that the condition on λ_j is satisfied, and so our results are valid under these conditions.

3.1. The Single-Period Model

We first consider the single-period model to motivate the multiperiod one because the single-period analysis can be effectively seen as the Nth period analysis in the multiperiod model and provides some useful insights. The demand function is given by $\lambda = \Lambda - \alpha p$, where Λ is the demand potential at the beginning of the period. The total expected profit $\mathcal{R}$ will consist of the revenue generated by the customers less the congestion cost,

$$\begin{aligned}\mathcal{R}(\mu, p) &= p\lambda - \gamma\frac{\lambda}{\mu-\lambda} + \theta\Lambda_f \\ &= p(\Lambda-\alpha p) - \gamma\frac{\Lambda-\alpha p}{\mu-\Lambda+\alpha p} + \theta\Lambda_f, \quad (5)\end{aligned}$$

where Λ_f is the demand potential at the end of the period and is given by $\Lambda_f = \Lambda - \delta\lambda(\mu - \hat{\mu})$. The parameter θ controls the importance of future profits (captured by ending market potential Λ_f) relative to current profits. The service provider's objective is to maximize her expected profit by choosing the price p and the service speed μ.

In the next theorem, we state the optimal pricing decision and speed of service. We use the superscript asterisk (*) for all the optimal quantities hereafter.

Theorem 1. *In the single-period model,*

$$p^* = \frac{\Lambda + 2\alpha\Lambda\delta\theta - \alpha\delta\theta\hat{\mu} + 2\alpha\sqrt{\gamma\delta\theta}}{2\alpha(1+\alpha\delta\theta)} \quad \text{and}$$
$$\mu^* = \frac{\Lambda + \alpha\delta\theta\hat{\mu} + 2\sqrt{\gamma/\delta\theta}}{2(1+\alpha\delta\theta)}. \quad (6)$$

At optimality, the expected profit of the service provider will be

$$\mathcal{R}^* = \frac{(\Lambda + \alpha\delta\theta\hat{\mu} - 2\alpha\sqrt{\gamma\delta\theta})^2}{4\alpha(1+\alpha\delta\theta)} + \theta\Lambda. \quad (7)$$

We note that when the benchmark speed is large and/or the demand potential is small, the firm is more likely to work at a speed lower than the benchmark speed. In particular, we can show that if $\Lambda < (>)\alpha\delta\theta\hat{\mu} - 2\sqrt{\gamma/\delta\theta} + 2\hat{\mu}$, then $\mu^* < (>)\hat{\mu}$. A further examination of the optimal price and service speed shows that when customers are more sensitive to the quality of the product (larger δ), the firm slows down the service process but charges higher prices and serves less demand (refer to the online appendix for a summary of the impact of the parameters on the optimal price and service speed decisions). When a firm works at a speed that is slower than benchmark speed, an increase in δ results in an increase in profits. In such a scenario, higher-quality sensitivity results in higher prices, lower speed, and higher profits, thus benefiting the firm. This is similar to the more customer-intensive or high-value services as described in the paper by Anand et al. (2011), where they reach similar conclusions. It is also reminiscent of the work of Chase and Tansik (1983), where they differentiate services depending on the customer contact required. They recommend that high-customer-contact services should maximize the quality offered, whereas low-customer-contact services should focus on efficiency and maximize productivity. The impact of θ is identical to that of δ.

4. The Multiperiod Model: Constant Service Speed

The constant speed setting is appropriate for scenarios where the process may be automated (e.g., manufacturing environments) or the process design cannot be changed easily and so speed is hard to change. However, the firm has the flexibility to vary prices so as to modulate demand and congestion levels. For instance, the cutting machine at Coverking is the bottleneck, and its speed cannot be changed easily, but they vary prices to modulate demand through promotions, etc. (Gupta and Rajagopalan 2012). Likewise, a printing press that caters to a market segment with certain print quality expectations may not be able to change the speed of its printing equipment but can vary prices.

The firm decides the price p_j to quote every period, whereas speed μ is a one-time decision made at the beginning of the horizon. The firm has to carefully consider the impact of its decisions on current period demand, congestion, and profits versus the impact on the demand potential in the next period and, thus, future profits. The demand potential Λ_{j+1} at the beginning of period $j+1$ is

$$\Lambda_{j+1} = \Lambda_j - \delta(\mu - \hat{\mu})(\Lambda_j - \alpha p_j) \quad j = 1, \ldots, N. \quad (8)$$

The total expected profit $\mathcal{R}$ is given by

$$\begin{aligned}\mathcal{R} &= \sum_{i=1}^{N}\left[p_i\lambda_i - \gamma\frac{\lambda_i}{\mu-\lambda_i}\right] + \theta\Lambda_{N+1} \\ &= \sum_{i=1}^{N}\left[p_i(\Lambda_i - \alpha p_i) - \gamma\frac{\Lambda_i - \alpha p_i}{\mu - \Lambda_i + \alpha p_i}\right] + \theta\Lambda_{N+1}. \quad (9)\end{aligned}$$

We now explore the behavior of prices and demand over time. The next theorem states that prices vary in a manner such that demand rate remains constant over time.

Theorem 2. $\lambda_1^* = \cdots = \lambda_N^* = \lambda^*$.

The rationale for this result can be understood by comparing the marginal impact on current and future profits of a change in λ_{t-1} versus a change in λ_t. The impact on current period profit is the same in both

scenarios ($(t-1)$ and t) except for the difference in starting market potential, Λ_{t-1} versus Λ_t, respectively. The impact of a change in λ_{t-1} on future profits is greater than the impact of λ_t on future profits by the amount $-(\delta/\alpha)(\mu-\hat{\mu})\lambda_t$ which is the impact of λ_{t-1} on period t profit. But this is exactly equal to the *difference* in the marginal impact on current profits in periods $(t-1)$ and t due to the differences in market potential Λ_{t-1} and Λ_t, respectively. The two effects offset each other, and the trade-off between the marginal impact of changing λ on current and future profits is the same in periods $(t-1)$ and t, and so the optimal λ is the same in both periods.

Thus, demand and utilization do not change over time, and neither does congestion cost. But demand potential increases over time (if $\mu^* < \hat{\mu}$), and so prices increase too, resulting in higher revenues over time. Thus, even though it may be tempting to increase demand and exploit demand potential to achieve higher current revenues, this will also increase congestion costs, but, more importantly, future demand potential will decrease, which will depress profits in all future periods. So, the firm should increase its price, but at a controlled pace so that demand remains constant over time if its speed cannot be changed easily. One of the notable consequences of this result is that utilization and therefore congestion are constant over time, and this makes it easier to manage the production system. Next, we explore the optimal price and the conditions under which the firm will choose to work at a slower or faster speed than benchmark speed.

THEOREM 3. (a) *The optimal pricing policy is*

$$p_i^* = \frac{\gamma\mu^*}{(\mu^*-\lambda^*)^2} + \frac{\lambda^*}{\alpha} + \delta\theta(\mu^*-\hat{\mu}) + \frac{(N-i)\lambda^*\delta(\mu^*-\hat{\mu})}{\alpha}, \tag{10}$$

where (λ^*, μ^*) *is the solution to*

$$\mu^* = \frac{\Lambda_1 - 2\lambda^* + \hat{\mu}\delta(\alpha\theta + (N-1)\lambda^*)}{\delta(2\alpha\theta + (3/2)(N-1)\lambda^*)}, \tag{11}$$

$$-\frac{\delta\lambda^*(N-1)}{2} + \frac{\alpha\gamma}{(\mu^*-\lambda^*)^2} - \alpha\theta\delta = 0. \tag{12}$$

(b) μ^* *and* λ^* *decrease with* θ, μ^* *and* λ^* *decrease with* N, *and* p_i^*, *increases with* θ *and* N.

(c) *Suppose that* $\hat{\hat{\lambda}}$ *and* $\hat{\hat{\mu}}$ *are defined as follows*:

$$\Lambda_1 - 2\hat{\hat{\lambda}} - \frac{\alpha\hat{\hat{\mu}}\gamma}{(\hat{\hat{\mu}}-\hat{\hat{\lambda}})^2} = 0, \tag{13}$$

$$\hat{\hat{\mu}} = \frac{2(\Lambda_1 - 2\hat{\hat{\lambda}})}{\delta(\hat{\hat{\lambda}}(N-1) + 2\alpha\theta)}. \tag{14}$$

Assume that the parameters are such that $\hat{\hat{\mu}} < 2/\delta$. *Then, the following scenarios occur*:

(1) *If* $\hat{\mu} = \hat{\hat{\mu}}$, *then* $\mu^* = \hat{\mu}$ *and price will remain constant over time with* $p^* = \gamma\hat{\mu}/2(\hat{\mu}-\lambda^*)^2 + \Lambda_1/2\alpha$; λ^* *is obtained by solving* $2\lambda^* + \alpha\gamma\hat{\mu}/(\hat{\mu}-\lambda^*)^2 = \Lambda_1$.

(2) *If* $\hat{\mu} > \hat{\hat{\mu}}$, *then* $\hat{\hat{\mu}} < \mu^* < \hat{\mu}$ *and prices will increase over time*.

(3) *If* $\hat{\mu} < \hat{\hat{\mu}}$, *then* $\hat{\hat{\mu}} > \mu^* > \hat{\mu}$ *and prices will decrease over time*.

Recall that θ captures the impact of the value of the ending market potential, i.e., future profits. The higher θ is, the less important current profits are, and so the firm will operate at a lower speed and satisfy less demand. The effect of N is similar to that of θ, and a longer time horizon results in lower speed and typically lower demand. In the scenario $\mu^* < \hat{\mu}$, if N is higher, the firm has more time to build up its market potential, and so it operates at a lower speed so that the market potential can be built up to higher levels, thus reaping greater profits in later periods; that is, given a problem with horizons N_1 and N_2 $(>N_1)$, the firm will have higher market potential at the end of period N_1 in the N_2-period problem compared to the N_1-period problem. The lower speed also implies that it can satisfy less demand, and so total profits in the first N_1 periods in the N_2-period problem will be lower. So, a firm that has a longer planning horizon or values the future more, i.e., is more far sighted, is more likely to operate at lower speeds in the initial periods. In the scenario $\mu^* > \hat{\mu}$, although market potential is declining over time and so are profits, the impact of higher θ and N are similar; i.e., speed is lower, which results in higher market potential.

We find that when benchmark service speed is high $(\hat{\mu} > \hat{\hat{\mu}})$, where $\hat{\hat{\mu}}$ is a threshold for the benchmark speed, the firm is more likely to work at a slower speed than benchmark $(\hat{\hat{\mu}} < \mu^* < \hat{\mu})$ and vice versa. However, independent of how high or low the benchmark speed is compared to $\hat{\hat{\mu}}$, the firm will work at a higher (lower) speed if the benchmark speed is higher (lower). Of course, the value of $\hat{\hat{\mu}}$ that determines which scenario in Theorem 3(c) is likely to occur is a function of many parameters, and although we do not have a closed-form solution, we can predict how $\hat{\hat{\mu}}$ varies with the problem parameters. For instance, we find that $\hat{\hat{\mu}}$ decreases with quality sensitivity δ.

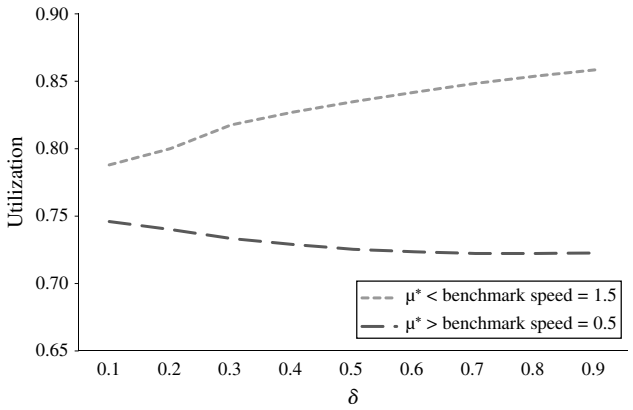
This result can be interpreted as follows. Consider a firm providing water quality testing service. Some tests, e.g., for certain high-fidelity industrial customers, may have very high-quality standards, and customers may care deeply about the quality of the service. In this case, δ will be high and $\hat{\hat{\mu}}$ is likely to be lower. Suppose the testing service has a unique process with a benchmark speed higher than the threshold, i.e., $\hat{\mu} > \hat{\hat{\mu}}$. Then, in this case, $\hat{\hat{\mu}} < \mu^* < \hat{\mu}$, and the firm should work at a slower speed than the benchmark and increase its market potential, which will enable it to charge higher prices and achieve

higher profits in the future. In effect, the unique process enables a high benchmark speed, but then the technicians work at a speed slower than this benchmark by carefully analyzing and evaluating the test reports so that quality is higher. This result illustrates the important role that operations can play in service industries—developing an efficient process provides a buffer that allows a firm to work at a slower speed. This is similar in spirit to the "quality buffer" mentioned in Hopp et al. (2007). Since service firms or make-to-order firms cannot hold inventory, the ability to operate at speeds below the benchmark speed is especially valuable because it serves as a buffer similar to the role of inventory in make-to-stock firms. When customers care more about the quality (high δ), the quality buffer is more valuable and is more likely to be built up by operating at a below-benchmark speed.

Therefore, when customers value quality highly, i.e., when δ is high and $\hat{\hat{\mu}}$ is low, the optimal pricing strategy is to begin with a low price and increase it over time. This strategy resembles a penetration strategy that is used to rapidly penetrate the market for a new product with an initial low price followed by increasing prices. Similarly, when δ is low and $\hat{\hat{\mu}}$ is high, it is optimal to begin with a high price and decrease it later on, as in a skimming strategy. However, the underlying drivers behind these pricing strategies are quite different in the new product diffusion literature. Penetration (and skimming) strategies are used when products are introduced in a competitive market with heterogeneous customers that make a one-time purchase (Mahajan et al. 1990), unlike in our model, or to stimulate demand when quality is unobservable (Kirmani and Rao 2000).

We also conducted numerical studies to understand the impact of the parameters δ, γ, Λ_1, and $\hat{\mu}$ on the behavior of speed, utilization, etc., and we highlight here their impact on utilization and queue length that is intriguing. As sensitivity to quality (δ) increases, one might expect congestion costs to be less important and hence queue length to increase. However, this depends on the scenario in Theorem 3(c), i.e., whether the optimal speed is higher or lower than the benchmark speed. When optimal speed is lower than benchmark speed, the utilization and queue length do increase with δ. However, if speed is higher than benchmark, then utilization actually decreases as δ increases (Figure 1). If δ is high, the firm is willing to have lower demand (resulting in lower utilization) and sacrifice current profits, so that demand potential in future periods does not decline too much and hurt profits. Moreover, when the δ is higher, i.e., the sensitivity of customers to the speed–quality trade-off is greater, not only are prices higher, but the price trajectory over time is steeper.

Figure 1 Optimal Utilization with Quality Sensitivity



Note. $\Lambda_1 = 6$, $\alpha = 0.5$, $\theta = 17$, $N = 6$, and $\gamma = 0.4$.

Thus, prices are modulated so as to carefully control demand, taking into account the constant speed, speed–quality sensitivity, and other factors. Another finding is that as $\hat{\mu}$ increases, one would expect the increase in benchmark speed to result in lower utilization, but we find the reverse to be true. The utilization and queue length actually increase as $\hat{\mu}$ increases because demand increases at a faster rate than speed. Although the increase in $\hat{\mu}$ allows the firm to increase its speed, the firm also has greater scale economies, and so the utilization can be higher. Although congestion costs are higher, these are offset by the increased revenue that comes from the faster growth in market potential due to the higher $\hat{\mu}$, which increases future profits substantially. In mass markets, the optimal speed and demand tend to be high, and, in fact, the firm may operate at a speed higher than $\hat{\mu}$ when Λ_1 is sufficiently high. In such cases, one is less likely to see a firm keeping its speed low and trying to build its reputation; it would rather operate at high speeds and be willing to lose some market potential.

5. The Multiperiod Model: Constant Price

Now we consider a setting where price cannot be changed but speed can be varied over time. The model we develop incorporates the key characteristics of several service settings (high-end restaurants, Starbucks). The restaurant industry is the largest service industry and one of the largest overall in Western countries, with Americans spending approximately \$970 million a day dining out (Lord 1999). In such a service environment, it is not common to observe frequent price changes due to administrative and other reasons. A study conducted by MacDonald and Aaronson (2000) shows that restaurant pricing remains constant on average for 10 months. A high-end restaurant determines the price p in a monopolistic setting because there are high barriers to entry.

Also, the capacity decision is fixed at the beginning of the horizon, and it depends on the actual space of the restaurant and the acclaimed chef. As a consequence, in the short run, one can only speed up the service process by rushing the customers. The market potential will evolve in each period $j+1$ according to

$$\Lambda_{j+1} = \Lambda_j - \delta(\mu_j - \hat{\mu})(\Lambda_j - \alpha p), \quad j = 1, \ldots, N. \quad (15)$$

The formulation is similar to §4 in other respects.

In a dynamic setting, the owner has to consider the impact on future profits of current decisions on price and speed. The *Starbucks experience* involves a warm atmosphere with background music and a relaxed pace where customers do not mind some waiting. On the contrary, they dislike fast-paced employees and mass drink preparation. In fact, an attempt to adopt lean techniques (Jargon 2009) annoyed its loyal customers, and Starbucks had to change its strategy to regain its customers (Carey 2009). Prices, however, hardly vary, and hence the initial price decision is critical because it impacts current period demand and congestion costs as well as future speed and demand potential. Next, we explore the behavior of the optimal speed over time and the optimal price.

THEOREM 4. (i) *If $\mu_i^* < \hat{\mu}$, then $\mu_{i+1}^* - \lambda_{i+1}^* \geq \mu_i^* - \lambda_i^*$ and $\mu_i^* < \mu_{i+1}^*$.*

(ii) *If $\mu_i^* > \hat{\mu}$, then $\mu_{i+1}^* - \lambda_{i+1}^* \leq \mu_i^* - \lambda_i^*$ and $\mu_i^* > \mu_{i+1}^*$.*

(iii) *If either $\mu_i^*(N) < \hat{\mu}\ \forall\, i < t$ or $\mu_i(N)^* > \hat{\mu}\ \forall\, i < t$, then $\lim_{t\to\infty} \mu_i^*(\infty) = \hat{\mu}$, where μ_i^* is denoted as $\mu_i^*(N)$ because it is a function of the time horizon N. Furthermore, there exists $\theta_N > 0$ for finite N such that $\mu_N^* = \hat{\mu}$ if $\Lambda_N > 2\hat{\mu}$.*

(iv) *The optimal price p^* will satisfy*

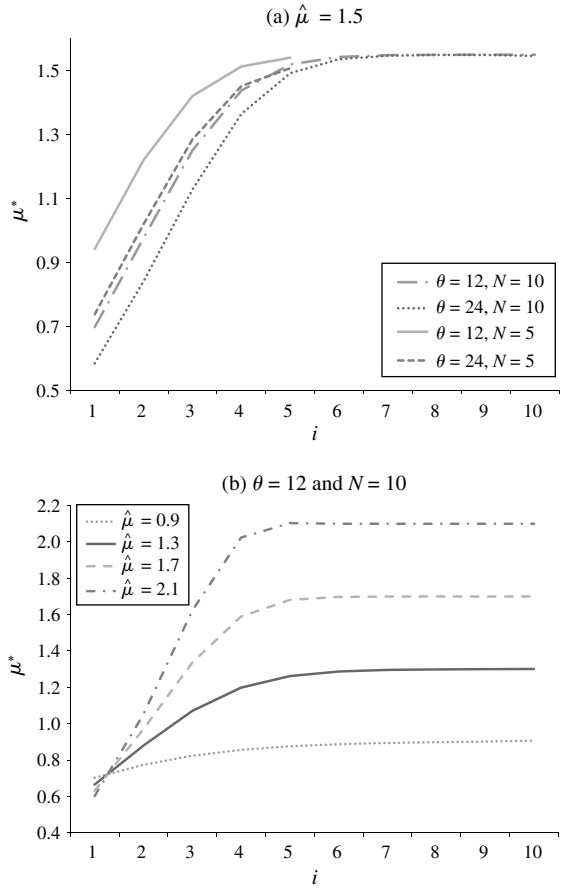
$$\frac{\Lambda_1}{\alpha}\left\{1 + \sum_{i=2}^{N}\prod_{k=1}^{i-1}[1-\delta(\mu_k^* - \hat{\mu})]\right\} + p^*\sum_{i=2}^{N}\prod_{k=1}^{i-1}[1-\delta(\mu_k^* - \hat{\mu})] - \frac{\gamma[1-\delta(2\mu_1^* - \hat{\mu})]}{\delta(\mu_1^* - \Lambda_1 + \alpha p^*)^2} = 0. \quad (16)$$

If speed is constant, $\mu_i^* = \hat{\mu}$, then the optimal price is $p^* = \gamma\mu/2(\hat{\mu} - \lambda^*)^2 + \Lambda_1/2\alpha$, where $\lambda^* = \Lambda_1 - \alpha p^*$.

(v) *The optimal price p^* increases with θ, λ_1^* and μ_1^* decrease with θ, and $(\mu_i^* - \lambda_i^*)$ decreases with θ.*

The optimal policy suggests that a restaurant may start by serving customers faster or slower than the benchmark speed (depending on the parameters as discussed later in this section), but in either case it will gradually approach the benchmark speed when N is large. If N is small, the service speed is sensitive to the value of θ and may fluctuate around $\hat{\mu}$ rather than smoothly approach $\hat{\mu}$ depending on the model parameters (based on numerical studies). Higher values of θ imply that the ending market potential Λ_{N+1} is more valuable and so the restaurant should serve customers more leisurely (see Figure 2(a)). From part (iv), we know that the optimal price is also higher when θ is higher, which dampens demand, consistent with the lower speed. Thus, at higher values of θ, the restaurant is willing to sacrifice current revenues for the future as in the constant speed case. The effect of N is similar; as N increases, the restaurant starts at a lower speed in period 1 because there are more periods in which the higher market potential achieved can be used to generate greater revenues.

Figure 2 **Optimal Service Speed Behavior with Respect to θ and N (a) and Benchmark Speed (b) When Price Is Constant and $\mu^* < \hat{\mu}$**



Note. $\Lambda_1 = 6$, $\alpha = 0.5$, $\gamma = 0.4$, and $\delta = 0.6$.

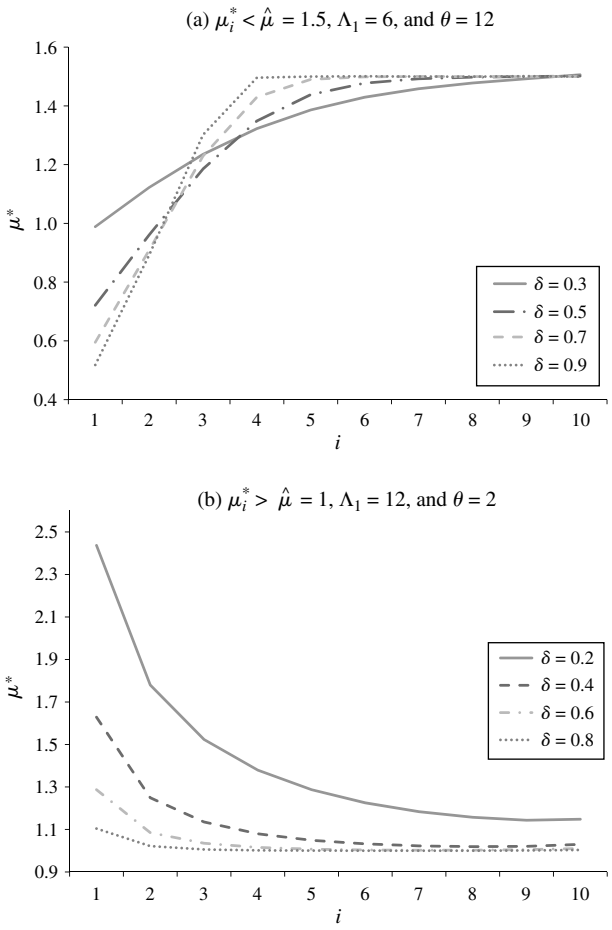
An implication of parts (i) and (ii) of Theorem 4 is that utilization and congestion costs decrease over time (because the gap between demand and speed increases) when $\mu_i^* < \hat{\mu}$, but increase over time when $\mu_i^* > \hat{\mu}$. When $\mu_i^* < \hat{\mu}$, Starbucks, for example, can tolerate higher congestion costs in earlier periods because it already serves customers slowly and is enhancing its reputation and market potential sufficiently. But later on, as it approaches the benchmark speed, it does not want high congestion because the *gain* in quality perception and demand potential is too small to offset it. Conversely, when $\mu_i^* > \hat{\mu}$, it is

willing to accept higher congestion only in later periods because it is operating at almost the benchmark speed and *losing* very little market potential.

Whether it is optimal to start at a speed faster or slower than the benchmark speed depends on the problem parameters. Numerical analysis suggests that it is more likely to start at a higher speed than benchmark when the starting demand potential Λ_1 is high, $\hat{\mu}$ is small, and δ is small. Consider a fast-food restaurant where starting market potential is large and δ is low because customers are not very sensitive to service speed and quality, unlike at a high-end restaurant. If this restaurant has a low benchmark speed due to, say, poor processes, the restaurant cannot satisfy much demand, so it starts at a higher speed than benchmark even though this means lower quality and loss of market potential. The optimal price is higher (correspondingly, demand is lower) than that charged by a single-period monopolist who acts myopically and decides on the optimal price that maximizes every period's profit separately. This is due to both the impact of high demand on current period congestion costs as well as the negative impact of high current demand and the resultant high speed on future market potential and, therefore, future profits. Even if the benchmark speed is lower than the speed required to meet a single-period monopolist's demand, the restaurant will start at a *slower* speed than benchmark provided δ is not too low.

We performed some numerical studies to understand the dynamic behavior of price, speed, and profits as a function of the parameters δ, γ, and $\hat{\mu}$. The impact of γ is as expected because it implies higher congestion costs and so speed is higher, demand is lower, and profits are lower. To understand the effect of $\hat{\mu}$, suppose Starbucks outlets place a pretty pattern on top of lattes in some regions and do not in other regions based on differing customer expectations about quality. The regions without the pattern will have a higher benchmark speed, and one might expect that correspondingly demand will be higher and prices will be lower. But this is not always the case. If the optimal speed is less than the benchmark speed, then the price is higher when $\hat{\mu}$ is higher. Starbucks starts out working at a *lower* speed when $\hat{\mu}$ is higher and the corresponding demand is also lower and prices higher. However, over time, the speed increases at a faster rate and reaches the benchmark speed, as shown in Figure 2(b). When $\hat{\mu}$ is higher, Starbucks can satisfy more demand, and so it is willing to start at a lower speed and exploit the increase in market potential and higher demand in the future. This results in higher overall profits. Thus, a higher $\hat{\mu}$ need not translate into a higher production rate or lower costs immediately. Instead, the firm benefits by

Figure 3 Optimal Service Speed Behavior When Price Is Constant

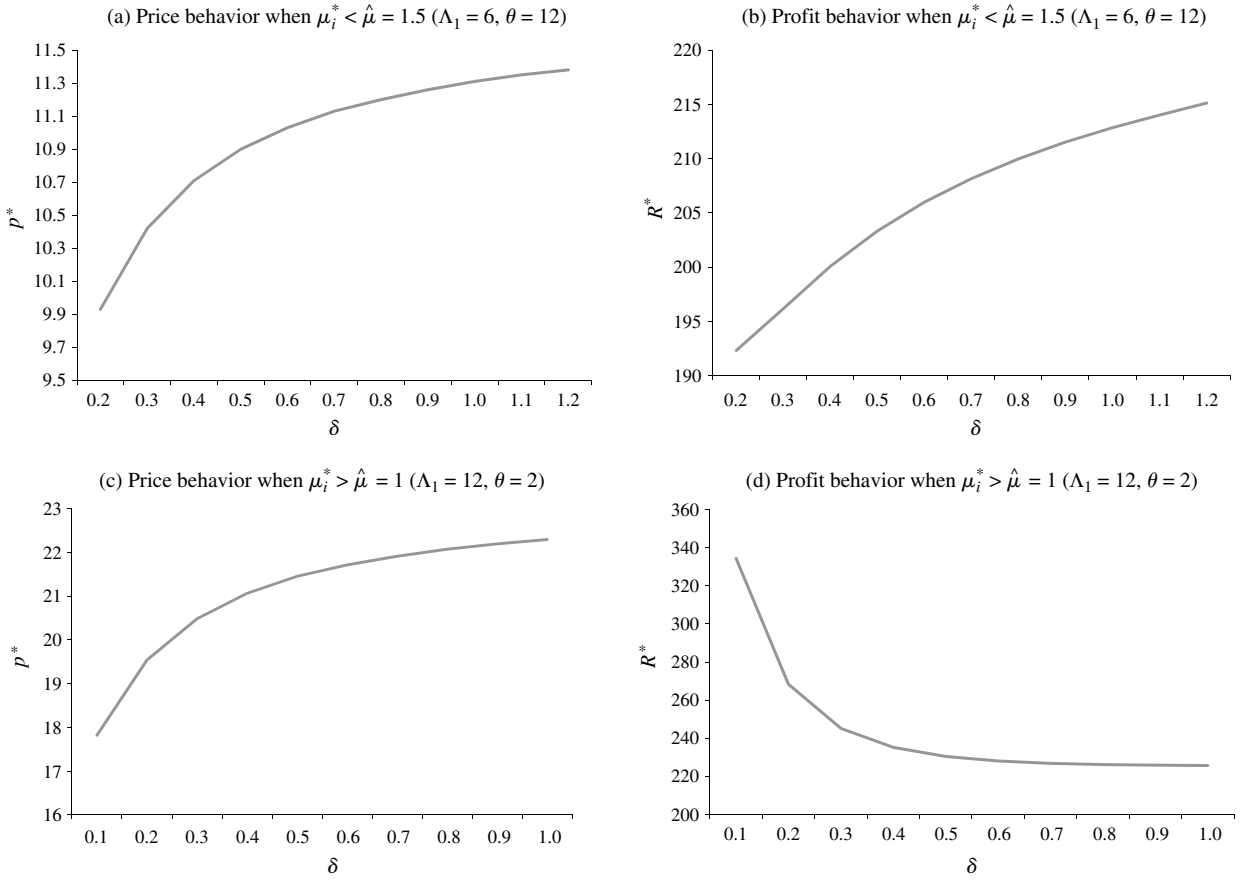


Note. $\alpha = 0.5$, $\gamma = 0.4$, and $N = 10$.

lowering its actual speed in the short run and deriving greater benefits in the future.

Now we consider the impact of δ. We start with the scenario where $\mu_i^* < \hat{\mu}$. As δ increases, speed or service rate is lower in the initial periods but increases at a faster rate (see Figure 3(a)), and demand varies corresponding to the speed. The speed is lower for larger δ when $\mu_i^* > \hat{\mu}$ (see Figure 3(b)), but speed decreases at a slower rate for higher δ. The impact on profits, however, depends on the speed scenario—if the restaurant works at a speed higher than benchmark, then profit *decreases* if customers are more sensitive to quality (Figure 4(d)). But if the restaurant works at a speed lower than benchmark (and reaches the benchmark speed over time), then higher customer sensitivity is actually beneficial in that profit *increases* with δ (Figure 4(b)). The rationale is as follows. When δ is large, quality sensitivity is high, and so the restaurant starts at a slower speed (and charges higher prices) and thus maximizes the increase in market potential and reaps the rewards in later periods. A further increase in quality sensitivity benefits

Figure 4 Optimal Price and Profit Behavior When Price Is Constant with Respect to δ



Note. $\alpha = 0.5$, $\gamma = 0.4$, and $N = 10$.

the restaurant in terms of increased market potential and the resultant higher profits. The sensitivity of price with respect to δ (i.e., price increasing with δ) is the same independent of whether $\mu_i^* <$ or $> \hat{\mu}$ (see Figures 4(a) and 4(c)).

Although we have focused on specific examples, our insights are applicable to other settings where the quality–speed trade-off exists and the price cannot vary over time. This is often the case, for instance, in hospitals, pharmacies, or chemical testing firms, where prices do not change frequently.

6. The General Multiperiod Model

We now consider the scenario where both price and speed are allowed to vary over time. For instance, a software testing company may have flexibility over both price and speed. The firm could carefully test the software using more resources and more tests so as to maximize detection of errors and improve software quality; so the testing will take longer, but they may be able to charge a higher price. The company has to decide on its positioning in the market initially as well as over time. Does it want to be known as a "budget" software testing firm that responds quickly and at reasonable prices or as a high-quality firm that targets high-end customers? Also, how will its strategy with respect to speed and price evolve over time as customers experience its services and its market potential changes? All of these factors will eventually impact its current and future profitability. Recall from §4 that when speed cannot be changed, the firm cannot take full advantage of the increased market potential and allow demand to increase due to the concern about congestion costs. Since the firm can change both price and speed, this gives the firm more flexibility in optimizing its operations, and we will explore how this can benefit the firm. The demand potential evolves as

$$\Lambda_{j+1} = \Lambda_j - \delta(\mu_j - \hat{\mu})\lambda_j = \Lambda_j - \delta(\mu_j - \hat{\mu})(\Lambda_j - \alpha p_j), \quad j = 1, \ldots, N. \quad (17)$$

We then have the following result on the evolution of price, speed, and demand over time.

Theorem 5. (i) $(\mu_{i+1}^* - \lambda_{i+1}^*) \geq (\mu_i^* - \lambda_i^*)$, $\forall\, i < N$.
(ii) *If* $\mu_i^* < \hat{\mu}$, *then* $\mu_{i+1}^* > \mu_i^*$, $\lambda_{i+1}^* > \lambda_i^*$ *and* $p_{i+1}^* > p_i^*$.

We see that, as in the constant price case, if $\mu_i^* < \hat{\mu}$ in the initial periods, then speed increases during

these periods. What is more interesting here is that the price and demand rate are both increasing over time. Thus, the firm increases price over time as we observed in the constant speed scenario. However, unlike in that scenario, it is allowed to change speed, and this lets the firm satisfy a higher demand rate over time. So, the firm exploits the increase in market potential by increasing both price and demand over time. An important difference relative to the constant price case is that speed increases at a faster rate than demand in all scenarios, i.e., whether $\mu_i^* <$ or $> \hat{\mu}$. Also, unlike in the constant price case, it is not always true that speed decreases when $\mu_i^* > \hat{\mu}$.

To illustrate the implications of the results, consider Minute Clinics, which provide routine, low-risk health-care services for minor illnesses and injuries, wellness tests, etc., and have grown to be the largest provider of retail health care in the United States, with over 500 clinics (Scott 2006). They have designed their processes for such services so that they are better designed than the ones in typical doctor offices and urgent care facilities in hospitals and use technology effectively. They have used electronic kiosks to help patients provide information and proprietary software systems to help nurse practitioners arrive at a quick and accurate diagnosis to speed up the process (Scott 2006). So, their *benchmark* speed is high for many routine tasks, but they may deliberately operate at a slower speed wherein their nurse practitioners spend more time with patients to provide a better customer experience. This, in turn, results in a better reputation and increased market potential, and allows them to raise both prices and demand over time. Although there are other factors that do impact their profits and performance over time, our model and results provide some insights into this phenomenon. In particular, our results suggest that it is better for them to start at a lower speed than their more efficient process would permit, but then increase speed, demand, and price over time.

Next, we consider the conditions under which the above scenarios ($\mu_i^* < \hat{\mu}$) may occur in terms of the δ and $\hat{\mu}$ based on numerical studies. If the benchmark service speed is large enough to cover the demand at optimum monopoly prices in a single period, then $\mu_i^* < \hat{\mu}$ even if δ is small. But if δ is high, then even if $\hat{\mu}$ is not very high (for instance, not enough to cover optimum single-period monopoly demand), we find that $\mu_i^* < \hat{\mu}$ and speed is increasing over time but typically staying below $\hat{\mu}$. Comparison of the scenario $\mu_i^* < \hat{\mu}$ with the constant price case numerically for the same set of parameters indicates that both speed and demand start at higher values in period 1, but increase at a slower pace when both price and speed can be changed. The ability to change both price and speed changes the dynamic trade-offs and allows the firm to reap the benefits of both higher quality and higher price earlier in the horizon, and thus achieve higher overall profits. We note that we do not observe the property of service speed identified in the constant price scenario when $\mu_i^* < \hat{\mu}$ (recall Figure 3(a)) where the service speed was higher at the beginning for smaller δ and was gradually approaching the benchmark speed.

Table 1 **Comparison of the Optimal Profit for the Constant Price, Constant Service Speed, and the General Case vs. the Static Policy When $\Lambda_1 = 6$, $\alpha = 0.5$, $\gamma = 0.4$, $\delta = 0.6$, $\theta = 12$, and $N = 12$**

$\hat{\mu}$	Static policy	Constant price	Constant service speed	General
0.5	109.35	110.20	113.41	113.81
1	152.91	154.79	154.37	155.77
1.5	193.54	205.98	218.71	222.61
2	229.66	260.26	313.83	322.94
2.5	262.14	316.43	447.09	466.44

Note. For these values of the parameters ($\hat{\hat{\mu}} = 0.83$, $\hat{\hat{\lambda}} = 0.64$), the optimal price is 10.72.

Finally, we compare the profits in the scenario where both speed and price vary over time with policies wherein either the price or speed is kept constant over time. Also, we evaluate the performance of a *static* policy wherein the firm operates at the benchmark speed and charges the same price every period to provide a basis for comparison and evaluate the value of a dynamic policy. In this case, the market potential and demand will not change over time. The optimal price in such a scenario was identified in the previous section (Theorem 3) and is given by $p^* = \lambda^*/\alpha + \gamma\hat{\mu}/(\hat{\mu} - \lambda^*)^2$. In cases where it is optimal to work slower than $\hat{\mu}$, the static price is higher than the optimal price charged under any of the dynamic policies. In the constant speed case in particular, the optimal service speed is slower than the corresponding static speed, meaning that a firm that decides to follow a simple static policy will tend to rush its customers. Table 1 provides a comparison of profits using the various policies. (Additional numerical experiments were performed but are not reported to keep the exposition concise, but are available from the authors.) It is clear that a static policy can perform quite poorly compared to the other policies. We see that the constant speed policy does quite well relative to the general policy and typically outperforms the constant price policy. Thus, pricing flexibility is often more valuable than speed flexibility and may be a better alternative than varying the speed to exploit the speed–quality trade-off.

7. Extension: Impact of Congestion on Demand

Whereas the models discussed so far, starting in §3, assumed that the firm incurs a congestion cost, in this section we assume that congestion delay impacts the demand rate, i.e., demand is a function of the full price paid by a consumer, equal to the congestion cost plus the nominal price. Our objective is to understand the difference in the optimal behavior of speed, demand, etc., in situations where the congestion delay negatively impacts the demand rate. Using the framework in the congestion pricing literature (Naor 1969), the full price, say, p_j^f, paid by a customer in period j is equal to the nominal price p_j plus the congestion cost $\kappa/(\mu_j - \lambda_j)$, and we assume that the average wait time $1/(\mu_j - \lambda_j)$ is known to the customers before they arrive for service. This may be because the firm announces the wait time as in the case of restaurants or call centers, or customers learn about it from public information sources. Then the downward sloping demand function λ_j in period j is given by

$$\lambda_j = \Lambda_j - \alpha p_j^f = \Lambda_j - \alpha\left(p_j + \frac{\kappa}{\mu_j - \lambda_j}\right) = \Lambda_j - \alpha p_j - \beta\frac{1}{\mu_j - \lambda_j}, \quad j = 1, \ldots, N, \tag{18}$$

where $\beta = \kappa\alpha$ captures the per unit impact of wait time on demand. The rest of the notation and definitions are identical to those in the original model setup in §3.

The firm maximizes the total revenue given by

$$\mathcal{R}(\mu, p) = \sum_{i=1}^{N} p_i\lambda_i + \theta\Lambda_{N+1} = \sum_{i=1}^{N}\left\{\lambda_i\frac{\Lambda_i - \lambda_i}{\alpha} - \frac{\beta}{\alpha}\frac{\lambda_i}{\mu_i - \lambda_i}\right\} + \theta\Lambda_{N+1}. \tag{19}$$

Note that this is almost identical to the objective function in §3 except we now have β/α instead of γ. But the equation for the demand rate is different. We do not consider each of the scenarios such as constant speed, constant price, etc., separately, but instead summarize the results below.

Theorem 6. (i) *If speed is constant over time, then the following are true*:

(a) *The demand rate remains constant in all periods, i.e.,* $\lambda_1^* = \cdots = \lambda_N^* = \lambda^*$.

(b) *The optimal pricing policy will be*

$$p_i^* = \frac{\beta\mu^*}{\alpha(\mu^* - \lambda^*)^2} + \frac{\lambda^*}{\alpha} + \delta\theta(\mu^* - \hat{\mu}) + \frac{(N-i)\lambda^*\delta(\mu^* - \hat{\mu})}{\alpha}, \tag{20}$$

where (λ^*, μ^*) *is the solution to*

$$\mu^* = \frac{\Lambda_1 - 2\lambda^* + \hat{\mu}\delta(\alpha\theta + (N-1)\lambda^*)}{\delta(2\alpha\theta + (3/2)(N-1)\lambda^*)}, \tag{21}$$

$$-\frac{\delta\lambda^*(N-1)}{2} + \frac{\beta}{(\mu^* - \lambda^*)^2} - \alpha\theta\delta = 0. \tag{22}$$

Moreover, suppose that $\hat{\hat{\lambda}}$ *and* $\hat{\hat{\mu}}$ *are defined as follows*:

$$\Lambda_1 - 2\hat{\hat{\lambda}} - \frac{\beta\hat{\hat{\mu}}}{(\hat{\hat{\mu}} - \hat{\hat{\lambda}})^2} = 0,$$

$$\hat{\hat{\mu}} = \frac{2(\Lambda_1 - 2\hat{\hat{\lambda}})}{\delta(\hat{\hat{\lambda}}(N-1) + 2\alpha\theta)}.$$

Assume $\hat{\hat{\mu}} < 2/\delta$. *Then the following scenarios occur*:

(1) *If* $\hat{\mu} = \hat{\hat{\mu}}$, *then* $\mu^* = \hat{\mu}$ *and price will remain constant over time with* $p^* = \gamma\hat{\mu}/2(\hat{\mu} - \lambda^*)^2 + \Lambda_1/2\alpha$; λ^* *is obtained by solving* $2\lambda^* + \beta\hat{\mu}/(\hat{\mu} - \lambda^*)^2 = \Lambda_1$.

(2) *If* $\hat{\mu} > \hat{\hat{\mu}}$, *then* $\hat{\hat{\mu}} < \mu^* < \hat{\mu}$ *and prices will increase over time.*

(3) *If* $\hat{\mu} < \hat{\hat{\mu}}$, *then* $\hat{\hat{\mu}} > \mu^* > \hat{\mu}$ *and prices will decrease over time.*

(ii) *If price is constant over time, then the following are true*:

(a) *If* $\mu_i^* < \hat{\mu}$, *the gap between optimal speed and demand rate increases in the next period, i.e.,*

$$\mu_{i+1}^* - \lambda_{i+1}^* \geq \mu_i^* - \lambda_i^*,$$

and $\mu_i^* < \mu_{i+1}^*$, *i.e., the optimal speed also increases in the next period.*

(b) *The optimal price* p^* *will satisfy*

$$\sum_{i=1}^{N}\lambda_i^* - \alpha\theta\left\{\prod_{j=1}^{N}\left(1 + \frac{\delta}{\beta}\lambda_j^*(\mu_j^* - \lambda_j^*)^2\right) - 1\right\} = 0. \tag{23}$$

Thus, many of the key insights and conclusions identified in §§4 and 5 also hold in this model where the congestion delay impacts demand rate rather than being internalized by the firm as a penalty cost. But some of the results obtained earlier (such as speed decreasing when $\mu^* > \hat{\mu}$) do not extend to this scenario. Even though the objective function is similar, the difference in the demand rate function makes a significant difference in some of these scenarios. In this model, a change in speed directly impacts the demand rate, and this influences the behavior of speed, demand, etc., over time. For example, when $\mu_t^* > \hat{\mu}$ in the constant price model in §5, the market potential Λ_{t+1} and demand λ_{t+1} have to decrease, and the firm correspondingly reduces its speed. However, this need not be the case here because a decrease in market potential Λ_{t+1} does not imply a decrease in demand, and the firm can offset this impact by

increasing the speed so that demand increases. So, the speed and even demand can increase for some parameter ranges when $\mu_t^* > \hat{\mu}$, unlike in §5. Since μ_t^* may not decrease even if $\mu_t^* > \hat{\mu}$, we cannot in turn guarantee that μ_t^* will converge to $\hat{\mu}$ unlike in §5. Thus, whether a firm internalizes its congestion cost or lets customers take into account waiting time in their purchase decisions can have an impact on the firm's optimal decisions with respect to speed and price under some settings.

8. Conclusions

We summarize the key results and insights from this paper and point out opportunities for further research. This paper addressed a fundamental trade-off between speed and quality faced by organizations when they optimize their operations. Speeding up a process enables a firm to meet more demand with less congestion, but a faster process can result in poorer quality, and this impacts customer satisfaction and future demand potential and profits. Using this basic trade-off, we find that it is optimal for a firm to keep demand constant over time when the speed has to be kept constant. This speed can be faster or slower than the benchmark speed, depending on how sensitive customers are to quality and other factors. When price has to be kept constant, the firm tends to approach the benchmark speed over time with the rate of convergence dependent upon congestion costs and quality sensitivity. Our results suggest that firms endowed with a high benchmark speed can operate at lower than benchmark speed and thus make customers happier and benefit greatly over time.

There are several possible directions along which the models in the paper can be extended. A natural extension is to consider a discounted version of the models especially when allowing for a longer time horizon. The key result about the convergence of speed to the benchmark speed in the constant price scenario holds here too (detailed results for these extensions are available from the authors). Moreover, the discount factor does not impact the monotone behavior of the service speed and the difference between service and demand rates. But when speed is constant over time, we find that demand is increasing (decreasing) over time when $\mu_i^* < (>)\hat{\mu}$, with a rate that depends on the discount factor. However, when the number of time periods becomes larger, the optimal demand does become constant over time in later periods. Since speed need not be equal to the benchmark speed, this implies that market potential may change over time, and price will change proportionally. Thus, discounting does not substantially change the main insights derived. We can also consider a general queuing system instead of an M/M/1 queue. In such a setting, we can show that all of the main results and insights discussed in §§4 and 5 still hold under this relaxation. Other variations that can be explored include a more complex and stochastic evolution of the demand potential that is not a linear and symmetric function of the actual speed relative to benchmark speed, and where quality may have a long-term impact and strategic customers are taken into consideration.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2013.0458.

Acknowledgments

The authors are grateful to the anonymous associate editor and the referees for their comments and suggestions, which improved this paper substantially. The authors thank Victor DeMiguel and Jérémie Gallien for helpful discussions relating to the issues discussed in this paper.

References

Aflaki S, Popescu I (2010) Managing satisfaction in relationships over time. Working paper, INSEAD, Fontainebleau, France.

Allon G, Federgruen A (2007) Competition in service industries. *Oper. Res.* 55(1):37–55.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* 57(1):40–56.

Armony M, Gurvich I (2009) When promotions meet operations: Cross-selling and its effect on call center performance. *Manufacturing Service Oper. Management* 12(3):470–488.

Carey R (2009) Starbucks' lean ruins the experiences. *Quality Digest* (August 9), http://www.qualitydigest.com/inside/twitter-ed/Starbucks-lean-ruins-experience.html.

Chase RB, Rajagopalan S (1993) A production planning and scheduling system at a chemical testing laboratory. *Internat. J. Production Econom.* 29(2):125–138.

Chase RB, Tansik DA (1983) The customer contact model for organization design. *Management Sci.* 29(9):1037–1050.

Churchill GA Jr, Surprenant C (1982) An investigation into the determinants of customer satisfaction. *J. Marketing Sci.* 19(4):491–504.

Cohen LB (2008) Production pressure in endoscopy: Balancing quantity and quality. *Gastroenterology* 135(6):1842–1844.

Croasmun J (2005) Overworked pharmacies could be causing medication errors. *Ergonomics Today* (January 10), http://www.ergoweb.com/news/detail.cfm?id=1042.

de Véricourt F, Zhou YP (2005) Managing response time in a call-routing problem with service failure. *Oper. Res.* 53(6):968–981.

Flynn EA, Barker KN (2000) Medication errors research. Cohen MR, ed. *Medication Errors: Causes, Prevention, and Risk Management*, Chap. 6 (Jones and Bartlett Publishers, Sudbury, MA).

Gans N (2002) Customer loyalty and supplier quality competition. *Management Sci.* 48(2):207–221.

Gronroos C (1983) *Strategic Management and Marketing in the Service Sector* (Marketing Science Institute, Cambridge, MA).

Gupta S, Rajagopalan S (2012) Managing operations and customer demand at Coverking. Case study, Marshall School of Business, University of Southern California, Los Angeles.

Gurvich I, Armony M, Maglaras C (2009) Cross-selling in a call center with a heterogeneous customer population. *Oper. Res.* 57(2):299–313.

Hall J, Porteus R (2000) Customer service competition in capacitated systems. *Manufacturing Service Oper. Management* 2(2): 144–165.

Hallowell R (1996) The relationships of customer satisfaction, customer loyalty, and profitability: An empirical study. *J. Service Management* 4(7):27–42.

Hasija S, Pinker EJ, Shumsky RA (2008) Call center outsourcing contracts under information asymmetry. *Management Sci.* 54(4):793–807.

Hasija S, Pinker EJ, Shumsky RA (2010) Work expands to fill the time available: Capacity estimation and staffing under Parkinson's Law. *Manufacturing Service Oper. Management* 12(1):1–18.

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Jargon J (2009) Latest Starbucks buzzword: "Lean" Japanese techniques. *Wall Street Journal* (August 4), http://online.wsj.com/news/articles/SB124933474023402611.

Jargon J (2010) At Starbucks, baristas told no more than two drinks. (October 13), http://online.wsj.com/news/articles/SB10001424052748704164004575548403514060736.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.

Kirmani A, Rao AR (2000) No pain, no gain: A critical review of the literature on signaling unobservable product quality. *J. Marketing* 64(2):66–79.

Lord M (1999) There's a fly in my soup. *U.S. News & World Report* (November 22), 53.

Lovejoy WS, Sethuraman K (2000) Congestion and complexity costs in a plant with fixed resources that strives to make schedule. *Manufacturing Service Oper. Management* 2(3):221–239.

Lu LX, Van Mieghem JA, Savaskan RC (2009) Incentives for quality through endogenous routing. *Manufacturing Service Oper. Management* 11(2):254–273.

MacDonald JM, Aaronson D (2000) *How Do Retail Prices React to Minimum Wage Increases?* (Federal Reserve Bank of Chicago, Chicago).

Mahajan V, Muller E, Bass FM (1990) New product diffusion models in marketing: A review and directions for research. *J. Marketing* 54(1):1–26.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K (2002) Nurse staffing levels and the quality of care in hospitals. *The New England J. Medicine* 326(22):1715–1722.

Nelson P (1970) Information and consumer behavior. *J. Political Econom.* 78(2):311–329.

Oliva R, Sterman JD (2001) Cutting corners and working overtime: Quality erosion in the service industry. *Management Sci.* 47(7):894–914.

Parasuraman A, Zeithaml VA, Berry LL (1985) A conceptual model of service quality and its implications for future research. *J. Marketing* 49(4):41–50.

Png IPL, Reitman D (1994) Service time competition. *RAND J. Econom.* 25(4):619–634.

Ren ZJ, Wang X (2009) Should patients be steered to high volume hospitals? An empirical investigation of hospital volume and operations service quality. Working paper, Boston University School of Management, Boston.

Ren ZJ, Zhou Y (2008) Call center outsourcing: Coordinating staffing level and service quality. *Management Sci.* 54(2):369–383.

Scott MJ (2006) Health care in the express lane: The emergence of retail clinics. Report, California HealthCare Foundation, Oakland, CA.

Smith RA, Houston MJ (1983) *Script-Based Evaluations of Satisfaction with Services* (American Marketing Association, Chicago).

So KC, Song JS (1998) Price, delivery time guarantees and capacity selection. *Eur. J. Oper. Res.* 111(1):28–49.

Urban TL (2009) Establishing delivery guarantee policies. *Eur. J. Oper. Res.* 196(1):959–967.

Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing Service Oper. Management* 11(4):543–562.

Whitt W (2006) Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* 15(1):88–102.

Zeithaml VA, Berry LL, Parasuraman A (1993) The nature and determinants of customer expectations of service. *J. Acad. Marketing Sci.* 21(1):1–12.