MANAGEMENT SCIENCE
Vol. 50, No. 4, April 2004, pp. 479–488
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5004 | 0479

informs

DOI 10.1287/mnsc.1040.0170

Setting Customer Expectation in Service Delivery: An Integrated Marketing-Operations Perspective

Teck H. Ho
Haas School of Business, University of California, Berkeley, Berkeley, California 94720, hoteck@haas.berkeley.edu
Yu-Sheng Zheng
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, zheng@wharton.upenn.edu

Service firms have increasingly been competing for market share on the basis of delivery time. Many firms now choose to set customer expectation by announcing their maximal delivery time. Customers will be satisfied if their perceived delivery times are shorter than their expectations. This gap model of service quality is used in this paper to study how a firm might choose a delivery-time commitment to influence its customer expectation, and delivery quality in order to maximize its market share. A market share model is developed to capture (1) the impact of delivery-time commitment and delivery quality on the firm's market share and (2) the impact of the firm's market share and process variability on delivery quality when there is a congestion effect. We show that the choice of the delivery-time commitment requires a proper balance between the level of service capacity and customer sensitivities to delivery-time expectation and delivery quality. We prove the existence of Nash equilibria in a duopolistic competition, and show that this delivery-time commitment game is analogous to a Prisoners' Dilemma.

Key words: customer expectation; delivery-time commitment; queueing theory; gap model of quality
History: Accepted by Christopher S. Tang, special issue editor; received January 2002. This paper was with the authors 9 months for 2 revisions.

1. Motivation and Integrative Framework

Increasingly, firms have been competing on the basis of response, delivery, or shipping time. Many firms now choose to announce a guarantee on their maximal service delivery time in order to entice customers. For example, several cable TV companies (e.g., Time Warner Cable) guarantee that they will be on time for installation—otherwise, their installation is free. Similarly, many product firms (e.g., Tradewinds Coffee) waive their shipping charges if they do not deliver their products on time. Some banks (e.g., IndyMac Bank) even offer handsome rebates on mortgage closing costs if they fail to respond to loan applications within a number of hours. The conventional wisdom is that such commitment can provide a powerful source of competitive advantage if the service guarantee represents a breakthrough in service and the firm is able to fulfill the guarantee at high reliability.

How does a firm choose a delivery-time commitment that will have the most significant marketing impact, and what factors determine this choice? In selecting a delivery-time commitment, the firm must consider not only how customers will react to the commitment, but also whether it has adequate service capacity (e.g., level of staffing) to fulfill the commitment with high reliability. A tight delivery-time commitment has both benefits and costs. It can attract impatient customers, but the performance of a congested system might deteriorate unless service capacity is expanded accordingly. Depending on the inherent random nature of the customer arrival and service delivery processes, an excessive capacity may be required to fulfill the tight service guarantee. Thus, the choice of a delivery-time commitment requires careful consideration of both marketing-related (i.e., customer) and operations-related (i.e., capacity) factors.

This paper presents an integrative framework that allows the analysis of the fundamental trade-off above. We consider a service firm that is interested in maximizing its demand rate (which is equivalent to its market share when the total demand rate for the industry is held fixed). While the firm's demand rate is potentially affected by other service attributes, we focus on the impact of the service delivery time and assume that customers would be attracted by a low expected maximal delivery time and a high delivery quality. Here the delivery quality is restricted in the time dimension. We define the delivery quality as conformance of the customer's perceived delivery time to the expected delivery time. More precisely, the delivery quality is the probability that the

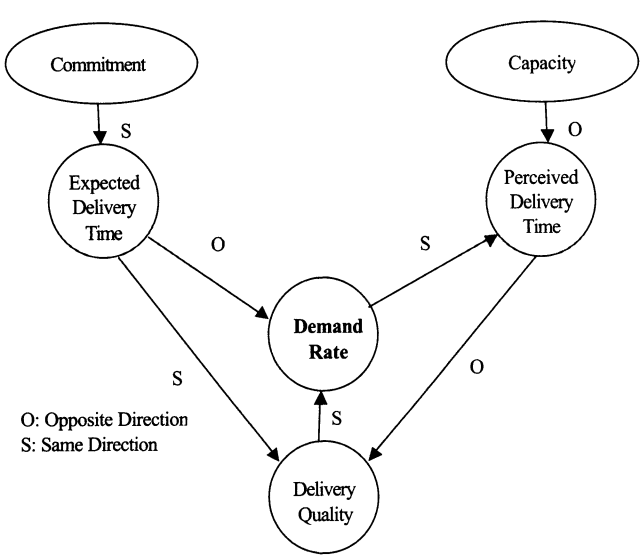
perceived delivery time is shorter than the expected maximal delivery time.

While the customer's expected delivery time can be influenced by other factors such as price, word of mouth, communications controlled by the company, and prior service experiences (Zeithaml et al. 1993), we naturally assume that an announced commitment sets the customer's expected delivery time. Larson (1987) has observed that the perceived delivery time can be influenced by many psychological and social factors. It is, however, reasonable to assume that the perceived delivery time is positively related to the actual delivery time, which is determined by both the demand rate and the level of capacity. A high demand rate increases the degree of congestion, and thus lengthens the perceived delivery time. The integrative framework is illustrated by an influence diagram, as shown in Figure 1.

This integrative framework builds on models and concepts from the marketing and operations literatures. The basic building block of the above integrative framework is the well-known gap model of service quality developed in the marketing literature (Anderson 1973, Oliver 1977, Parasuraman et al. 1985, Boulding et al. 1994). The gap model suggests that if a customer expects a certain level of service, and perceives the service received to be higher, she will be a satisfied customer. This stream of literature points to the importance of managing customer expectation and perception for improving service quality. In addition, it is empirically shown that purchase intention (and hence demand rate) increases as service quality improves (Boulding et al. 1994). We contribute to this literature in three ways:

- Our definition of delivery quality captures the impact of process variability on quality explicitly, a critical dimension that is often ignored in the literature.
- We model the impact of congestion explicitly by incorporating the influence of the demand rate on the perceived delivery time.
- Service capacity is considered explicitly so that the ability of the firm to meet the expected delivery time can be investigated.

Figure 1 The Integrative Framework



Delivery time in a congested system is the central topic of the vast queueing-theory literature (for comprehensive reviews see, for example, Kleinrock 1975 and Cooper 1990), which provides us with a good understanding of service system performance for various customer arrival and service processes. Typical system performance measures of interest include server (manpower or facility) utilization, queue length, and delivery time. Delivery time is further classified into the so-called delay (which is the waiting time in queue before entering service) and total waiting time (i.e., sum of the delay and service time). The level of capacity is usually modeled by the system configuration (e.g., number of servers and the service-time distribution). Our model framework employs the well-understood relationships between delivery time and demand rate, as well as level of capacity, developed in this body of literature. Our framework differs from the traditional queueing literature in the following ways:

- Customer's expectation on delivery time is explicitly considered in the analysis of the system.
- Delivery quality, which is also the level of customer satisfaction, is used to measure system performance. Average queue length and waiting time have impact on delivery quality, but they are not equivalent to it.[1]
- The demand rate here is endogeneous rather than exogeneous.

Based on the integrative framework, we develop a normative model to study the impact of a delivery-time commitment. A simple graphical representation is used throughout in our model analysis, which begins with the establishment of demand-rate equilibrium. When congestion effect is negligible (in systems with ample capacity), we obtain a closed-form solution for the optimal delivery-time commitment. Under congestion, we derive optimality conditions for the delivery-time commitment and use them to design an algorithm for computing the optimal commitment. We also analyze a duopolistic delivery-time commitment game and establish conditions for the existence

[1] The traditional queueing literature has not always paid attention to the human aspect of the service encounter. One important exception is Hall (1991), which considers balking and reneging behavior in queues. Other exceptions include Larson (1987), Rothkopf and Rech (1987), Green and Kolesar (1987), and Carmon et al. (1995).

of *Nash equilibria*. We illustrate with an example of how this game is analogous to a Prisoners' Dilemma.

This paper is organized as follows. Section 2 develops the mathematical model. Optimality conditions for delivery-time commitment are derived and a computational scheme for calculating the commitment is outlined in §3. Section 4 extends the analysis to a duopolistic competition. Section 5 concludes and suggests future research directions.

2. The Basic Model

2.1. Delivery Time and Delivery Quality

We consider a firm that serves a population of homogeneous customers who are impatient and sensitive to service delivery time.[2] The firm's objective is to maximize its demand rate, which is affected by customers' expectation for the delivery time as well as the probability that this expectation is being fulfilled. Let the service delivery time be denoted by t, which is a random variable because customer arrival and the service processes are inherently random. Let T be the customers' expected maximal delivery time. We define $Q = \text{Prob}(t \le T)$, which is the probability that a service delivery meets the customers' expectation, as the delivery quality. Ceteris paribus, a customer is more likely to use the service if it has a tighter delivery-time commitment and a higher delivery quality.

Here we assume that the customer population is homogeneous. The delivery quality, as defined above, is equivalent to the fraction of satisfied customers.[3] In the context of managing service firms for customer satisfaction, we believe that this definition of quality seems more relevant than the commonly used system performance measures such as average waiting time and queue length. We also note that the above definition of quality can be easily extended to other service attributes.

The delivery time depends on the demand rate and service capacity. Fix the firm's process capacity. Let λ be the firm's demand rate and $F(s, \lambda)$ be the probability distribution function of the delivery time. Thus,

$$Q = F(T, \lambda). \tag{2.1}$$

We model the congestion effect by assuming that $F(s, \lambda)$ is decreasing (nonincreasing) in λ. The service system is referred to as an uncongested system if F is independent of λ.

[2] In some service contexts, customer may actually prefer delay (Greenleaf and Lehmann 1995), possibly enjoying the anticipation of the event.

[3] A more refined model may capture the extent of delay experienced by customer. Our model basically assumes that customer satisfaction is a binary variable: Customer is happy iff the firm meets its promised delivery time.

For a given demand rate λ, $F(\cdot, \lambda)$ is the distribution function (cdf) of delivery time, which is readily available either in exact closed-form expressions or in good approximations for F for *actual* delivery times for various classes of customer arrival and service delivery processes. Depending on the application, the delivery time can be referred to as either the waiting time in queue or the total system time (the waiting time + the service time). In bank teller and telephone-ordering/enquiry services, the waiting time in queue is more relevant. For these applications, the classic $M/M/c$ queueing system is an appropriate model. Let μ be the service rate of a server and $a = \lambda/\mu$. The waiting-time distribution can be expressed as follows (Kleinrock 1975):

$$F(s, \lambda) = 1 - A(\lambda)e^{-(c\mu-\lambda)s}, \tag{2.2}$$

where

$$A(\lambda) = \frac{a^c/\big(c!(1-(a/c))\big)}{a^c/(c!(1-(a/c))) + \sum_{i=0}^{c-1}(a^i/i!)}$$

is the probability that an incoming customer has to wait. In many repair, mailing, and fast food delivery services, however, customers are interested in the total system time. In this case, for simplicity we model the whole service delivery process as an $M/M/1$ queueing system in our analysis. The distribution of the total system time of the $M/M/1$ system is:

$$F(s, \lambda) = 1 - e^{-(\mu-\lambda)s}. \tag{2.3}$$

It has been observed that the perceived delivery time may not be the same as the actual delivery time. Katz et al. (1991) showed empirically that customers visiting bank tellers tended to overestimate the amount of time they spent waiting in line and that the difference between perceived and actual waiting times is approximately normal, with a mean overestimation of one minute and a standard deviation of 2.5 minutes.

We assume that a delivery-time commitment will narrow the gap between perceived and actual waiting times because customers will become more conscious about the actual time and may monitor it more closely as a result of the firm's service commitment. It remains, however, an empirical question as to how delivery-time commitment will impact the gap between perceived and actual delivery times. In our numerical examples, we use the actual delivery-time distribution function. In particular, we will use (2.2) and (2.3) as the delivery-time distribution of service systems that can be represented by $M/M/c$ systems and $M/M/1$ systems, respectively.

2.2. Demand-Rate Equilibrium

We model the firm's demand rate by the following general formulation.

$$\lambda = \Lambda \cdot S(U), \tag{2.4}$$

where:

Λ: total demand rate of the market,

S: firm's market share,

U: customer's utility for the firm's service.

Note that Λ is assumed to be fixed, and $S \in (0, 1)$ may be any continuous, increasing function. The customer's utility for the firm's service depends on the expected delivery time and service quality:

$$U(T, Q) = \beta_0 - \beta_T T + \beta_Q Q, \tag{2.5}$$

where β_0, β_T, and β_Q are nonnegative constants. β_T and β_Q reflect customer sensitivity to the delivery-time expectation and to the service quality, respectively, and β_0 summarizes her utility for all the firm's other attributes. The model says that the firm's market share is decreasing in the delivery-time expectation and increasing in the service quality. We note, in general, that β_Q could also depend on T. If more-impatient consumers care more about delivery quality, then we have $\beta_Q = \beta_Q(T)$ being decreasing in T.

A distinction has recently been made between two types of expectations: (1) the "will" expectation (i.e., a level that is expected to occur) and (2) the "should" expectation (i.e., a level that ought to happen) (Tse and Wilton 1988, Boulding et al. 1994). These researchers show empirically that the higher the customer's "will" expectation and the lower their "should" expectation of the service, the more satisfied she is likely to find the service. A slightly more general version of our model can capture this distinction. If we rewrite $U(T_W, T_S, Q)$ as $\beta_0 - \beta_T T_W + \beta_Q F(T_S, \lambda)$, the customers utility is increasing in "will" expectation (as higher "will" expectation is indicated by a lower T_W) and decreasing in "should" expectation (i.e., higher T_S).[4] In this paper, we implicitly assume that announcement of a delivery-time commitment will close the gap between the "will" (T_W) and the "should" (T_S) expectations so that $T_W \approx T_S$ (c.f. Green et al. 1992).

The market share function S can take various functional forms (see Lilien et al. 1992 for a review). We use the multilogit model in our numerical examples throughout the paper. This market share model is widely used in marketing and operations research (McFadden 1980, Lee and Cohen 1985, Cooper 1993). Here, the market share of firm i, S_i, in an industry with m firms is given by:

$$S_i = \frac{e^{U_i}}{\sum_{j=1}^{m} e^{U_j}}, \tag{2.6}$$

where U_j is the customer's utility for firm j's service. In §3, we concentrate on analyzing a passive competitive environment and assume that the customer's utility for other firms' services is not significantly affected by the firm's decision. In the multilogit model, this implies that the firm's market share is given by (dropping the subscript i):

$$S = \frac{e^{U}}{e^{U} + A}, \tag{2.7}$$

where $A = \sum_{j \neq i} e^{U_j}$. In §4, we explore how the customer's utilities for different firms interact with each other in an duopolistic market.

For notational convenience, let

$$\phi(T, \lambda) = \Lambda S\big[U(T, F(T, \lambda))\big].$$

Equation (2.4) becomes:

$$\lambda = \phi(T, \lambda). \tag{2.8}$$

We note that, due to the congestion effect, the customer's utility is decreasing (nonincreasing) in λ for a given T, and so is $\phi(T, \lambda)$. Because the demand rate is endogeneous and appears in both sides of Equation (2.8), the existence of equilibrium, which is the solution of (2.8), must be established first. For convenience, we intuitively refer to $\phi(T, \lambda)$ as "tomorrow's demand rate" given todays' demand rate λ. A market equilibrium is reached when tomorrow's demand rate is the same as today's.

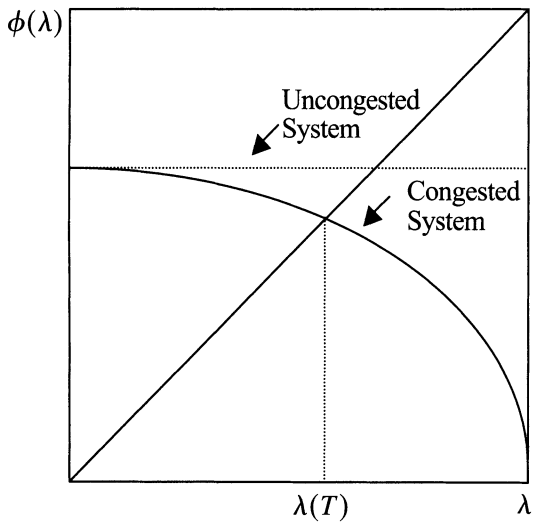
Proposition 1. *For any given T, there exists a unique $\lambda(T) \in [0, \Lambda]$ that satisfies* (2.8).

Proof. Since $\phi(T, 0) > 0$, $\phi(T, \Lambda) < \Lambda$, and ϕ is continuous in λ, $\phi(T, \lambda) - \lambda = 0$ has a solution $\lambda(T)$ due to the *mean-value* theorem. The uniqueness follows from the fact that $\phi(T, \lambda) - \lambda$ is strictly decreasing in λ (because $\phi(T, \lambda)$ is decreasing in λ). □

In Figure 2, we let the horizontal axis represent today's demand rate and the vertical axis tomorrow's. Then, for a given T, the ϕ function is represented by a continuous curve that has a negative slope (because of the congestion effect). The equilibrium demand rate is the intersection of the ϕ curve and the 45-degree straight line. For a given T, the slope of the ϕ curve provides a measure for the degree of congestion: A higher slope (in absolute value) indicates a more congested system. A horizontal line indicates an uncongested system (or a system with ample capacity) since tomorrow's demand rate is not affected by today's demand rate.

[4] Boulding et al. (1994) indicate that ideally one would want to simultaneously increase customer's "will" expectation and decrease their "should" expectation. They suggest that such activity seems impossible. In our model, delivery-time commitment is a marketing activity that will increase both the customer's "will" and "should" expectations.

Figure 2 $\phi(T, \lambda)$



3. Maximization of Demand Rate

We consider the following optimization problem:

$$\max_T \ \lambda$$
$$\text{subject to} \quad \lambda = \phi(\lambda, T). \tag{3.1}$$

Let T^* be the optimal delivery-time commitment, and λ^* be the maximum demand rate the firm can obtain by making a proper choice of T; that is, $T^* = \arg\max_T \lambda(T)$ and $\lambda^* = \lambda(T^*)$. Next, we show that T^* can be identified efficiently.

3.1. Uncongested Systems

We say a system is uncongested if an incoming customer never needs to wait. Such a system may be modeled nicely as an $M/G/\infty$ queueing system with an infinite number of servers and an arbitrary service-time distribution. For an uncongested service system, the delivery time is of course referred to as the service time for each individual customer, which is independent of the demand rate. Letting $G(T)$ be the distribution function of the service time, we have $F(T, \lambda) = G(T)$.

Because in this case the right-hand side of (2.8) is independent of λ, and the S-function is increasing, the optimal time commitment T^* also maximizes U. Let $g(\cdot)$ be the density function of the delivery time. Note that $\partial U(T, \lambda)/\partial T = -\beta_T + \beta_Q g(T)$, which is decreasing in the range of T where $g(T)$ is decreasing. Here we make the further assumption that the relevant range of T is the range where $g(T)$ is decreasing. Within this range, $g(T^*) = \beta_T/\beta_Q$ is a necessary and sufficient condition because then $\partial^2 U/\partial T^2 = \beta_Q g'(T) < 0$. For most well-behaved probability distributions, the density function $g(t)$ declines for the range of t when $G(t)$ is close to 1 (say, ≥ 0.8). This assumption is plausible because a delivery-time commitment makes sense only if the service quality is high enough. For example, our assumption means that the quality of the commitment must be at least 50% if g is of normal density. Under these assumptions, and for appropriate β_T and β_Q, we may write:

$$T^* = g^{-1}\left(\frac{\beta_T}{\beta_Q}\right). \tag{3.2}$$

Thus, the choice of a delivery-time commitment requires a proper understanding of both customer attitudes and service delivery process. The form of the optimal time commitment suggests that only the tail distribution of the service delivery process matters. This result should not be surprising because the "tail" region is where the firm does not fulfill its service commitment. Consequently, a fat-tail process must be accompanied by a looser commitment. Thus, a competitive marketing strategy that is based on a tight delivery-time commitment must be matched by a first-class service process that has a thin tail.

The optimal time commitment should be tighter if the customers are highly impatient (high β_T). This may explain why many service firms are pushing for a tight delivery-time commitment. However, as indicated in (3.2), this is just one of the three factors that determine the level of service commitment. The level of commitment is also affected by the customers' sensitivity to service quality. A looser commitment should be adopted if the customers are very conscious about service quality. Failing to meet customers' expectation in a service delivery can hurt the firm's future market share. Indeed, we suspect that many service firms "overcommit" and ignore the ramifications of failing to keep a service guarantee.

In an uncongested system, the optimal level of time commitment is not a function of the competitive attraction level A. This is so because the assumed ample capacity decouples the firm from its environment. We shall show, in §3.2, that this is not true in a congested system. This suggests that if a particular firm in an industry has ample capacity (acquired perhaps through a new technology), it needs only to ensure that its service guarantee matches the speed of the service process, and may ignore the level of service of the competitors.

Example 1. Exponential service time $g(t) = \mu e^{-\mu t}$, where μ is the mean service rate.

$$T^*(\mu) = \frac{1}{\mu} \ln\left(\frac{\beta_Q \mu}{\beta_T}\right). \tag{3.3}$$

Note that $T^* > 0$ only if $\mu > \beta_T/\beta_Q$. Note that T^* is convex in μ for positive T^*. Also note that $\partial T^*/\partial \mu < 0$ for $\mu\beta_Q/\beta_T < e$ or the delivery quality is below $1 - e^{-\mu T^*} > 1 - e^{-1} = 63.2\%$. Under the high delivery quality assumption, a faster service process should be accompanied by a tighter delivery-time commitment.

EXAMPLE 2. Normal service time

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/\sigma^2},$$

where m and σ are the mean and variance of the service time.

$$T^* = m + \sigma\sqrt{\ln \frac{\beta_Q}{\sqrt{2\pi}\sigma\beta_T}}. \tag{3.4}$$

The expression is valid only for $\sigma < \beta_Q/(\sqrt{2\pi}\beta_T)$. Note that T^* increases linearly with m. In the high-service region (above 78%), T^* is increasing and concave in σ. Thus, a decrease in σ will not lead to a proportionate decrease in time commitment. A numerical example will make this point clear. Let $\beta_T = 0.01$, $\beta_Q = 4.0$, $m = 5$, and $\sigma = 0.1$. Then $T^* = 8.995$. If σ is reduced to 0.05 (i.e., by 50%), then T^* becomes 7.825 (i.e., by 15%).

3.2. Congested Systems

For general congested systems, maximizing the demand rate is more involved. Let $f(t, \lambda)$ be the density function of the delivery time. The necessary optimality condition is characterized by the following pair of equations:

$$f(\lambda^*, T^*) = \frac{\beta_T}{\beta_Q}, \tag{3.5}$$

$$\lambda^* = \phi(\lambda^*, T^*). \tag{3.6}$$

Besides β_Q, β_T, and the tail distribution f, T^* is also a function of the total demand rate Λ and the competitive attraction level A (since it depends on the ϕ function). In general, there is no closed-form solution for T^*. We propose a procedure to compute T^* below.

LEMMA 1. *For any given* T^0, *let* $\lambda^0 = \lambda(T^0)$, *let* $T^1 = \arg\max_T \phi(T, \lambda^0)$, *and let* $\lambda^1 = \lambda(T^1)$. *We have* $\lambda^1 \geq \lambda^0$.

PROOF. By definition, $\phi(T^1, \lambda^0) \geq \phi(T^0, \lambda^0) = \lambda^0$. Since $\phi(T^1, \lambda^1) = \lambda^1$ and $\phi(T^1, \cdot)$ is a decreasing function, we have $\lambda^1 \geq \lambda^0$. □

Algorithm

Step 1. For an initial T^0, find $\lambda^0 = \lambda(T^0)$.

Step 2. Find $T^1 = \arg\max_T \phi(T, \lambda^0)$.

Step 3. If $T^1 = T^0$, then $T^* := T^1$ and stop, else $\lambda^0 := \lambda(T^1)$ and repeat Step 2.

PROPOSITION 2. *The algorithm finds* T^*.

PROOF. We prove by contradiction. Let T^* be the solution generated by the algorithm, and $\lambda^* = \lambda(T^*)$. Suppose, on the contrary, that there was a better time commitment, say T', under which the firm could have a larger demand rate λ'. Then, since $\phi(T', \lambda)$ is decreasing in λ, $\phi(T', \lambda^*) \geq \phi(T', \lambda') = \lambda' > \lambda^* = \phi(T^*, \lambda^*)$, which is a contradiction. □

Similar to the assumption made in the uncongested system, we assume that T^* is always at the declining tail of the density function of the actual delivery time, i.e., $T^* \in \{t: g(t, \lambda^*) \text{ is decreasing}\}$. We further assume that for any pair of λ_i, $i = 1, 2$ with $\lambda_2 > \lambda_1$, if $g(t, \lambda_2)$ is decreasing in $t \in [a, b]$, so is $g(t, \lambda_1)$. Under these assumptions, the algorithm is rather efficient because it follows logic similar to that in §2.1, that in Step 2, T^1 is the inverse function of $f(\cdot, \lambda)$. For a given density function such as the exponential or the normal, T^1 can be solved by a closed-form expression.

Using the above algorithm, we conduct an extensive numerical simulation. We observe that a larger A will result in a smaller λ^*, other things being equal. If the process has an exponential tail and a high-enough service level is assumed, T^* will be tighter for a higher competitive attraction level A. The total demand rate (Λ) has an opposite effect. It will lead to a higher λ^* and a looser time commitment. The parameters β_T and β_Q affect T^* in ways similar to those in uncongested systems.

3.3. Service Delivery Capacity

In this subsection, we study how the firm's optimal market share depends upon the firm's capacity. Assume that the firm's capacity level can be characterized by a variable C. In this subsection, we denote the delivery-time probability distribution function by $F(T, \lambda, C)$. Naturally, F is assumed to be increasing in C. We also augment all of the other notation by adding the argument C. For example, $\lambda(T, C)$ would be the demand rate satisfying (3.1) for given T and C, $\lambda^*(C)$ the maximum demand rate achievable for given C.

PROPOSITION 3. $\lambda^*(C)$ *is increasing in* C.

PROOF. For any C' and C with $C' > C$, we show $\lambda^*(C') \geq \lambda^*(C)$.

$$\begin{aligned}\lambda^*(C') &= \max_T \lambda(T, C') \geq \lambda(T^*, C') \\ &= S[U(T^*, \lambda(C', T^*))] \\ &\geq S[U(T^*, \lambda(C, T^*))] = \lambda^*(C). \quad \square\end{aligned}$$

Proposition 3 suggests that the optimal market share is increasing in capacity. A stronger result that we have been unable to prove but will be very useful is that the optimal market share is concave in C. In this case, the capacity-planning problem ($\max_C \lambda^*(C)$) will reduce to solving the first-order condition. An extensive numerical analysis shows that under reasonable parameter values, the optimal market share is a concave function of C.

4. Competitive Interactions

In this section, we consider a duopolistic setting in which two firms compete for a fixed market.

Throughout this section, we assume the logit model for market share function. For each firm, the model framework studied in §§1 and 2 remains valid with one exception: The effect of one firm's decision on the other can no longer be ignored. This is because one firm's gain must be the other's loss when the total demand rate of the market is fixed.

We add subscript i ($i = 1, 2$) to all the notation to represent firm is. Let $\mathbf{T}$ and $\boldsymbol{\lambda}$ be the vectors of $\{T_1, T_2\}$ and $\{\lambda_1, \lambda_2\}$, respectively. Then, for any given $\mathbf{T}$, the market equilibrium is reached under the following conditions:

$$\lambda_i = \phi_i(\mathbf{T}, \boldsymbol{\lambda}), \qquad i = 1, 2, \tag{4.1}$$

$$\Lambda = \lambda_1 + \lambda_2, \tag{4.2}$$

where

$$\phi_i(\mathbf{T}, \boldsymbol{\lambda}) = \Lambda \cdot \frac{e^{U_i}}{e^{U_1} + e^{U_2}}, \qquad i = 1, 2, \tag{4.3}$$

and

$$U_i(T_i, \lambda_i) = \beta_{0i} - \beta_T T_i + \beta_Q F_i(T_i, \lambda_i), \qquad i = 1, 2. \tag{4.4}$$

Note that we allow the sensitivity parameter β_0 to be different between the two firms to reflect the differences in their other service attributes, but assume that β_T and β_Q are the same.

For fixed $\mathbf{T}$, since $\Lambda = \lambda_1 + \lambda_2$, ϕ_i can be viewed as a function of λ_i only. Due to symmetry, we may further focus on analyzing one firm, say, Firm 1. The market equilibrium equation for Firm 1 (the first equation of (4.1)) can be written as

$$\lambda_1 = \bar{\phi}(\mathbf{T}, \lambda_1), \tag{4.5}$$

where $\bar{\phi}(\mathbf{T}, \lambda_1) \equiv \phi_1(\mathbf{T}, \{\lambda_1, \lambda_2\}) = \phi_1(\mathbf{T}, \{\lambda_1, \lambda - \lambda_1\})$. It follows from (4.4) and the congestion effect that $\bar{\phi}$ is continuous and decreasing in λ_1. Therefore, as before, we know that there exists a unique market equilibrium $\lambda_1(\mathbf{T})$ that satisfies (4.1) and (4.2).

We note that $\bar{\phi}$ decreases faster than ϕ in §2. To see this, we compare (4.3) with (2.7). The difference is that while previously A was assumed to be a constant independent of λ, e^{U_2} in (4.3) is an increasing function of λ_1. In other words, in the duopolistic competition, when Firm 1's demand rate increases, Firm 2's must decrease by the same amount. Due to the congestion effect, Firm 1's delivery quality deteriorates and Firm 2's improves. In turn, customers' utility of Firm 1's service decreases and that of Firm 2's increases. Both of these changes contribute to the decrease of Firm 1's tomorrow's demand rate; see (4.3). Thus, the benefits of a tight delivery-time commitment is less than that without competitive interaction. This is an important point. Even if Firm 2 does not respond to Firm 1's move to a tighter time commitment, Firm 2's delivery quality will improve as a result of less congestion. This challenges the wisdom that a drop in delivery time will lead to a quantum leap in market share.

Next, we address the question of whether a Nash equilibrium exists in this duopolistic competition. A set of time commitments is in equilibrium if, given time commitments of other firms, a firm cannot increase its own market share by choosing a time commitment other than the equilibrium time commitment. To show its existence, it suffices to show that $\lambda_1(\mathbf{T}) = \lambda_1(\{T_1, T_2\})$ is unimodal in T_1 for any given T_2 (i.e., $\lambda_1(\mathbf{T})$ is quasi-concave in T_1 given T_2). Fix T_2. Focus on Firm 1 so that we may drop the subscript 1 whenever this would not cause confusion. Let $\bar{\lambda}(T) \equiv \lambda_1(\{T, T_2\})$. It is geometrically clear (see Figure 3) that in order to show the unimodality of $\bar{\lambda}(T)$, it suffices to show that for any $T^c > T^b > T^a$,

(i) $\bar{\phi}(\{T^b, T_2\}, 0) < \bar{\phi}(\{T^a, T_2\}, 0)$;

(ii) $\bar{\phi}(\{T^b, T_2\}, \lambda)$ and $\bar{\phi}(\{T^a, T_2\}, \lambda)$ cross at most once; and

(iii) $\phi(\{T^c, T_2\}, \lambda)$ and $\phi(\{T^b, T_2\}, \lambda)$ do not cross before $\phi(\{T^b, T_2\}, \lambda)$ and $\phi(\{T^a, T_2\}, \lambda)$ do.

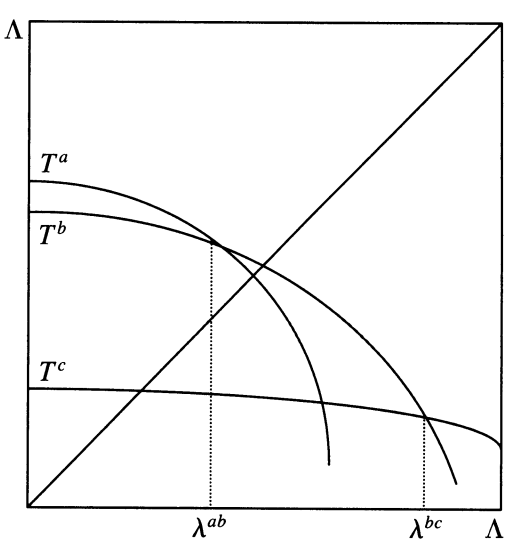
Since for fixed λ, U_2 is the same for $\bar{\phi}$s with different T_1 values, and because $\bar{\phi}$ is strictly increasing in U_1, it suffices to show the above properties (i)–(iii) for U_1 instead of $\bar{\phi}$.

Proposition 4. *Nash equilibrium exists if the service processes of the two firms in the duopolistic competition are $M/M/1$ systems and the delivery time of interest is the total system time.*

Proof.

1. When $\lambda \simeq 0$, the delivery quality is $F(T, 0) = 1 - e^{-\mu T}$ (see Equation (2.3)), and $U(T, 0) = \beta_0 - \beta_T T + \beta_Q(1 - e^{-\mu T})$, which is the utility function of the corresponding uncongested system. Note that $U(T, 0)$

Figure 3 Proof of Existence of Nash Equilibrium



is not monotone in T. In fact, from the discussion in §3.1, we know that $U(T,0)$ is concave and reaches its maximum at $T_o = (1/\mu)\ln(\beta_Q\mu/\beta_T)$. We also know that when the delivery quality is sufficiently high, the optimal time commitment is decreasing in the service rate. For any positive λ, $U(T,\lambda) = \beta_0 - \beta_T T + \beta_Q(1 - e^{-(\mu-\lambda)T})$, which can be viewed as the utility function of the uncongested system with an *effective* service rate $\hat{\mu} = \mu - \lambda$ $(<\mu)$. Therefore, the firm's choice of time commitment must be larger than T_o. So, we may restrict our discussion within the range of $\{T\colon T > T_o\}$ without loss of rigor. Clearly, $U(T,0)$ is decreasing in T.

2. We show that for any $T^b > T^a$, $U(T^b,\lambda) - U(T^a,\lambda)$ is monotonically increasing in the relevant range of λ, which in turn implies that $U(T^a,\lambda)$ and $U(T^b,\lambda)$ cross at most once. Since

$$\frac{d[U(T^b,\lambda) - U(T^a,\lambda)]}{d\lambda} = T^a e^{-(\mu-\lambda)T^a} - T^b e^{-(\mu-\lambda)T^b},$$

it reduces to show that $Te^{-(\mu-\lambda)T}$ is decreasing in T. Differentiating it with respect to T, we have $\partial(Te^{-(\mu-\lambda)T})/\partial T = e^{-(\mu-\lambda)T} - T(\mu-\lambda)e^{-(\mu-\lambda)T}$. It is negative when $T(\mu-\lambda) > 1$, or the delivery quality is higher than 63.2%. Therefore, we have shown that $U(T^a,\lambda)$ and $U(T^b,\lambda)$ cross at most once in the range of λ where the delivery quality is higher than 63.2%.

3. For any $T^c > T^b > T^a$ $(>T_o)$. Let λ^{ab} be an intersection of $U(T^a,\lambda)$ and $U(T^b,\lambda)$, and λ^{bc} be an intersection of $U(T^b,\lambda)$ and $U(T^c,\lambda)$. By definition, we have $U(T^a,\lambda^{ab}) = U(T^b,\lambda^{ab})$ and $U(T^b,\lambda^{bc}) = U(T^c,\lambda^{bc})$. From these two equalities, we obtain

$$\frac{F(T^b,\lambda^{ab}) - F(T^a,\lambda^{ab})}{T^b - T^a} = \frac{\beta_T}{\beta_Q} = \frac{F(T^c,\lambda^{bc}) - F(T^b,\lambda^{bc})}{T^c - T^b}.$$

We need to show that $U(T^c,\lambda^{ab}) < U(T^b,\lambda^{ab})$ and $U(T^a,\lambda^{bc}) < U(T^b,\lambda^{bc})$. The first inequality is true because

$$\begin{aligned} &U(T^c,\lambda^{ab}) - U(T^b,\lambda^{ab}) \\ &\quad = \beta_Q(T^c - T^b)\left[\frac{F(T^c,\lambda^{ab}) - F(T^b,\lambda^{ab})}{T^c - T^b} - \frac{\beta_T}{\beta_Q}\right], \\ &\quad = \beta_Q(T^c - T^b)\left[\frac{F(T^c,\lambda^{ab}) - F(T^b,\lambda^{ab})}{T^c - T^b} \right. \\ &\qquad\qquad \left. - \frac{F(T^b,\lambda^{ab}) - F(T^a,\lambda^{ab})}{T^b - T^a}\right], \end{aligned}$$

and $F(T,\lambda)$ is (exponentially) concave in T. Similarly, $U(T^a,\lambda^{bc}) < U(T^b,\lambda^{bc})$ because

$$\begin{aligned} &U(T^b,\lambda^{bc}) - U(T^a,\lambda^{bc}) \\ &\quad = \beta_Q(T^b - T^a)\left[\frac{F(T^b,\lambda^{bc}) - F(T^a,\lambda^{bc})}{T^b - T^a} \right. \\ &\qquad\qquad \left. - \frac{F(T^c,\lambda^{bc}) - F(T^b,\lambda^{bc})}{T^c - T^b}\right]. \end{aligned}$$

Thus, $\bar{\lambda}(T)$ is unimodal, and the proposition is proven. □

For firms with $M/M/c$ service systems and the queueing time being of interest, it is easy to show that (i) and (iii) holds: (i) is true because when $\lambda \simeq 0$, the service quality $F(T,0) = 1$ (a customer hardly needs to wait). Therefore, $U(T,0) = \beta_0 - \beta_T T$, and it is obvious that $U(T,0)$ is decreasing in T; (iii) holds because the waiting-time distribution is also exponential, and hence concave in T. For the general $M/M/c$ case, we have not been able to prove analytically that (ii) holds. We conjecture this on the basis of numerical examples. We prove this for the special case of $M/M/1$ with queueing time.

Proposition 5. *Nash equilibrium exists if the service processes of the firms in the duopolistic competition are $M/M/1$ systems and the delivery time of interest is the waiting time in queue.*

Proof. The service quality of the $M/M/1$ system (in waiting time) is $F(T,\lambda) = 1 - (\lambda/\mu)e^{-(\mu-\lambda)T}$. As before, we need to show that $\partial F/\partial\lambda$ is an increasing function of T, since

$$\frac{\partial F}{\partial\lambda} = -\frac{1}{\mu}e^{-(\mu-\lambda)T} - \frac{\lambda}{\mu}Te^{-(\mu-\lambda)T},$$

where the first term is clearly increasing in T, and the second has been shown (in the proof of the previous proposition) to be increasing in the range of $(\mu-\lambda)T > 1$. For λ in the range of $(\mu-\lambda)T < 1$ and $F(T,\lambda)$ is high, we need to show that

$$\frac{\partial^2 F}{\partial\lambda\partial T} = \mu - 2\lambda + \lambda T(\mu-\lambda) \geq 0,$$

which holds true for $\mu \geq 2\lambda$. Note, however, that for $(\mu-\lambda)T > 1$, the service quality $Q < 1 - (\lambda/\mu)e^{-1}$. The service level would be lower than $1 - (1/2)e^{-1} \simeq 0.82$, which is not relevant in today's service environment.

Example 3: $M/M/1$ Systems (Total System Time). Let $\beta_0 = 0$, $\beta_T = 0.01$, $\beta_Q = 4.0$, and $\Lambda = 1$. We determine the optimal market shares for four games with service rate for both firms varying at two levels (2.0 and 4.0). The optimal market shares are summarized in Figure 4.[5] In Figure 4, the first number in each cell is the payoff (in units of market share) to Firm 1 and the second number to Firm 2. For instance, when both firms have the same capacity of 2.0, each firm receives a payoff of 0.5. The value of ϵ captures the additional cost of high service capacity. If ϵ is

[5] Equilibrium delivery-time commitments are as follows. When both firms have the same service rate, the symmetric equilibrium delivery-time commitment is 4.3 ($\mu_1 = \mu_2 = 2$) and 2.1 ($\mu_1 = \mu_2 = 4$). When the service rates for the firm are different, the equilibrium time commitment for the faster firm is 2.1 and for the slower firm is 4.1.

Figure 4 The Duopolistic Game

Firm 1's Service Rate (μ_1) \ Firm 2's Service Rate (μ_2)	2.0	4.0
2.0	0.5, 0.5	$0.49, 0.51-\varepsilon$
4.0	$0.51-\varepsilon, 0.49$	$0.5-\varepsilon, 0.5-\varepsilon$

smaller than 0.01, then the game is similar to a Prisoners' Dilemma. Both firms will choose a service rate of 4.0 and receive a pareto-inferior payoff of $0.5-\epsilon$. If $\epsilon > 0.01$, then the equilibrium is (2.0, 2.0), with both firms receiving 0.5. Finally, it is worth noting that while both firms will experience a lower payoff for investing in additional service capacity, the resulting levels of delivery quality experienced by the customers will be higher.

5. Conclusion and Future Research

In this paper, we have presented a simple model for studying how a firm should set its delivery-time guarantee in managing service delivery. The model integrates the gap model of service quality from marketing with the classical queueing models from operations. We obtain a closed-form solution for the optimal delivery-time commitment when the firm has an ample capacity. Under congestion, we characterize the optimal delivery-time commitment with a set of conditions and use it to design a computational scheme. We prove the existence of Nash equilibria in a duopolistic game and show that the delivery-time game is similar to a Prisoners' Dilemma when the cost of adding capacity is small.

The model allows us to study several marketing-operations interface issues. First, if there exist multiple classes of customers that have significantly different β_T and β_Q values, then the delivery-time commitment for each class may be different, and it would be interesting to examine how the delivery-time commitment decision is tied to the pricing in each service class. So and Song (1998) and So (2000) considers the pricing issue in a single market segment case. Second, if a firm has service outlets in multiple locations with different total demand intensity and level of competition, it will be interesting to analyze the choice of a delivery-time commitment and the capacity design problem for each service outlet. Finally, if the reputation of a firm's service quality takes time to spread through the population, it is worthwhile to see how this might impact the choice of a delivery-time commitment (c.f. Gans 2002).

Acknowledgments

The authors thank seminar participants at MIT, Wharton, Stanford, and Berkeley for their helpful comments. Taizan Chan made excellent comments and suggestions and Grace Ho provided excellent research assistance.

References

Anderson, R. E. 1973. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *J. Marketing Res.* **10**(February) 38–44.

Boulding, W., R. Staelin, A. Kalra, A. Zeithaml. 1994. A dynamic process model of service quality: From expectations to behavioral intentions. *J. Marketing Res.* **30**(February) 7–27.

Carmon, Z., G. Shanthikumar, T. F. Carmon. 1995. A psychological perspective on service segmentation: The significance of accounting for consumers' perceptions of waiting and service. *Management Sci.* **41** 1806–1815.

Cooper, L. G. 1993. Market-share models. J. Eliashberg, G. L. Lilien, eds. *Marketing, Handbooks in Operations Research and Management Science*, Vol. 3. North-Holland, Amsterdam, The Netherlands.

Cooper, R. B. 1990. Queueing theory. D. P. Heyman, M. J. Sobel, eds. *Stochastic Models, Handbooks in Operations Research and Management Science*, Vol. 2. North-Holland, Amsterdam, The Netherlands.

Gans, N. 2002. Customer loyalty and supplier quality competition. *Management Sci.* **48** 202–221.

Green, L., P. Kolesar. 1987. On the validity and utility of queueing models of human service systems. *Ann. Oper. Res.* **9** 469–479.

Green, L., D. Lehmann, B. Schmitt. 1992. Time perceptions in service systems: An overview of the TPM framework. *Adv. Services Marketing Management* **5** 85–107.

Greenleaf, E., D. R. Lehmann. 1995. A typology of reasons for substantial delay in consumer decision making. *J. Consumer Res.* **22**(September) 186–199.

Hall, R. W. 1991. *Queueing Methods for Services and Manufacturing*. Prentice Hall, Englewood Cliffs, NJ.

Katz, K. L., B. M. Larson, R. C. Larson. 1991. Prescription for the waiting-in-line blues: Entertain, enlighten, and engage. *Sloan Management Rev.* **32**(2) 44–53.

Kleinrock, L. 1975. *Queueing Systems, Vol. 1: Theory*. John Wiley and Sons, New York.

Larson, R. C. 1987. Perspectives on queues: Social justice and the psychology of queuing. *Oper. Res.* **35**(6) 895–905.

Lee, H., M. Cohen. 1985. Equilibrium analysis of disgregate facility choice systems subject to congestion-elastic demand. *Oper. Res.* **33**(2) 293–311.

Lilien, G. L., P. Kotler, K. S. Moorthy. 1992. *Marketing Models*. Prentice Hall, Englewood Cliffs, NJ.

McFadden, D. 1980. Econometric models for probabilistic choice among products. *J. Bus.* **53**(3) 513–530.

Oliver, R. L. 1977. Effect of expectation and disconfirmation of post-exposure product evaluation: An alternative interpretation. *J. Appl. Psych.* **62**(April) 480–486.

Parasuraman, A., V. A. Zeithaml, L. Berry. 1985. A conceptual model of service quality and implications for future research. *J. Marketing* **64**(Spring) 12–40.

Rothkopf, M., P. Rech. 1987. Perspectives on queues: Combining queues is not always beneficial. *Oper. Res.* **35** 906–909.

So, R. 2000. Price and time competition for service delivery. *Manufacturing Services Oper. Management* **2**(4) 392–409.

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

Tse, D. K., P. C. Wilton. 1988. Models of consumer satisfaction formation: An extension. *J. Marketing Res.* **25** 203–212.

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