Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support

Ety Zohar • Avishai Mandelbaum • Nahum Shimkin
Faculty of Electrical Engineering, Technion, Haifa 32000, Israel
Faculty of Industrial Engineering, Technion, Haifa 32000, Israel
Faculty of Electrical Engineering, Technion, Haifa 32000, Israel
ety@tx.technion.ac.il • avim@tx.technion.ac.il • shimkin@ee.technion.ac.il

We address the modeling and analysis of abandonments from a queue that is invisible to its occupants. Such queues arise in remote service systems, notably the Internet and telephone call centers; hence, we refer to them as *tele-queues*. A basic premise of this paper is that customers adapt their patience (modeled by an abandonment-time distribution) to their service expectations, in particular to their anticipated waiting time. We present empirical support for that hypothesis, and propose an $M/M/m$-based model that incorporates adaptive customer behavior. In our model, customer patience depends on the *mean* waiting time in the queue. We characterize the resulting system equilibrium (namely, the operating point in steady state), and establish its existence and uniqueness when changes in customer patience are bounded by the corresponding changes in their anticipated waiting time. The feasibility of multiple system equilibria is illustrated when this condition is violated. Finally, a dynamic learning model is proposed where customer expectations regarding their waiting time are formed through accumulated experience. We demonstrate, via simulation, convergence to the theoretically anticipated equilibrium, while addressing certain issues related to censored-sampling that arise because of abandonments.

(*Exponential Queues; Abandonment; Invisible Queues; Tele-Queues; Adaptive Customer Behavior; Tele-Services; Call Centers*)

1. Introduction

Customer characteristics in service systems are largely dependent upon the system performance characteristics as perceived by its users. For example, the arrival rate is likely to increase as the typical waiting time decreases. This dependence interacts with the queueing process to determine the system operating point, and may have a considerable effect on performance.

Our focus in this paper is on the modeling of customer abandonments and their interplay with the system performance. We consider a queueing system with impatient customers, who may abandon the queue if not admitted to service soon enough. We assume that the queue is *invisible*, in the sense that waiting customers do not obtain any information regarding the queue size or their remaining waiting time before admitted to service. Queues of this type are especially relevant to *remote* service systems, such as telephone call centers or Internet-based services; hence, we refer to them as *tele-queue*. For a discussion of the central role that customer patience plays in tele-queues see Garnett et al. (1999).

The foundation for our model is the *hypothesis* that customers' patience significantly depends on their expectations regarding the waiting time in the system. These expectations, in turn, are formed through

0025-1909/02/4804/0566$5.00
1526-5501 electronic ISSN

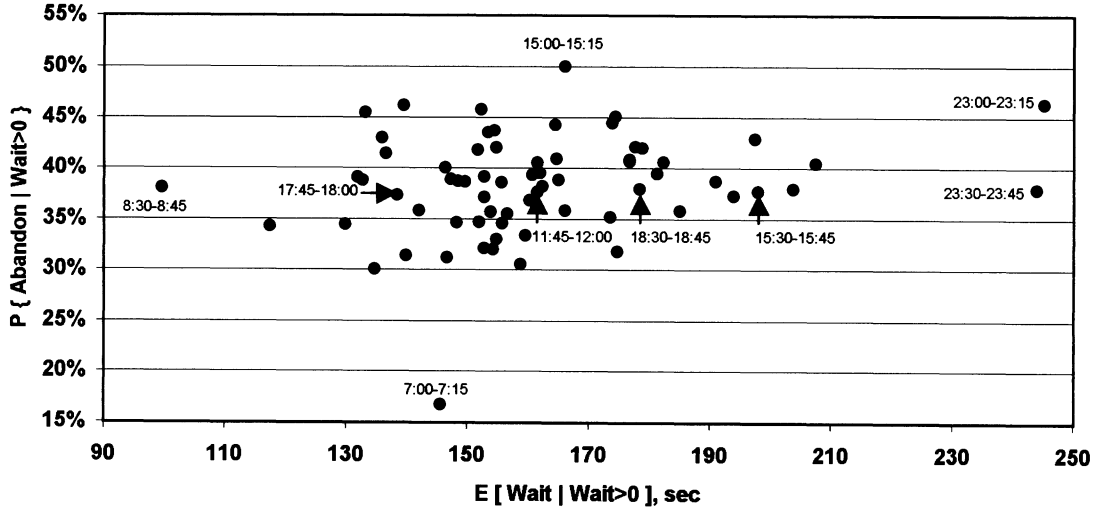
accumulated experience and affected by subjective factors—time perception, the importance of the service being sought, and so on. As an example, customers who expect to wait a few seconds will behave differently, in terms of their abandonment time, in case they expect to wait several minutes or even hours. These expectations, in turn, conceivably differ if past experience consists of short waits, or long waits, or short and long waits intertwined. Patience is obviously influenced by numerous factors related to customer profiles and environment characteristics (see, for example, Maister 1985, Zakay and Hornik 1996, Levine 1997). However, for the purpose of performance analysis, most of these factors can be taken as a priori given and fixed. The waiting time distribution is singled out in this respect since it is the outcome of the queueing process (hence, in fact, itself is influenced by the patience profile).

Empirical Support—A Preview. Inconsistent with the above adaptivity hypothesis, the prevalent assumption in traditional queueing theory is that patience (the time-to-abandon or its probability distribution) is "assigned" to individual customers independently of any system performance characteristic (see Garnett et al. 1999 for a recent literature review). In particular, patience is unaltered by possible changes in congestion. Such models, however, *cannot* accommodate the scatterplot in Figure 1 that exhibits remarkable patience-adaptivity.

The data is from a bank call center as reported in Mandelbaum et al. (2000); see also §4. We are scatterplotting abandonment fraction against average delay, for delayed customers (positive waiting time) who seek technical Internet support. It is seen that average delay during 8:30–8:45 A.M., 17:45–18:00 P.M., 18:30–18:45 P.M., and 23:30–23:45 P.M. is about 100, 140, 180, and 240 seconds, respectively. Nonetheless, the fraction of abandoning customers (among those delayed) is remarkably stable at 38%, for *all* periods. This stands in striking contrast to traditional queueing models, where patience is assumed unrelated to system performance: Such models would predict a strict increase of the abandonment fraction with the waiting time, as in Figure 3. The behavior indicated in Figure 1 clearly suggests that customers do adapt their patience to system performance.

A Descriptive Approach. Several recent papers have proposed an optimization-based model for customer patience, where abandonment decisions are based on a personal cost function that balances service utility against the cost associated with the expected remaining time to service. In particular, Hassin and Haviv (1995) and Haviv and Ritov (2001) analyze

Figure 1 **Adaptive Behavior of IN (Experienced) Customers—Abandonment Probability vs. Average Wait (of Customers Who Waited a Positive Time)**



Note. Each point corresponds to a 15-minute period of the weekdays, starting at 7:00 am, ending at midnight, and averaged over the whole year of 1999.

systems with a single customer type, and Mandelbaum and Shimkin (2000) consider a heterogeneous customer population, in terms of utility functions and the resulting abandonment profiles. In these models, the optimal abandonment decision depends on the *entire* waiting-time distribution offered by the system.

Unlike this prescriptive approach, we consider here a *descriptive* model, where the dependence of patience on system performance is explicitly specified within the model primitives, in much the same way that a demand function is assumed to be given in economic models. Such an explicit model can be more directly related to experimental data, and is not restricted by the assumption and consequences of strictly rational behavior of the customers.

Our model is highly simplified by assuming that customers' patience depends on the waiting time in the queue only through its *average*, namely the mean wait; thus, the patience depends on a single performance parameter rather than an entire distribution. The motivation for this simplified model is threefold. First, the mean arguably presents a natural parameter that summarizes customers' expectations regarding their waiting time; indeed, a typical customer can hardly be expected to form a clear estimate of the entire waiting time distribution based on limited experience. Second, the dependence on a single parameter makes it much easier to relate the model to empirical data; see §4. And third, it offers a considerable simplification in performance analysis (compared, say, with Mandelbaum and Shimkin 2000).

Outline of the Paper. Section 2 presents the basic queueing model, which incorporates the dependence of the patience profile on the average waiting time, and defines the system equilibrium point.[1] We distinguish between the average waiting time *assumed* by the customers (denoted x), which determines the patience profile, and between the actual quantity, namely the *offered* expected wait that results from this patience profile. Simply put, equilibrium is achieved when the two coincide.

[1] The term *equilibrium* in this paper refers to an operating point of the system, as used in standard market and supply-demand models, and should not be confused with the Nash equilibrium or other game-theoretic concepts.

In §3, we analyze the equilibrium and its properties, focusing first on existence and uniqueness. Assuming that customer patience *decreases* as the (assumed) average wait x increases, existence and uniqueness of equilibrium follow from basic monotonicity considerations, as shown in §3.1. The more interesting case is when patience is allowed to increase with x (§3.2). Here customers adjust their behavior to comply with their expectations. When patience can grow not more than proportionally with x, existence and uniqueness of the equilibrium can still be established and the equilibrium point may be calculated. When this growth condition is violated, multiple equilibria are feasible, as we explicitly demonstrate there.

In §3.3, we apply the proposed model to address the following question: What is the required dependence of customer patience, so that the abandonment fraction is kept constant despite varying congestion conditions. This question is motivated by the relative insensitivity of the abandonment fraction that was revealed in Figure 1.

Section 4 presents additional empirical support for the dependence of customer patience on the anticipated waiting time. Section 5 provides a brief survey of the literature on patience modeling.

Our basic equilibrium model assumes that the system is in steady state, in the sense that the system characteristics are stationary and the customers are well acquainted with those characteristics that are relevant to their behavior. In §6, we complement the static equilibrium viewpoint with a dynamic learning model, which incorporates the additional ingredient of learning by the customers, and traces the system evolution towards a possible equilibrium. Indeed, the average waiting time parameter x is not initially known, but may be estimated by the customers based on their accumulated experience. We briefly address the issue of *censored sampling* that arises here: In those customer's visits that end up with abandonment, the offered wait itself is not observed but rather a lower bound on it, namely the abandonment time. As consistent estimation of the mean is quite complicated in this case, we also consider a simpler nonconsistent estimator and its effect on the equilibrium point. The dynamics of the queueing system which incorporates the proposed learning process is examined

via simulation, and its convergence to the anticipated equilibrium is demonstrated. We conclude in §7 with a brief summary and comments concerning future work.

2. Model Formulation

Consider an $M/M/m$ queue with Poisson arrivals at rate λ, and an exponential service time with mean μ^{-1} at each of the m servers. The service discipline is first-come-first-served. Waiting customers may abandon the queue at any time before admitted to service. Potential abandonment times of individual customers are assumed independent and identically distributed, according to a probability distribution $G(\cdot)$ over the nonnegative real line. We shall refer to G as the *patience* distribution function. Let $\overline{G} = 1 - G$ denote the survival function; thus $\overline{G}(t)$ is the probability that a waiting customer will not abandon within t time units. We allow G to depend on a parameter x to be specified below, so that $G(t) = G(x, t)$. When convenient, we shall suppress the dependence on x. While we assume here for simplicity that the arrival rate λ is constant, our model and analysis easily extend to the case where λ depends on the same parameter x; see the remark at the end of §3.

Let V denote the *offered* waiting time, or offered wait, which is the time that a (nonabandoning) customer would have to wait until admitted to service. We assume throughout that the system is in steady state, so that the distribution of V is the same for all customers. Under the stability condition $m\mu > \lambda\overline{G}(\infty)$, the density F_V' of V is given by (Baccelli and Hebuterne 1981)

$$F_V'(t) = \lambda P_{m-1} \exp(J(t)), \qquad t > 0, \tag{1}$$

with P_{m-1} specified below, and

$$J(t) = -\int_0^t (m\mu - \lambda\overline{G}(s))\, ds. \tag{2}$$

Let P_j denote the stationary probability for exactly j occupied servers; thus, V has an atom at 0, with $P(V = 0) = \sum_{j=0}^{m-1} P_j$. The normalization condition is

$$\sum_{j=0}^{m-1} P_j + \int_0^\infty F_V'(t)\, dt = 1, \qquad P_j = \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} P_0.$$

It follows that

$$F_V'(t) = \frac{\exp(J(t))}{\frac{K_m}{\lambda} + \int_0^\infty \exp(J(s))\, ds}, \tag{3}$$

where

$$K_m = \sum_{j=0}^{m-1} \frac{(m-1)!}{j!} \left(\frac{\lambda}{\mu}\right)^{j-m+1}. \tag{4}$$

We shall also refer to the distribution F_0 of $(V|V > 0)$, namely the distribution of the waiting time V given that the customer is not immediately admitted to service; the corresponding density is obviously given by the expression (3) with K_m set to zero.

Consider next the dependence of the patience function G on system performance. As discussed in the introduction, we focus here on a simplified model which assumes that this dependence is expressed through a single parameter x, corresponding to the average offered wait in the system. Specifically, we shall consider the following two alternatives:

1. $x = E(V)$, the expected wait.
2. $x = E(V|V > 0)$, the expected wait given that the wait is nonzero (all servers busy upon arrival).

These two options correspond to slightly different evaluations of the waiting time, and lead to some differences in the analysis. The expected waiting time may be the most natural single parameter that comes to mind as a summary of waiting time performance. Still, the probability of finding a vacant server upon arrival becomes irrelevant to customers who are required to wait, and therefore the second option may turn out to be more appropriate.

We remark that for modeling purposes, it may be useful to specify the dependence of G on x in two steps. First, let G_η be some parameterized family of probability distributions. For example, G_η may be the set of exponential distributions, with η the expected value. Or it may the set of degenerate distributions, where now η is the deterministic time of abandonment. Further, let the parameter η be determined by the value of the performance parameter x, namely $\eta = \eta(x)$. The actual patience distribution G is thus selected out of the family G_η and it depends on x according to $G = G_{\eta(x)}$. This parameterization will be employed in some of our examples.

We have thus parameterized the patience distribution G in terms of the performance parameter x, which may be one of the two options itemized above. This completes the model description. We can now consider the ensuing operating point of the system in equilibrium. Note that the operating point is fully specified once the value of the parameter x has been determined.

We proceed to characterize the equilibrium conditions explicitly. Of the two options specified above, first consider the case of $x = E(V)$. For each $x > 0$, define

$$v_1(x) = E_x(V),$$

where E_x is the expectation induced by the distribution (3), with $G = G(x, \cdot)$. Thus, $v_1(x)$ is the expected waiting time that would be induced by the patience distribution associated with x. The equilibrium condition requires that the customers' evaluation of the expected waiting time (x) coincide with the actual value, namely

$$x = v_1(x). \tag{5}$$

This gives a scalar equation in the single variable x. The questions of existence and uniqueness of an equilibrium point are thus equivalent to the existence and uniqueness of a fixed point in Equation (5).

Similarly, when the performance parameter x is taken as the conditional waiting time $E(V|V > 0)$, define

$$v_2(x) = E_x(V|V > 0).$$

The equilibrium condition is then

$$x = v_2(x). \tag{6}$$

We assume throughout that the stability condition $\overline{G}(x, \infty) < m\mu$ holds for *some* x. Both expected values $v_i(x)$ are finite at these values of x.

3. Equilibrium Analysis

We now turn to examine the system equilibrium and analyze its properties—focusing first on the questions of existence and uniqueness of the equilibrium point. We shall then employ the model to address some performance analysis issues, related to the feasibility of maintaining a constant abandonment fraction despite different load conditions, as depicted in Figure 1.

The equilibrium analysis proceeds in two steps. Recall that the customer patience distribution depends on a performance parameter x, which represents the expected wait in the queue. In §3.1, we address the relatively simple case where patience is *decreasing* in the performance parameter x (Assumption 1). This dependence may be interpreted as intolerance of the customer population to service degradation: When the waiting time becomes longer, customers find it less appealing to keep waiting and react by abandoning earlier. This behavior can also be explained within a "rational" model for abandonments as presented in Mandelbaum and Shimkin (2000), since the expected return per unit wait becomes smaller as time progresses. Still, in practice one often observes an opposite tendency of customers who adapt their patience to comply with the expected waiting time in the system. This was indeed observed in the empirical results of §4. In §3.2, we extend our analysis to the "increasing patience" case.

3.1. Decreasing Patience

We assume first that the customer patience is decreasing in the performance parameter x, in the sense of stochastic ordering. Recall the following definitions (Shaked and Shanthikumar 1994). Given two real-valued random variables Y_1 and Y_2 with distributions F_1 and F_2, we say that Y_1 stochastically dominates Y_2, denoted $Y_1 \geq_{st} Y_2$, if $\overline{F}_1(t) \geq \overline{F}_2(t)$ for all t (here $\overline{F}_i = 1 - F_i$). Y_1 *strictly* dominates Y_2, denoted $Y_1 >_{st} Y_2$, if, in addition, $\overline{F}_1 \neq \overline{F}_2$. We shall also adopt the corresponding notations $\overline{F}_1 \geq_{st} \overline{F}_2$ and $\overline{F}_1 >_{st} \overline{F}_2$ to denote these relations. Note that $E(Y_1) \geq E(Y_2)$ is implied in the former case, and $E(Y_1) > E(Y_2)$ in the latter. A set of random variables $\{T(x)\}$ in the real parameter x is said to be *decreasing* in stochastic order if $x_1 < x_2$ implies $T(x_1) \geq_{st} T(x_2)$, and is *strictly decreasing* if the latter dominance relation is strict.

ASSUMPTION 1. *The set of patience distribution functions $\{G(x, \cdot)\}$ is decreasing in x in stochastic order. That is, $x_1 > x_2$ implies that $\overline{G}(x_1, t) \leq \overline{G}(x_2, t)$, for all $t \geq 0$.*

PROPOSITION 3.1. *Let Assumption 1 hold.*

(i) *Let G_1 and G_2 be two patience distributions, with F_1 and F_2 the corresponding distributions of the offered waiting time V, specified in (3). Then $\overline{G}_1 \leq_{st} \overline{G}_2$ implies $\overline{F}_1 \leq_{st} \overline{F}_2$.*

(ii) *A similar implication holds for F_0, the distribution function corresponding to the conditional waiting times $(V|V>0)$ as specified following* (3).

PROOF. For each G_i, $i=1,2$, denote:

$$J_i(t) = -\int_0^t (m\mu - \lambda \overline{G}_i(s))\, ds, \tag{7}$$

and let $D(t) = \overline{G}_2(t) - \overline{G}_1(t)$. By our assumption, $D \geq 0$. Thus,

$$J_2(t) = J_1(t) + \lambda \int_0^t D(s)\, ds \geq J_1(t). \tag{8}$$

The hazard rate functions H_i corresponding to these waiting time distributions are given by

$$H_i(t) = \frac{F_i'(t)}{\overline{F}_i(t)} = \frac{\exp(J_i(t))}{\int_t^\infty \exp(J_i(v))\, dv}, \quad t \geq 0. \tag{9}$$

To establish $\overline{F}_1 \leq_{st} \overline{F}_2$, we shall in fact prove the stronger property that $\overline{F}_1(t)/\overline{F}_2(t)$ is (weakly) decreasing in t. The latter is equivalent to dominance in the hazard rate order; see Shaked and Shanthikumar (1994, Chapter 1). To establish that $\overline{F}_1/\overline{F}_2$ is a decreasing function, it suffices to show that $H_1(t) \geq H_2(t)$ for all $t \geq 0$, and that at the discontinuity point at $t = 0$, we have $\overline{F}_1(0)/\overline{F}_2(0) \leq 1$. By substituting (8) in the expression for H_1, we obtain:

$$H_2(t) = \frac{\exp(J_1(t)) \exp(\lambda \int_0^t D(s)\, ds)}{\int_t^\infty [\exp(J_1(v)) \exp(\lambda \int_0^v D(s)\, ds)]\, dv}. \tag{10}$$

But by the assumed positivity of D, we have that $\exp(\lambda \int_0^v D(s)\, ds) \geq \exp(\lambda \int_0^t D(s)\, ds)$ for all $v \geq t$, which immediately implies

$$H_2(t) \leq \frac{\exp(J_1(t))}{\int_t^\infty \exp(J_1(v))\, dv} = H_1(t).$$

It remains only to show that $\overline{F}_1(0)/\overline{F}_2(0) \leq 1$, or equivalently that $F_1(0) \geq F_2(0)$. This follows from $J_1(t) \leq J_2(t)$ by noting from (3) that

$$F_i(0) = \frac{K_m}{\lambda} \Big/ \left[\frac{K_m}{\lambda} + \int_0^\infty \exp(J_i(t))\, dt\right].$$

The proof of (ii) follows similarly to the first part of the proof above, since V and $(V|V>0)$ have identical hazard rate functions for $t \geq 0$, while $\overline{F}_0(0) = 1$ by definition. □

Uniqueness of the equilibrium follows easily from the last result, as shown next. For existence, some basic continuity and stability conditions are naturally required. The parameterized family of distributions $G(x,\cdot)$ is *weakly continuous* in x if $g(x) := \int \phi(t)\, dG(x,t)$ is continuous in x for every bounded continuous function ϕ. Note that this allows the distributions G to contain point masses which depend continuously on x.

THEOREM 3.2. *Let Assumption* 1 *hold. Assume further that the patience distributions $G(x,\cdot)$ are weakly continuous in x. Then for either one of the equilibrium equations* (5) *or* (6)*, a solution exists and is unique.*

PROOF. Recall that $X \leq_{st} Y$ implies $E(X) \leq E(Y)$. From the last proposition, we therefore obtain that both functions $v_1(x)$ and $v_2(x)$ are decreasing in x, and uniqueness of the solution follows immediately. As for existence, the assumed continuity condition is easily shown to imply the continuity of v_1 and v_2. Since our model assumes that both functions are finite for some x, existence follows. □

3.2. Increasing Patience

We shall now relax the decreasing-patience assumption, and replace it by a bound on the growth rate of the patience distribution (Assumption 2). The main result here is Theorem 3.3, which extends the results of the previous section while relying on them for the proof.

Assumption 2 allows an increase in the customer's patience with the performance parameter x, but essentially requires that the rate of increase of the former does not exceed that of the latter. That is, when x (the anticipated average wait) increases by δ, the patience (willingness to wait) of the customer population will increase by δ at the most. Some growth condition of that nature is essential to guarantee uniqueness, as demonstrated by the example that closes this subsection.

ASSUMPTION 2. *Let $T(x)$ be a random variable with distribution $G(x,\cdot)$. Then the family of random variables $\{T(x) - x\}$ is decreasing in x, in stochastic order.*

An equivalent statement of the last condition is that $T(x+y) \leq_{st} T(x) + y$ for every $y \geq 0$. In terms

of the distribution functions, it may be expressed as $\overline{G}(x+y,\cdot) \leq_{st} \overline{G}(x,\cdot+y)$. It implies, in particular, that $E(T(x))-x$ is decreasing in x.

We establish below that under Assumption 2, the functions $v_i(x)-x$ $(i=1,2)$ are strictly decreasing in x. This immediately implies uniqueness of the corresponding equilibria defined in (5) or in (6). To establish existence, it is further required to show that $v_i(x)-x \leq 0$ for x large enough (note that $v_i(0) > 0$). However, Assumption 2 alone may not suffice here (as may be verified via a simple example, e.g., with a deterministic $T(x)=x$). The existence claim will thus require an additional condition, which is either a system stability requirement or a slight strengthening of Assumption 2, as specified below.

THEOREM 3.3. *Let Assumption 2 hold. Consider the equilibrium defined in* (5) *or in* (6).

(i) Uniqueness: *The equilibrium point, if one exists, is unique.*

(ii) Existence: *Assume, in addition, that the patience distribution functions $G(x,\cdot)$ are weakly continuous in x, and that either one of the following conditions hold:*

a. $\lambda < m\mu$, *or*

b. $[T(x)-(1-\epsilon)x]$ *is decreasing in x in stochastic order, for some $\epsilon > 0$.*

Then the equilibrium exists.

The proof proceeds through some lemmas. We start by establishing the uniqueness of the equilibrium defined through v_2 in (6), which turns out to be simpler, and follows directly from the next proposition. In the following, W stands for the random variable $(V|V>0)$ with distribution F_0.

LEMMA 3.4. *Let Assumption 2 hold. Then $\{W(x)-x\}$ is strictly decreasing in stochastic order. In particular, the function $[v_2(x)-x]$ is strictly decreasing in x.*

PROOF. For any x and $y>0$, we need to show that $W(x+y) \leq_{st} W(x)+y$. Our basic Assumption 2 is that $T(x+y) \leq_{st} T(x)+y$. Since W is increasing in T, as established in Proposition 3.1(ii), it is clearly sufficient to prove the lemma under the assumption that $T(x+y)=T(x)+y$.

Assume, then, that the latter holds. In terms of the distribution functions, our assumption is that $\overline{G}(x+y,t)=\overline{G}(x,t-y)$, and we wish to show that $\overline{F}_0(x+y,t) \leq \overline{F}_0(x,t-y)$ for all t. As in the proof of Proposition 3, it is convenient to work here with the corresponding hazard rate functions. Since the distributions F_0 are absolutely continuous, namely the density F_0' exists at every point, it suffices to show that for all t,

$$\frac{F_0'(x+y,t)}{\overline{F}_0(x+y,t)} \geq \frac{F_0'(x,t-y)}{\overline{F}_0(x,t-y)}. \tag{11}$$

Now, from (1),

$$F_0'(x,t-y) = C(x)\exp\left(\int_0^{t-y} K(x,s)\,ds\right), \qquad t \geq y,$$

where $K(x,t) := \mu\overline{G}(x,t) - m\lambda$, and $C(x)$ is a normalization constant. Note that $F_0'(x,t-y)=0$ for $t<y$. On the other hand,

$$F_0'(x+y,t) = C(x+y)\exp\left(\int_0^{t} K(x+y,s)\,ds\right), \quad t \geq 0.$$

But our assumption on G implies that $K(x+y,s) = K(x,s-y)$. We thus obtain

$$\begin{aligned} F_0'(x+y,t) &= C(x+y)\exp\left(\int_{-y}^{t-y} K(x,s)\,ds\right) \\ &= C(x+y)\exp\left(\int_{-y}^{0} K(x,s)\,ds\right) \\ &\quad \times \exp\left(\int_0^{t-y} K(x,s)\,ds\right). \end{aligned}$$

Comparing the expressions above, it is apparent that (11) holds with equality for $t \geq y$. For $t<y$ the right-hand side of (11) is null, so that inequality holds trivially. Moreover, since the left-hand side is nonzero for $0<t<y$, then strict inequality holds on that interval. This implies that $\overline{F}_0(x+y,t) \leq \overline{F}_0(x,t-y)$, with strict inequality holding on some interval; hence $\overline{F}_0(x+y,\cdot) <_{st} \overline{F}_0(x,\cdot)$. This establishes the main claim of this lemma. Since $v_2(x)=E(W(x))$, the second claim follows immediately. □

We proceed to establish the uniqueness of the equilibrium defined in (5), with $v_1(x)=E_x(V)$. To relate this case to the previous one, observe that $v_1(x) = \bar{p}_0(x)v_2(x)$, where $\bar{p}_0(x)=P\{V>0\}$ is the probability that an arriving customer does not find an available server. It was shown above that $v_2(x+y) \leq v_2(x)+y$. However, as $G(x,\cdot)$ increases so does $\bar{p}_0(x)$, and we

cannot infer from the above equality a similar relation for $v_1(x)$. On the technical side, the distribution $F_V(x,\cdot)$ of V obviously contains a jump at $t=0$ (with magnitude $p_0(x)$), and this prevents the application of the hazard-rate comparison argument which was used in Lemma 3.4. We therefore resort in the analysis below to direct calculation of $v_1(x)$ and its derivative.

LEMMA 3.5. *Let Assumption 2 hold. Then* $[v_1(x)-x]$ *is strictly decreasing in* x.

PROOF. It is required to establish the assertion under Assumption 2, namely $\overline{G}(x+y,t) \le \overline{G}(x,t-y)$ for $y>0$. By the monotonicity result in Proposition 3.1, it is sufficient to consider the extreme case where $\overline{G}(x+y,t)=\overline{G}(x,t-y)$, which we henceforth enforce.

We introduce some further notations. From (3), we have that $v_1(x)=\frac{A(x)}{B(x)}$, with

$$\begin{aligned} A(x) &= \int_0^\infty t\exp[J(x,t)]\,dt,\\ B(x) &= k_m+\int_0^\infty \exp[J(x,t)]\,dt\\ J(x,t) &= \int_0^t K(x,s)\,ds,\\ K(x,s) &= \lambda\overline{G}(x,s)-m\mu, \end{aligned}$$

and $k_m=K_m/\lambda$. Note that our assumption concerning G implies that $K(x+y,t)=K(x,t-y)$. We proceed to evaluate $v_1(x+y)$ for $y>0$. First,

$$\begin{aligned} J(x+y,t) &= \int_0^t K(x,s-y)\,ds\\ &= \int_{-y}^0 K(x,s)\,ds+\int_0^{t-y} K(x,s)\,ds\\ &= by+J(x,t-y), \qquad t>y, \end{aligned}$$

since $K(x,s)=b$ for $s<0$, with $b=\lambda-m\mu$. Similarly, $J(x+y,t)=bt$ for $0\le t\le y$. Thus,

$$\begin{aligned} A(x+y) &= \int_0^\infty t\exp[J(x+y,t)]\,dt\\ &= \int_0^y te^{bt}\,dt+e^{by}\int_0^\infty (t+y)\exp[J(x,t)]\,dt\\ &= g(y)+e^{by}[A(x)+y(B(x)-k_m)], \end{aligned}$$

where $g(y)$ stands for the first integral. Note that $\lim_{y\to 0} g(y)/y=0$, which we denote by $g(y)=o(y)$. Similarly,

$$\begin{aligned} B(x+y) &= k_m+\int_0^y e^{bt}\,dt+e^{by}\int_0^\infty \exp[J(x,t)]\,dt\\ &= k_m+ye^{by}+o(y)+e^{by}[B(x)-k_m]\\ &= e^{by}[B(x)+(1-bk_m)y]+o(y). \end{aligned}$$

It follows that

$$\begin{aligned} &v_1(x+y)-v_1(x)\\ &= \frac{A(x+y)}{B(x+y)}-\frac{A(x)}{B(x)}\\ &= \frac{A(x)+y[B(x)-k_m]+o(y)}{B(x)+(1-bk_m)y+o(y)}-\frac{A(x)}{B(x)}\\ &= y\left(1-\frac{k_mB(x)+(1-bk_m)A(x)}{B(x)^2}\right)+o(y), \end{aligned}$$

which implies

$$\frac{d}{dx}[v_1(x)-x]=-\frac{k_mB(x)+(1-bk_m)A(x)}{B(x)^2}.$$

Obviously, the proof may be concluded if we show that the latter is negative. Since $A(x)$, $B(x)$, and k_m are all positive, we need only verify that $(1-bk_m)\ge 0$. Using the definition of k_m and b, this inequality is equivalent to $(1-m\mu/\lambda)K_m\le 1$. This obviously holds when $m\mu/\lambda\ge 1$. Otherwise, we have from (4),

$$\begin{aligned} K_m &\le \sum_{j=0}^{m-1} m^{m-1-j}\left(\frac{\lambda}{\mu}\right)^{j-m+1}\\ &= \sum_{j=0}^{m-1}\left(\frac{m\mu}{\lambda}\right)^{m-1-j} < \left(1-\frac{m\mu}{\lambda}\right)^{-1}, \end{aligned} \tag{12}$$

which again implies the required inequality. □

PROOF OF THEOREM 3.3. Uniqueness of the equilibrium under either definition follows from the last two lemmas. As for existence of the equilibrium defined in (6), since $v_2(0)>0$ and $v_2(x)$ is continuous by the Theorem's continuity assumption, it suffices to show that $v_2(x)-x<0$ for x large enough. If (a) holds then the system is stable even without abandonments so that $v_2(\cdot)$ is bounded. If (b) holds, then by rescaling in x it follows from Proposition (3.4) that $v_2(x)-(1-\epsilon)x$

is decreasing in x, hence $v_2(x) - x \le C - \epsilon x$ for some finite constant C, which clearly implies the required inequality. Existence of the equilibrium (5) follows similarly since $v_1(x) \le v_2(x)$. □

We conclude this section with a simple example that shows that multiple equilibria are feasible when Assumption 2 is violated.

EXAMPLE 1. MULTIPLE EQUILIBRIA. Consider an $M/M/1$ queue with $\lambda = 1$, $\mu = 1$, and a deterministic abandonment time $T(x)$ which is the same for all customers. Thus $\overline{G}(x,t) = 1$ for $t \le T(x)$ and $\overline{G}(x,t) = 0$ for $t > T(x)$. By (3) we have

$$v_2(x) := E_x[V|V>0] = \frac{\int_0^\infty t \exp(J(t))}{\int_0^\infty \exp(J(t))\,dt}.$$

Substituting $\overline{G}$ and $m = \lambda = \mu = 1$ gives by explicit calculation

$$v_2(x) = \frac{T^2/2 + T + 1}{T+1} = \frac{1}{2}\left(T + 1 + \frac{1}{T+1}\right), \qquad (13)$$

where $T = T(x)$. It is now simple to verify that the choice $T(x) = x - 1 + \sqrt{x^2 - 1}$ gives $v_2(x) = x$ for all $x \ge 1$. According to the definition of the equilibrium in (6), this implies that *every* value $x \ge 1$ corresponds to equilibrium point, hence there is a continuum of equilibria. It may be seen that by slightly perturbing the above expression for $T(x)$, we can also induce any discrete number of equilibria.

REMARK. So far we have assumed a constant arrival rate λ. It stands to reason that the arrival rate would also depend on the system performance. In our model, we may assume that λ depends on the system performance parameter x, and is naturally decreasing as x increases. It may be verified that the offered waiting time V (possibly conditioned on $V > 0$) is stochastically decreasing in λ, so that the previous results hold in this case as well.

3.3. Maintaining a Constant Abandonment Fraction

We shall briefly examine here certain aspects of system performance using the adaptive patience model and the related equilibrium framework. As has been observed in §4, one possible effect of customer adaptation is to keep the abandonment fraction approximately constant, even under varying congestion conditions. It may thus be of interest to find the precise patience variation that would keep the abandonment fraction constant. A reasonable conjecture in this regard, which we verify below, is that patience should be approximately proportional to the offered waiting time in order to keep the abandonment fraction fixed. This indeed conforms well with the empirical relation that will be observed between these quantities in Figure 4.

We shall consider as before an $M/M/m+G$ queue, with $m\mu$ fixed (normalized to 1), and let the arrival rate λ serve as a parameter that controls the system load. We require $P_{ab} = \beta$, with β a specified constant (taken as 0.3 below), and P_{ab} is the fraction of abandoning customers out of those that are not immediately admitted to service. The patience distribution G depends on a system performance parameter x, taken as $x = v_2 := E(V|V>0)$. We are thus considering the system equilibrium defined in Equation (6). We specify G as a member of some parametric family $\{G_\eta\}$, where the parameter η is also the mean of G_η, and depends on x according to some relation $\eta = \eta(x)$, which is determined below. We shall consider two parametric families:

1. Deterministic: $G_\eta(t) = 1\{t \ge \eta\}$. Thus, $T \equiv \eta$.
2. Exponential: $G_\eta(t) = 1 - \exp(-t/\eta)$.

We now wish to compute the required dependence of η on x so that the abandonment fraction is fixed at $P_{ab} = \beta$, for all feasible λ. This is done as follows. For each fixed λ, P_{ab} is a function of η, and one may solve (possibly numerically) for the value of η that gives $P_{ab} = \beta$. Given η, namely G_η, we can now compute the corresponding $x = E(V|V>0)$. This procedure yields x and η, parameterized by λ, and hence obtains the required function $\eta(x)$.

For concreteness, let us outline the computation of η. We have

$$P_{ab} := P\{abandon|V>0\} = P\{T \le V|V>0\} = \int_{v=0}^\infty F_0'(v)G(v)\,dv,$$

where F_0' is the density of $(V|V>0)$ obtained from (1). In the deterministic case, substituting $G(t) = 1\{t \ge \eta\}$

and using (1) gives, after some calculations,

$$P_{ab} = \int_{v=\eta}^{\infty} F_0'(v)\,dv = \frac{\int_{\eta}^{\infty} e^{J(t)}\,dt}{\int_{0}^{\infty} e^{J(t)}\,dt}$$
$$= \frac{\frac{1}{m\mu} e^{-\eta(m\mu-\lambda)}}{\frac{1}{m\mu-\lambda}(1-e^{-\eta(m\mu-\lambda)}) + \frac{1}{m\mu} e^{-\eta(m\mu-\lambda)}}.$$

Solving $P_{ab} = \beta$ for η gives

$$\eta = \frac{1}{m\mu-\lambda} \log\left[1 + \frac{1-\beta}{\beta}\left(1 - \frac{\lambda}{m\mu}\right)\right].$$

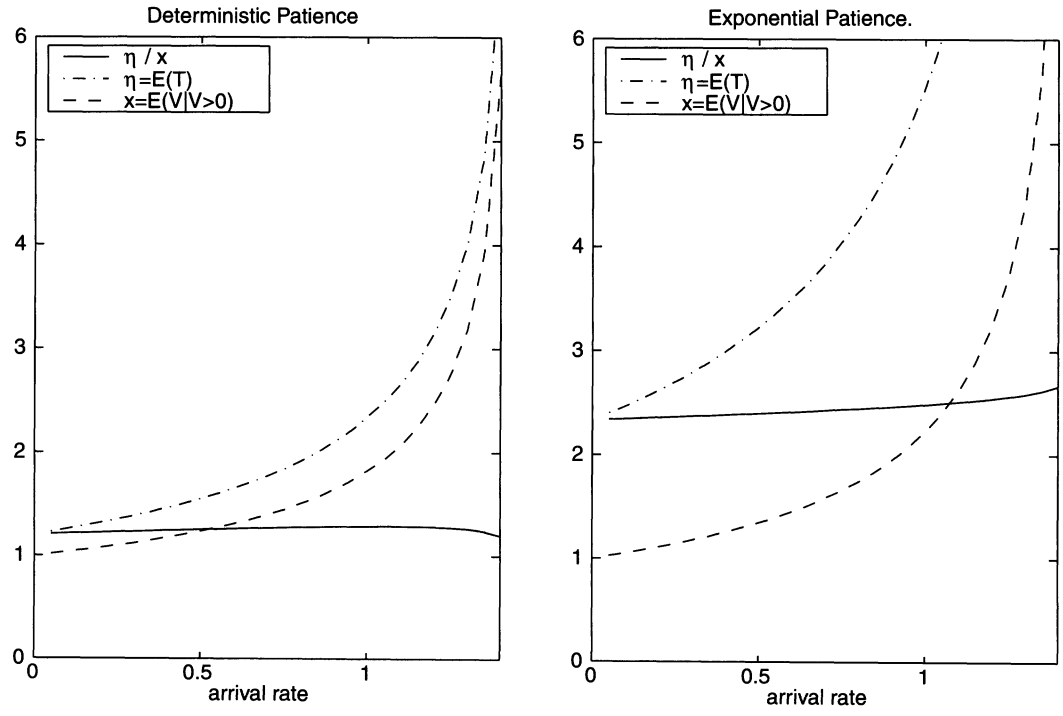
In the exponential case a numeric computation is required.

The results obtained for $m\mu = 1$ and $\beta = 0.3$ for deterministic and exponential patience, respectively, are shown in Figure 2. It depicts both $\eta := E(T)$, $x := E(V|V > 0)$ and their ratio η/x as a function of λ. (Observe that λ beyond $m\mu/(1-\beta) = 1.43$ is not feasible since it implies a service rate which is higher than the server capacity.) It may be seen that the ratio is approximately constant over the entire range of λ, which means that indeed η should be approximately proportional to x to obtain a fixed abandonment rate. It is interesting to note that the required ratio of η to x is significantly lower for the deterministic case.

4. Empirical Support

Traditional queueing theory has been naive in its modeling of abandonment. To wit, from the classical Palm (1953), Riordan (1962), Daley (1965) to the state-of-the-art Baccelli and Hebuterne (1981), Garnett et al. (1999), Brandt and Brandt (2000), it has always been assumed that patience is assigned to customers only upon arrival to the system, independently and identically distributed among customers, and unrelated to experiences of the past or anticipation of the future. In practical applications of the theory, furthermore, the distribution of patience, if at all acknowledged, has been assumed exponential; see, e.g., Garnett et al. (1999). (The papers Palm 1953 and Roberts 1979 are notable, but perhaps outdated, exceptions.) This is despite the fact that theory has actually accommodated *general* patience (Daley 1965, Baccelli and Hebuterne 1981). A main reason for that, one deduces,

Figure 2 **Patience Profiles That Keep $P_{ab} = 0.3$, with Patience That Is Deterministic (Left) and Exponentially Distributed (Right)**



is the lack of empirical evidence that either supports or refutes exponentiality. More fundamentally, we believe that there is simply sufficient understanding of human patience in general, and of the distribution of the time to abandon while waiting in tele-queues in particular.

A comprehensive empirical analysis of a telephone call center has been recently documented in Mandelbaum et al. (2000). This center provides banking teleservices of various types, for example balance inquiries, information to prospective customers, technical Internet support, stock management, and more. The event history of each *individual* call during 1999 was recorded, starting at the Voice Response Unit (VRU) and culminating in either a service by an agent or an abandonment from the tele-queue.

Part of the analysis in Mandelbaum et al. (2000) focuses on customer patience while waiting, and among its relevant findings we single out the following three observations:

(1) Patience definitely need not be exponential, and it varies significantly with service type, customer priority, and information provided during waiting; see §6.2 in Mandelbaum et al. (2000). We note that the heterogeneity of patience among customers has already been confirmed convincingly; for example, in Thierry (1994), Friedman and Friedman (1997), Diekmann et al. (1996) it is shown that patience, or value of time as its proxy, is affected by factors such as goal (service) motivation, mood, social status, and others.

(2) The *waiting time* distribution, over customers who actually got served, is found to be remarkably exponential (Mandelbaum et al. 2000, Figure 11). Note that this result is theoretically exact for the M/M/m queue in steady state only when there are no abandonments (cf. (1)).

(3) Experienced callers seem to adapt their patience to system performance (congestion), as exhibited in Figure 1. Patience of novice callers, on the other hand, is less sensitive to system performance.

For the rest of the section, we substantiate this last observation with further empirical evidence, first for novice and then for experienced callers.

Calls by novice customers are denoted in Mandelbaum et al. (2000) by type NW (for New). An example of such calls is inquiries by potential customers on marketing campaigns. In analogy to Figure 1, the scatterplot in Figure 3 relates the fraction of NW abandonment to their actual wait (restricted to delayed customers). As in Figure 1 and throughout the figures below, each scatterpoint corresponds to 15-minute periods of a day (Sunday to Thursday), starting at 7:00 A.M., ending at midnight, and averaged over the whole year of 1999.

The plotted relation in Figure 3 seems linearly increasing, with a positive intercept through the y-axis. (The line in the figure, as well as those below, are standard least-square fits.) We take this linearity as supporting the independence between patience and system performance. Indeed, for the $G/G/m$ queue in steady state, with abandonment times that are i.i.d. exponential (θ), the relation is exactly linear through the origin:

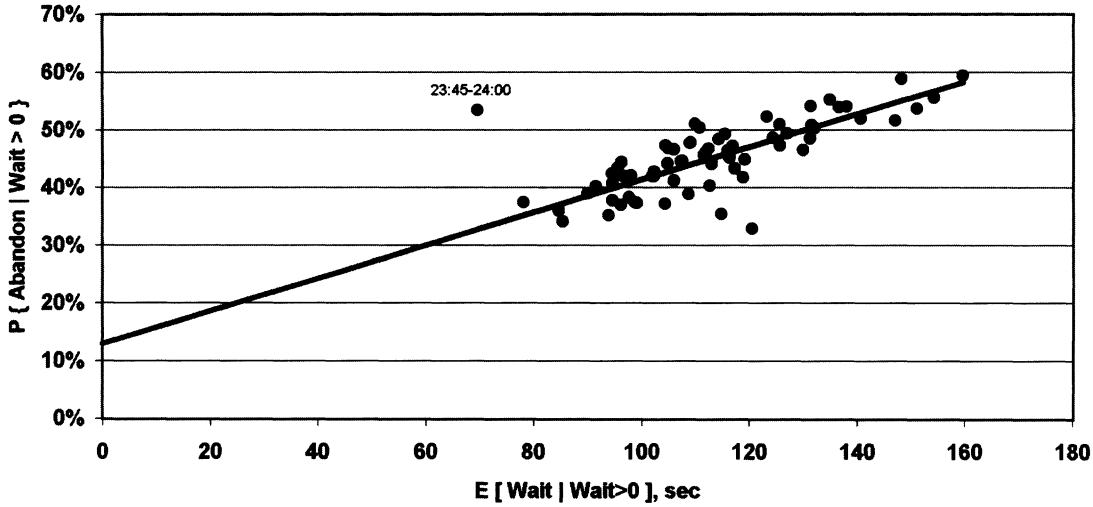
$$P\{abandon|wait > 0\} = \theta \times E[wait|wait > 0]. \quad (14)$$

For a verification, start with the fact that the abandonment rate equals either $\lambda \times P\{abandon\}$ or $E[queue\text{-}length] \times \theta$. Equating these last two expressions, using Little's law $E[queue\text{-}length] = \lambda \times E[wait]$, and dividing by $P\{wait > 0\}$, yields the above linearity. (For nonexponential patience, linearity holds asymptotically, as demonstrated in Theorem 4.2 of Brandt and Brandt 2000). To allow for a positive y-intercept, assume further that, among the abandoning customers, some abandon immediately upon arrival if forced to wait—which is commonly referred to as "balking." We then have $P\{abandon\} = P\{balk\} + \theta \times E[wait]$. Letting V denote the offered wait, one deduces the relation

$$P\{abandon|V > 0\} = P\{balk|V > 0\} + \theta \times E[wait|V > 0]. \quad (15)$$

(Note that here we condition on $V > 0$ rather than wait > 0 since balking is inconsistent with the latter.) One can now interpret Figure 2 as portraying customers whose patience seems unaffected by varying conditions of congestion. For example, an increase in $E[Wait|Wait > 0]$ from 80 to 120 seconds has the same effect as an increase from 120 to 160 seconds: Both accompany an increase of about 12.5% in abandonment, out of those delayed.

Figure 3 **Novice (NW) Customers**



Note. $P\{\text{abandon}|\text{wait} > 0\}$ vs. $E\{\text{wait}|\text{wait} > 0\}$.

We now turn to experienced callers, denoted IN (technical INternet support) in Mandelbaum et al. (2000). As already demonstrated in the Introduction (Figure 1), the patience of experienced callers may exhibit remarkable adaptivity to system performance. The difference between NW customers (Figure 3) and IN customers (Figure 1) is clearly manifested (note the different time scales is the two figures).

Finally, we examine the relation between patience and perceived system performance. To this end, *Patience* will be represented by $E[\textit{time-to-abandon}]$, while system performance will be measured by $E[\textit{offered-wait}|\textit{wait} > 0]$. For *experienced* callers, we expect that actual performance, represented by this measure, coincides with anticipated performance, the latter being forged through previous experience. In other words, with enough service (sampling) experience, the distribution of the offered wait would be unraveled to experienced customers; they summarize this distribution via its mean, which in turn approximates their anticipation.

Figure 4 covers IN (experienced) customers. Each point corresponds to a pair (patience, anticipation), during a 15-minute period of a day. We see that y (patience) increases with x (anticipation). The slope of the least-square line fit is somewhat over unity. We take this as a confirmation for the adaptivity of patience to variations in anticipated system performance.

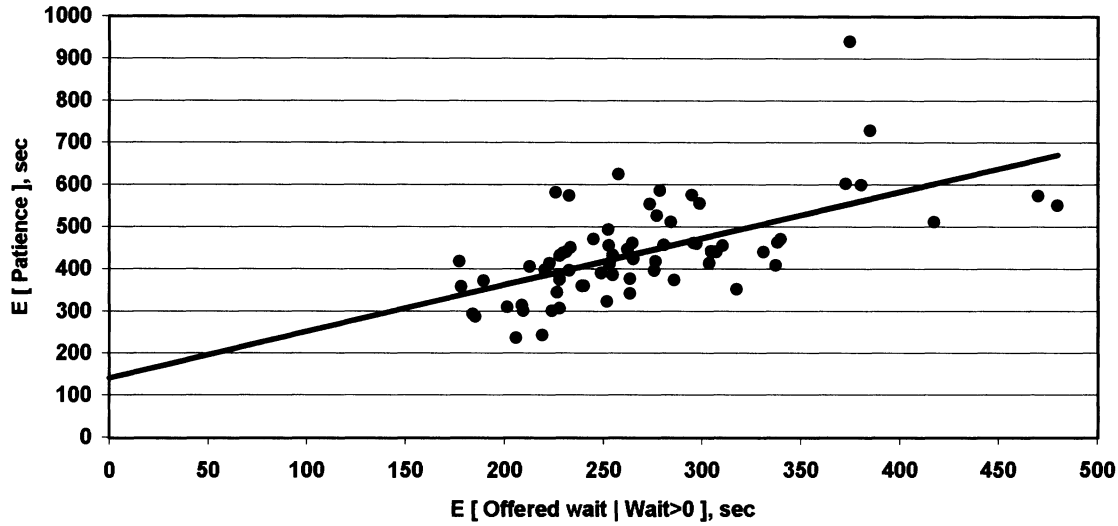
REMARK. *On Censoring:* The data in Figures 1 and 3 are directly observable. In Figure 4, on the other hand, both coordinates have to be "uncensored," since what is actually observed for each customer i is the actual wait $W_i = \min\{V_i, T_i\}$, which equals T_i (the patience, or time-to-abandon) only when i abandons, and V_i (the offered wait) only if i survives to be served. We use for this purpose the classical Kaplan-Meier estimator (Kaplan and Meier 1958), for specific details see Zohar et al. (2001).

REMARK. An analogue of Figure 4 for NW (novice) customers is not displayed. The reason is a lack of statistical confidence which is associated with data censoring. Some comments on the issue of robustness in censored estimation may be found in Zohar et al. (2001).

5. Modeling Patience

Abandonments of waiting customers are a common and important factor in service systems, and most people personally experience potential abandonment situations on a daily basis. Still, there appears to be little work concerning the modeling of the abandonment decision process and its contributing factors. We

Figure 4 **IN Customers**



Note. E[patience] vs. E[Offered wait |wait > 0]; $E[\cdot]$ stands for the mean of the Kaplan-Meier estimator for the corresponding distribution.

present here a brief discussion of some of the literature that seems relevant to abandonment modeling.

Abandonment decisions are predominantly a psychological process, which is triggered by negative feelings that build up while waiting. These are coupled with various factors such as the service utility and urgency, observed queue status, time perception, and exogenous circumstances. The exact trigger for abandonment remains largely unexplored. In an early work, Palm (1953) assumed that the abandonment *rate* is proportional to the momentary dissatisfaction, or annoyance, of the customers. An alternative model could specify an abandonment when annoyance (or another measure of negative feelings) reaches a certain threshold. A central ingredient in either case is the subjective disutility (or cost) of waiting, that has been addressed in a number of papers. A distinction can be made between the economical (opportunity) component of that cost and the psychological cost. The latter relies on both the sense of waste of invested time, and the stress caused by the remaining waiting time and associated uncertainty. Major factors that affect the waiting experience and its effect on service evaluation have been discussed in Maister (1985) and Larson (1987). A mathematical model for stress that has been introduced in Osuna (1985), and further developed in several papers, for example, Suck and Holling (1997), explicitly models the dependence of stress on the distribution of the remaining waiting time. However, this model does not directly address the effect of customer service expectations. Empirical studies include Taylor (1994), Leclerc et al. (1995), Hui and Tse (1996), and Carmon and Kahneman (1998). The latter, in particular, studies the evolution of the *momentary* affect in a queue and its relation to (observed) queue length.

The dependence of the subjective waiting cost on service expectations, and particularly on the expected waiting time, has been addressed qualitatively from several perspectives. The "first law of service" in Larson (1987) postulates that "satisfaction equals perception minus expectation." A reasonable consequence is that stress picks up when the expected wait has been surpassed. Hueter and Swart (1998) point out that customer perception of waiting time in a fast-food establishment increases steeply beyond an actual wait of several minutes (with a corresponding increase in the likelihood of abandonment). The effect of expectations and their disconfirmation on the momentary affective response is discussed and indicated empirically in Carmon and Kahneman (1998).

A normative, utility-maximizing model for abandonments has been considered in several recent

papers (Hassin and Haviv 1995, Mandelbaum and Shimkin 2000, Haviv and Ritov 2001). The abandonment time of each customer is chosen to maximize a personal utility function, which balances the service utility and the expected cost of waiting. We note that in the basic form of these models, the customer choice relies on the entire distribution of the offered waiting time, rather than just on its average (x) as was assumed in the present paper. Still, the model may be appropriately reduced by allowing the customers to assume an exponentially distributed waiting time. The reduced model is presented in Zohar et al. (2001), and related there to the Assumptions of §3.

Further work is required to establish analytical abandonment models that are based on the integration of a psychological framework with experimental and empirical data.

6. Modeling the Learning Process

Our equilibrium model assumes that customers know the average waiting time in the system. The model is thus static with respect to the customer's knowledge. In practice, however, the customer assessment of the waiting may be evolve through experience.

In this section, we consider a simple model for such a learning process, where each customer estimates the average waiting time based on personal experience, namely his own waiting times in previous visits. He then goes own to modify his abandonment decision according to the current estimate. Of prime interest to us here is the long-term or steady-state behavior of this learning process, which serves to validate our equilibrium analysis and examine some of its hypotheses. The transient behavior of the process may also be of considerable importance, for example to assess the time it takes to reach the steady operating point after the system is considerably modified, but we shall not address this aspect here.

Learning processes of similar nature have been considered in Altman and Shimkin (1998), Ben-Shachar et al. (2000) in the context of bulking decisions. In our case, abandonments complicate the estimation process, since the observations of the offered waiting time are *censored* by abandonment; that is, a customer who abandons the queue before being admitted to service does not observe the required wait but rather a lower bound on it. We are thus faced again, as in §4, with the need to estimate the mean of a distribution based on censored data.

We first employ a standard nonparametric estimator for censored data, namely the Kaplan-Meier (KM) estimator mentioned before, which provides a consistent estimator of the mean. It will be demonstrated that when each simulated customer uses KM, the system does indeed converge to its unique equilibrium point.

The KM estimator relies on complex computations, and in practice the customers' estimates are likely to be formed by much simpler procedures. It is therefore of interest to examine the consequences of using simpler estimators. The estimator we consider here is a (parametric) maximum likelihood estimator, which is derived based on the assumption that the estimated quantity (the virtual waiting time in our case) is exponentially distributed (or equivalently that the hazard rate of entering service is constant). This assumption, while false in the presence of abandonments, is a reasonable starting point from the customer's viewpoint, and leads to a simple estimator. It is given by (Miller 1981, p. 22):

$$\widehat{E(T)} = \frac{1}{N_s} \sum_{i=1}^{N} W_i, \tag{16}$$

where $\{W_1, W_2, \ldots, W_N\}$ are the collection of all the perceived waiting times, both from abandoned trials and successful ones, and N_s is the number successful trials, namely those that ended up with a service and were not censored by abandonment. We shall refer to this estimator as the Censored MLE. If T is not exponential, the estimator is biased enough to be inconsistent. Since the exponential assumption is false in our system, the Censored MLE turns out to be biased, and thus leads to a steady state of the learning system that *differs* from the previously postulated equilibrium. Our simulations will demonstrate convergence to this alternative steady state.

The online learning model that we propose is based on the following scenario. Each customer initially pos-

sesses some estimate x of the average waiting time, and his abandonment time (or distribution) is given by a function $T(x)$. The queueing system is that of §2, with the specific customer to enter the queue at each arrival is chosen randomly from a finite population. When the customer leaves the queue, either through service completion or abandonment, he updates his estimate x, and returns to the pool of idle customers.

6.1. Simulation Results

We describe here the results of two simulation experiments: The first employs the KM-based estimator, while the second employs the simpler Censored MLE. In both, the system is a single-server ($M/M/1$) queue, with $\lambda = \mu = 1$. Each customer maintains a personal estimate x of the average waiting time, and determines his abandonment time in the next trial as $T(x) = 0.8 \cdot x$. The estimated waiting time is taken here as $v_2 = E(V|V > 0)$ (see (6)). Note that the customer population is homogeneous in terms of the patience function. Simulation results for heterogeneous customer populations may be found in Zohar (2000), and lead to similar conclusions. This reference also contains a more complete description of the present simulations.

The specific customer who enters the queue is randomly and uniformly selected out of a pool of idle customers. If the pool is empty, a new customer is created. The initial knowledge base of a new customer is "inherited" from one of the existing customers, chosen at random. The first customer who initializes the simulation is arbitrarily initialized with ten "observations" of waiting times with duration $w_0 = 1.5$ each.

For reference, let us first calculate the equilibrium point for this system as per the analysis of §3. Note that the specified patience function $T(x)$ satisfies the requirements of Theorem 3.3, and hence the equilibrium is unique. The equilibrium condition (6) is $v_2(x) = x$. An expression for $v_2(x)$ is terms of $T(x)$ has been obtained in (13) for this system, which gives:

$$\frac{T(x)^2/2 + T(x) + 1}{T(x) + 1} = x.$$

With $T(x) = 0.8 \cdot x$, this equation indeed has a single positive solution at $x = 1.25$, which is the equilibrium value.

A slight modification was implemented in these simulations regarding the choice of abandonment times. Every once in a while (on each 30th trial), each customer was allowed to stay in the queue until admitted to service, instead of abandoning at $T(x)$. This allowed customers with low patience to sample the actual waiting time more fully, and turned out to be important for a reasonable convergence of the estimators.

Simulation 1: Kaplan-Meier Estimator. The system was simulated with the KM-based estimator. Recall that this estimator calculates an estimate of the entire waiting-time distribution (from which the mean is extracted). The results of the simulation are shown in Figures 5 and 6. The number of customers created in this example was 8; this is just the number that was required in this run to prevent starvation in the arrival process. The simulation was run for over 40,000 arrivals, which amounted to about 5,200 arrivals for each customers. Figure 5 shows the estimates of Customers 1 and 8 for the distribution of $(V|V > 0)$, as obtained at the end of the simulation. The graphs also depict for reference the theoretical distribution at the equilibrium point according to (1), and an exponential distribution with the same mean. The results for the other customers were similar (Zohar 2000). Figure 6 shows the estimated mean $v_2 = E(V|V > 0)$ of the offered waiting time for these two customers, as a function of their "iteration number" (the number of times they visited the queue). We can see that the estimates tend to converge. At the end of the simulation the mean estimate of the waiting time across the eight customers was 1.2007, with a standard deviation of 0.0672. This agrees well with the theoretical equilibrium value of $x = 1.25$ as calculated above.

Simulation 2: Censored MLE. The same system was simulated with the Censored MLE estimator (16). The number of customers created in this simulation was 11. The results are depicted in Figure 7. We can see that the estimated waiting time converges. The simulation yields a much higher mean waiting time of 1.6452 across 11 customers with standard deviation of 0.0218. This deviation may be attributed to the bias of this estimator, as discussed in the previous subsection, since the waiting time distribution here is not exponential.

Figure 5 **Simulation 1: Estimates of the Waiting Time Distribution for Customers 1 and 8 Using the Kaplan-Meier Estimator**

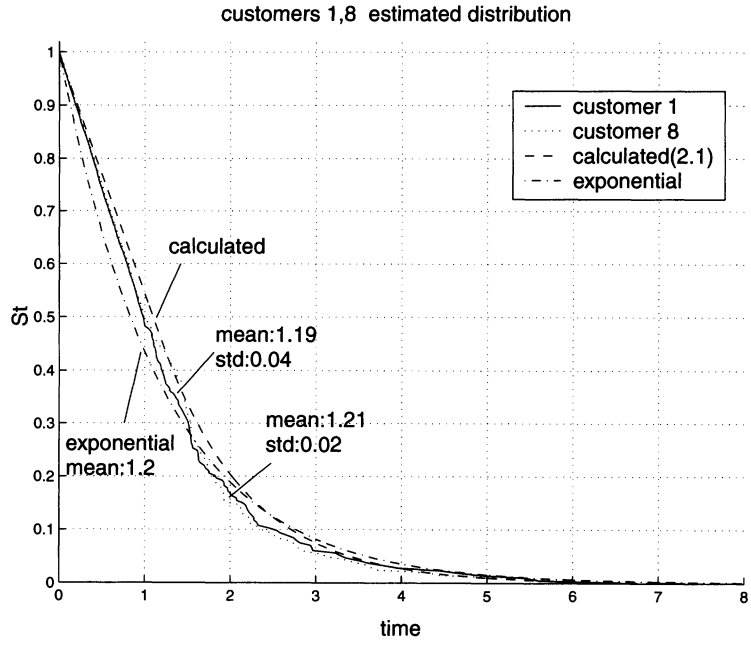


Figure 6 **Simulation 1: Estimates of the Mean Waiting Time $E(V|V > 0)$ for Customers 1 and 8**

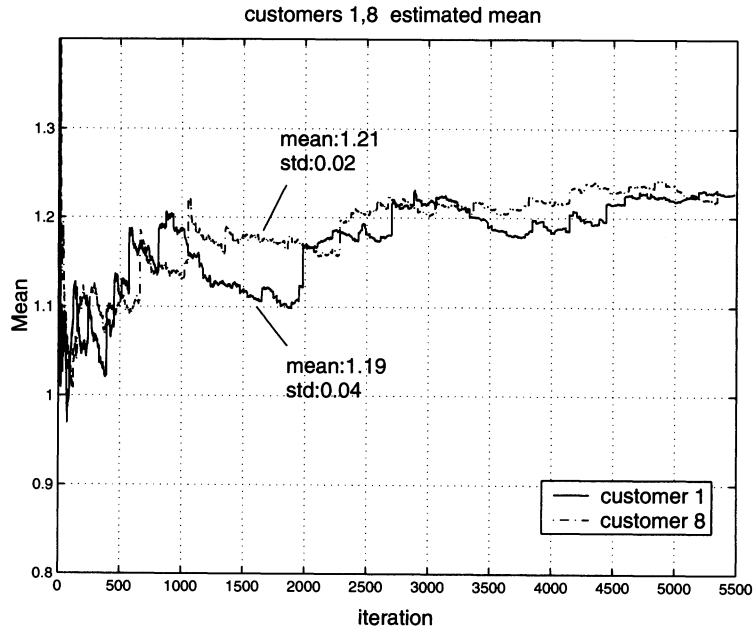
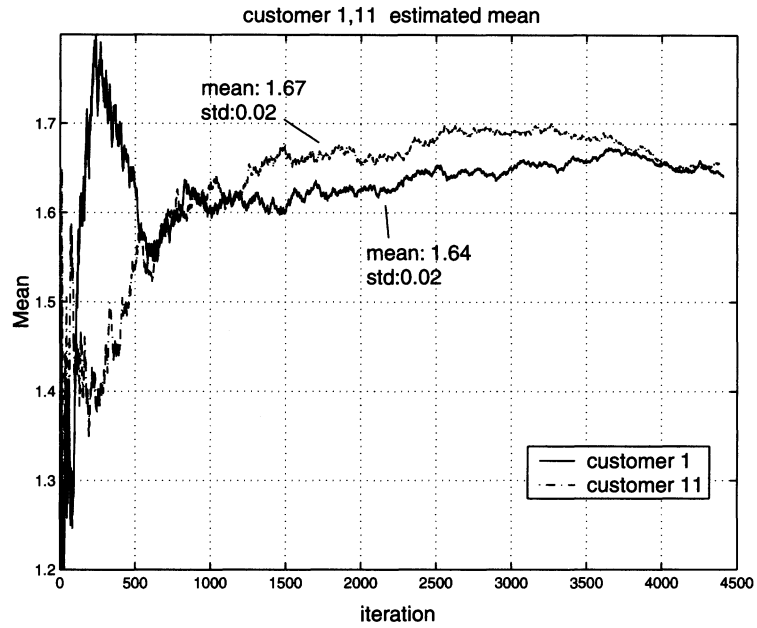


Figure 7 **Simulation 2: Estimates of the Mean Waiting Time $E(V|V > 0)$ for Customers 1 and 8**



The theoretical value of the equilibrium in the last example can in fact be recalculated with an appropriate consideration of the Censored MLE. As shown in Zohar et al. (2001), this calculation gives $x = 5/3 \simeq 1.66$. This is in close agreement with the estimated value that was obtained in the simulation.

7. Conclusion

This paper focused on certain adaptive aspects of customer behavior, namely the dependence of the customers' patience on the anticipated waiting time, and its effect on the performance of queues with invisible state. We have shown how the steady-state operating point (or equilibrium) can be characterized and computed, and demonstrated the applicability of the proposed model for performance analysis. We have shown how the static equilibrium concept can be interpreted as the steady state of a dynamic learning process; while highly idealized, this lends in our opinion considerable credibility to the proposed equilibrium solution. At the same time, the learning process examples demonstrate how the way that customers evaluate their experience can have a significant effect on the resulting equilibrium.

Our model allows considerable freedom in the specific dependence of patience on system performance (i.e., the dependence of G on x). To extend its usefulness in queueing practice, further characterization of this dependence is required, specifying both trends and quantitative relations that hold in given classes of systems. This calls for further research into the abandonment process. Such research must combine empirical analysis, as in Mandelbaum et al. (2000), with further understanding of the triggers of abandonment, as in Zakay and Hornik (1996).

Acknowledgments

The authors would like to thank the two referees and the associate editor for their careful comments which helped to improve the exposition of the paper. We thank Sergey Zeltyn for carefully handling the data analysis and for his very useful feedback. This research was partially supported by the Israeli Science Foundation, Grant 388/99-2, by the Technion V.P.R. fund for the promotion of sponsored research, and by the Fund for Promotion of Research at the Technion.

References

Altman, E., N. Shimkin. 1998. Individual equilibrium and learning in processor sharing systems. *Oper. Res.* **46** 776–784.

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. F. Kylstra, ed. *Performance '81*. North Holland, Amsterdam, The Netherlands, 159–179.

Ben-Shachar, I., A. Orda, N. Shimkin. 2000. Dynamic service sharing with heterogeneous preferences. *Queueing Sys.* **35** 83–103.

Brandt, A., M. Brandt. 2000. Asymptotic results and a markovian approximation for the $M(n)/M(n)/s + GI$ system. Preprint, SC00-12, Konrad-Zuse-Zentrum, Berlin, Germany.

Carmon, Z., D. Kahneman. 1998. The experienced utility of queuing: experience profiles and Retrospective Evaluations of Simulated Queues. Working paper, Fuqua School of Business, Duke University, Durham, NC.

Daley, D. J. 1965. General customer impatience in the queue $G/G/1$. *J. Appl. Probab.* **2** 186–205.

Diekmann, A., M. Jungbauer-Gans, H. Krassnig, S. Lorenz. 1996. Social status and aggression: A field study analyzed by survival analysis. *J. Social Psych.* **136** 761–768.

Friedman, H. H., L. W. Friedman. 1997. Reducing the "wait" in waiting-line systems: Waiting line segmentation. *Bus. Horizons* **40** 54–58.

Garnett, O., A. Mandelbaum, M. I. Reiman. 1999. Designing a telephone call-center with impatient customers. *M&SOM* Submitted for publication ⟨http://ie.technion.ac.il/serveng⟩.

Hassin, R., M. Haviv. 1995. Equilibrium strategies for queues with impatient customers. *Oper. Res. Lett.* **17** 41–45.

Haviv, M., Y. Ritov. 2001. Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Sys.* **38** 495–508.

Hueter, J., W. Swart. 1998. An integrated labor-management system for Taco Bell. *Interfaces* **28**(1) 75–91.

Hui, M. K., D. K. Tse. 1996. What to tell customers in waits of different lengths: An iterative model of service evaluation. *J. Marketing* **60** 81–90.

Kaplan, E. L., P. Meier. 1958. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Association* **53** 457–481.

Larson, R. C. 1987. Perspectives on queues: social justice and the psychology of queueing. *Oper. Res.* **35** 895–905.

Leclerc, F., B. H. Shmitt, L. Dube. 1995. Waiting time and decision making: Is time like money? *J. Consumer Res.* **22** 110–119.

Levine, R. 1997. *A Geography of Time*. Harper Collins Publishers, New York.

Maister, D. H. 1985. The psychology of waiting lines. J. A. Czepiel, ed., *The Service Encounter*. Lexington Books, Lexington, MA, 322–331.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Sys.* **36** 141–173.

——, A. Sakov, S. Zeltyn. 2000. Empirical analysis of a call center. Technical report, Technion. Haifa, Israel.

Miller, R. G. 1981. *Survival Analysis*. Wiley, New York.

Osuna, E. E. 1985. The psychological cost of waiting. *J. Math. Psych.* **29** 82–105.

Palm, C. 1953. Methods of judging the annoyance caused by congestion. *Tele* **2** 1–20.

Riordan, J. 1962. *Stochastic Service Systems*. Wiley, New York.

Roberts, J. W. 1979. Recent observations of subscriber behavior. *9th Internat. Teletraffic Conf. (ITC-9)* (Vol. III). Torremolinos, Spain.

Shaked, M., J. G. Shanthikumar. 1994. *Stochastic Orders and their Applications*. Academic Press, Boston, MA.

Suck, R., H. Holling. 1997. Stress caused by waiting: a theoretical evaluation of a mathematical model. *J. Math. Psych.* **41** 280–286.

Taylor, S. 1994. Waiting for service: The relationship between delays and evaluations of service. *J. Marketing* 56–69.

Thierry, M. 1994. Subjective importance of goal and reactions to waiting in line. *J. Social Psych.* 819–827.

Zakay, D., J. Hornik. 1996. Psychological time: The case of time and consumer behavior. *Time Soc.* **5**(3) 385–397.

Zohar, E. 2000. Adaptive behavior of impatient customers in invisible queues. M.Sc. thesis, Technion, Haifa, Israel.

——, A. Mandelbaum, N. Shimkin. 2001. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. Technical report, Department of Electrical Engineering, Technion, Haifa, Israel.

Accepted by Paul Glasserman; received November 28, 2000. This paper was with the authors $4\frac{1}{2}$ months for 1 revision.