DISCUSSION PAPER SERIES

No. 4036

RESPONSIVE PRICING

Pascal Courty and Mario Pagliero

INDUSTRIAL ORGANIZATION

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP4036.asp**

ISSN 0265-8003

RESPONSIVE PRICING

Pascal Courty, London Business School (LBS) and CEPR
Mario Pagliero, London Business School (LBS)

Discussion Paper No. 4036
August 2003

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 4036

August 2003

ABSTRACT

Responsive Pricing*

We study the efficiency property of responsive pricing ? a scheme first proposed by Vickrey ? that increases prices as a function of capacity utilization. We show that although responsive pricing implements allocations that are arbitrarily close to market clearing, these allocations are not always efficient. We identify conditions under which efficiency occurs and discuss implications for the use of responsive pricing.

JEL Classification: D45 and L97
Keywords: dynamic pricing, real time pricing and responsive pricing

Pascal Courty
London Business School
Sussex Place
Regent's Park
LONDON
NW1 4SA
Tel: (44 20) 7262 5050 x3317
Fax: (44 20) 7402 0718
Email: pcourty@london.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=134679

Mario Pagliero
London Business School
Sussex Place
Regent's Park
LONDON
NW1 4SA
Email: mpagliero@london.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=153261

*We would like to thank seminar participants at the LSE and Marcus Asplund, Gary Becker, and Marco Ottaviani for useful comments. All opinions and any errors are ours.

Submitted 01 August 2003

1 Introduction

Economists have long recognized the necessity to vary prices to allocate congestible resources efficiently when demand changes over time. Peak load pricing, which deals with the simplest case where demand changes are predictable, constitutes the most celebrated application.[1] In this paper, we investigate the extent to which responsive pricing, a pricing scheme introduced by Vickrey in 1971 that proposes to vary prices in real time as a function of capacity utilization, can increase efficiency when demand changes are unpredictable.[2] The class of applications that are relevant include:

- Telephone use: This was the original application used by Vickrey to motivate responsive pricing. Vickrey proposed to quote each new user a charge that would vary as a function of the level of network congestion.

- Some economists have proposed to vary price in real time in electricity markets (Borenstein, 2001), Internet pricing (Varian and MacKie-Mason, 1994) and road pricing to provide better consumption incentives and reduce congestion and wasted capacity.[3]

- Internet café: Lines and empty terminals prevail at Internet cafés that use fixed pricing or peak load pricing. easyEverything, the largest chain of Internet café in the world, followed Vickrey's proposal and gives discounts that are a function of the number of vacant terminals (Courty and Pagliero, 2003).[4]

- Ski resorts: Prices could vary in real time to give an incentive to ski less during high demand periods thus reducing lines, and to ski more when demand is low thus

[1] See Boiteux (1956 and 1960) and Crew, Fernando and Kleindorfer (1995) for a recent review.

[2] Vickrey's main message was to "call attention to the possibilities that arise if one attempts seriously to promote efficiency through causing prices to fluctuate so as to clear the market [...] even in response to those fluctuations that can not be fully predicted in advance."

[3] In an experiment, the San Diego's Regional Planning Agency has used responsive pricing to allocate fast track lanes in highways. See http://argo.sandag.org/fastrak/. Cars that want to use the fast track lanes have to pay a fee that varies in real time as a function of congestion. Consumers face a trade-off between the amount of time they want to save and the fees they are willing to pay.

[4] See http://www.easyeverything.com/

achieving a more efficient use of capacity. The same principle could be applied to price access to other sport facilities and theme parks.[5]

Other examples can easily be found. In these applications, measures of congestion (i.e. utilization rate) can be used to compute congestion-contingent prices that are communicated to consumers in real time. Responsive pricing proposes to increase access prices as utilization rates increase – that is, as capacity gets closer to congestion.

To understand why prices have to respond to demand shocks, consider what happens under un-responsive pricing. If prices are set according to the expected level of demand at a given time, as predicated under peak load pricing, the very nature of the randomness of the arrival process implies that there are times when the number of new arrivals exceeds or falls short of available capacity. If prices do not vary as a function of realized demand, some potential buyers are denied access when there is a sudden arrival of consumers and capacity is wasted when there is an unexpected decrease in the flow of consumers.

The set of applications where responsive pricing could be used have the characteristics that although demand variations are to some extent impossible to predict, it may be possible to influence the length of time consumers use the service. When this is the case, one can seriously think of using prices to achieve more efficient allocations of the congestible resource between users. The welfare gains from using responsive pricing are potentially great since congestion and/or unused capacity otherwise prevail. For example, electricity blackouts sometimes occur as a result of temporary excess demand, although costly capacity is available and unused most of the year. Lines in ski resorts and unused telephone capacity are also common.

Our analysis proceeds in two steps. First, we show that responsive pricing achieves market clearing in a limit sense – where the price is extremely responsive to occupancy. In the limit, sudden demand shocks trigger immediate changes in prices and consumers adjust their length of use, resulting in an elimination of rationing or unused capacity. When demand is high, for example, the price increases and consumers have an incentive

[5]To deal with waiting on popular rides, some theme parks sell fast track passes that enables holders to bypass queues (http://www.sixflags.com/parks/wyandotlake/parkinfo/fastlane.asp) while others offer reservation systems which replace waits with virtual lines assigning ride times (http://www.themeparksonline.org/).

to satisfy only their most urgent needs (e.g., important email in the Internet café example). When demand is low, consumers are incentivized to satisfy marginal needs (e.g. less urgent emails or web surfing). In both cases, prices adjust so that demand equals capacity.

Second, we ask whether the limit responsive scheme that achieves market clearing implements an efficient allocation. We show that the limit outcome is efficient under a simple condition on consumer demand, called the no-crossing condition. When this condition holds, consumers leave when their marginal willingness to pay is equal to the instantaneous price. This decision rule implies that the consumers who retain access always value consumption more than the consumers who decide to stop consumption. When the no-crossing condition does not hold, however, responsive pricing does not always aggregate consumer's private information efficiently and does not always achieve the (information constrained) efficient outcome.

This works stresses the distinction between the concepts of market clearing and efficiency. These two concepts are equivalent in the standard textbook model of supply and demand. In our application, however, these two concepts are not always equivalent. Although responsive pricing achieves outcomes that are arbitrarily close to market clearing, these outcomes are not always efficient because market clearing prices sometimes communicate inefficient consumption incentives.

As mentioned above, the closest work to our analysis is Vickrey (1971). Vickrey introduced the concept of responsive pricing and speculated that it may achieve efficiency although he expressed (in the context of an application to telephone pricing) the reservation that "one significant imperfection would remain with such a system: a user upon being informed of the current rate may still be unclear as to whether he should let the call go through at the current rate or defer the call until later, since he has no assurance of what the rate would be at the later time." Our model formalizes Vickrey's conjecture that consumer strategic behavior may impede efficiency. In addition, we identify a condition under which the efficient outcome is always achieved.[6]

[6]Interestingly, Vickrey focused on consumers' decision to strategically postpone the start of consumption while our model focuses on the decision to strategically postpone the decision to end consumption. The logic for inefficiency is the same in both cases and rests on the ideas that instantaneous prices may inefficiently aggregate consumers' private information.

The paper is organized as follows. The next section presents a simple steady state example that introduces the main themes of the paper. In this example, however, prices do not change once they reach the steady state. Section 3 introduces a model with dynamic pricing and section 4 presents the main results. Section 5 discusses some important extensions. Section 6 presents some concluding remarks.

2 Steady State Arrival Rate

For expositional clarity, we frame the model in the context of an Internet café, but the analysis applies equally to all the applications discussed earlier. Let Q represent the store capacity. We treat Q as exogenously given and we assume that all costs are fixed. The marginal cost of serving an additional consumer is zero up to capacity Q and infinite once capacity is reached. On the demand side, we assume that consumers have identical demands but the argument presented in this section would easily extend to heterogeneous demands. A consumer who has already consumed n unit and consumes one more unit gets utility $v(n)dt$ for that additional unit where $v'(n) < 0$. This demand function captures the fact that consumers first satisfy their most urgent needs and value each additional unit of consumption less than the previous one.

The number of consumers who join the store per unit of time is $\tilde{\epsilon}dt$ where $\tilde{\epsilon}$ is a random variable that captures the role of unpredictable demand shocks. The interpretation we have in mind for $\tilde{\epsilon}$ is the following. Each day, a demand shock is realized and the demand shock is constant throughout the day. Without loss of generality, we focus on the efficiency properties of the equilibrium steady state.[7]

The instantaneous price as a function of occupancy q is $p(q)$ where where $p(.)$ is a non-negative, continuous, and increasing function with support $[0, Q]$. This captures the spirit of Vickrey's proposition that "it seems entirely satisfactory to base rates on levels of activity." In the case where $p(q)$ is linear, one should interpret $p(Q)$ as the congestion charge and the slope of p as a measure of how responsive the pricing function is.

We solve for the equilibrium price, length of stay, and occupancy level given pricing

[7] In this simple framework, the result that responsive pricing implements the efficient outcome would still hold if one also takes into account the transition phase as will become clear in the next section.

scheme $p(q)$ and arrival realization $\tilde{\epsilon}$. Occupancy and the instantaneous price are jointly determined in equilibrium. An equilibrium is a triplet (n, q, p) such that consumption decisions maximize consumer utility, occupancy is determined by consumers consumption decisions, and the price is given by the pricing rule. Given equilibrium price p, consumers consume up to the point where their willingness to pay for a unit of consumption equals the price $v(n) = p$. Given consumers consumption decisions, the store occupancy is $q = n\tilde{\epsilon}$, and the price is determined by the pricing curve $p = p(q)$. After replacement, equilibrium occupancy in state $\tilde{\epsilon}$ must satisfy

$$v(q/\tilde{\epsilon}) = p(q).$$

This gives the equilibrium occupancy level if $q \leq Q$; otherwise $q = Q$ and rationing occurs. Higher arrival rates imply that consumers consume less ($dn/d\tilde{\epsilon} < 0$), store occupancy is higher ($dq/d\tilde{\epsilon} > 0$), and the price is higher ($dp/d\tilde{\epsilon} > 0$). Figure 1 illustrates these properties. To simplify, the figure assumes that the arrival rate can only be high or low. The equilibrium occupancy is located at the point where the inverse demand ($v(q/\tilde{\epsilon})$) and the pricing curve intercept. One can think of $p(q)$ as a supply curve. The realized price is higher in the high state, that is, when capacity is scarcer, and consumers respond by sharing the capacity available more (lower n).

To understand what is specific to responsive pricing, we contrast the outcome under responsive pricing with the outcome under fixed pricing. Under fixed price ($p(q) = p$) consumers stay n periods such that $v(n) = p$. Length of stay does not vary because consumers do not have any incentive to vary consumption as a function of store congestion. It is easy to show that fixed pricing is welfare-dominated by responsive pricing. Assume for example that the price is fixed at p_0 and consider a responsive pricing scheme such that $p(Q) = p_0$ and $p(q) < p_0$ for $q \leq Q$. Fixed pricing and responsive pricing generate the same outcome (i.e. same level of congestion) in states of the world where $\tilde{\epsilon} \geq Q/v^{-1}(p_0)$. In states of the world where $\tilde{\epsilon} \leq Q/v^{-1}(p_0)$, however, consumers consume more under responsive pricing, that is, less capacity is wasted.

In this simple steady state framework responsive pricing can achieve the efficient outcome. To show that, consider a social planner who chooses the consumption rule to

maximize expected steady-state surplus

$$E\tilde{\epsilon}\int_0^{n(\tilde{\epsilon})} v(x)dx,$$

where the expectation is taken with respect to $\tilde{\epsilon}$, subject to the constraint that occupancy is feasible $\tilde{\epsilon}n(\tilde{\epsilon}) \leq Q$. The efficient consumption rules specifies that consumers should equally share the capacity

$$n(\tilde{\epsilon}) = \frac{Q}{\tilde{\epsilon}}.$$

Under that consumption rule no capacity is wasted and it is not possible to reallocate capacity to increase welfare. To show that responsive pricing can implement the efficient outcome in a limit sense, consider the class of pricing schemes $p(q) = v(0) - \beta(Q - q)$. This class of pricing scheme rules out rationing since $p(Q) = v(0)$. Occupancy in state $\tilde{\epsilon}$ is $p(q(\tilde{\epsilon})) = v(q(\tilde{\epsilon})/\tilde{\epsilon}) > 0$ and this implies that

$$Q - q(\tilde{\epsilon}) < v(0)/\beta.$$

More responsive schemes (higher β) increase occupancy and therefore efficiency (see Figure 2). Occupancy converges to capacity as β converges to infinity. But this corresponds to the allocation that maximizes social welfare. If one interprets the pricing curve under responsive pricing as a supply curve, then the limit scheme (arbitrarily large β) corresponds to a supply curve that is vertical at Q.

This simple steady state example shows that responsive pricing endogenously set prices in response to demand realizations and implements an outcome that both achieves market clearing and is efficient. The steady state case, however, does not capture the dynamic nature of the consumer arrival process. As a result, prices do not vary in steady state and consumers face a simple decision problem. When the arrival rate changes over time, however, prices will also change and consumers will face a more complex decision problem because they will have to anticipate future prices to decide whether to retain access or quit. The rest of this paper generalizes the analysis to general arrival processes and asks whether the results on efficiency carry through.

3 Model

The model makes several assumptions to simplify the exposition and to focus the analysis on the most important issues. We discuss in the extension section those issues that the core of the analysis ignores. $\epsilon_t(\omega)dt$ consumers walk by the store in interval dt for $t \geq 0$ in state ω where $\epsilon_t(\omega)dt$ is an integrable continuous stochastic process on some probability space with increments distributed over $[\epsilon_l, \epsilon_h]$ $(0 < \epsilon_l < \epsilon_h < \infty)$. Beyond this, we do not make any assumption on ϵ_t.[8] There could be a seasonal component (distribution of ϵ_t depends on t) and also a random component that could be correlated over time. Occupancy in period t is denoted by $q_t(\omega)$ and we fix $q_0(\omega) = 0$.

The store capacity and the consumer demand are the same as before. We assume first that consumers have identical demands and then consider the heterogeneous demand case. We focus on the simplest consumer decision problem where consumers only decide when to stop consumption. Consumers start consuming as soon as they arrive and they never temporarily interrupt consumption. To simplify, we assume that consumers are risk neutral and do not discount the future.

The instantaneous spot price is defined by the same function $p(.)$ as before. We assume that $p(0) = 0$ and $p(Q) \geq v(Q/\epsilon_h)$. Lemma 2 shows that the later assumption rules out rationing and it is without loss of generality as will be argued later. We say that pricing scheme p is more responsive than pricing scheme p' if $p(q) < p'(q)$ for $q < Q$ and $p(Q) = p'(Q)$.

We look for rational expectations equilibrium: consumers maximize their expected future surplus given expected future prices, and the equilibrium price path is determined by consumer behavior and the pricing function $p(.)$. More specifically, let $e_t(\omega)$ represent the optimal stopping time for the consumer who arrived at time t. This decision is conditional on the information revealed up to time $e_t(\omega)$. To simplify notations, we denote $e^{-1}(t, \omega)$ the inverse of $e_t(\omega)$, that is, $e^{-1}(e_t(\omega), \omega) = t$. Consumer t chooses $e_t()$

[8] The assumption that the increments $\epsilon_t(\omega)dt$ are positive and bounded greatly simplifies the derivations because it guarantees that all equilibrium outcomes are bounded and continuous functions of time. Without the assumption $\epsilon_l > 0$ we would have to keep track of the periods when no consumers arrive/leave. Without the assumption that $\epsilon_h < \infty$ we would have to consider unbounded pricing functions. In both cases, the analysis follows.

to maximize

$$E_t\left(\int_0^{e_t(\omega)-t} v(x)dx - \int_t^{e_t(\omega)} p_s(\omega)ds\right)$$

where E_t is the expectation conditional on the information revealed up to time t. An equilibrium is a set $(e_t(\omega))$ such that for each t and ω, equilibrium occupancies are feasible, $q_t(\omega) \leq Q$, equilibrium prices are given by the pricing rule, $p_t(\omega) = p(q_t(\omega))$, and the termination decisions maximize consumers' utility.

4 Analysis

To get a flavour for our analysis, it helps to understand the first best allocation. Although we postpone a formal proof until subsection 4.2, the intuition for the first best allocation in the case of homogeneous demands is simple. Once store occupancy has reached capacity, it is efficient to share the capacity so that for every new consumer who arrives, the consumer who has been in the store the longest leaves. The rest of this paper studies whether responsive pricing can achieve efficiency. The analysis proceeds in three steps. First, we characterize the equilibrium when consumers have identical demands. Second, we investigate how efficiency changes in response to changes in the pricing function and then identify the properties that a pricing function must have to achieve efficiency. Finally, we consider the case where consumers have heterogeneous demands.

4.1 Equilibrium

We start by characterizing consumer behavior. The next lemma shows that consumers leave the store in a first-in, first-out fashion in any equilibrium.

Lemma 1 *In any equilibrium, $e_t(\omega)$ is increasing in t. Equilibrium store occupancy at time $e_t(\omega)$ is*

$$q_{e_t(\omega)}(\omega) = \int_t^{e_t(\omega)} \epsilon_s(\omega)ds, \tag{1}$$

Proof We show that the consumer who started to consume at $t' > t$ consumes whenever the consumer who started to consume at t consumes. To do that, let $U_t(s, e_t(\omega)) = E_s(\int_0^{e_t(\omega)-s} v(s-t+x)dx - \int_s^{e_t(\omega)} p_s(\omega)ds)$ represent the expected utility at time s of a

consumer who joined at $t < s$ and leaves at $e_t(\omega) \geq s$. If $e_t(\omega) > s$ then $U_t(s, e_t(\omega)) > 0$ and $U_{t'}(s, e_t(\omega))$ is also positive for any $t' > t$ since $U_t(.,.)$ is increasing in t. This proves the first claim of the lemma. The second claim follows immediately since all consumers who arrive between t and $e_t(\omega)$ consume at time $e_t(\omega)$ and only those consumers consume. □

All consumers who started to consume after a given consumer never leave the store before that consumer. This property implies that the optimal stopping time determines store occupancy. Next we show that there is always capacity available. To simplify notations, let $q_t(t', \omega) = \int_{t'}^{t} \epsilon_s(\omega) ds$ denote the mass of consumers who arrive between t' and t in state ω.

Lemma 2 *In any equilibrium,* $q_t(\omega) < Q$.

Proof The proof goes by contradiction. Assume that capacity is reached in state ω, and denote t^Q the first time when it is reached. The store has reached capacity if the cumulated arrival rate between $e^{-1}(t^Q, \omega)$ and t^Q is such that $q_{t^Q}(e^{-1}(t^Q, \omega), \omega) = Q$.[9] But $\int_{e^{-1}(t^Q,\omega)}^{t^Q} \epsilon_s(\omega) ds < (t^Q - e^{-1}(t^Q, \omega))\epsilon_h$ implies that $t^Q - e^{-1}(t^Q, \omega) > Q/\epsilon_h$ and this implies that $v(t^Q - e^{-1}(t^Q, \omega)) < v(Q/\epsilon_h) \leq p(Q)$ where the last inequality holds by assumption. But this is a contradiction, since consumer $e^{-1}(t^Q, \omega))$ gets negative instantaneous utility at t^Q and her utility can only decrease from then on, as Lemma 1 implies that no consumer will leave until she leaves. □

The assumption $p(Q) \geq v(Q/\epsilon_h)$ implies that the congestion charge is sufficiently high so that there is always some capacity available. If this assumption does not hold, however, consumers may have to be rationed in equilibrium. One would have to model both the rationing rule and consumers' decision to wait in line. This greatly complicates the exposition but adds no additional insight on the question of whether responsive pricing implements the efficient outcome. In fact, we will see that congestion never occurs under the efficient outcome. Therefore, a pricing scheme that lets congestion occur is a pricing

[9] If no consumer leaves the store at t^Q, $e^{-1}(t^Q, \omega)$ is defined as the consumer who has used the service more than any other consumer at time t^Q.

scheme that does not increase prices enough when occupancy reaches full occupancy. Such a scheme cannot be efficient. The next Lemma characterizes the equilibrium stopping rule $e_t()$.

Lemma 3 *In any equilibrium, the optimal stopping rule is a continuous and increasing function of t defined by*

$$v(\mathbf{e}_t(\omega) - t) = p(q_{\mathbf{e}_t(\omega)}(t, \omega)), \quad for\ all\ t \geq 0. \tag{2}$$

Proof To start, we show that $\mathbf{e}_t(\omega)$ is well-defined by 2. By the intermediate value theorem, the equation $v(x - t) = p(q_x(t, \omega))$ has a unique solution $x \in (t, t + Q/\epsilon_l]$ since v, p, and q are continuous, v is decreasing, p is increasing, $q_x(t, \omega)$ is increasing and unbounded in x and the boundary conditions are implied by $v(0) > p(0)$ and $v(Q/\epsilon_l) < v(Q/\epsilon_h) \leq p(Q) = p(q_{t+Q/\epsilon_l}(t, \omega))$.[10] By the implicit function theorem, there exist a continuously differentiable $\mathbf{e}_t(\omega)$ such that $v(\mathbf{e}_t(\omega) - t) = p(q_{\mathbf{e}_t(\omega)}(t, \omega))$. Differentiating 2 with respect to t,

$$\frac{d}{dt}\mathbf{e}_t(\omega) = \frac{p'(q_{\mathbf{e}_t(\omega)}(t, \omega))\epsilon_t - v'(\mathbf{e}_t(\omega) - t)}{p'(q_{\mathbf{e}_t(\omega)}(t, \omega))\epsilon_{\mathbf{e}_t(\omega)} - v'(\mathbf{e}_t(\omega) - t)} > 0.$$

The function $\mathbf{e}_t(\omega)$ defined by 2 is increasing.

Next, we show the stopping rule $\mathbf{e}_t(\omega)$ defined by 2 is an equilibrium. If all consumers follow stopping rule $\mathbf{e}_t(\omega)$, the price that will result is equal to $p_t(\omega) = v(t - \mathbf{e}^{-1}(t, \omega))$. We prove that given that price, it is optimal for consumers to follow stopping rule $\mathbf{e}_t(\omega)$. The proof goes by contradiction. Assume that consumer t stops consuming at $t' < \mathbf{e}_t(\omega)$. The price at t' is $p_{t'}(\omega) = v(t' - \mathbf{e}^{-1}(t', \omega))$. But $v(t' - t) > v(t' - \mathbf{e}^{-1}(t', \omega))$ implies that $v(t' - t) > p_{t'}(\omega)$. The consumer gets positive surplus. A contradiction. Assume that consumer t stops at $t' > \mathbf{e}_t(\omega)$. The price at t' is $p_{t'}(\omega) = v(t' - \mathbf{e}^{-1}(t', \omega))$. But $v(t' - t) < v(t' - \mathbf{e}^{-1}(t', \omega))$ implies that $v(t' - t) < p_{t'}(\omega)$. The consumer gets negative surplus and this surplus is decreasing from t' onward since $\mathbf{e}_t$ is increasing. A contradiction.

Finally, we show that $\mathbf{e}_t(\omega)$ defined by 2 is the only equilibrium. Assume that there exists another equilibrium stopping time $e'_t(\omega)$ with equilibrium price $p'_t(\omega)$. $e'_t(\omega)$ is an

[10] Without loss of generality, we can assume that $p(q) = p(Q)$ for $q \geq Q$.

increasing function of t. Consider the first t such that $e'_t(\omega) \neq \mathbf{e}_t(\omega)$. Assume $e'_t(\omega) < \mathbf{e}_t(\omega)$. $p'_{e'_t(\omega)}(\omega) \leq p_{e'_t(\omega)}(\omega)$ and $v(e'_t(\omega) - t) - p'_{e'_t(\omega)}(\omega) \geq v(e'_t(\omega) - t) - p_{e'_t(\omega)}(\omega) > 0$. A contradiction. Assume $e'_t(\omega) > \mathbf{e}_t(\omega)$. This implies that $p'_s(\omega) \geq p_s(\omega)$ for $s > \mathbf{e}_t(\omega)$. Consumer t gets negative utility from staying beyond $\mathbf{e}_t(\omega)$. A contradiction. □

To simplify notations, we denote the function $\mathbf{e}_t(\omega)$ defined by 2 simply $e_t(\omega)$. Consumers leave as soon as their willingness to pay for a unit of consumption falls below the price. For any $t > \tilde{t}(\omega)$, where $\tilde{t}(\omega)$ is such as $v(\tilde{t}(\omega)) = p(q_{\tilde{t}(\omega)}(0, \omega))$, the price is equal to the marginal valuation of the oldest consumer, that is, the consumer who is still in the store at time t and who has used the service more than any other consumer. The oldest consumer is indifferent between leaving the store and staying. One may argue that consumers should stay in the store even if they get negative instantaneous utility, if they expect that prices will decline fast enough such that expected future surpluses eventually outweigh short-term losses. This, however, cannot happen in equilibrium. In fact, equation 2 can be rewritten as $p_t(\omega) = v(t - e^{-1}(t, \omega))$, and this implies that the consumer who arrived at $e^{-1}(t, \omega)$ should not stay beyond t since for any $t' > t$, $v(t' - e^{-1}(t, \omega)) - p_{t'}(\omega) < v(t' - e^{-1}(t', \omega)) - p_{t'}(\omega) = 0$.
We can now characterize the equilibrium.

Proposition 1 *There is a unique rational expectations equilibrium $e_t(\omega)$ and it is characterized by (1,2).*

In equilibrium, the instantaneous price communicates all the information that is necessary for consumers to make optimal consumption decisions. The optimal dynamic consumption rule simplifies to a simple rule specifying that consumers will leave as soon as the instantaneous utility falls below the instantaneous price. An implication is that a consumer cannot benefit by acquiring superior information about the process $\epsilon_t(\omega)$. Although that consumer can predict future prices more accurately, s/he cannot benefit from this knowledge. A corollary is that the outcome under responsive pricing does not depend on consumers' information about $\epsilon_t(\omega)$.

4.2 Pricing Responsiveness, Market Clearing, and Efficiency

Now that the equilibrium is characterized, we show that market clearing can be achieved in a limit sense. Many classes of pricing schemes implement market clearing in the limit. Since our goal is to show only that this is possible, we focus on a very simple class. Define the class of schemes $p_\alpha(q)$ such that $p_\alpha(q) = 0$ for $q \leq Q - \alpha$ and $p_\alpha(q) = v(Q/\epsilon_h)(1 - \frac{Q}{\alpha} + \frac{q}{\alpha})$ otherwise.[11] This function is equal to zero up to $q - \alpha$ and then is linear with $p_\alpha(Q) = v(Q/\epsilon_h)$.

Proposition 2 *As α converges to 0, $q_{e_t^\alpha(\omega)}(t, \omega)$ converges to Q.*

Proof Define $e_t^\alpha(\omega)$ such that $v(e_t^\alpha(\omega) - t) = p_\alpha(q_{e_t^\alpha(\omega)}(t, \omega))$. Since $v > 0$, $Q - \alpha \leq q_{e_t^\alpha(\omega)}(t, \omega) \leq Q$ and $q_{e_t^\alpha(\omega)}(t, \omega)$ converges to Q as α converges to 0. □

This proposition says that responsive pricing achieves market clearing in the limit. This result is more general in the following sense. Let $p_t^*(\omega)$ represent the limit equilibrium price as α converges to zero and let $t^*(\omega)$ represent the first time when market clearing is reached $q_{t^*(\omega)}(0, \omega) = Q$. Responsive pricing implements one price schedule among the class of price schedules $p_t(\omega)$ that charge an instantaneous price for each unit of consumption. $p_t^*(\omega)$ is the only price schedule that achieves market clearing for $t \geq t^*$.[12] Next, we conduct comparative statics with respect to the pricing function $p(q)$ and analyse the efficiency properties of responsive pricing.

Proposition 3 *Consider two schemes p and p' such that p is more responsive than p'. Welfare under p is greater than under p'.*

Proof Let $e_t(\omega)$ and $e'_t(\omega)$ represent the optimal stopping time under pricing regime p and p'. $e_t(\omega) > e'_t(\omega)$. The proof goes by contradiction. Assume $e_t(\omega) \leq e'_t(\omega)$. Equilibrium

[11] There are several ways to define the limit of pricing scheme p_α. One can define the limit as a vertical line at Q. This pricing scheme, however, is not implementable in the sense that it does not identify a unique price when the store is at congestion. (Another way to define the limit is that the price is equal to zero for $q < Q$ and $p(Q) = v(Q/\epsilon_h)$. The interpretation of that pricing scheme is simple. The price is zero as long as congestion is not reached and a congestion charge is imposed once congestion is reached. The proofs presented in this work do not hold because the pricing function is discontinuous.) Furthermore, any class of continuous increasing functions such that $p_\alpha(Q) = v(Q/\epsilon_h)$ and $\int_0^Q p_\alpha(x)dx$ converges to zero will achieve market clearing.

[12] The proof follows the proof that the equilibrium stopping rule is unique.

condition (2) implies that

$$v(e_t(\omega) - t) = p(q_{e_t(\omega)}(t, \omega))$$

and similarly $v(e'_t(\omega) - t) = p'(q_{e'_t(\omega)}(t, \omega))$. But $e_t(\omega) \leq e'_t(\omega)$ implies that $v(e_t(\omega) - t) \geq v(e'_t(\omega) - t)$ so $p(q_{e_t(\omega)}(t, \omega)) \geq p'(q_{e'_t(\omega)}(t, \omega))$, $q_{e_t(\omega)}(t, \omega) > q_{e'_t(\omega)}(t, \omega)$, and $e_t(\omega) > e'_t(\omega)$. A contradiction. This implies that all consumers consume more under p than under p'. Efficiency is higher under p than under p'. □

Proposition 3 shows that more responsive pricing schemes achieve more efficient outcomes, but can the first best outcome be achieved? To answer this question we need to first characterize the efficient outcome.

Proposition 4 *In the efficient allocation, the consumer who arrives at t consumes from t until $e^*_t(\omega)$ such that*

$$q_{e^*_t(\omega)}(t, \omega) = Q.$$

Proof The efficient stopping time has to be increasing. Otherwise, one could increase welfare by swapping any two consumers who do not leave in a first-in, first-out fashion. Any stopping rule such that full occupancy is not reached for $t \geq t^*(\omega)$ is sub-efficient, since welfare could be increased by letting the last consumer who left since t consume more. It has to be true that $q_{e^*_t(\omega)}(t, \omega) = Q$ for all t. By the implicit function theorem, this equation defines a unique function $e^*_t(\omega)$. □

Efficiency occurs for any $t < t^*(\omega)$, when all consumers who arrived between 0 and t consume, and for any $t \geq t^*(\omega)$, when only those consumers who arrived between $t - e^{*,-1}(t, \omega)$ and t consume. Under that scheme, the store reaches capacity as early as possible and stays at capacity once it has reached it. In addition, new consumers always replace older consumers, who value the service less. We show that efficiency can be achieved in a limit sense.

Proposition 5 *As α converges to 0, $e^\alpha_t(\omega)$ converge to $e^*(\omega)$.*

Proof As α converges to 0, $q_{e^\alpha_t(\omega)}(t, \omega)$ converges to Q but since $\epsilon_s(\omega) > 0$ for all $s \geq 0$, this implies that $e^\alpha_t(\omega)$ converges to $e^*_t(\omega)$. □

In the limit case where the function p is close to a vertical line, there is almost no wasted capacity and responsive pricing approaches the efficient outcome. The price at date $e_t^*(\omega)$ converges to $v(e_t^*(\omega) - t)$ which corresponds to the valuation of the oldest consumer under the efficient outcome. If one considers the outcome at a point in time t, the allocation problem is similar to a Q units auction. Theoretically, efficiency may be reached using a multi-unit auction mechanism (e.g., simultaneous ascending non-discriminatory auction). Under such a mechanism, new consumers bid up to the point where the marginal consumers leave. The difference between such mechanisms and responsive pricing is that responsive pricing is much simpler. Under responsive pricing, consumers do not have to send bids. They have to decide only whether they want to consume. Consumption decisions determine the price, which in turn determines the equilibrium. The seller, in turn, only needs to be able to measure congestion and to update prices in real time.

The result on efficiency holds for a general class of arrival process, since we have not made any assumption on ϵ_t besides the support of the instantaneous increments. The increments could be time dependent or correlated over time. Our results apply equally for arrival processes with unexpected demand shocks and for processes with predictable demand shocks. The efficiency result, however, assumed that consumers had identical demands. We relax this assumption in the next two subsections.

4.3 Generalization to Heterogeneous Demands

We assume that there are I consumer types. Type $i = 1, .., I$ has inverse demand $v^i(n)$ where v is continuous and $v' < 0$. The arrival process is now a vector $\epsilon_t = (\epsilon_t^i)_{i=1..I}$, where $\epsilon_t^i dt$ represents the mass of type i consumers who arrive in interval dt and we assume as before that it is distributed over $[\epsilon_l^i, \epsilon_h^i]$. For now, we do not impose any further restrictions on the arrival process.

We say that the set of demands $(v^i)_{i=1..I}$ satisfies the no-crossing condition if for any (i, i') there do not exist $n, n', \delta \geq 0$ such that

$$v^{i'}(\delta + n) > v^i(n) \;\; and \;\; v^{i'}(\delta + n') < v^i(n').$$

The no-crossing condition has a clear economic interpretation. It says that no two con-

sumers who arrive at different points in time in the store can have residual demands that cross. This condition imposes a fairly strong restriction on the set of demands v^i.[13] In fact, we will see that it is equivalent to say that demands are horizontal shift of one another. The efficiency result generalizes when the v^i satisfy the no-crossing condition. To show that, we assume without loss of generality that $v^1(0) \geq v^2(0) \geq ... \geq v^I(0)$ and we define a^i such that $v^i(a^i) = v^{i+1}(0)$ and $A^i = a^1 + ... + a^i$ with $A^0 = 0$. For $\delta \geq 0$, define the function $i(\delta)$ such that $A^{i(\delta)-1} < \delta \leq A^{i(\delta)}$ and $i(\delta) = I$ if $\delta > A^I$. Define also the function $q_t(t', \omega) = \sum_{j=1}^{i(t-t')} \int_{t'+A^{j-1}}^{t} \epsilon_s^j(\omega) ds$.

Proposition 6 *Define $e_t^{1,*}(\omega)$ such that $q_{e_t^{1,*}(\omega)}(t, \omega) = Q$. In the efficient allocation, consumer of type i who arrives at t consumes if $e_{t-A^{i-1}}^{1,*}(\omega) \geq t$ from t until $e_t^{i,*}(\omega) = e_{t-A^{i-1}}^{1,*}(\omega)$.*

Proof By the implicit function theorem, $e_t^{1,*}(\omega)$ is well-defined and it follows that $e_t^{i,*}(\omega)$ is also well defined for $1 < i \leq I$. The allocation presented in the proposition never wastes capacity, so to demonstrate efficiency we need to show only that it is not possible to increase welfare by reallocating consumers. To prove that, we will show that no consumer leaving the store ever values consumption more than any consumer in the store.

Before proceeding, we need to establish a preliminary result. The no-crossing condition implies that $v^i(n) = v^1(A^{i-1} + n)$ for $i = 1...I$. The proof goes by contradiction. Assume that there exist $i \neq 1$ and n such that $v^i(n) \neq v^1(A^{i-1} + n)$. Assume for example that $v^i(n) > v^1(A^{i-1} + n)$. (The proof is similar if the inequality if reversed.) Then, by continuity $v^i(n) > v^1(A^{i-1} + n - \epsilon)$ for ϵ small. But $v^1(A^{i-1}) = v^i(0)$ implies that $v^1(A^{i-1} - \epsilon) > v^i(0)$. But these two inequalities contradict the assumption that v^1 and v^i satisfy the no-crossing condition.

Next, we show that any consumer in the store in period t values consumption more than $v^1(t - e^{1,*,-1}(t, \omega))$ where $e^{1,*,-1}$ is the inverse function of $e^{1,*}$, $e^{1,*,-1}(e_t^{1,*}(\omega), \omega) = t$. A consumer of type 1 who is still in the store in t had to arrive after $e^{1,*,-1}(t, \omega)$. That consumer values consumption more than $v^1(t - e^{1,*,-1}(t, \omega))$. Consider a consumer of type

[13] An example of a class of demands that satisfy the no-crossing condition is the class $v^i(n) = a^i - bn$ where a^i are positive numbers.

$i \geq 1$. Any consumer of type i had to arrive at $e^{1,*,-1}(t,\omega) + A^{i-1}$ or after. The lowest valuation among type i consumers is $v^i(t - (e^{1,*,-1}(t,\omega) + A^{i-1})) = v^1(t - e^{1,*,-1}(t,\omega))$. To conclude, we show that any consumer out of the store in period t values consumption $v^1(t - e^{1,*,-1}(t,\omega))$ or less. Consider a consumer of type $i \geq 1$. That consumer had to arrive at $e^{1,*,-1}(t,\omega) + A^{i-1}$ or before. The highest valuation among these consumers is $v^i(t - (e^{1,*,-1}(t,\omega) + A^{i-1})) = v^1(t - e^{1,*,-1}(t,\omega))$. $\square$

Under the no-crossing condition, the efficient allocation changes slightly. For any $t \geq t^*(\omega)$, where $q_{t^*(\omega)}(0,\omega) = Q$, the consumers with the lowest demands in the store are replaced by new consumers, starting with those consumers with highest demands up to the point where no new consumer values consumption more than the marginal consumer in the store.

Similarly, the derivation of the equilibrium still holds after straightforward generalizations. Lemma 1 must take into account the fact that the stopping rule for consumers of type one, call it $e^1_t(\omega)$, will determine the stopping rule for all other types,

$$e^i_t(\omega) = e^1_{t-A^{i-1}}(\omega).$$

Although consumers of a same type leave in a first-in, first-out fashion, consumers of different types may not do so. For example, a consumer of type i who arrived at t will leave before a consumer of type $i-1$ who arrived between $t - a_i$ and t. Lemma 3 says that the equilibrium stopping rule is

$$v(e^1_t(\omega) - t) = p(q_{e^1_t(\omega)}(t,\omega)).$$

The equilibrium occupancy is

$$q_t = q_t(e^{1,-1}(t,\omega),\omega).$$

Lemma 2 holds as long as $p(Q) \geq v^1(Q/\epsilon_h)$ where $\epsilon^h = \sum_i \epsilon^i_h$. As before, this condition rules out congestion. Propositions 1-5 follow, and the equilibrium real-time price at time t converges to $v^1(t - e^{1,*,-1}(t,\omega))$ as α converges to 0, so that efficiency is again achieved in a limit sense.

Although the no-crossing condition is restrictive, the results would still hold under more general demands if one were willing to impose some restrictions on the arrival process ϵ_t. Stated loosely, the main message of this section is that the results generalize as long as no two consumers who can overlap in the store have residual demands that cross over the length of time over which they overlap. For example, the demand of two consumers who never overlap could intercept. Similarly, the demand of two consumers could intercept after one has left the store. To formalize this idea, define n^- as the maximum possible length of stay in any equilibrium.[14] The analysis follows as long as the (weaker) no-crossing condition holds: for any $\epsilon_t^i > 0$, there does not exist $\epsilon_{t'}^{i'} > 0$ with $0 \leq t' - t \leq n^-$ such that $v^{i'}(n) > v^i(n + (t' - t))$ and $v^{i'}(n') < v^i(n' + (t' - t))$ for $n, n' \leq n^- - (t' - t)$. Under this condition, consumers can be ordered in the sense that if consumer i values consumption more consumer j at time t, then i values consumption more than j at any $t' > t$ such that both consumers are in the store. The analysis follows and the main results hold. This more general interpretation of the no-crossing condition is important because the analysis does not always hold when this condition is not met, as we show in the next section.

4.4 An Example of Inefficiency

The following example illustrates that inefficiencies can occur when the no-crossing condition does not hold. (See also Table 1.) There are only two periods $t \in [0, 2]$ and we will use the terminology period one (two) to mean $t \in [0, 1]$ ($[1, 2]$).[15] The store capacity is $2 + \beta$ where β is an arbitrarily small positive number that will be used to nail down the equilibrium price under responsive pricing. To simplify the exposition, we consider step-function demands. A demand is a pair of numbers. A consumer with demand (a, b) who arrives at t, for example, is willing to pay a from t to $t + 1$ and b from $t + 1$ to $t + 2$ and nothing after $t + 2$. There are three types of demands $v^1 = (20, 20)$, $v^2 = (25, 0)$, and

[14] An upper bound for n^- is Q/ϵ_l where $\epsilon_l = \sum_i \epsilon_l^i$.

[15] Although the arrival rate in the example does not satisfy the assumptions of the model since it has zero arrival rates ($\epsilon_t = 0$ for $t \neq 0$ and $t \neq 1$) and positive mass at $t = 0$ and $t = 1$, this is not what causes inefficiencies. In fact, one could extend the results for arrival increments that take zero and positive mass values but this would greatly complicate the exposition as explained in footnote 8. As we will demonstrate later, inefficiencies occur in the example because the no-crossing condition is violated.

$v^3 = (25, 25)$.

The arrival process is the following. Consumers arrive only at $t = 0$ or $t = 1$. At $t = 0$, there are two possible states of the world, state π and state $1 - \pi$, which occur with respective probability π and $1 - \pi$ with $\pi \in [0, 1]$. In state π the arrival realization at date 0 is $\epsilon_0^{\pi} = (1, 2, 0)$ while in state $1 - \pi$ the arrival realization is $\epsilon_0^{1-\pi} = (1, 1, 1)$. At date one, the arrival realization is $\epsilon_1 = (0, 1, 0)$. Arrival realization ϵ_1, for example, means that one unit of consumer of type v^2 joins the store at date 1. We will denote v_t^i the consumer of demand type $i = 1, 2, 3$ who arrived at date $t = 0, 1$.

Table 1: Consumer Preferences

	State π		State $1 - \pi$	
Type	$t = 0$	$t = 1$	$t = 0$	$t = 1$
$v^1 = (20, 20)$	1	0	1	0
$v^2 = (25, 0)$	2	1	1	1
$v^3 = (25, 25)$	0	0	1	0

The efficient allocation maximizes total surplus subject to feasibility constraints and subject to the constraint that the allocation rule at time t is conditional on the information revealed up to that time. In state π, all consumers v_0^1 should consume in both periods, $1+\beta$ consumers v_0^2 should consume in period one, and all consumers v_1^2 should consume in period two. In state $1 - \pi$, all consumers v_0^2 should consume in period one, all consumers v_0^3 and β consumers v_0^1 should consume in both periods, and all consumers v_1^2 should consume in period two. The expected consumer surplus in the first-best allocation is $\pi(90 + \beta 25) + (1 - \pi)(100 + \beta 40)$.

Consider next responsive pricing. The information structure is common knowledge but consumers privately know their types. Consider scheme p_α defined in subsection 4.2 where $\alpha < \beta$ and $p(Q) > 25$. To understand the construction of the equilibrium, note first that prices will change only at $t = 0$ and $t = 1$, since these are the only two dates when either new consumers arrive or when consumers' willingness to pay change.[16] Next, consider consumers' consumption decisions. Consumers v_0^3 will consume in state $1 - \pi$ because their demand (weakly) dominates any other consumer. Consumers v_0^2's

[16]It is easy to show that it cannot be optimal for a consumer to leave at any time $t \in (0, 1)$ or $t \in (1, 2)$.

consumption decision is also simple. They are willing to pay 25 and no more than 25 at date 0. Solving the decision problem of consumers v_0^1 is more tricky. How much are consumers v_0^1 willing to pay at date 0? This decision depends on their expectations about the second period price. Assume that all consumers v_0^1 decide to stay. In state π, the demand at $t = 1$ is composed of one unit of v_0^1 who is willing to pay at most 20 and one unit of v_1^2 who is willing to pay at most 25. The capacity is $2 + \beta$ and since $\alpha < \beta$, the responsive price is 0 during period 2. In state $1 - \pi$, v_0^3 have joined at date 0 and all these consumers have to be in the store at date 1, since it has to be optimal for them to stay if it is so for v_0^1. At $t = 1$, two units of consumers are willing to pay 25 and one unit of consumers is willing to pay 20. The equilibrium price is 20 during period 2. To summarize, a consumer v_0^1 should expect at date 1 a surplus of $20 - 0$ with probability π and $20 - 20$ with probability $1 - \pi$. Consumers v_0^1 are willing to pay $20 + \pi 20 + (1 - \pi)0 = (1 + \pi)20$ at $t = 0$. Since $\pi > 0$ consumers v_0^1 are willing to pay more than their period 1 valuation. When $(1 + \pi)20 > 25$, the equilibrium price is 25 for $t \in [0, 1]$ and all consumers v_0^1 stay in the store while only β consumers v_0^2 stay. When $(1 + \pi)20 < 25$, the price is $(1 + \pi)20$ for $t \in [0, 1]$ and all consumers v_0^2 stay in the store while only β consumers v_0^1 stay. The allocation is independent of α as long as $0 < \alpha < \beta$.

An inefficiency occurs because consumer v_0^1's decision to stay in the store does not depend on the state of the world as it should under the first best outcome. Because consumers v_0^1 do not know the composition of consumers in the store, they are willing to pay too little (respectively much) when they believe that it is very likely that consumers v_0^3 (respectively v_0^2) are in the store. They end up over- (respectively under-) consuming with probability $1 - \pi$ (respectively π). The efficiency loss is $\pi(1 - \beta)15$ when $(1 + \pi)20 < 25$ and $(1 - \pi)(1 - \beta)5$ when $(1 + \pi)20 > 25$. The realized price does not reveal sufficient information for consumers v_0^1 to make efficient decisions at $t = 0$. To achieve efficiency, consumer v_0^1 would need to know whether consumers v_0^2 or consumers v_0^3 have arrived in the store at $t = 0$. This information, however, is not revealed by the price.

The problem identified in this example is general and can be summarized as follows. The no-crossing condition does not hold for consumer v_0^1 and v_0^2. It is not optimal for consumer v_0^1 to leave when the price is equal to her instantaneous valuation 20. More

generally, a consumer with high long-term demand may prefer to stay and bear negative instantaneous utility if she believes that (a) there are some consumers with weak long-term demands who are about to leave, and (b) few consumers are likely to arrive.[17] If this is the case, prices are likely to collapse and it may be optimal to wait. But because the consumer makes this decision without knowing all the information that is available at the time the decision is made, she will sometimes make inefficient (although privately optimal) decisions.

Consumers' decision problems differ dramatically when the no-crossing condition holds and when it doesn't. Under no-crossing, consumers need to know only the current price to decide whether to stay or leave. They do not need to know, or to guess, the realization of ϵ_t. The fact that consumers do not know who is in the store when they arrive does not prevent efficiency from being achieved. That is, incomplete information is not a problem because the current price contains all the information consumers need to know to make efficient decisions. When the no-crossing condition does not hold, however, consumers do not decide when to leave only on the basis of the current price. Consumers use past realizations of prices to update their beliefs about future prices.

One way to interpret our result is using a general equilibrium framework. If markets were open for consumption in all future dates and if consumers could continuously trade in these markets, then the competitive equilibrium that would result would be efficient. In our applications, however, such trading is not feasible because by definition consumers only find out at the last minute whether they want access so opening future markets in the absence of those consumers who have not yet requested access would be meaningless.

5 Extensions

We return to some of key assumptions of the model and discuss how the main results on efficiency generalize.

[17] The problem identified in the example does not rest on the assumption that one type of consumers is imperfectly informed. Imperfect information alone does not imply that there will be inefficiencies. Recall that consumers were equally imperfectly informed in the previous section and that this did not prevent efficiency. One needs in addition to imperfect information that the no-crossing condition fails to hold for inefficiencies to arise.

5.1 Delaying Consumption

The analysis assumes that consumers never postpone consumption. They start consuming as soon as they arrive and never temporarily interrupt consumption. This assumption is valid, for example, if there is an opportunity cost of waiting and if this opportunity cost is high enough relative to the benefit from waiting which corresponds to the expected savings from lower prices. A sufficient condition is that the per-unit-of-time opportunity cost of waiting is greater than $v(Q/\epsilon_h) - v(Q/\epsilon_l)$. This implies that it is never efficient for consumers to wait and the analysis follows.

When the cost of postponing consumption is low, however, consumer waiting may occur both under responsive pricing and in the first best allocation. To make this point clear, consider the extreme case where the opportunity cost of waiting is zero. Under responsive pricing, consumers will prefer to delay consumption if they anticipate that prices are likely to decrease in the future. But it is not efficient anymore that a consumer leave the store for every new consumer who joins the store, since there is no welfare cost associated with consumers waiting. More generally, even when consumers have a low but positive cost of waiting, it is not efficient anymore to rule out waiting, since there is a trade-off between the welfare cost of waiting and the opportunity cost of cutting off some consumers and asking them to leave.[18]

5.2 Endogenous Arrival Rate

To simplify the analysis, we assumed that consumers' decisions to request access were independent of the pricing scheme. In general, however, it is possible that some consumers' decisions to request access will depend on the pricing scheme, since this determines the equilibrium distribution of price and therefore the expected surplus from consumption. We present a very simple extension of the model that suggests that responsive pricing actually gives an efficient incentive to request access.

[18]Positive but low cost of waiting may explain why country clubs and ski resorts do not use prices to allocate capacity although waiting is often observed in equilibrium. In these situations, consumers may have a low cost of waiting and it would be suboptimal to cut some consumers short to free up capacity when there is a sudden arrival flow of new consumers. This conclusion is consistent with the analysis of ski lifts presented in Barro and Romer (1987).

To keep the point as simple as possible, we return to the simplest case where all consumers have the same demand and where the arrival flow is constant over time. The arrival flow is composed of two types of consumers. As before $\tilde{\epsilon}dt$, where $\tilde{\epsilon}$ is a random variable, impulse consumers join the store independently of the pricing curve. In addition to impulse consumers, planners join the store only if they receive an expected utility that is greater than their outside utility. To keep the point simple, we assume that impulse and planner have the same demand but the argument generalizes to heterogeneous demands. $F(U)dt$ planners join the store per unit of time if the expected utility from joining the store is U with $F' > 0$. Planners do not know the realization of $\tilde{\epsilon}$ when they decide whether to join. Given the total arrival flow $(\tilde{\epsilon} + F(U))dt$ occupancy is

$$q = (F(U) + \tilde{\epsilon})n$$

where n is the equilibrium length of stay. For expositional reasons, we solve for the equilibrium length of stay, rather than equilibrium occupancy, keeping in mind that the two are uniquely related by the above relation. The equilibrium price is $p(q) = v(n)$ and after replacing q

$$p((F(U) + \tilde{\epsilon})n) = v(n) \tag{3}$$

Planners' expected utility from joining the store is

$$U = E_{\tilde{\epsilon}} \left(\int_0^{n(\tilde{\epsilon})} v(x)dx - p((F(U) + \tilde{\epsilon})n(\tilde{\epsilon})) \right). \tag{4}$$

An equilibrium is a $(n(\tilde{\epsilon}), U)$ that solves 3 and 4. The equilibrium exists and is unique. To see that, define the implicit function $n(\tilde{\epsilon}, U)$ from 3. This function determines the equilibrium length of stay for each realization of $\tilde{\epsilon}$ given U. The equilibrium U exists and is unique since $n(\tilde{\epsilon}, U)$ is decreasing in U so that the right-hand side in 4 is decreasing in U.

To compute the first best allocation, we consider an information structure that maps the information structure under responsive pricing. The social planner first decides who joins the store without knowing the realization of $\tilde{\epsilon}$ and then decides how much consumers should consume after the realization of $\tilde{\epsilon}$ is observed. The allocation $(n^*(\tilde{\epsilon}), U^*)$ such that

$n^*(\tilde{\epsilon})\tilde{\epsilon} = Q$ and 4 holds is efficient because $n^*(\tilde{\epsilon})$ is efficient conditional on the threshold U and the threshold U^* is efficient conditional on the allocation rule.

To show that responsive pricing achieves the efficient allocation, consider the class of pricing schemes p_α defined in subsection 4.2. The equilibrium (n^α, U^α) defined by 3 and 4 converges to the efficient allocation $(n^*(\tilde{\epsilon}), U^*)$. Although very simplistic this illustration suggests that real time pricing gives not only efficient ex-post consumption incentives but also efficient ex-ante incentives to request access.

6 Summary and Conclusions

This paper investigates the efficiency properties of responsive pricing, a simple and easily implementable scheme initially proposed by Vickrey to eliminate inefficiencies that result from last minute demand shocks. Responsive pricing changes prices in real time in response to demand realizations, increasing prices when the resource gets close to congestion and decreasing prices when unused capacity increases, thus promoting market clearing.

We show that responsive pricing implements the efficient outcome in a limit sense when consumer demands satisfy a no-crossing condition. When this condition holds, the equilibrium characterization dramatically simplifies because consumers stop consuming as soon as their willingness to pay for a marginal unit falls below the instantaneous price. However, when this condition is violated the analysis does not follow, and responsive pricing sometimes fails to achieve efficiency. The problem with responsive pricing is that consumers can bid only for the current unit of consumption, and the equilibrium price does not always aggregate consumers' private information efficiently. An implication for policymaking is that responsive pricing will work well when consumer demands satisfy the no-crossing condition, such as among homogenous populations of consumers.

One could easily conceive more sophisticated information revelation schemes than responsive pricing. We believe, however, that one should focus on simple schemes, such as the one proposed by Vickrey and considered in this work, because such schemes are more likely to be used in applications where highly unpredictable last-minute demand shocks play an important role. If one accepts this view, a relevant question for future research is

to generalize the class pricing mechanisms, possibly incorporating more information than just current utilization rates, such that the efficient outcome can be implemented.

References

1. Barro, Robert and Paul Romer. "Ski-Lift Pricing, with Applications to Labor and Other Markets." *American Economic Review*, 77, 5, (1987): 875-890.

2. Boiteux, M. "Sur la Gestion des Monopoles Publics Astreints a l'Equilibre Budgetaire." *Econometrica*, 24, 1, (1956): 22-40.

3. Boiteux, M. "Peak Load Pricing", *Journal of Business*, 33, (1960):157-179.

4. Borenstein, Severin. "Frequently Asked Questions about Implementing Real-Time Electricity Pricing in California for Summer 2001." Mimeo, University of Berkeley, 2000.

5. Crew, Michael A., Chitru S. Fernando, and Paul R. Kleindorfer. "The Theory of Peak-Load Pricing: A Survey." *Journal of Regulatory Economics*, 8, 3 (November 1995): 215-48.

6. Courty, Pascal and Mario Pagliero. "Estimating the Welfare Gains From Real Time Pricing: Evidence from an Internet café." Mimeo, London Business School, 2003.

7. MacKie-Mason, Jeffrey K. and Hal R. Varian. "Some Economics of the Internet. Mimeo, University of Michigan, 1994.

8. Vickrey, William. "Responsive Pricing of Public Utility Services." *The Bell Journal of Economics and Management Science*, 1, 2, (1971):337-346.

Figure 1: The 2-States Steady State Case

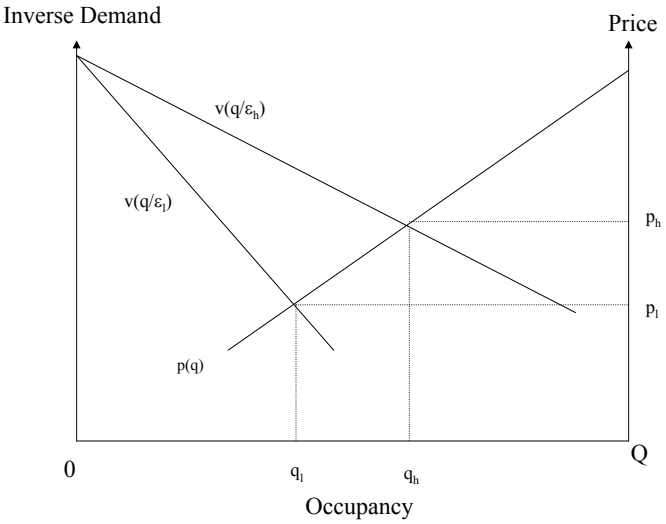

Figure 2: Increase in Responsiveness