Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support[1]

Ety Zohar[2], Avishai Mandelbaum[3] *and* Nahum Shimkin[4]

October 14, 2001 (Extended Version)

Abstract

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.

Key words: Exponential Queues; Abandonment; Invisible Queues; Tele-queues; Adaptive Customer Behavior; Tele-Services; Call Centers

---

[1]To Appear In *Management Science*, April 2002

[2]Faculty of Electrical Engineering, Technion, Haifa 32000, Israel. e-mail: ety@tx.technion.ac.il

[3]Faculty of Industrial Engineering, Technion, Haifa 32000, Israel. e-mail: avim@tx.technion.ac.il

[4]Faculty of Electrical Engineering, Technion, Haifa 32000, Israel. e-mail: shimkin@ee.technion.ac.il

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 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 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, can *not* accommodate the following scatterplot, that exhibits remarkable patience-adaptivity.

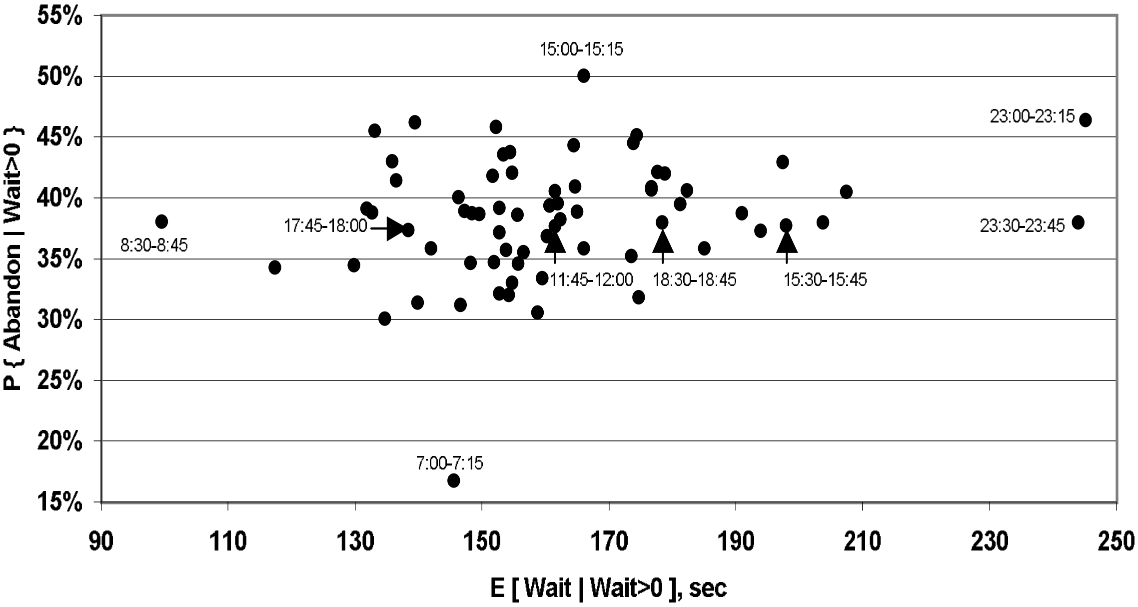

Figure 1: *Adaptive behavior of IN (experienced) customers — abandonment probability vs. average wait (of customers who waited a positive time). Each point corresponds to a 15-minute period of the weekdays, starting at 7:00am, ending at midnight, and averaged over the whole year of 1999.*

The data is from a bank call center as reported in Mandelbaum et al. (2000); see also Section 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 8:30-8:45am, 17:45-18:00, 18:30-18:45 and 23:30-23:45pm 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), Haviv and Ritov (2001) analyze systems with a single customer type, and Mandelbaum and Shimkin (2000) considers 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 Section 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[5]. We distinguish between the average waiting time *assumed* by the customers (denoted x), which determines the patience profile, and between the actual quantity, namely

[5] 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.

the *offered* expected wait that results from this patience profile. Simply put, equilibrium is achieved when the two coincide.

In Section 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 Subsection 3.1. The more interesting case is when patience is allowed to increase with x (Subsection 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 Subsection 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 Section 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 Section 7 with a brief summary and comments concerning future work. The Appendix describes some methods of censored sampling that are used in the paper to estimate means of censored data.

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 non-negative real line. We shall refer to G as the *patience* distribution function. Let $\bar{G} = 1 - G$ denote the survival function; thus $\bar{G}(t)$ 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 Section 3.

Let V denote the *offered* waiting time, or offered wait, which is the time that a (non-abandoning) 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 $m\mu > \lambda\bar{G}(\infty)$, the density F'_V of V is given by (Baccelli and Hebuterne, 1981)

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

with P_{m-1} specified below, and

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

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:

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

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 $\bar{G}(x, \infty) < m\mu$ holds for *some* x. Both expected values $v_i(x)$ are finite when this condition holds.

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 Subsection 3.1, we address the relatively simple case where patience is *decreasing* in the performance parameter x (Assumption A1). 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 Section 4. In Subsection 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 $\bar{F}_1(t) \geq \bar{F}_2(t)$ for all t (here $\bar{F}_i = 1 - F_i$). Y_1 *strictly* dominates Y_2, denoted $Y_1 >_{st} Y_2$, if, in addition, $\bar{F}_1 \neq \bar{F}_2$. We shall also adopt the corresponding notations $\bar{F}_1 \geq_{st} \bar{F}_2$ and $\bar{F}_1 >_{st} \bar{F}_2$ 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 A1. The set of patience distribution functions $\{G(x,\cdot)\}$ is decreasing in x in stochastic order. That is, $x_1 > x_2$ implies that $\bar{G}(x_1,t) \leq \bar{G}(x_2,t)$, for all $t \geq 0$.

Remark. The condition above on the patience distribution, which relates to the customer population as a whole, can be usefully interpreted in terms of individual customer patience. Suppose that each arriving customer is randomly and independently assigned a type, denoted z, which determines his personal abandonment time as a function of x. That is, the abandonment time of a customer of type z is $T_z(x)$. Then assumption A1 is satisfied if $T_z(x)$ is decreasing in x, for each z. Note further that assumption A1 can always be translated to this form, by the standard construction which translates stochastic dominance relations between distributions to almost-sure relations between random variables, cf. Shaked and Shanthikumar (1994).

Proposition 3.1 *Assume A1.*

(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 $\bar{G}_1 \leq_{st} \bar{G}_2$ implies $\bar{F}_1 \leq_{st} \bar{F}_2$.

(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:

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

and let $D(t) = \bar{G}_2(t) - \bar{G}_1(t)$. 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:

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

To establish $\bar{F}_1 \leq_{st} \bar{F}_2$, we shall in fact prove the stronger property that $\bar{F}_1(t)/\bar{F}_2(t)$ is (weakly) decreasing in t. The latter is equivalent to dominance in the hazard rate order; see Chapter 1 of Shaked and Shanthikumar (1994). To establish that $\bar{F}_1/\bar{F}_2$ 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 $\bar{F}_1(0)/\bar{F}_2(0) \leq 1$. 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 $\bar{F}_1(0)/\bar{F}_2(0) \leq 1$, 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 $F_i(0) = \frac{K_m}{\lambda}/[\frac{K_m}{\lambda} + \int\limits_0^\infty \exp(J_i(t)\,)dt]$.

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 $\bar{F}_0(0) = 1$ by definition. $\square$

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 *Assume A1. 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 we our model assumes that both functions are finite for some x, existence follows. $\square$

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 A2). The main result here is Theorem 3.3, which extends the results of the previous section while relying on them for the proof.

Assumption A2 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 A2. 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 $\bar{G}(x+y,\cdot) \leq_{st} \bar{G}(x,\cdot+y)$. It implies, in particular, that $E(T(x))-x$ is decreasing in x.

We establish below that under assumption A2, 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 A2 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 $A2$, as specified below.

Theorem 3.3 *Assume A2. 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 *Assume A2. 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 assumption in A2 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 $\bar{G}(x+y,t) = \bar{G}(x,t-y)$, and we wish to show that $\bar{F}_0(x+y,t) \leq \bar{F}_0(x,t-y)$ 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,

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

Now, from (1),

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

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

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

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

$$F_0'(x+y,t) \;=\; C(x+y)\exp(\int_{-y}^{t-y} K(x,s)ds)$$

$$= \quad C(x+y)\exp(\int_{-y}^{0} K(x,s)ds)\exp(\int_{0}^{t-y} K(x,s)ds)\,.$$

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 $\bar{F}_0(x+y,t) \leq \bar{F}_0(x,t-y)$, with strict inequality holding on some interval; hence $\bar{F}_0(x+y,\cdot) <_{st} \bar{F}_0(x,\cdot)$. This establishes the main claim of this lemma. Since $v_2(x) = E(W(x)\,)$, the second claim follows immediately. $\square$

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 *Assume A2. Then $[v_1(x) - x]$ is strictly decreasing in x.*

Proof: It is required to establish the assertion under Assumption A2, namely $\bar{G}(x+y,t) \leq \bar{G}(x,t-y)$ for $y > 0$. By the monotonicity result in Proposition 3.1 it is sufficient to consider the extreme case where $\bar{G}(x+y,t) = \bar{G}(x,t-y)$, which we henceforth enforce.

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

$$A(x) = \int_0^\infty t\exp[J(x,t)]dt\,, \qquad B(x) = k_m + \int_0^\infty \exp[J(x,t)]dt$$

$$J(x,t) = \int_0^t K(x,s)ds\,, \qquad K(x,s) = \lambda\bar{G}(x,s) - m\mu\,.$$

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

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

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 $(1-\frac{m\mu}{\lambda})K_m \le 1$. This obviously holds when $\frac{m\mu}{\lambda}\ge 1$. Otherwise, we have from (4)

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

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 re-scaling 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 A2 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 $\bar{G}(x,t) = 1$ for $t \le T(x)$ and $\bar{G}(x,t) = 0$ 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 $\bar{G}$ and $m = \lambda = \mu = 1$ gives by explicit calculation

$$v_2(x) = \frac{T^2/2 + T + 1}{T+1} = \frac{1}{2}(T + 1 + \frac{1}{T+1}) \tag{13}$$

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 will be observed in Section 4, one possible effect of customer adaptation is to keep the abandonment fraction approximately constant, even under varying congestion conditions (cf. Figures 1 and 4 and the related discussion). 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 has been observed between these quantities in Figure 5.

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: $G_\eta(t) = 1\{t \geq \eta\}$. 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

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

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

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

In the exponential case a numeric computation is required.

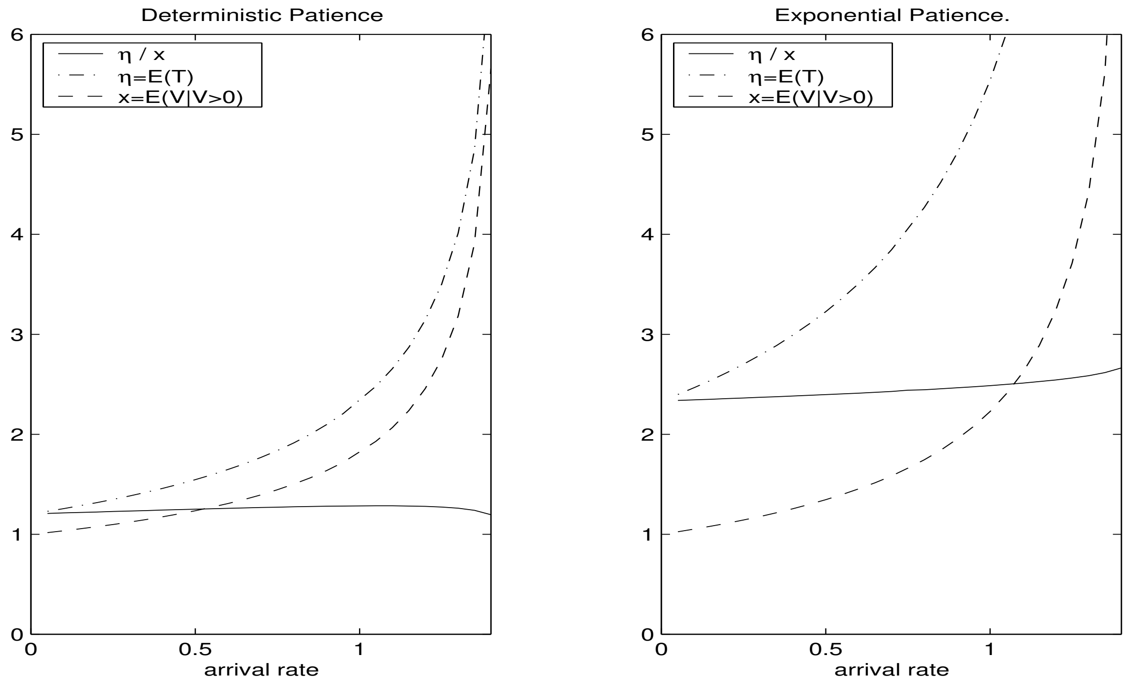

Figure 2: *Patience profiles that keep $P_{ab} = 0.3$, with patience that is deterministic (left) and exponentially distributed (right).*

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, 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 tele-services 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 VRU (Voice Response Unit) 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 Section 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 (Figure 11 in Mandelbaum et al., 2000). 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 following scatterplot 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 7:00am, 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:

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

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 $E[queue\text{-}length] =$

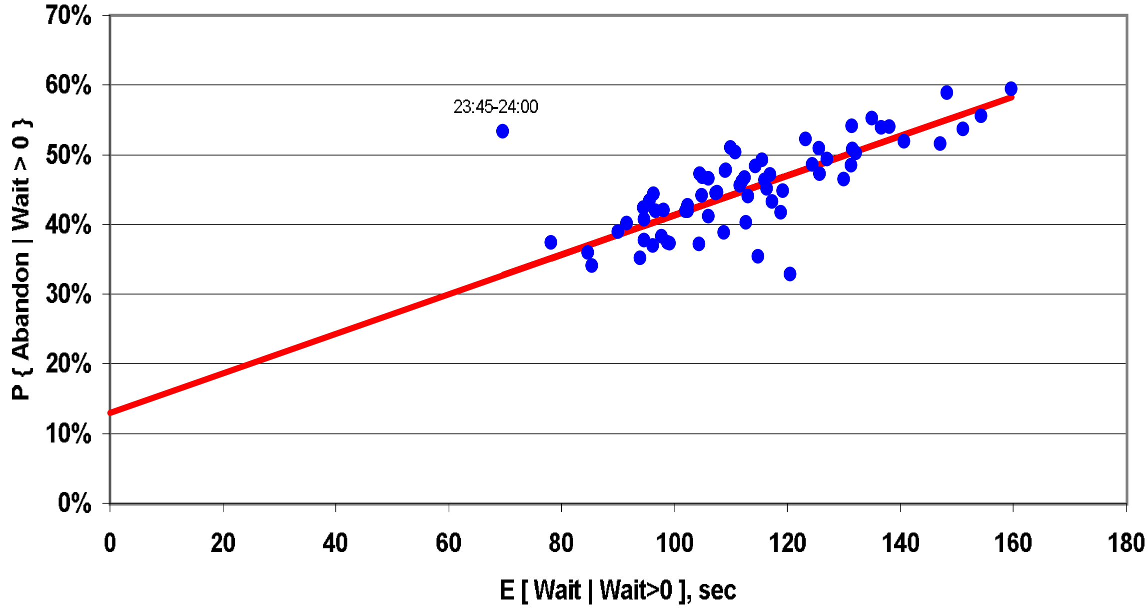

Figure 3: *Novice (NW) customers.* $P\{abandon|wait > 0\}$ *vs.* $E[wait|wait > 0]$.

$\lambda \times E[wait]$, and dividing by $P\{wait > 0\}$, yields the above linearity. (For non-exponential 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

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

(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 affect as an increase from 120 to 160 seconds: both accompany an increase of about 12.5% in abandonment, out of those delayed.

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. This is first rediscovered through Figure 4 (of which Figure 1 is simply a zoom). The difference between

NW customers (Figure 3) and IN customers (Figure 4) is clearly manifested.

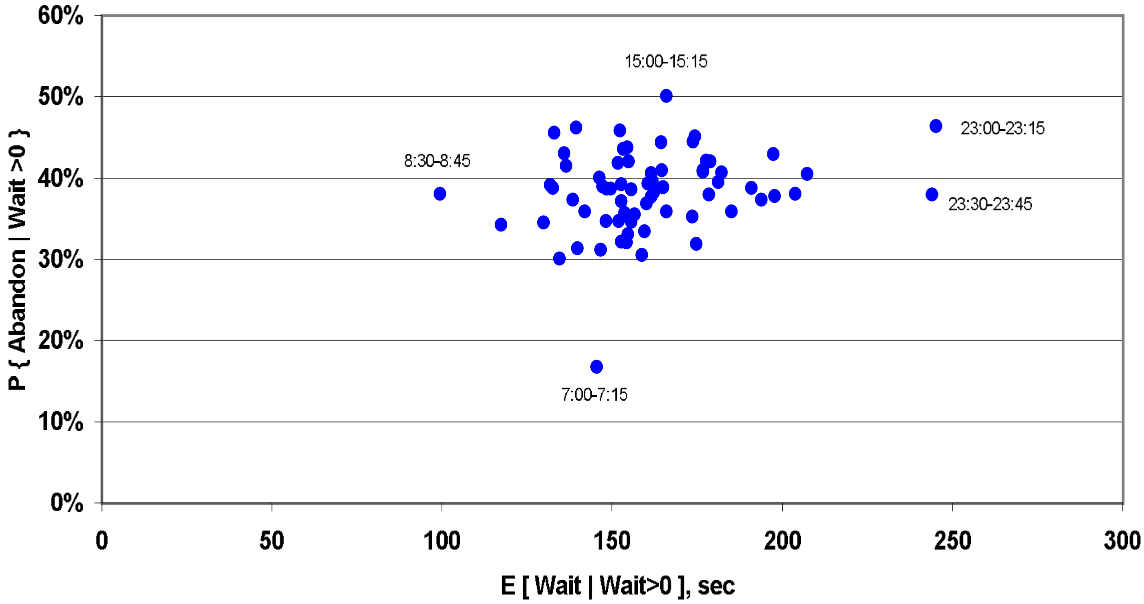

Figure 4: *IN customers.* $P\{abandon|wait > 0\}$ *vs.* $E[wait|wait > 0]$.

Finally, we examine the relation between patience and perceived system performance. To this end, *Patience* will be represented by $E[time\text{-}to\text{-}abandon]$, while system performance will be measured by $E[offered\text{-}wait \; |wait > 0]$. 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 5 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 to 4 is directly observable. In Figure 5, 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

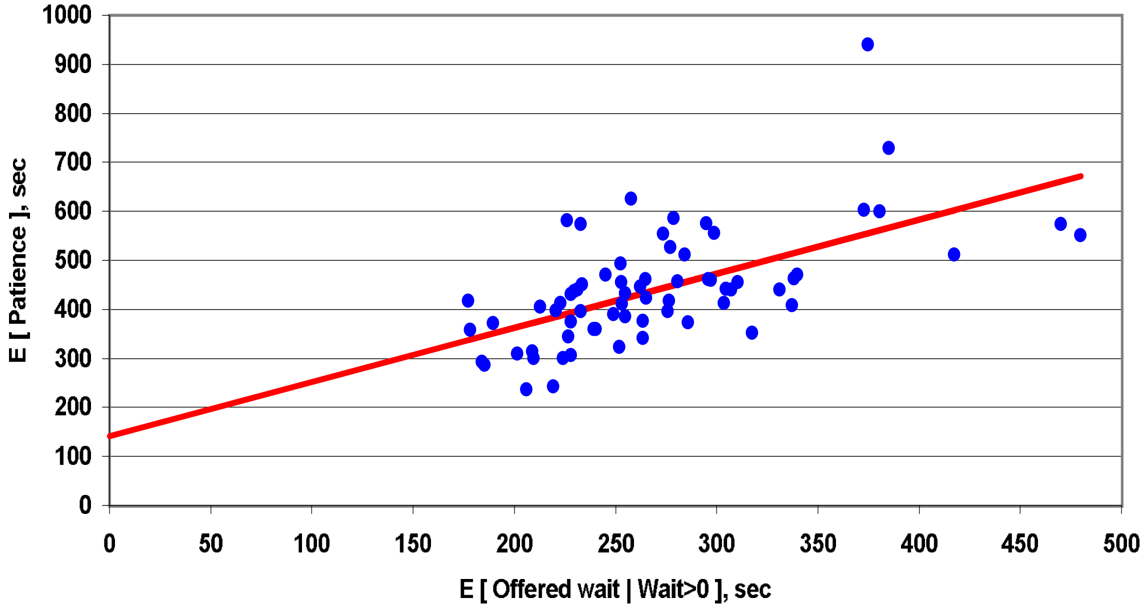

Figure 5: *IN customers. $E[patience]$ vs. $E[offered\ wait|wait > 0]$; $E[\cdot]$ stands for the* mean *of the Kaplan-Meier estimator for the corresponding distribution.*

served. We use for this purpose the classical Kaplan-Meier estimator which is described in the Appendix.

Remark. An analogue of Figure 5 for NW (novice) customers is not displayed. The reason is a lack of statistical confidence – see the remark on robustness in the Appendix, especially Figure 9.

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 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 Section 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.

In this subsection we consider a normative model (decision-theoretic) that has been considered in the literature, relate it to our basic descriptive model, and discuss some of its limitations. The abandonment time of each customer is chosen to maximize a personal utility function. In essence, the optimal choice strikes an optimal balance between the utility of the requested service and the disutility of the (remaining) wait.

The model we consider here is a simplification of the one proposed in Mandelbaum and Shimkin (2000). In the latter, the customer's choice relies on the knowledge of the entire distribution of the offered waiting time. Here, in keeping with the simplifying model assumption of Section 2, we only allow the customer to observe the *mean* waiting time, namely the performance parameter x specified above. We shall briefly present the main elements of the model in Mandelbaum and Shimkin (2000), and then consider its reduction to the present framework.

Consider a specific customer, or customer type, indexed by z. The relative occurrence of customer types is specified by some probability distribution on z. The elements of the decision model for each type-z customer:

- $C_z(t)$, a waiting cost function which specifies the cost for waiting t time units in the queue, and its derivative $c_z(t)$.
- r_z, the expected service utility.
- $F_z(t)$, the probability distribution of the offered waiting time V in the queue.

We shall assume here that the marginal waiting cost $c_z(t)$ is strictly increasing and continuous in t. The optimal abandonment time T_z maximizes the utility function

$$u_z(T) = E_z(r_z 1\{V \le T\} - C_z(\min\{V,T\}))\,,$$

where T is the abandonment time (to be chosen), $\min\{V,T\}$ is the actual time in queue, $\{V \le T\}$ denotes the event that the customer enters service before abandoning, and E_z is the expectation with respect to the distribution F_z of V. It may be verified by differentiation that a stationary point of u_z satisfies

$$H_z(T) = \gamma_z(T)$$

where $H_z = F'_z/\bar{F}_z$ is the hazard rate function, and $\gamma_z(T) = c_z(T)/r_z$ is the cost-benefit ratio.

To utilize the above relation using the single parameter x, we shall employ the above decision model under the (subjective) presumption that the virtual waiting time distribution F_z is exponential, with expected value x. Besides its associated model simplification, this assumption may be justified on the ground that typically a customer will not be aware of the details of the waiting time distribution, but rather summarize his beliefs concerning the expected wait by its average. Under this assumption, we have that $H_z(T) = 1/x$, and the optimality condition reduces to $1/x = \gamma_z(T)$. Note that this equation has at most one solution since c_z (hence γ_z) is increasing by assumption. It may now be verified that the optimal solution which maximizes $u_z(T)$ is $T_z = 0$ if $1/x$ is below the range of γ_z, $T_z = \infty$ if $1/x$ is above the range of γ_z, and

$$T_z = \gamma_z^{-1}(\frac{1}{x}) \tag{16}$$

otherwise. Here γ_z^{-1} is the inverse function of γ_z. To illustrate, if $\gamma_z(T) = a_z T$, then $T_z = 1/(a_z x)$.

Let us now relate this model to the assumptions made in Section 3, concerning the dependence of the patience distribution on x. Since γ_z is increasing, it follows immediately that $T_z = T_z(x)$ is decreasing in x. Since this holds for each customer type, it immediately implies assumption A1 of Section 3.1.

As argued in Section 3.2, it is plausible that the patience of certain customers will increase, rather than decrease, when they anticipate a longer wait. To include such a trend in the above model, it is required to allow the cost function itself to depend on the customer expectations, namely on x. Such dependence is reasonable since in a typical situation the waiting cost is subjective and includes a predominant psychological component, as discussed at the beginning of the present section.

To be specific, assume that as the expected waiting time x increases, the waiting cost may decrease, but in a manner bounded by a shift of the original cost; that is $c_z(t) = c_z(x,t)$, and $c_z(x+y,t) \geq c_z(x,t-y)$. Then (16) implies that $T_z(x+y) < T_z(x) + y$, which corresponds to assumption A2 of section 3.2.

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 that 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 Section 4, with the need to estimate the mean of a distribution based on censored data.

We first employ a standard non-parametric estimator for censored data, namely the Kaplan-Meier (KM) estimator discussed in the Appendix, 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 – see (18). We shall refer to it as the Censored MLE. 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 one-line learning model that we propose is based on the following scenario. Each customer initially possesses 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 Section 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 Section 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 untill 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 6 and 7. 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 40000 arrivals, which amounted to about 5200 arrivals for each customers. Figure 6 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 7 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 8 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.

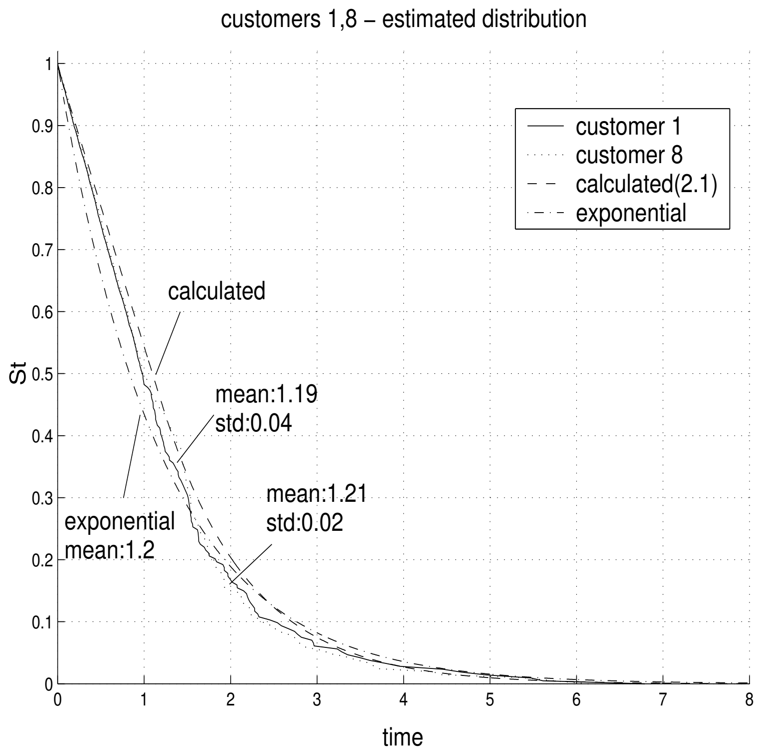

Figure 6: *Simulation 1: Estimates of the waiting time distribution for customers 1 and 8 using the Kaplan–Meier estimator*

Simulation 2: Censored MLE. The same system was simulated with the Censored MLE estimator (18). The number of customers created in this simulation was 11. The results are depicted in Figure 8. We can see that the estimated waiting time converges. The simulation

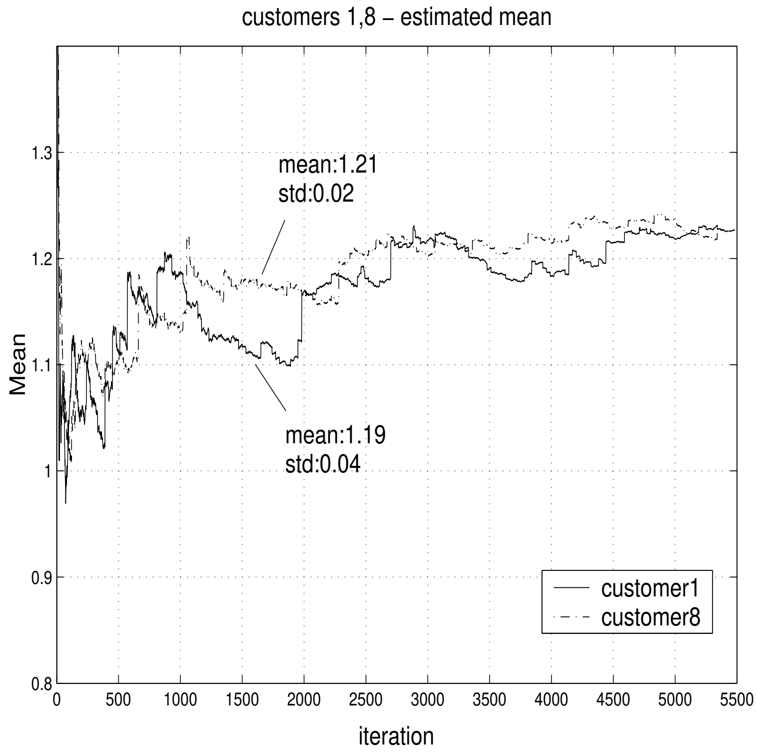

Figure 7: *Simulation 1: Estimates of the mean waiting time $E(V|V>0)$ for customers 1 and 8*

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.

The theoretical value of the equilibrium in the last example can in fact be recalculated with an appropriate consideration of the Censored MLE. Based on (18), the asymptotic value $\hat{x}$ of the Censored-MLE for $E(V|V>0)$ may be written as

$$\hat{x} \;=\; \frac{E(W|V>0)}{P(no\ abandon|V>0)} = \frac{E(\min(V,T)|V>0)}{P(V<T|V>0)}\,,$$

where we suppress the dependence of T on x. Letting p_0 denote the distribution of $(V|V>0)$, this gives

$$\hat{x} = \frac{\int_0^T tp_0(t)dt + T\int_T^\infty p_0(t)dt}{\int_0^T p_0(t)dt}$$

and from (3) we have $p_0(v) = K_0 \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(s)\right) ds$, with K_0 a normalization constant. Recall that the abandonment time T is assumed here deterministic and identical for

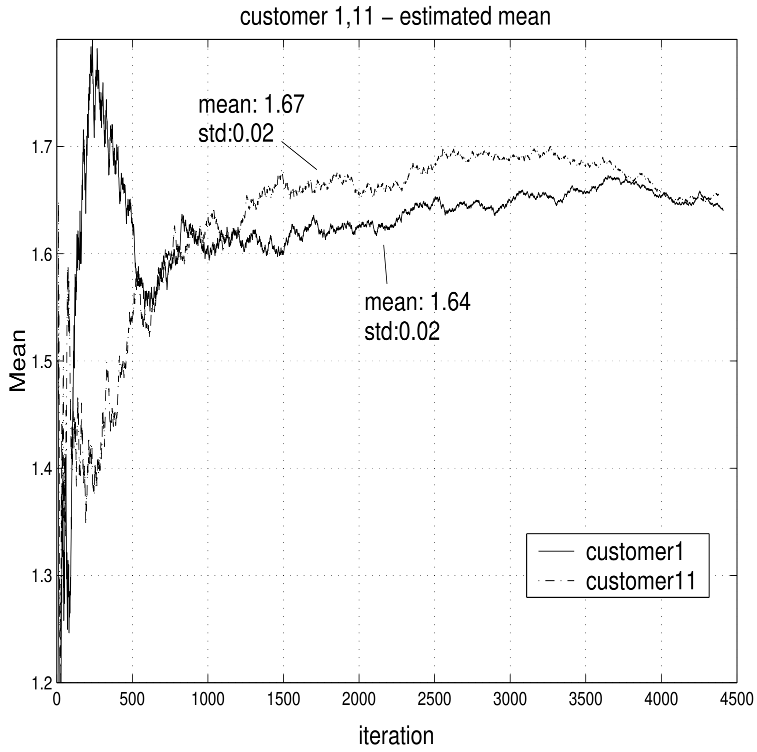

Figure 8: *Simulation 2: Estimates of the mean waiting time $E(V|V > 0)$ for customers 1 and 11, using the biased MLE estimator*

all customers, so that $\bar{G}(t) = 1$ for $t \leq T$ and $\bar{G}(t) = 0$ for $t > T$. With $m = \mu = \lambda = 1$ we obtain $p_0(v) = K_0$ for $t < T$, and $p_0(v) = K_0 \exp(T - t)$ for $t > T$. It follows that

$$\hat{x} = \frac{\int_0^T t dt + T \int_T^\infty \exp(T - t) dt}{\int_0^T 1 dt} = \frac{T}{2} + 1\,.$$

The required equilibrium equation now is $\hat{x} = x$, with $T = T(x) = 0.8 \cdot x$. This gives $x = \frac{0.8x}{2} + 1$, or $x = \frac{5}{3} \simeq 1.66$. 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., and Shimkin, N. (1998). Individual equilibrium and learning in processor sharing systems. *Operations Research*, *46*, 776–784.

Baccelli, F., and Hebuterne, G. (1981). On queues with impatient customers. In F. Kylstra (Ed.), *Performance '81* (pp. 159–179). North Holland, Amsterdam.

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

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

Carmon, Z., and Kahneman, D. (1998). *The experienced utility of queuing: experience profiles and retrospective evaluations of simulated queues.* Working paper, Fuqua School of Business, Duke University.

Daley, D. J. (1965). General customer impatience in the queue $G/G/1$. *Journal of Applied Probability*, *2*, 186–205.

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

Fleming, T. R., and Harrington, D. (1991). *Counting Processes and Survival Analysis.* Wiley-Interscience, New York.

Friedman, H. H., and Friedman, L. W. (1997). Reducing the 'wait' in waiting-line systems: waiting line segmentation. *Business Horizons*, *40*, 54–58.

Garnett, O., Mandelbaum, A., and Reiman, M. I. (1999). *Designing a telephone call-center with impatient customers.* Under revision to MSOM; preprint available at http://ie.technion.ac.il/serveng.

Hassin, R., and Haviv, M. (1995). Equilibrium strategies for queues with impatient customers. *Operations Research Letters*, *17*, 41–45.

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

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

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

Larson, R. C. (1987). Perspectives on queues: social justice and the psychology of queueing. *Operation Research*, *35*, 895-905.

Leclerc, F., Shmitt, B. H., and Dube, L. (1995). Waiting time and decision making: is time like money? *Journal of Consumer Research*, *22*, 110–119.

Levine, R. (1997). *A Geography of Time.* Harper Collins Publishers, New-York, NY.

Maister, D. H. (1985). The psychology of waiting lines. In J. A. Czepiel (Ed.), *The service encounter* (pp. 322–331). Lexington Books, Lexington, Mass.

Mandelbaum, A., Sakov, A., and Zeltyn, S. (2000). *Empirical analysis of a call center.* Technical report, Technion, August 2000.

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

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

Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology*, *29*, 82–105.

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

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

Roberts, J. W. (1979). Recent observations of subscriber behavior. In *9th international teletraffic conference (ITC-9)* (Vol. III). Torremolinos, Spain.

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

Suck, R., and Holling, H. (1997). Stress caused by waiting: a theoretical evaluation of a mathematical model. *Journal of mathematical psychology*, *41*, 280–286.

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

Thierry, M. (1994). Subjective importance of goal and reactions to waiting in line. *Journal of Social Psychology*, 819–827.

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

Zohar, E. (2000). *Adaptive behavior of impatient customers in invisible queues.* M.Sc. Thesis, Technion, October 2000.

Zohar, E., Mandelbaum, A., and Shimkin, N. (2001). *Adaptive behavior of impatient customers in tele-queues: Theory and empirical support.* Technical Report, Department of Electrical Engineering, Technion, October 2001.

Appendix – Censored Sampling

The need for accommodating censored data arose first in Section 4. Based on the call center data in Mandelbaum et al. (2000), we sought to estimate *patience* – the distribution of the time a customer is willing to wait, and relate it to *offered wait* - the time a customer if forced to wait. As explained in Section 4, these two quantities actually censor each other. Then, in Section 6, censored data arose again. Simulated customers sought to estimate the system's offered wait, based on their individual service history where some samples of the offered wait were censored by abandonment. In both Sections 4 and 6, one is required actually to estimate only means, as opposed to the full fledged distribution. (The latter is needed, for example, to support our first observation in Section 4, regarding the non-exponentiality of patience. See Mandelbaum et al. (2000), Section 6, especially Figures 12 and 14, for interesting hazard-rate estimators of patience and offered wait.)

Techniques for analyzing censored data have been developed within the well-established Statistical branch of Survival Analysis (Miller, 1981 is an elementary exposition, and Fleming and Harrington (1991) is advanced measure-theoretic). As will be explained in the sequel, our needs for such techniques vary from the rudimentary to the unexplored.

In Section 4 we estimated mean patience and mean offered-wait via the means of the corresponding classical Kaplan-Meier (KM) estimator (17). KM generalizes the empirical distribution function to accommodate censored samples (see page 46 in Miller, 1981, or page 4 in Fleming and Harrington, 1991). It is a *non-parametric* estimator, proven to have desirable properties, and common enough to be incorporated in essentially all respectable statistical packages. In Section 6 we used again KM, and then continued with a simpler *parametric* estimator, namely the maximum-likelihood estimator (MLE) of the mean of an exponential distribution; it is defined in (18) and referred to in our paper as the censored MLE (CMLE). The rest of the Appendix is devoted to a description of KM and CMLE, tailored to the estimation of patience and offered wait.

The KM setup for estimating patience is as follows. We are given a sample $\{W_i\}$ of N waiting times from a call center. Some of the calls end up with abandonment ($W_i = T_i$) and the others with a service ($W_i = V_i$). Denote by $M \leq N$ the number of *distinct* abandonment times in the sample. Let $T^1 < T^2 < ... < T^M$ be the ordered observed abandonment times, and A_k the number of abandonment at T^k, namely those who abandon after exactly T^k units of time. The *Kaplan-Meier estimator* $\hat{S}(t)$, $t \geq 0$, estimates the survival function $\bar{F}(t) = P(T > t)$, where T is the time to abandon (patience). It is given by

$$\hat{S}(t) = \prod_{k:T^k \leq t} (1 - \frac{A_k}{\bar{B}_k}),$$

where $\bar{B}_k$ denotes the number of customers still present at T^k, that is neither served nor abandoned before T^k. The estimator for mean patience is then based on the tail-formula

$$\widehat{E[T]} = \int_0^\infty \hat{S}(t)dt. \tag{17}$$

In the above we estimated patience, which was censored by offered wait. Similarly, KM can be used to estimate the offered wait, by switching the roles of V_i and T_i. This estimate was used both in Section 4 and 6, in the latter by individual customers in order to estimate the system's offered wait that affects their patience.

A simpler alternative for estimating offered wait takes a parametric approach. As above, let $\{W_1, W_2, ..., W_N\}$ denote the collection of all waiting times, both abandoning and served. Assuming that offered wait is exponentially distributed, the standard parametric (maximum likelihood) estimator for its mean is given by (Miller, 1981, page 22)

$$\widehat{E(T)} = \frac{1}{N_s} \sum_{i=1}^{N} W_i \,, \tag{18}$$

where N_s is the number of service experiences that ended up with a service, i.e. were not censored by abandonment. If T is not exponential, the estimator (18) is biased enough to be inconsistent.

Remark. *On Independence:* KM assumes independence for the observations whose distribution is to be estimated. Such an independence is plausible for patience (T_k's). It also applies for offered wait (V_i's), if these are sampled during independent sparsely-timed visits to the queue, as in Section 6. Such independence can *not* hold for successive offered loads, that are in fact highly dependent. In this case one is taken out of the KM paradigm. The effect of such dependence has been ignored in Section 4, as well as in Mandelbaum et al. (2000), and it is the subject of ongoing research.

Remark. *On Robustness:* The KM (Kaplan-Meier) estimator is very sensitive to censored data at the upper tail of the sample. For example, if the longest wait in a customer's history ended up with an abandonment, the KM estimator of the offered wait has a positive mass at infinity, hence its mean is infinity; similarly if one is interested in patience, and the longest wait ended up with a service. The consequence is that in estimating patience and offered wait, one of the resulting two KM's must be defective, and common practice is to simply truncate it at its last observation. (There are some parametric tail-smoothing techniques, but to the best of our knowledge they are ad-hoc.)

Another alternative is to use medians, rather than means, as more robust estimators of a location-parameter. For example, the analogue of Figure 5 for NW customers, but with medians rather than means, is the following:

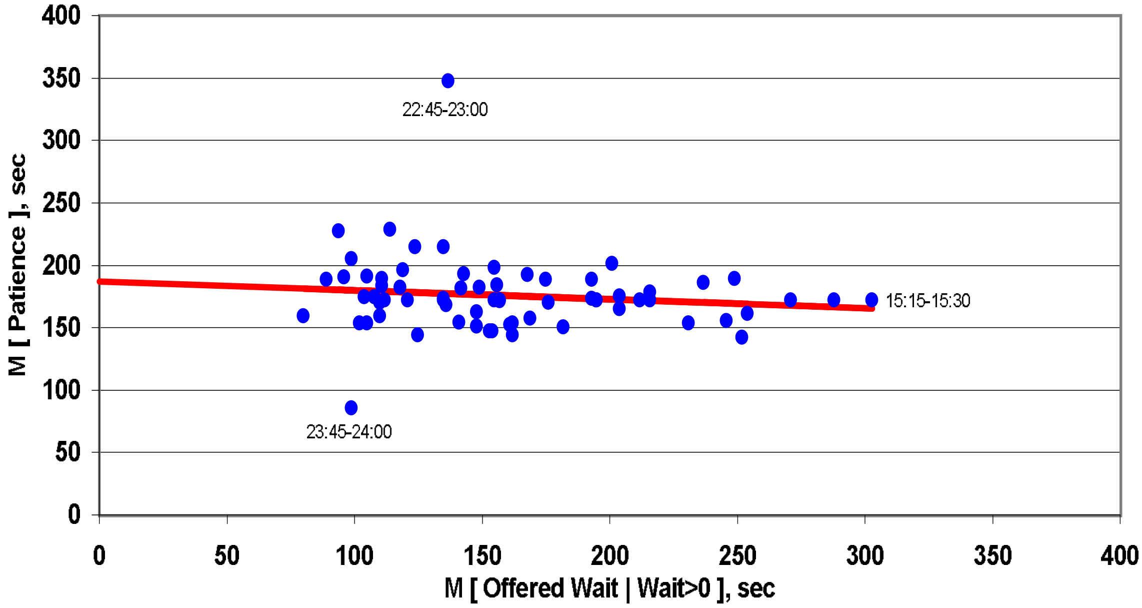

Figure 9: *NW customers.* $M[patience]$ *vs.* $M[offered\ wait|wait > 0]$; $M[\cdot]$ *stands for the* median *of the Kaplan-Meier estimator for the corresponding distribution.*

The flatness, to be compared against the slope in Figure 5, can be attributed to insensitivity of NW patience to congestion, due to their unfamiliarity with the system. As mentioned in Section 4, replacing the medians in Figure 9 with means yields statistically unreliable scatterplots – this is, in fact, the subject of ongoing research.

Two final comments (or reservations) on the use of medians. First, in the context of this paper the mean seems to be a more natural descriptor of human perception of past performance, and is also more amenable for analysis. Hence the median is not appropriate as a basis for an adaptive theory as developed here. On the technical side, one should note that with ample censoring it is also possible for the KM median to be undefined; this happens, for example, when the whole upper half of the sample consists of customers who were patient enough to get served, hence their patience is censored. This phenomenon does occur in Mandelbaum et al. (2000), but not for the customer types that are discussed here.