NBER WORKING PAPER SERIES

ON THE EFFICIENCY OF COMPETITIVE ELECTRICITY MARKETS WITH TIME-INVARIANT RETAIL PRICES

Severin Borenstein
Stephen P. Holland

Working Paper 9922
http://www.nber.org/papers/w9922

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2003

Borenstein: Director of the University of California Energy Institute (www.ucei.org) and E.T. Grether Professor of Business Administration and Public Policy at the Haas School of Business, U.C. Berkeley (www.haas.berkeley.edu). Email: borenste@haas.berkeley.edu. Holland: Visiting Researcher, University of California Energy Institute. Email: sholland@uclink.berkeley.edu. For helpful comments and discussions, we thank Jim Bushnell, Joe Farrell, Morten Hviid, Erin Mansur, Michael Riordan, Lawrence White and seminar participants at UC Berkeley, the UC Energy Institute, Columbia University/ NYU, the Econometric Society Summer Meetings, the International IO Conference, and the CRRI Western Conference. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices
Severin Borenstein and Stephen P. Holland
NBER Working Paper No. 9922
August 2003
JEL No. L9, L8, L5

ABSTRACT

The standard economic model of efficient competitive markets relies on the ability of sellers to charge prices that vary as their costs change. Yet, there is no restructured electricity market in which most retail customers can be charged realtime prices (RTP), prices that can change as frequently as wholesale costs. We analyze the impact of having some share of customers on time-invariant pricing in competitive electricity markets. Not only does time-invariant pricing in competitive markets lead to outcomes (prices and investment) that are not first-best, it even fails to achieve the second-best optimum given the constraint of time-invariant pricing. We then show that attempts to correct the level of investment through taxes or subsidies on electricity or capacity are unlikely to succeed, because these interventions create new inefficiencies. In contrast, increasing the share of customers on RTP is likely to improve efficiency, though surprisingly, it does not necessarily reduce capacity investment, and it is likely to harm customers that are already on RTP.

Severin Borenstein
Haas School of Business
University of California
Berkeley, CA 94720-1900
and NBER
borenste@haas.berkeley.edu

Stephen P. Holland
The Bryan School of Business and Economics
University of North Carolina at Greensboro
Visiting Researcher
University of California Energy Institute
sholland@uclink.berkeley.edu

In many industries, retail prices do not adjust quickly to changes in costs or market conditions. Restaurants keep stable menu prices even when ingredient prices fluctuate. Service providers, from house cleaners to veterinarians, regulate fluctuating demand with non-price mechanisms (usually queuing) rather than by adjusting price to clear the market in times of excess demand.

Perhaps nowhere is the disconnect between retail pricing and wholesale costs so great as in restructured electricity markets. In the last decade, it has become apparent that wholesale electricity price fluctuations can be extreme, but retail prices have in nearly all cases been adjusted only very gradually. Typically, wholesale electricity prices vary hour by hour, while retail prices are adjusted two or three times per year. Because electricity is not economically storable and fixed retail prices create price-inelastic wholesale demand, it is not uncommon for wholesale prices within one day to vary by 100% or more while retail prices do not adjust at all.

Economists, recognizing the potential inefficiencies when prices do not reflect incremental production or wholesale acquisition costs, have been among the most vocal proponents of realtime pricing (RTP) of electricity, under which retail prices can change very frequently, usually hourly. With the 2000-01 California electricity crisis, many market participants also expressed support for more responsive retail prices. RTP has been explored in economics in what is commonly referred to as the peak-load pricing literature.[2] That literature, however, has focused almost entirely on time-varying pricing in a regulated market. Much of what is known from that literature carries over immediately to a deregulated market if *all* customers are on RTP, but that situation is unlikely to occur in any electricity system in the near future.

While many deregulated (and some regulated) electricity markets are considering implementing RTP for some customers, nowhere is RTP likely to encompass all, or even most, of the retail demand. In all cases, the outcome is likely to be a hybrid in which some customers see realtime prices and others see time-invariant prices, more commonly called flat-rate service. In this paper, we examine such a structure under deregulation, where competitive generation markets develop time-varying wholesale prices, but competitive retail sellers still charge some customers flat retail rates.[3]

---

[2] See Steiner (1957), Boiteaux (1960), Wenders (1976), Panzar (1976), Williamson (1966), Williamson (1974) and Bergstrom and MacKie-Mason (1991). For a survey of the literature on peak-load pricing see Crew, Chitru and Kleindorfer (1995).

[3] During the debate over electricity restructuring and in the aftermath of the California electricity crisis, analysis has focused on market power and on the design of efficient auction mechanisms. We

Closely tied to time-invariant retail pricing is the issue of investment adequacy. Many participants in the electricity industry have argued, generally without much economic explanation, that deregulated electricity markets will result in inadequate investment in production capacity. While this clearly is not the case with peak-load pricing under regulation—as explained by the earlier literature—and similarly does not result from a model of competitive electricity markets in which all customers are on RTP, we show that capacity investment is not efficient in competitive markets when some customers are on flat retail rates. Not only is the level of investment not the first-best level that results when all customers are on RTP, it is not even the second-best optimal level of capacity investment *given the constraint that some customers cannot be charged realtime prices.*

Those who have argued that capacity investment will be suboptimal under deregulation have generally then advocated for capacity subsidies in order to support greater capacity investment. We analyze a number of possible proposals for capacity subsidies and demonstrate that the very limited cases might be able to overcome the inefficiency caused by suboptimal investment.

We then analyze the impact of expanding the use of RTP. We show that if customers have homogenous demand patterns, expansion of RTP actually harms customers who are already on RTP, but benefits customers who remain on flat rates. We demonstrate that incremental changes in the use of RTP have impacts on the efficiency of the market that are not captured by those changing to RTP, an externality that implies the incentive to switch to RTP will not in general be optimal. We also show, surprisingly, that increasing use of RTP will not necessarily reduce the equilibrium amount of installed generation capacity.

We focus in this paper on the electricity industry, but the results have implications well beyond electricity. Due to technologies or institutions, retail prices in many markets are smoothed representations of underlying wholesale costs. Our results demonstrate that this sort of pricing has significant implications for capital investment and long-run efficiency, particularly in service industries and others markets with little or no ability to carry inventories.[4]

We begin in section I by presenting a model of competitive wholesale and retail electricity markets in which some share of customers is able to be charged realtime electricity prices. We demonstrate the short-run pricing and long-run investment inefficiency that

---

analyze the efficiency of the competitive markets in the absence of these other potential distortions.

[4] An important attribute of electricity that is not present in most other industries is the potentially extremely high costs of using non-price methods to accommodate a shortage of the product.

results from the inability to charge all customers realtime prices. In section II, we explore the possible use of subsidies or taxes to overcome the inefficiency from such "inaccurate" retail pricing. In section III, we examine the welfare effects of changing the proportion of customers on RTP and the customer's incentives to switch to RTP. We conclude in section IV.

I. Competition in wholesale and retail electricity markets

In deregulated electricity markets, wholesale prices are envisioned to result from competition among generators, and retail prices would result from competition among retail service providers serving the final customers. To understand these competitive interactions, consider the following simple model of electricity markets.

Since electricity cannot be stored economically, demand must equal supply at all times. Assume there are T periods per day with retail demand in period t given by $D_t(p)$ where $D_t' < 0$.[5,6] A fraction, α, of the customers pay realtime prices, *i.e.,* retail prices that vary hour to hour. The remaining fraction of customers, $1-\alpha$, pay a flat retail price $\bar{p}$. We assume that $\alpha \in (0,1]$ is exogenous and that customers on realtime pricing do not differ systematically from those on flat-rate pricing. Aggregate (wholesale) demand from the customers is then $\tilde{D}_t(p,\bar{p}) = \alpha D_t(p) + (1-\alpha)D_t(\bar{p})$ which implies that $\tilde{D}_t$ is decreasing in $\bar{p}$ and p. Note that $\tilde{D}_t(\bar{p},\bar{p}) = D_t(\bar{p})$. For $p > \bar{p}$, the flat-rate customers do not decrease consumption in response to the higher realtime price so $\tilde{D}_t(p,\bar{p}) > D_t(p)$, and for $p < \bar{p}$, the flat-rate customers do not increase consumption in response to the lower realtime price so $\tilde{D}_t(p,\bar{p}) < D_t(p)$. Finally, $\tilde{D}_t(p,\bar{p})$ is decreasing in α for $p > \bar{p}$, and $\tilde{D}_t(p,\bar{p})$ is increasing in α for $p < \bar{p}$. That is, increasing alpha increases the elasticity of wholesale demand by rotating $\tilde{D}_t$ around the point $(D_t(\bar{p}),\bar{p})$.

Figure 1 illustrates the demand curves if everyone were on RTP, D_t, and the wholesale demand curves with $1-\alpha$ share on flat rate service, $\tilde{D}_t$, where there are only two periods: peak, p, and off-peak, op. Note that the less elastic curves are the wholesale demand of the realtime and flat-rate customers. For prices above $\bar{p}$, wholesale quantity demanded is greater than the quantity demanded if everyone were on realtime prices since the flat-rate customers do not decrease consumption in response to the higher realtime price. Similarly,

---

[5] Following the literature on peak-load pricing, we also assume that cross price elasticities between demands in different periods are zero. Bergstrom and MacKie-Mason (1991) allow non-zero cross price elasticities, but assume homothetic preferences across hours.

[6] Analysis of stochastic demand in competitive markets is beyond the scope of this paper. For analysis of peak-load pricing with stochastic demand see Carlton (1977), Panzar and Sibley (1978), and Chao (1983).

for prices below $\bar{p}$, wholesale quantity demanded is less than the quantity demanded if everyone were on realtime prices since the flat-rate customers do not increase consumption in response to the lower realtime price.

Generators install capacity and sell electricity in the wholesale market. Assume that each generator is small relative to the market and has access to identical technology. Assume marginal costs of generation, which depend on the installed capacity, are continuous and are increasing. Since marginal costs are increasing and each generator has the identical technology, industry costs are minimized when production from each generator is identical. Let $C(q, K)$ be the short-run industry cost of generating q units of electricity given that K units of capacity are installed. Assume that the partial derivatives, C_q and C_k, are continuous and that
(a) $C_q > 0$, increasing generation output increases costs;
(b) $C_k < 0$, generating a given quantity of electricity is cheaper with more installed capacity;
(c) $C_{qq} > 0$, short-run marginal costs are increasing in quantity;
(d) $C_{kk} > 0$, the reduction in short-run generation costs from installing additional capacity is smaller at higher levels of installed capacity, *i.e.,* $-C_k$ is downward sloping in K; and
(e) $C_{qk} < 0$, additional investment reduces the marginal cost of generating.

Profit maximization implies that each firm would equate its short-run marginal cost of generation with the wholesale price, *i.e.,* $w = C_q(q, K)$ where w_t is the wholesale price in period t. Thus, the short-run industry supply curve is upward sloping.[7] Figure 2 illustrates demand curves for six different time periods and two short-run industry supply curves for capacities K and K' where $K < K'$. Market clearing prices for each time period are given by the intersection of the demand curves with the relevant short-run supply curve. In the long-run, investment in capacity from K to K' lowers the marginal cost of generation and lowers the market clearing price in each period.

In the long run, generators can add or retire capacity. Assume that the cost per unit of capacity is r per day. If q_t MW of electricity is generated in period t, industry profits for the generators are $\sum_{t=1}^{T}[w_t q_t - C(q_t, K)] - rK$ per day. Since each firm has identical technology and generates the same amount per unit of capacity, firm profit is simply a fraction of industry profit.

---

[7] The profit maximization condition $w_t = C_q(q, K)$ can be inverted to derive an industry supply curve. The assumption of identical technologies implies that the industry supply curve is proportional to the supply from a single unit, *i.e.,* industry supply can be written $KS(w)$ where $S(w)$ is the unit supply curve. We will occasionally use this equivalent characterization of the generation technology.

The retail sector purchases electricity from generators in the wholesale market and distributes it to the final customers. Firms in the retail sector are assumed to have no costs other than the wholesale cost of the electricity that they buy for their retail customers.[8] The retail firms choose realtime retail prices, p_t, and the flat retail rate, $\bar{p}$, engaging in Bertrand competition over these prices. Bertrand competition represents accurately the competition among retail electricity providers, because they would be price takers in the wholesale market, would be selling a nearly homogeneous product in the retail market, and would face no real capacity constraints. Profit of the retail sector is given by $\sum_{t=1}^{T}(\bar{p} - w_t)(1-\alpha)D_t(\bar{p}) + (p_t - w_t)\alpha D_t(p_t)$ per day. Since electricity cannot be stored economically, demand greater than capacity in any period would require non-price rationing. The flat retail price, $\bar{p}$, is *feasible* if there exists some p_t such that $C_q(\tilde{D}_t(p_t, \bar{p}), K) < \infty$ for all t, *i.e.,* if the marginal cost of producing the quantity demanded is finite. In other words, $\bar{p}$ is feasible if enough customers are on RTP to allow the wholesale market to clear at some finite price.

A. Competitive equilibrium in wholesale and retail markets

Equilibrium prices in the retail sector are determined by competition among retailers. First, consider the customers on RTP. If a realtime price, p_t, were greater than the wholesale price, a competing retailer could make profits by undercutting p_t and attracting more customers. Since charging a price less than w_t would imply losses, the equilibrium short-run retail realtime price is $p^e_{t_{SR}} = w_t$ for every t. In other words, competition among retailers drives retail prices for RTP customers to be equal to wholesale prices in each period.

Similarly, competition forces the flat retail rate to be set to cover exactly the cost of providing electricity to the flat-rate customers. Since this implies zero profits for the retail sector, the condition $\sum_{t=1}^{T}(\bar{p}^e_{SR} - w_t)(1-\alpha)D_t(\bar{p}^e_{SR}) = 0$ determines the short-run equilibrium flat retail price $\bar{p}^e_{SR}$. Note that this zero profit condition can be written $\bar{p}^e_{SR} = \sum_{t=1}^{T} w_t D_t(\bar{p}^e_{SR}) / \sum_{t=1}^{T} D_t(\bar{p}^e_{SR})$. In other words, the equilibrium flat retail price is a weighted average of the realtime wholesale (and retail) prices where the weights are the relative quantities demanded by the customers facing a flat retail price. Thus, competition among retailers drives $\bar{p}^e_{SR}$ to be equal to the demand-weighted average wholesale price.[9]

---

[8] Extending the analysis to include retailer costs of billing or distribution does not alter the analysis in any significant way.

[9] Existence of the equilibrium can be shown since (i) retail profits are continuous in $\bar{p}$, (ii) retail profits are negative for $\bar{p} = c$, and (iii) retail profits are positive if $\bar{p}$ is equal to the highest wholesale price that occurs during the time period.

In the short run, equilibrium prices in the wholesale market are determined by the intersection of the demand curve and the short-run supply curve in each period. Since generators equate the marginal cost of generation with the wholesale price in every period, supply equals demand when $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$.[10] The short-run competitive equilibrium can now be characterized:

Characterization of Short-run Competitive Equilibrium — *For a given capacity, K, and a given share of customers on realtime pricing, α, the short-run competitive equilibrium is characterized by realtime retail prices $p_t^e = w_t^e$ and flat-rate retail price $\bar{p}^e = \sum_{t=1}^T w_t^e D_t(\bar{p}^e) / \sum_{t=1}^T D_t(\bar{p}^e)$. The equilibrium wholesale (realtime) prices are determined by $w_t^e = C_q(\tilde{D}_t(p_t^e, \bar{p}), K)$ for every t.*

The equilibrium characterized above is illustrated in Figure 3 for two demand periods: peak and off peak. Since not all customers face the realtime prices, wholesale demand is given by the less elastic demand curves $\tilde{D}_p$ and $\tilde{D}_{op}$. The realtime prices, p_p and p_{op}, are then determined by the intersection of these demand curves with the short-run supply $C_q(Q, K)$. The equilibrium flat rate $\bar{p}^e$ is the demand-weighted average of p_p and p_{op}. The demand-weighted average $\bar{p}^e$ is closer to p_p than to p_{op} since the flat-rate customers demand more in the peak than off peak.

In the long-run, generation capacity will enter (exit) the wholesale market as long as profits are positive (negative). Thus, competitive investment drives long-run profits to zero. The zero profit condition on the wholesale sector is $\sum_{t=1}^T [w_t q_t - C(q_t, K)] - rK = 0$. Thus,

Characterization of Long-run Competitive Equilibrium — *For a given share of customers on RTP, α, the long-run competitive equilibrium wholesale prices are characterized by the conditions characterizing a short-run competitive equilibrium plus the additional condition $\sum_{t=1}^T [w_t^e \tilde{D}_t(p_t^e, \bar{p}) - C(\tilde{D}_t(p_t^e, \bar{p}), K)] = rK$.*

The long-run competitive equilibrium can also be illustrated in Figure 3. In the long run, capacity will enter or exit depending on whether investment is profitable. The short-run profits in each period are illustrated by the area bounded above by the realtime price and bounded below by $C_q(q, k)$. If the total short-run profits exactly equal the long-run cost of capital, then the industry is in long-run equilibrium.[11]

---

[10] This condition can alternately be written: $\tilde{D}_t(p_t, \bar{p}) = KS(w_t)$.

[11] A question remains about the feasibility of the competitive equilibrium in the short and long run. To see that the equilibrium flat price is always feasible in the short run, define $\bar{p}^{min}(K)$ as the

B. (In)efficiency of competitive equilibrium

The First Welfare Theorem ensures efficiency of the competitive equilibrium under certain conditions. However, the requirements of the welfare theorems are not met if $\alpha < 1$, since there is a missing market. Customers on flat retail prices cannot trade with customers on realtime prices or with producers since all electricity transactions must occur at the same price for flat-rate customers. This missing market implies that the competitive equilibrium discussed above may not be efficient.

However, if all customers face the realtime prices, *i.e.,* $\alpha = 1$, then the competitive equilibrium is Pareto efficient. Pareto efficiency follows immediately once $\alpha = 1$ because there is no missing market and all of the conditions of the First Welfare Theorem are satisfied. This implies that there is short-run allocative efficiency and long-run efficiency of capacity investments.

To see this in our particular application, consider first the short-run equilibrium. Since $\alpha = 1$, $\tilde{D}_t = D_t$ for every t. The equilibrium condition $w_t^e = C_q(D_t(p_t^e), K)$ implies that the marginal cost of production is equal to the wholesale price in every period. Since all customers are on realtime pricing, w_t is equal to the marginal utility of consumption for each customer. Since the marginal cost of generation equals the marginal utility of each customer in each time period, the short-run equilibrium is Pareto efficient.

For the long run, the marginal social value of capacity is given by the decrease in costs resulting from an increment to installed capacity. In period t, this decrease in costs is given by $-C_k(D_t(p_t), K)$. Since installing capacity decreases costs in all periods, the social optimum would dictate installing additional capacity as long as $\sum_{t=1}^{T} -C_k(D_t(p_t), K) > r$ and stopping investment when $\sum_{t=1}^{T} -C_k(D_t(p_t), K) = r$.[12] Recall that competition will lead to more investment as long as profits are positive, *i.e.,* $\sum_{t=1}^{T}[w_t D_t(p_t) - C(D_t(p_t), K)] > rK$, and investment ceases when $\sum_{t=1}^{T}[w_t D_t(p_t) - C(D_t(p_t), K)] = rK$. By differentiating the zero profit condition with respect to K, we see that competition leads to additional investment if and only if it is efficient. Thus, private incentives for investment accurately reflect social incentives and the long-run competitive equilibrium is efficient when all cus-

---

greatest lower bound of the set of feasible flat retail prices. If $\bar{p}^{min}(K)$ is not feasible, then note that $\max_t\{p_t\}$ goes to infinity as $\bar{p}$ decreases to $\bar{p}^{min}(K)$. This implies that the purchase costs of the retailer can be made arbitrarily large for flat rates above $\bar{p}^{min}(K)$, so $\bar{p}^e > \bar{p}^{min}(K)$, *i.e.,* the short-run equilibrium flat rate is feasible. On the other hand, if $\bar{p}^{min}(K)$ is feasible, then note that $\max_t\{p_t\}$ can be arbitrarily large for the flat rate $\bar{p}^{min}(K)$. This implies that $\bar{p}^e \geq \bar{p}^{min}(K)$, *i.e.,* the short-run equilibrium flat rate is feasible. Feasibility of the long-run equilibrium price is implied by feasibility of the short-run equilibrium price.

[12] This condition can be derived by solving the social planner's problem for the long run.

tomers are on realtime pricing.

If some customers do not face the realtime prices, $\alpha < 1$, the competitive equilibrium is not Pareto efficient, *i.e.,* does not attain the first-best electricity allocation and capacity investment. To see this, consider the short run in which K is fixed. Recall that competition among retailers drives retail prices for RTP customers to be equal to wholesale prices in each period and drives $\bar{p}$ to be equal to the demand-weighted average wholesale price. Equilibrium wholesale prices are determined by supply and demand ($\tilde{D}_t$) in every period. This short-run equilibrium is clearly not first best because in almost all hours flat-rate customers are not charged a price equal to the industry marginal cost.

While it is clear that flat-rate retail pricing will not yield first-best resource allocation, there is still a question of what flat rate minimizes the resulting deadweight loss. In particular, does the competitive equilibrium flat rate, $\bar{p}^e_{SR}$, attain a second best by minimizing the deadweight loss associated with having flat-rate customers? To answer this question, consider the flat retail rate, $\bar{p}^*_{SR}$, and realtime prices $p^*_{t_{SR}}$ that minimize deadweight loss in the short run. $\bar{p}^*_{SR}$ and $p^*_{t_{SR}}$ can be found from the optimization:[13]

$$\max_{p_t,\bar{p}} \sum_{t=1}^{T} [\tilde{U}_t(p_t,\bar{p}) - C(\tilde{D}_t(p_t,\bar{p}),K)] - rK \qquad [1]$$

where the consumer surplus measure $\tilde{U}_t$ is defined by $\tilde{U}_t(p,\bar{p}) \equiv \alpha U_t(D_t(p)) + (1-\alpha)U_t(D_t(\bar{p}))$ and U_t maps quantities into the usual consumer surplus.[14] We refer to the result of this optimization as the *second-best optimal allocation.*[15] The optimization can be described by two first-order conditions.

For the optimal realtime price in period t, the first-order condition is

$$\alpha\{U'_t(D_t(p_t)) \cdot D'_t(p_t) - C_q(\tilde{D}_t(p_t,\bar{p}),K) \cdot D'_t(p_t)\} = 0, \qquad [2]$$

which, since $U'_t(D_t(p_t)) = p_t$, implies that $p_t = C_q(\tilde{D}_t(p_t,\bar{p}),K)$.

---

[13] This optimization is equivalent to a social planner's problem where the planner is constrained to choose a vector of quantities that satisfies the demands of both the flat-rate and realtime customers at the chosen prices.

[14] As usual, the marginal utility and demand are inverse functions, *i.e.,* $U'_t(D_t(p)) = p$.

[15] This optimization is the sum of consumer surplus, $\sum \tilde{U}_t(p_t,\bar{p}) - \alpha p_t D_t(p_t) - (1-\alpha)\bar{p}D_t(\bar{p})$, retail profits, $\sum \alpha p_t D_t(p_t) + (1-\alpha)\bar{p}D_t(\bar{p}) - w_t\tilde{D}_t(p_t,\bar{p})$, and generator profits, $\sum[w_t\tilde{D}_t(p_t,\bar{p}) - C(\tilde{D}_t(p_t,\bar{p}),K)] - rK$. Note that w_t is simply a transfer and does not affect deadweight loss.

For the optimal flat rate, the first-order condition is

$$\sum_{t=1}^{T}[\bar{p}^*_{SR} - C_q(\tilde{D}_t(p_t,\bar{p}),K)](1-\alpha)D'_t(\bar{p}^*_{SR}) = 0. \qquad [3]$$

Substituting $p^*_{t_{SR}}$ for $C_q(\tilde{D}_t(p_t,\bar{p}),K)$ for all t in [3] yields

$$\sum_{t=1}^{T}[\bar{p}^*_{SR} - p^*_{t_{SR}}]D'_t(\bar{p}^*_{SR}) = 0 \qquad [4]$$

which implies

$$\bar{p}^*_{SR} = \sum_{t=1}^{T} p^*_{t_{SR}} D'_t(\bar{p}^*_{SR}) / \sum_{t=1}^{T} D'_t(\bar{p}^*_{SR}). \qquad [5]$$

Thus, the flat retail price that minimizes the deadweight loss is a weighted average of the realtime prices where the weights are the relative slopes of the demand curves.[16] Since $\bar{p}^e_{SR}$ is also a weighted average of the realtime prices but with different weights, we have the first result:

Result 1: Non-attainment of the Second Best in the Short Run — *The short-run competitive equilibrium does not attain the second-best optimal electricity allocation. Furthermore, the equilibrium flat rate, $\bar{p}^e_{SR}$, can be either higher or lower than optimal.*

Proof: Since both $\bar{p}^e_{SR}$ and $\bar{p}^*_{SR}$ are weighted averages of the p_t but their weights are not necessarily equal, comparison of the two weighted averages implies that $\bar{p}^e_{SR}$ does not necessarily equal $\bar{p}^*_{SR}$. We can construct an example where $\bar{p}^e_{SR}$ is higher (lower) than optimal by making the $D'_t(\bar{p})$ arbitrarily large (small) for all t such that $D_t(\bar{p}) > KS(\bar{p})$ and the $D'_t(\bar{p})$ arbitrarily small (large) for all t such that $D_t(\bar{p}) < KS(\bar{p})$. ∎

To illustrate that the equilibrium flat retail price may be either too high or too low, consider a simple example with two time periods: peak and off-peak. Clearly, the competitive equilibrium flat rate is less than the peak realtime price and greater than the off-peak price. If peak demand were perfectly inelastic, *i.e.,* if $D'_p = 0$, then the optimal flat rate would place no weight on the peak period price and all weight on the off-peak price. The competitive equilibrium flat rate is then higher than optimal since decreasing the flat rate does not change consumption on peak, but reduces the consumption distortion off peak.[17]

---

[16] For example, if the demands all have the same slope, $\bar{p}^*_{SR}$ is simply the arithmetic mean of the wholesale prices.

[17] In this special case, the first-best and second-best optimal allocations are identical.

Conversely, if $D'_{op} = 0$ and $D'_p < 0$, then the optimal flat rate places no weight on the off-peak price and the competitive flat rate is too low. Now increasing the flat rate does not change consumption off peak, but reduces the consumption distortion on peak. This illustrates a case where the equilibrium flat rate is too low.

Figure 4 illustrates the case where off-peak demand is perfectly inelastic and $\bar{p}^*_{SR} > \bar{p}^e_{SR}$. The equilibrium flat rate, $\bar{p}^e_{SR}$, is a demand-weighted average of the peak wholesale price p^e_p and off-peak wholesale price p^e_{op}. Since off-peak demand is perfectly inelastic, there is no inefficiency off-peak. Note, however, that increasing the flat rate reduces the peak-period deadweight loss between flat-rate and RTP customers. Clearly, setting $\bar{p}^*_{SR} = p^*_p$ eliminates the peak-period misallocation since flat-rate and RTP customers both face the same prices. Note also that increasing the flat rate from $\bar{p}^e_{SR}$ to $\bar{p}^*_{SR}$ decreases the peak demand from $\tilde{D}_p$ to $\tilde{D}'_p$ and lowers the peak realtime price.

Interestingly, if all demands have the same elasticity at $\bar{p}^e$, then the $\bar{p}^e = \bar{p}^*$. To see this, note that if demands in two periods, i and j, have the same elasticity at $\bar{p}$, then

$$\frac{\bar{p}}{D_i(\bar{p})}D'_i(\bar{p}) = \frac{\bar{p}}{D_j(\bar{p})}D'_j(\bar{p}) \qquad \Longleftrightarrow \qquad \frac{D'_i(\bar{p})}{D_i(\bar{p})} = \frac{D'_j(\bar{p})}{D_j(\bar{p})} \qquad \Longleftrightarrow \qquad \frac{D'_i(\bar{p})}{D'_j(\bar{p})} = \frac{D_i(\bar{p})}{D_j(\bar{p})}.$$

Thus, a weighted average of wholesale prices using as weights the flat-rate quantities will be the same as a weighted average using as weights the demand slopes at those flat-rate quantities, *i.e.,* $\bar{p}^e = \bar{p}^*$. Furthermore, this shows that if the elasticity at $\bar{p}$ in period i is greater than the elasticity in period j then $\frac{D'_i(\bar{p})}{D'_j(\bar{p})} > \frac{D_i(\bar{p})}{D_j(\bar{p})}$. Therefore, the weighted average with slopes as weights puts more relative weight on the more elastic periods. Thus if the high demand periods are relatively more (less) elastic, then the equilibrium flat rate is lower (higher) than optimal.

Although competition distorts the consumption of the flat-rate customers relative to the second best, competition does not introduce additional distortions into the realtime market for a given flat rate. For a given $\bar{p}$, the optimal realtime prices are determined by the first-order conditions from the planner's problem, which imply that $p_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$ for every t. Note that these optimal prices are exactly the realtime prices that would result from competition, given a $\bar{p}$, namely, the prices such that supply equals demand. Thus, if a planner were to force the retail sector to charge $\bar{p}^*_{SR}$ to flat-rate customers, the realtime prices resulting from retail competition would be second-best optimal. In this manner, the second-best optimal allocation could be achieved in the short run. Note however that this would imply profits or losses in the retail sector.[18]

---

[18] Policies for improving the efficiency of the competitive equilibrium will be discussed in Section II.

C. Inefficiency in the long run

In the long run, supply and demand are equated by the realtime wholesale prices; retail competition forces $p_t = w_t$ for every t; the equilibrium flat retail price, $\bar{p}^e_{LR}$, is determined by retail competition; and equilibrium capacity, K^e_{LR} is determined by wholesale competition. Because of the flat retail price, the first-best outcome is not achieved in either capacity investment or production. In light of our short-run results from the previous subsection, it is not surprising that the long-run outcome is not second-best optimal given the existence of flat-rate customers.

To determine the second-best optimum in the long run, consider the flat retail rate, $\bar{p}^*_{LR}$, realtime prices, $p^*_{t_{LR}}$, and capacity, K^*_{LR}, that minimize deadweight loss. The optimum can be found from the maximization in equation [1] where now optimization is also with respect to capacity.[19] The first-order conditions for p_t and $\bar{p}$ are given by [2] and [3] and the first-order condition for K is

$$\sum_{t=1}^{T} -C_k(\tilde{D}_t(p_t, \bar{p}), K) = r \qquad [6]$$

As in the short run, the second-best price, $\bar{p}^*_{LR}$, is a weighted average of the realtime prices where the weights are the relative slopes of the demand curves. The optimal realtime prices are determined by $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$ for every t. Note that equation [6] implies that at the second-best optimal capacity, the marginal cost reduction from an additional unit of investment is exactly equal to the daily cost of capital, *i.e.,* that there are zero profits net of capital costs. This implies that, given the second-best flat rate, competition in investment would lead to the second-best capacity investment.

As in the short run, $\bar{p}^e_{LR}$ and $\bar{p}^*_{LR}$, are different weighted averages of the realtime prices. Therefore, $\bar{p}^e_{LR}$ is not generally equal to $\bar{p}^*_{LR}$, and the equilibrium flat price can be either too high or too low relative to the second best. This implies that the competitive equilibrium may lead to suboptimal installation of capacity as well. Therefore,

Corollary 1: Non-attainment of the Second Best in the Long Run — *The long-run competitive equilibrium does not attain the second-best optimal electricity allocation and capacity investment. Furthermore, the equilibrium flat rate, $\bar{p}^e_{LR}$, is higher than optimal, if and only if the equilibrium capacity investment, K^e_{LR}, is smaller than optimal.*

Proof: To see that K^e_{LR}, can be either larger or smaller than K^*_{LR}, suppose that slopes of the demand curves are such that $\bar{p}^*_{LR} > \bar{p}^e_{LR}$, *i.e.,* the equilibrium flat price is too

[19] As above, the planner regards the wholesale prices as transfers which do not affect efficiency.

low. Further suppose that the market is in long-run equilibrium, and the planner tries to improve efficiency in the short run by increasing the flat retail price to $\bar{p}^*_{LR}$. In the short run, this would decrease demand $\tilde{D}_t$ in every period so prices and consumption would fall. Since consumption has fallen, this implies that the cost reduction from an additional unit of capacity has decreased, *i.e.,* $-C_k$ has decreased since $C_{qk} < 0$. But this implies that $\sum -C_k$ is now less than r, so to improve investment efficiency the planner would have to reduce capacity. This implies that the equilibrium long-run capacity was too large relative to the second-best optimal long-run capacity. A symmetric argument shows that $K^*_{LR} > K^e_{LR}$ iff $\bar{p}^*_{LR} < \bar{p}^e_{LR}$.[20] ∎

As in the short run, the distortion in the competitive equilibrium stems from the flat retail price. In particular, if a planner were to impose the optimal flat rate, $\bar{p}^*_{LR}$, then competition would lead to the second-best optimal realtime prices and capacity investment, K^*_{LR}. As above, the conditions $p_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$ imply that supply equals demand in every period and the condition $\sum_{t=1}^{T} -C_k(\tilde{D}_t(p_t, \bar{p}), K) = r$ implies that there are no profits in investment. Thus, competitive investment and retail markets would attain the second-best optimum if the planner were to impose the second-best optimal flat retail price.

II. Subsidies/Taxes on Capacity or Electricity

In restructured wholesale electricity markets, many parties have suggested that in order to assure sufficient investment in generation, "capacity payments" to producers are necessary. These payments directly subsidize the holding of capacity, generally without a commitment on the producer's part to offer any certain quantity of energy or any certain price.[21] Such payments can be seen as part of a general category of market interventions designed to move the equilibrium outcome closer to the (constrained) social optimum. In this section, we consider such policies.

Among such interventions, there are two characteristics that are central to the economic analysis of the policy. First, the subsidy/tax can be directed at the retail price of electricity or it can be directed at capacity. Second, the revenues from a subsidy/tax can flow to or from an external source (such as the government's general fund) or the scheme

[20] The result can also be proved by defining the welfare function, W, from [1] where $\bar{p}$, p_t, and K are a long-run competitive equilibrium. It is easy to show that $\frac{dW}{dK} = \frac{\partial W}{\partial \bar{p}} \frac{d\bar{p}}{dK}$. Since the equilibrium flat rate falls when capacity increases, *i.e.,* $\frac{d\bar{p}}{dK} < 0$, it follows that $\frac{dW}{dK} > 0$ if and only if $\frac{\partial W}{\partial \bar{p}} < 0$.

[21] In some markets, capacity payments are contingent on a minimum level of capacity availability.

can operate on a balanced-budget basis with all revenues flowing to or from electricity customers. Finally, for any adjustment to retail rates, RTP and flat-rate customers may be treated symmetrically or the tax/subsidy can apply to only one group, generally the flat-rate group because the RTP group begins from a second-best optimum.

Analytically, the simpler cases are those in which no balanced-budget requirement is imposed; all net funds flow to/from an external source. We begin with those.

A. Externally financed subsidies or taxes on retail electricity or on capacity

The simplest policy intervention to analyze is a tax or subsidy on flat-rate retail electricity prices. Since the retailers receive no surplus in equilibrium, adding such a tax would drive up the retail price paid by the flat-rate customers thereby decreasing wholesale demand during all periods. The decrease in wholesale quantity demanded would cause wholesale prices to decrease, generators to exit in the long run, and industry generation capacity to decrease. A subsidy to the flat-rate retail price would have the opposite effect.

We can characterize the long-run competitive equilibrium with a retail tax τ on the flat-rate customers. As in the of the competitive equilibrium above, RTP customers pay the wholesale prices, *i.e.,* $p_t = w_t$; wholesale demand equals supply, *i.e.,* $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$; and wholesale profits cover capacity costs, *i.e.,* $\sum_{t=1}^{T} w_t \tilde{D}_t(p_t, \bar{p}) - C(\tilde{D}_t(p_t, \bar{p}), K) = rK$. In the flat-rate retail market, however, there is now a tax wedge between the flat-rate price paid by the customers $\bar{p}$ and the flat rate received by the retail sector, $\bar{p} - \tau$. Thus, the equilibrium flat rate is determined by $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$.

Given this characterization of the equilibrium, it is straightforward to show that the optimal tax or subsidy will be $\tau^* = \bar{p}^*_{LR} - \sum_{t=1}^{T} p^*_t D_t(\bar{p}^*_{LR}) / \sum_{t=1}^{T} D_t(\bar{p}^*_{LR})$ charged to all customers paying a flat retail rate. The second term is the quantity-weighted average price of buying wholesale power for flat-rate customers when the flat rate is $\bar{p}^*_{LR}$. Thus, τ^* is the tax or subsidy that allows the retailer to break even while charging $\bar{p}^*_{LR}$.[22,23] Therefore, we have:

Result 2: Optimality of Retail Tax/Subsidy on Flat-Rate Retail Customers — *With external financing, a tax/subsidy* $\tau^* = \bar{p}^*_{LR} - \sum_{t=1}^{T} p^*_t D_t(\bar{p}^*_{LR}) / \sum_{t=1}^{T} D_t(\bar{p}^*_{LR})$ *on the flat-rate customers achieves the second-best optimal allocation and capacity investment.*

---

[22] It is worth pointing out that the optimal tax/subsidy is not, in general, equal to the difference between the second-best optimal flat rate and the equilibrium flat rate, $\bar{p}^e_{LR} - \bar{p}^*_{LR}$.

[23] The tax or subsidy, τ^*, is like a Pigouvian tax or subsidy on an externality. However, τ^* only allows the second best to be attained by competition.

The optimal policy, τ^*, may be a tax or a subsidy.

Proof: Since the retailers break even when charging $\bar{p}^*_{LR}$ and paying the retail tax, $\bar{p}^*_{LR}$ is the equilibrium flat rate. Therefore consumption of the flat-rate customers is at the second-best optimal level. The competitive equilibrium for a given $\bar{p}$ does not introduce any additional distortions in consumption of the realtime customers or in investment since realtime prices and investment costs are not distorted. ∎

Result 2 can be illustrated with Figure 4 in the short-run. If the retail sector were to charge the flat rate $\bar{p}^*_{SR}$, profits would be positive since its margin on the flat-rate customers in the off-peak, $\bar{p}^*_{SR} - p^*_{op}$, would be positive, but its margin in the peak, $\bar{p}^*_{SR} - p^*_{p}$, would be zero. Taxing the flat-rate customers would force the equilibrium flat rate up and improve efficiency. The optimal tax, described in Result 2, would leave the retail sector with zero profit when it charged the second-best flat rate $\bar{p}^*_{SR}$. Note that in the short run, this would decrease wholesale demand so capacity will exit in this example in the long run.

While a tax/subsidy on flat-rate customers can achieve the second-best optimal price, a tax/subsidy on all retail customers (flat-rate and RTP) cannot. If all retail customers are taxed, there are tax wedges in both the realtime and flat-rate markets. The equilibrium is then characterized by the equality of wholesale demand and supply, *i.e.,* $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$; and wholesale profits covering capacity costs, *i.e.,* $\sum_{t=1}^{T} w_t \tilde{D}_t(p_t, \bar{p}) - C(\tilde{D}_t(p_t, \bar{p}), K) = rK$; plus the two conditions on the distorted markets: $p_t - \tau = w_t$ and $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$.

A tax/subsidy on all retail customers cannot achieve the second-best optimum because the RTP customers are served optimally absent the tax/subsidy, as was shown in the previous section. Setting τ to achieve the second-best optimal price for flat-rate customers distorts the prices for RTP customers away from the second-best optimal level for them that is achieved if no tax/subsidy is applied to RTP customers.[24]

Though retail taxes/subsidies may seem the natural policy instrument to address the efficiency problem caused by flat retail pricing, the public policy debate has focused on taxes or subsidies (actually, just subsidies) for capacity. However, a tax/subsidy to capacity also cannot attain the second-best optimum. If capacity is subsidized, the cost of capital is lowered leading to excessive capacity installation. If σ is the capacity subsidy, the equilibrium is characterized by the equality of wholesale demand and supply, *i.e.,*

[24] An optimal retail tax/subsidy imposed on all customers would not equate the flat rate with $\bar{p}^*_{LR}$, but would instead allow some distortion in the flat-rate market in order to lessen the distortion in the realtime market.

$w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$; realtime prices equal to wholesale prices, *i.e.,* $p_t = w_t$; zero retail profits, *i.e.,* $\bar{p} = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$; and the zero profits in the wholesale sector where now the capital cost is $r - \sigma$, *i.e.,* $\sum_{t=1}^{T} w_t \tilde{D}_t(p_t, \bar{p}) - C(\tilde{D}_t(p_t, \bar{p}), K) = (r - \sigma)K$.

The subsidy to capacity lowers the cost of capital and induces installation of additional capacity. Installation of additional capacity lowers marginal costs and drives down the wholesale prices. Competition among the retailer then drives down the equilibrium flat retail rate. Note that reducing the flat rate improves efficiency if the equilibrium flat rate was higher than optimal, *i.e.,* if $\bar{p}^e > \bar{p}^*$. In fact, there may be a capacity subsidy which would cause the equilibrium flat rate to be driven to $\bar{p}^*$. However, since $\sum_{t=1}^{T} -C_k(\tilde{D}_t(p_t, \bar{p}), K) = r - \sigma < r$, this capacity subsidy leads to installation of capacity greater than K^*_{LR}. This implies that the second-best optimum is not attained by the capacity subsidy. Further note, that the realtime prices are driven below the second-best level by the installation of the additional capacity.

Result 3: Non-optimality of Retail Tax/Subsidy on All Retail Customers and of Capacity Tax/Subsidy— *The following policies cannot attain the second-best optimal allocation and capacity investment:*

(i) an externally funded tax/subsidy on all retail customers, and

(ii) an externally funded tax/subsidy on capacity.

Proof: An externally funded tax/subsidy on all retail customers or on capacity for which $\bar{p}^*_{LR}$ is an equilibrium does not attain the second best because consumption of RTP customers and/or investment is distorted. On the other hand, any policy for which $\bar{p}^*_{LR}$ is not an equilibrium does not attain the second best because consumption of the flat rate customers is distorted. ∎

B. Capacity subsidies/taxes financed by retail taxes/subsidies

Most of the public policy debates regarding investment in electricity markets have not actually considered capacity subsidies from outside the industry. Instead, the recommended policy tool has usually been capacity subsidies financed by fees collected from retail electricity providers. In most cases, the collection mechanism suggested has been a retail electricity tax that does not vary over time.

The retail electricity tax used to fund the capacity payments can be administered in a number of ways. First, we analyze the simpler case where the tax is levied only on the flat-rate customers. Combining the analyses above, the long-run competitive equilibrium can be characterized by: $p_t = w_t$; $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$; $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$;

and $\sum_{t=1}^{T} w_t \tilde{D}_t(p_t, \bar{p}) - C(\tilde{D}_t(p_t, \bar{p}), K) = (r - \sigma)K$. The balanced-budget condition is $\tau \sum_{t=1}^{T}(1-\alpha)D_t(\bar{p}) = \sigma K$. The balanced-budget condition ensures that the tax revenue collected by the retail sector exactly funds the capacity payments made to the wholesale sector.[25]

Such a capacity payment has two off-setting effects. The scheme includes a tax on the retail sector, which increases the equilibrium flat retail price, and a capacity subsidy to the wholesale sector, which decreases wholesale prices and, thereby, decreases the equilibrium flat retail price. Though at first it may seem that these effects would be offsetting, that isn't generally true if sufficient customers are on RTP.[26]

When some customers face the realtime prices, the effects of the capacity payments do not in general offset one another since the capacity payment lowers prices in the wholesale market. The lower wholesale prices increase consumption of the customers facing the realtime price. Effectively, the capacity payment raises the flat retail price (harming customers facing the flat price) but lowers the wholesale prices (benefiting customers facing the realtime price). If the flat-rate market is distorted the former effect may improve efficiency.

With the above characterization of the equilibrium, it can be shown that the second-best optimal price $\bar{p}^*_{LR}$ is the equilibrium outcome from a capacity payment of $\sigma^* = (1-\alpha)[\bar{p}^*_{LR} \sum_{t=1}^{T} D_t(\bar{p}^*_{LR}) - \sum_{t=1}^{T} \tilde{p}_t D_t(\bar{p}^*_{LR})]/\tilde{K}$ where the $\tilde{p}_t$ and $\tilde{K}$ are such that $\tilde{p}_t = C_q(\tilde{D}_t(\tilde{p}_t, \bar{p}^*_{LR}), \tilde{K})$; and $\sum_{t=1}^{T} \tilde{p}_t \tilde{D}_t(\tilde{p}_t, \bar{p}^e_{LR}) - C(\tilde{D}_t(\tilde{p}_t, \bar{p}^e_{LR}), \tilde{K}) = (r - \sigma^*)\tilde{K}$.[27] Note that σ^* is the retail profit per unit of capacity that would result if the retail sector charged $\bar{p}^*_{LR}$. Thus σ^* is positive if $\bar{p}^e_{LR} < \bar{p}^*_{LR}$, *i.e.,* if the equilibrium flat price is too low. This is equivalent to the effect of a tax on retail electricity. However, note also that although σ^* minimizes the deadweight loss in the flat-rate market, it leads to excessive investment if $\sigma^* > 0$ since $\sum_{t=1}^{T} -C_k(\tilde{D}_t(\tilde{p}_t, \bar{p}^e_{LR}), \tilde{K}) = (r - \sigma^*) < r$.[28] This implies that there is deadweight loss in the realtime market since the realtime prices will be too low relative to

---

[25] In what follows, we assume that σ is the policy instrument and that τ is determined endogenously such that the capacity payments are fully funded. Clearly, τ could be the policy instrument and σ could be determined endogenously.

[26] These two effects can be exactly offsetting under certain conditions if there are few or no customers on RTP. See Borenstein and Holland (2003) for an example.

[27] Since σ^* is defined implicitly by highly non-linear equations, it is difficult to prove that a general solution exists to the system of equations. See Borenstein and Holland (2003) for an example where σ^* can be easily derived.

[28] Alternatively there would be insufficient investment if $\sigma^* < 0$.

the second best. Thus the capacity payment scheme which results in $\bar{p}^e_{LR}$ as the equilibrium flat rate does not attain the second best.

Policy makers have generally proposed capacity payments to be funded by payments from all retail customers and not just the flat-rate customers. Combining the analyses of a retail tax on all customers from above with a capacity subsidy, the long-run competitive equilibrium can be characterized by: $p_t - \tau = w_t$; $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$; $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$; and $\sum_{t=1}^{T} w_t \tilde{D}_t(p_t, \bar{p}) - C(\tilde{D}_t(p_t, \bar{p}), K) = (r - \sigma)K$. The balanced-budget condition is now $\tau \sum_{t=1}^{T} \tilde{D}_t(p_t, \bar{p}) = \sigma K$. As above, the balanced-budget condition ensures that the revenue collected exactly covers the capacity subsidy where now revenue is collected from all customers.

Capacity payments funded by all retail customers have a number of effects. Consider a positive capacity payment. The capacity payment has two components: a tax on all retail electricity customers and a subsidy to capacity. If the equilibrium flat rate was too low, the tax on the flat-rate customers may improve efficiency. However, it distorts RTP consumption in all periods by driving up the realtime prices. The capacity subsidy decreases the cost of capital which increases capacity and drives down the wholesale prices. As above, this partially offsets the effect of the retail tax but will not completely offset the effect. In fact, we could define a capacity payment σ^* which would lead to an equilibrium flat rate of $\bar{p}^*_{LR}$, but this capacity payment scheme would lead to too much investment and suboptimal realtime consumption relative to the second best.

Corollary 2: Non-optimality of Balanced Budget Capacity Payments — *Capacity payments, funded by an excise tax only on electricity sold to flat-rate customers or funded by an excise tax on electricity sold to all retail customers, cannot achieve the second-best optimal allocation and capacity investment.*

Proof: Any capacity payment scheme—whether funded by all retail customers or only the flat-rate customers—that leads to $\bar{p}^*_{LR}$ as the equilibrium flat rate does not attain the second best because it distorts investment and realtime consumption. Any capacity payment scheme that does not result in $\bar{p}^*_{LR}$ as the equilibrium flat rate also does not attain the second best. ∎

III. Changing Proportion of Customers on Realtime Pricing

While it is clear that, absent metering costs, charging real-time prices to all customers would be Pareto efficient, in reality any changes towards RTP are likely to be incremental, with an increasing share of customers moving to RTP over time. This section examines

the effect of changing the proportion of customers on RTP. Following the assumptions of the previous sections, we first examine effects when all customers have the same demand patterns and α is set exogenously. Even in this relatively uncomplicated case, we reach some surprising conclusions. In the final subsection, we examine the outcomes when customers choose whether or not to switch to RTP in a market context, recognizing both the costs of metering and the fact that customers are heterogeneous.

A. The Effect on Prices of Increasing RTP Customers

Increasing the proportion of customers on RTP increases the elasticity of demand by rotating $\tilde{D}_t$ around $\bar{p}$. This has two effects on wholesale prices. For periods in which the wholesale price is above the flat rate, increasing α decreases demand since more customers face the higher realtime price. This decrease in demand drives down the wholesale price in these periods. Conversely, for periods in which the wholesale price is below the flat rate, demand *increases* with α since more customers face the lower realtime price. This drives up the wholesale prices in these periods. Thus, some wholesale prices increase and some decrease when more customers are put on realtime pricing.

The effect on the flat retail rate in the long run, however, is not ambiguous.

Result 4: Effect of Increasing RTP Customers on Flat Retail Rate — *In the long run, an increase in the proportion of customers on RTP reduces* $\bar{p}^e_{LR}$.

Proof: See appendix.

The key to this result is recognizing that retail profits on the flat-rate customers depend on covering losses when the retail margin is negative (peak periods) with gains when the margin is positive (off-peak periods). Since flat-rate customers demand more in peak periods, the retailer cares more about price changes in the peak period. Increasing the proportion of customers on RTP will decrease the peak prices and increase off-peak prices. This can be beneficial for the retailers if the peak prices decrease sufficiently relative to the increases in the off-peak prices. However, these price changes also affect wholesale profits and investment. In the appendix, we show that for a given $\bar{p}$ the decreased retail losses in the peak periods offset the decreased off-peak retail gains if capacity adjusts such that wholesale profits are unchanged. Thus, if customers were moved to RTP and $\bar{p}$ did not decline, retailers would be earning positive profits on flat-rate customers. Competition in the retail market would then force down retail prices.[29]

[29] Here and throughout this section, we do not state results as being *weakly* true, though there are special cases where changes are zero. We note here that zero-change cases can be constructed, but

B. The Effect on Capacity of Increasing RTP Customers

Investment in the regulated electricity industry was determined primarily by projections of annual peak loads. Additional generation was deemed necessary if reserve margins during peak hours were insufficient. Since putting additional customers on RTP would reduce peak loads, this could reduce the need for investment.[30]

In competitive markets, investment in generation capacity is driven by profit opportunities rather than by a planning process. Since putting more customers on RTP leads to decreased realtime prices in peak periods, this effect implies decreased wholesale profits in peak periods and reduced incentives for investment. However, in periods when the marginal cost is below the flat rate, increasing α would lead to increased demand. If the industry marginal cost has positive slope, this would increase prices and profits in these periods. New investment occurs if the additional wholesale profit off-peak is greater than the decline in profit during the peak periods.

To see this in our model, recall that equilibrium wholesale profits in the short-run are given by $\pi^w_{SR} = \sum_t p_t \tilde{D}(p_t, \bar{p}) - C(\tilde{D}(p_t, \bar{p}), K)$. Since $p_t = C_q$ in equilibrium, it follows that $\frac{\partial \pi^w_{SR}}{\partial \alpha} = \sum_t p_t \frac{\partial \tilde{D}(p_t,\bar{p})}{\partial \alpha} - C_q \frac{\partial \tilde{D}(p_t,\bar{p})}{\partial \alpha} = \sum_t (p_t - C_q) \frac{\partial \tilde{D}(p_t,\bar{p})}{\partial \alpha} = 0$. By similar reasoning, $\frac{\partial \pi^w_{SR}}{\partial \bar{p}} = 0$ and $\frac{\partial \pi^w_{SR}}{\partial p_t} = \tilde{D}_t$, so

$$\frac{d\pi^w_{SR}}{d\alpha} = \frac{\partial \pi^w_{SR}}{\partial \alpha} + \frac{\partial \pi^w_{SR}}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha} + \sum_t \frac{\partial \pi^w_{SR}}{\partial p_t} \frac{dp_t}{d\alpha} = \sum_t \tilde{D}_t(p_t, \bar{p}) \frac{dp_t}{d\alpha}. \qquad [7]$$

Since [7] is a weighted average of the $\frac{dp_t}{d\alpha}$, which may be positive or negative, the short-run wholesale profits may increase or decrease. This implies that investment may increase or decrease:

Result 5: Indeterminant Effect of Increasing RTP Customers on Capacity — *An increase in the proportion of customers on RTP can increase or decrease long-run equilibrium capacity* K^e_{LR}.

Proof: See appendix.

The proof of Result 5 depends on the convexity of the marginal costs across the relevant range. If the marginal cost curve is relatively flat at off-peak demand levels, then

they appear to be extreme or degenerate cases, so we do not emphasize them.

[30] Bergstrom and MacKie-Mason (1991) argue against the conventional wisdom by showing that peak-load pricing could lead to increased investment in a regulated industry. Our analysis is of competitive markets and does not assume homothetic preferences. In a related model of airline competition, Dana (1999) shows that equilibrium price dispersion, *i.e.,* stochastic peak-load pricing, can lead to lower capacity costs.

putting additional customers on RTP will not increase the off-peak prices very much.[31] If the marginal cost curve is relatively steep at peak demand levels, then increasing α will cause relatively large decreases in the peak prices. These relatively large price decreases on peak imply that wholesale profits decrease in the short run and equilibrium capacity decreases when α increases. It is easy to construct examples in which capacity decreases, and this is likely the policy relevant case.[32]

Conversely, if the marginal cost curve is relatively steep at off-peak demand levels and relatively flat at peak demand levels, *e.g.,* if C_q were concave, then the off-peak price increase would be greater than the peak price decrease, and wholesale profits and capacity would increase. Since this is the surprising case, we present in the appendix a simple example where putting more customers on RTP leads to increased investment. Note that in this example, the marginal cost curve is not concave.

C. The Effect on Efficiency of Increasing RTP Customers

As shown above, if all customers are on RTP, allocation and investment are efficient. When some customers are not on RTP, electricity is allocated inefficiently between the flat-rate and RTP markets. The question remains about the welfare effects of a marginal increase in the proportion of customers on RTP when $\alpha < 1$. This question is more subtle than it may appear at first glance since the welfare theorems are not applicable.[33]

To analyze the long-run welfare effects of increasing the proportion of customers on RTP, we analyze the surplus accruing to different groups: the generators, the retail service providers, the customers on RTP, the customers on flat-rate pricing, and the customers who switch from flat rates to RTP. First, the generators and retail service providers receive no surplus in the long run, so their surplus is unaffected by increasing α. Second, Result 4 shows that $\bar{p}^e_{LR}$ decreases in α. Therefore, the customers on flat-rate pricing consume more at a lower price. Thus, the flat-rate customers are better off with an increase in α.

Third, the customers who switch from the flat rate to RTP receive higher surplus. This can be shown by a revealed preferences argument. Since $\sum_{t=1}^{T} p_t D_t(\bar{p}) = \sum_{t=1}^{T} \bar{p} D_t(\bar{p})$, the

---

[31] As a limiting example, consider the case of L-shaped supply curves which would have no increase in off-peak prices if capacity is not fully utilized. See Borenstein and Holland (2003) for details of the model with L-shaped supply curves.

[32] Borenstein (2003) has calculated the effect on capacity of putting more customers on RTP in the California electricity market and found a decrease in equilibrium capacity levels.

[33] Since the competitive equilibrium is not efficient, we cannot rely on comparative statics results from a constrained optimization problem.

switchers could consume exactly the same electricity quantities as the flat rate customers choose at the exact same total bill. Since they choose to consume different quantities, they must be better off.[34]

Finally, the surplus to the customers on RTP decreases in α. To see this, first note that the envelope theorem implies that the change in consumer surplus to an RTP customer in period t is given by $-\frac{dp_t}{d\alpha}D_t(p_t)$. Thus, the change in surplus to RTP customers is

$$\alpha\frac{\mathrm{d}CS_{RTP}}{\mathrm{d}\alpha} = \sum_{t=1}^{T} -\frac{dp_t}{d\alpha}\alpha D_t(p_t) = \sum_{t=1}^{T}\frac{dp_t}{d\alpha}(1-\alpha)D_t(\bar{p}). \qquad [8]$$

where the second equality follows from [7], recognizing $\frac{d\pi^w}{d\alpha} = 0$ in the long run and recalling that $\tilde{D}(p_t, \bar{p}) = \alpha D(p_t) + (1-\alpha)D(\bar{p})$.

We can show that [8] is negative by differentiating the zero-profit retail condition: $\pi^r = (1-\alpha)\sum_{t=1}^{T}(\bar{p} - p_t)D(\bar{p}) = 0$. Differentiation implies that

$$0 = \frac{\partial \pi^r}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} + \sum_{t=1}^{T}\frac{\partial \pi^r}{\partial p_t}\frac{dp_t}{d\alpha} = \frac{\partial \pi^r}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} - \sum_{t=1}^{T}\frac{dp_t}{d\alpha}(1-\alpha)D_t(\bar{p}). \qquad [9]$$

Since the competitive equilibrium $\bar{p}$ results from Bertrand competition over the flat rates, the derivative $\frac{\partial \pi^r}{\partial \bar{p}}$ must be greater than or equal to zero. Since $\frac{d\bar{p}}{d\alpha} \leq 0$ by Result 4, $\sum_{t=1}^{T}\frac{dp_t}{d\alpha}(1-\alpha)D_t(\bar{p})$ must be less than zero. Combining this with [8] shows that the consumer surplus to the RTP customers is decreasing in α.

We have shown the long-run impact of increasing α on the four affected groups—incumbent RTP customers, "switchers," remaining flat-rate customers, and sellers. Since each group, except the incumbent RTP customers, is no worse off, the overall welfare impact depends on the ability of these groups to compensate the potential losses of the incumbent RTP customers.

Define W from [1] as the welfare attained in competitive equilibrium. The change in welfare from increasing customers on RTP is then given by

$$\frac{\mathrm{d}W(K, p_t, \bar{p}, \alpha)}{\mathrm{d}\alpha} = \frac{\partial W}{\partial K}\frac{dK}{d\alpha} + \sum_{t=1}^{T}\frac{\partial W}{\partial p_t}\frac{dp_t}{d\alpha} + \frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} + \frac{\partial W}{\partial \alpha}. \qquad [10]$$

We have shown that in the competitive equilibrium, $\frac{\partial W}{\partial K} = 0$, *i.e.,* capacity is set efficiently given the equilibrium prices. Likewise, $\frac{\partial W}{\partial p_t} = 0$ for all t, since we have explained earlier

[34] Samuelson (1972) uses a similar revealed preferences argument to show that consumers always benefit from price stabilization that leaves producers equally well off.

that realtime prices are set efficiently given the equilibrium $\bar{p}$. Thus [10] reduces to: $\frac{dW}{\mathrm{d}\alpha} = \frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} + \frac{\partial W}{\partial \alpha}$.

The last term, $\frac{\partial W}{\partial \alpha}$, is the direct welfare gain from customers switching from flat-rate to RTP and can be written as

$$\frac{\partial W}{\partial \alpha} = \sum_{t=1}^{T}[U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - p_t D_t(\bar{p})], \qquad [11]$$

which is positive by the revealed preference argument made above. Result 4 shows that $\frac{d\bar{p}}{d\alpha} \leq 0$, and in section I, we showed that $\frac{\partial W}{\partial \bar{p}}$ can be positive or negative depending on whether $\bar{p}^e$ is greater or less than $\bar{p}^*$.[35] Thus if decreasing $\bar{p}$ improves welfare, then increasing α improves efficiency. However, if decreasing $\bar{p}$ decreases efficiency, then the welfare effects depend on whether or not the gains to the switchers are greater than the losses from decreasing $\bar{p}$.[36] To summarize,

Result 6: Welfare Effects of Increasing RTP Customers — *In the long run, an increase in the proportion of customers on RTP (i) increases consumer surplus of customers remaining on flat-rate service, (ii) increases consumer surplus of customers switching from flat-rate to RTP, and (iii) decreases consumer surplus of incumbent RTP customers, and (iv) has no effect on generator or retailer profits, . Total welfare increases with an increase in the proportion of customers on RTP if $\bar{p}^e > \bar{p}^*$, but welfare may decrease if $\bar{p}^e < \bar{p}^*$, the case in which lowering the equilibrium flat rate reduces efficiency. Welfare always increases (and is maximized) by putting all customers on RTP.*

Proof: *i-iv* are proved in the text. Since $\frac{dW}{\mathrm{d}\alpha} = \frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} + \frac{\partial W}{\partial \alpha}$ and $\frac{\partial W}{\partial \alpha} > 0$, if $\bar{p}^e > \bar{p}^*$, so that $\frac{\partial W}{\partial \bar{p}} < 0$, then $\frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} > 0$ and increasing α increases total welfare. If $\bar{p}^e < \bar{p}^*$, then $\frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} < 0$. Since $\frac{\partial W}{\partial \alpha} > 0$, the net impact on welfare in this case is ambiguous. In the appendix, we demonstrate how examples with $\frac{dW}{d\alpha} < 0$ can be constructed.∎

The welfare effects of Result 6 depend on the ability of the switchers and flat-rate customers to compensate the losses of the customers on RTP. If lowering $\bar{p}$ increases welfare, then $\frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} > 0$ and the flat-rate customers can compensate the RTP customers. On the other hand, if $\frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} < 0$, then the flat-rate customers cannot compensate the RTP customers, and the welfare effects depend on the ability of the switchers to compensate the net loss. The surprise of Result 6 is that sometimes the switchers cannot compensate the other customers.

---

[35] This assumes that the profit function is single-peaked.

[36] As explained earlier, if all demands have the same elasticity at $\bar{p}$, then $\bar{p}^e = \bar{p}^*$, so $\frac{\mathrm{d}W}{\mathrm{d}\alpha} > 0$.

In the appendix, we construct an example in which increasing α lowers welfare. We know, however, from section I that increasing α to 1 from any lower value increases welfare. Moreover, we know that the welfare attained in competitive equilibrium is continuous in α even at $\alpha = 1$. So, the example in the appendix demonstrates the increase in welfare need not always be monotonic as it moves to the maximum welfare at $\alpha = 1$.[37]

D. RTP Adoption in Competitive Markets

Thus far, we have assumed that α is set exogenously, ignoring the incentives customers would have to adopt RTP if such programs were voluntary. In a voluntary system, each customer would balance the potential gains from RTP against the metering costs.[38] We now consider the incentives of customers to adopt RTP under competition.

We assume that customers adopting RTP must pay, directly or indirectly, for the additional metering and billing costs and that these costs are independent of quantity consumed.[39] Let M be the additional daily cost (variable plus amortized fixed cost) of metering and billing one customer when that customer switches to RTP.[40]

Assume, for now, that each customer constitutes a share γ of the total demand, where γ is very small. If we assume that customers can avoid this metering cost by choosing the flat rate service, then customers will switch to RTP until in equilibrium we have:

$$\gamma \sum_{t=1}^{T} [U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - \bar{p} D_t(\bar{p})] = M. \qquad [12]$$

[12] determines the equilibrium share of customers on RTP, α. The long-run competitive equilibrium with customer choice over rate structure is then fully described by [12] plus the conditions described in the characterization of the long-run equilibrium above.

As above we can define W as the welfare attained in the competitive equilibrium where now W incorporates the metering cost $\frac{\alpha}{\gamma} M$, *i.e.,* the costs of metering the RTP customers.

---

[37] Simulations with linear demands (in which $\frac{\partial W}{\partial \bar{p}} \gg 0$) and simulations presented in Borenstein (2003), which use actual California system load profiles and assume constant elasticity demand with higher elasticities in peak periods, showed no cases in which welfare declined with an increase in α.

[38] Here and throughout the analysis, we have ignored the potential for price volatility to lower the welfare of RTP customers due to risk aversion. Borenstein (forthcoming) explains how forward contracts can be used to mitigate these risks.

[39] Though costs do vary slightly with the size of customer demand, this is a reasonable approximation. See Jaske (2002).

[40] We continue to assume that the available flat retail rate, realtime prices and investment are all set competitively as described in the characterization of competitive equilibrium in section I.

From a long-run competitive equilibrium, differentiating W as in [10] yields the following result:

Corollary 3: Non-optimality of Competitive RTP Selection — *If metering costs are positive and customers choose between flat rates or realtime prices, then competition leads to excessive use of RTP if $\bar{p}^e < \bar{p}^*$. If $\bar{p}^e > \bar{p}^*$, RTP is used less than is optimal.*

Proof: The partial derivative of W with respect to α is now:

$$\frac{\partial W}{\partial \alpha} = \sum_{t=1}^{T}\{[U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - p_t D_t(\bar{p})]\} - \frac{M}{\gamma}$$

$$= \sum_{t=1}^{T}\{[U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - \bar{p} D_t(\bar{p})]\} - \sum_{t=1}^{T}(\bar{p} - p_t)D_t(\bar{p}) - \frac{M}{\gamma} = 0. \quad [13]$$

The first equality follows from differentiation of W as in [11], the second equality is algebra, and the third equality follows from [12] and the condition on retail profit. Since $\frac{\partial W}{\partial \alpha} = 0$, it follows that $\frac{dW}{\mathrm{d}\alpha} = \frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha}$. Since $\frac{d\bar{p}}{d\alpha} < 0$ from Result 4, increasing metering beyond the competitive level increases welfare if and only if $\frac{\partial W}{\partial \bar{p}} < 0$. ∎

Corollary 3 obtains because customers do not recognize that by switching to RTP they drive down the flat rate for the remaining customers. If the flat rate is higher than optimal, this externality is beneficial, and too few customers switch to RTP. On the other hand, if the flat rate is too low, then the externality is harmful, and too many customers switch to RTP.[41]

If customers differ in their size, but still have identical demands up to a scale parameter, we can represent each customer i as constituting γ_i of total demand. Since metering costs are independent of the scale parameter, [12] implies that the customers with the largest γ_i would be the first to switch to RTP. However, the marginal customer still would not consider the effect on $\bar{p}$ of his decision to switch, and switching could be excessive or insufficient.

We leave for future research an in-depth analysis of outcomes when customers have different demand profiles, but it seems clear that the incentive to switch is further complicated in two ways. First, an *elasticity effect* will cause customers with more elastic demand to be more inclined to switch. For instance, if two customers on flat-rate service demand

[41] Brennan (2002) and Doucet and Kleit (2003) do not recognize this externality in their analyses of competitive adoption of realtime pricing.

the same quantities in each period, but one has much more elastic demand in all periods, then that customer has a much stronger incentive to switch to RTP. This welfare effect, however, seems to be fully captured by the switcher. Second, the *adverse-selection effect* will cause customers who have relatively lower demands at peak times and relatively higher demands at off-peak times to be more inclined to switch. For these customers, even if they made no change to their purchasing, they would pay less on RTP. The adverse-selection effect, however, is just a transfer from customers with "peakier" demands who are subsidized under flat-rate pricing. This transfer, which doesn't by itself change total surplus, gives some customers inefficiently large incentives to pay M in order to move to RTP.[42] The adverse-selection effect will tend to cause $\bar{p}$ to rise as these customers switch, which may or may not outweigh the tendency for $\bar{p}$ to fall with switching when customers are identical. If $\bar{p}$ were to rise, this likely could raise or lower welfare, depending on whether $\bar{p}^e$ is greater or less than $\bar{p}^*$.

IV. Conclusion

Electricity deregulation has proceeded with support from many economists on the belief that competitive electricity markets will produce more efficient outcomes than regulation. That still may turn out to be true, though in many locations, most notably California, there is significant evidence that the markets have not been sufficiently competitive. Even if market changes succeed in making the markets competitive, however, we have shown that flat-rate pricing of a significant share of retail customers will remain a barrier to achieving efficient outcomes. Not only does flat-rate retail pricing have the obvious problem of preventing hour-by-hour prices that reflect wholesale costs, flat-rate pricing in a competitive market fails to achieve even the second-best optimum of the welfare-maximizing flat-rate price. As a result, we have shown that capacity investment will in general differ from the second-best optimal level.

In order to assure adequate capacity investment, many market participants and advisors have argued for "capacity payments," which are effectively subsidies that reduce the cost of owning capacity and, thus, increase equilibrium investment. We have demonstrated that capacity subsidies (or taxes) cannot achieve the second-best optimum, because they create other distortions as they address the distortion caused by flat-rate customers. Furthermore, capacity investment distortion under flat-rate pricing can lead to either excessive or insufficient investment. We also examine taxes or subsidies on retail electricity

[42] Borenstein (1989) develops a similar argument for why competitive insurance markets will use some costly risk-screening tests whose net effect is to lower total welfare.

as a policy response to the inefficiency caused by flat-rate pricing. A tax or subsidy on the flat-rate customers alone can indeed achieve the second-best optimal flat-rate price and capacity investment, but a tax or subsidy that applies to all customers—flat-rate and those on RTP—will distort the RTP market, so it will not achieve the second-best optimum.

Many economists and some industry participants have argued strongly for increasing the proportion of customers on RTP. We have shown that while increasing the proportion of customers on RTP is likely to increase market efficiency, exceptions are possible at least for some (locally) extreme shapes of demand functions. We have also demonstrated that increases in the share of customers on RTP can harm customers who are already on RTP, while benefitting those who remain on flat rates. The net effect of such a change on the level of equilibrium capacity, we demonstrate, is ambiguous.[43]

We've modeled the flat-rate retail price problem in the context and institutions of deregulated electricity markets, but the application is much broader.[44] In many markets, retail prices cannot or at least do not fluctuate to reflect changes in market and cost conditions. This is broadly recognized, but there seems to be a view that competitive determination of some sort of smoothed or average retail price allows the welfare analysis of competitive markets to go through at least approximately. Our results suggest that this isn't the case, that competitive determination of retail prices that are constrained not to adjust as frequently as costs will not achieve a second-best optimum.

In the general context of sticky prices, we have presented a different view of how markets may operate than presented by Carlton (1986) and others who examine non-price rationing. In those models, all prices are sticky and therefore non-price rationing is used to distribute the product. In our approach, prices are sticky to some customers and the remaining customers face a residual supply for which price is very volatile. Which model is more appropriate will depend on the specific institutions of a market.

---

[43] Like much of the peak-load pricing literature, we have made certain restrictive assumptions to simplify this analysis. We have assumed that there is no cross-elasticity of demand across periods, that all generation technology is identical, that all customers have identical distributions of demands across periods, and that demand has no stochastic component. Relaxing these assumptions, as we intend to do in future work, will introduce other influences on equilibrium outcomes, but is unlikely to alter the basic insights of this analysis.

[44] The flat rate we've studied is not specific to electricity markets and can represent any requirements contract, *i.e.,* a contract where a firm agrees to supply any quantity demanded at a specified price. Our results suggest use of such requirements contracts may have greater adverse efficiency effects than is generally recognized.

Appendix

Result 4: Effect of Increasing RTP Customers on Flat Retail Rate — *In the long run, an increase in the proportion of customers on RTP (α) reduces $\bar{p}^e_{LR}$.*

Proof: We demonstrate this proposition by evaluating the long-run change in retail profits, π^r, caused by a change in α, holding $\bar{p}$ constant. We show that retailer profits would increase, if $\bar{p}$ did not drop. Thus, competition in the retail sector reduces $\bar{p}$.

We wish to evaluate $\frac{d\pi^r}{d\alpha}$ holding $\bar{p}$ constant. Since $\bar{p}$ is constant, $\frac{d\pi^r}{d\alpha} = (1-\alpha)\sum_t -D_t(\bar{p})\frac{dw_t}{d\alpha}$ is a weighted average of $\frac{dw_t}{d\alpha}$.

First note that competitive investment implies that in the long run

$$0 = \frac{d\pi^w}{d\alpha} = \sum_t \tilde{D}(p_t, \bar{p})\frac{dw_t}{d\alpha} = K\sum_t S(w_t)\frac{dw_t}{d\alpha}$$

where $S(w_t)$ is the unit supply curve.

Next note that

$$\alpha D_t(w_t)+(1-\alpha)D_t(\bar{p}) = KS(w_t) \qquad \Longleftrightarrow \qquad D_t(w_t)-D_t(\bar{p}) = \frac{KS(w_t)-D_t(\bar{p})}{\alpha}. \quad (A1)$$

Differentiating the left-hand equation in (A1) with respect to α gives:

$$D_t(w_t) - D_t(\bar{p}) + \alpha D_t'(w_t)\frac{dw_t}{d\alpha} + (1-\alpha)D_t'(\bar{p})\frac{d\bar{p}}{d\alpha} = KS'(w_t)\frac{dw_t}{d\alpha} + S(w_t)\frac{dK}{d\alpha}. \quad (A2)$$

Recognizing that $\frac{d\bar{p}}{d\alpha} = 0$ by assumption and substituting using the right-hand equation in (A1), (A2) can be rearranged as:

$$\alpha[KS'(w_t) - \alpha D_t'(w_t)]\frac{dw_t}{d\alpha} = [K - \alpha\frac{dK}{d\alpha}]S(w_t) - D_t(\bar{p}). \quad (A3)$$

Since $[KS'(w_t)-\alpha D_t'(w_t)] > 0$, it follows that $\frac{dw_t}{d\alpha} > 0$ if and only if $[K-\alpha\frac{dK}{d\alpha}]S(w_t)-D_t(\bar{p}) > 0$, and that the product $\{[K-\alpha\frac{dK}{d\alpha}]S(w_t) - D_t(\bar{p})\}\frac{dw_t}{d\alpha}$ is positive for all t. This implies that their sum is also positive. But this implies that

$$0 < \sum_t \{[K-\alpha\frac{dK}{d\alpha}]S(w_t) - D_t(\bar{p})\}\frac{dw_t}{d\alpha} = \sum_t -D_t(\bar{p})\frac{dw_t}{d\alpha}$$

where the equality holds because $[K-\alpha\frac{dK}{d\alpha}]\sum_t S(w_t)\frac{dw_t}{d\alpha} = 0$. ∎

Result 5: Indeterminant Effect of Increasing RTP Customers on Capacity — *An increase in the proportion of customers on RTP can increase or decrease long-run equilibrium capacity* K^e_{LR}.

Proof: Consider a long-run competitive equilibrium with two time periods, peak and off-peak. Note that the short-run equilibrium does not depend on the shape of the marginal cost curve C_q but only on the equilibrium marginal costs. Similarly, the long-run equilibrium would not change if we perturbed C_q without changing the equilibrium marginal costs or the sum of the total costs. Thus, if we increased the convexity of C_q such that $C_q(\tilde{D}_{op}(p_{op},\bar{p}),K)$, $C_q(\tilde{D}_p(p_p,\bar{p}),K)$, and $C(\tilde{D}_{op}(p_{op},\bar{p}),K)+C(\tilde{D}_p(p_p,\bar{p}),K)$ did not change, then the long-run equilibrium would not change.

Note that $\frac{dp_t}{d\alpha}=C_{qq}\frac{d\tilde{D}_t}{d\alpha}$ in the short run, which implies that $\frac{dp_{op}}{d\alpha}>0$ and $\frac{dp_p}{d\alpha}<0$. Starting from a long-run equilibrium, we can increase the convexity of C_q without changing the long-run equilibrium if $C_{qq}(\tilde{D}_{op}(p_{op},\bar{p}),K)>0$. By increasing the convexity of C_q without changing the long-run equilibrium, we can make $C_{qq}(\tilde{D}_{op}(p_{op},\bar{p}),K)$ smaller and $C_{qq}(\tilde{D}_p(p_p,\bar{p}),K)$ larger. But this implies that $\frac{dp_{op}}{d\alpha}$ is less positive and that $\frac{dp_p}{d\alpha}$ is more negative. Thus [7] can be negative. Similarly, an example can be constructed where [7] is positive by increasing the concavity of C_q. An example of a capacity increase follows. ∎

Example of an Increase in RTP Customers that Increases Capacity

To show that increasing the proportion of customers on RTP can lead to increased investment, consider a parallel linear demand model with linear marginal costs. Let $D_t(p)=A_t-Bp$ and $C_q=q/K$. Since supply equals demand in every period, $p_t=[A_t-B(1-\alpha)\bar{p}]/(K+B\alpha)$, which implies that

$$\bar{p}-p_t=[(K+B)\bar{p}-A_t]/(K+B\alpha)=Y_t/(K+B\alpha) \qquad (A4)$$

where $Y_t\equiv(K+B)\bar{p}-A_t$. This implies retail profits can be written $\pi^r=f(\alpha)\sum Y_tD_t(\bar{p})$ where $f(\alpha)=\frac{(1-\alpha)}{K+B\alpha}$. Since $f(\alpha)\neq 0$, in short-run equilibrium, $\sum Y_tD_t(\bar{p})$ must equal zero. But since $\sum Y_tD_t(\bar{p})$ does not depend on α, it is also zero when α increases, *i.e.,* putting more customers on RTP does not change the short-run equilibrium flat rate.

Now consider how the short-run wholesale profits change with changes in α. Differentiating (A4) and noting that the short-run flat rate and Y_t do not depend on α implies that $\mathrm{d}p_t/\mathrm{d}\alpha=BY_t/(K+B\alpha)^2$. By the envelope theorem, the change in wholesale profits is $\sum(\mathrm{d}p_t/\mathrm{d}\alpha)\tilde{D}_t$ which implies that wholesale profits increase or decrease depending on whether $\sum Y_t\tilde{D}_t$ is positive or negative. From (A4), Y_t is positive iff $\bar{p}>p_t$ which occurs if and only if $\tilde{D}_t(p_t,\bar{p})>D_t(\bar{p})$. Therefore $\sum Y_t\tilde{D}_t>\sum Y_tD_t(\bar{p})$ since the first weighted

average of the Y_t puts more weight on each positive Y_t and less weight on each negative Y_t. Since $\sum Y_t D_t(\bar{p}) = 0$, the first weighted average is positive and the short-run wholesale profits increase with α.

Example of an Increase in RTP Customers that Decreases Welfare

We have shown that $\frac{dW}{\mathrm{d}\alpha} = \frac{\partial W}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha} + \frac{\partial W}{\partial \alpha}$. We construct an example in which $\frac{dW}{d\alpha}$ can be negative by showing that the second term, which is positive, can be made arbitrarily small while holding the first term, which can be negative, constant.

First, recall that the competitive equilibrium is characterized completely by p_t, $\bar{p}$, α, r, K, the unit supply function S, and the demand functions D_t. Note, however, that the equilibrium does not depend on the entire demand functions, but rather only on two points, $D_t(p_t)$ and $D_t(\bar{p})$, of each demand function. Thus, any system of demand equations which does not change these $2T$ points (nor α, S, or r) will have an equilibrium with the same prices and capacity.

Next, consider $\frac{d\bar{p}}{d\alpha}$, $\frac{dp_t}{d\alpha}$, and $\frac{dK}{d\alpha}$. By the Implicit Function Theorem, these derivatives can be found by totally differentiating the system of equations that characterize the competitive equilibrium. This implies that $\frac{d\bar{p}}{d\alpha}$ can be written as a function of the $7T + 4$ parameters: p_t, $D_t(p_t)$, $D_t'(p_t)$, $D_t(\bar{p})$, $D_t'(\bar{p})$, $\bar{p}$, α, $S(p_t)$, $S'(p_t)$, r, and K. Since $\frac{\partial W}{\partial \bar{p}}$ can also be written in terms of these $7T + 4$ parameters, the product $\frac{\partial W}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha}$ would not change if we were to perturb the demand curves such that the demands and slopes at p_t and $\bar{p}$ were unchanged.

Now consider $\frac{\partial W}{\partial \alpha}$. [11] can be written

$$\frac{\partial W}{\partial \alpha} = \sum_{t=1}^{T} [U_t(D(p_t)) - U_t(D(\bar{p}))] - p_t[D_t(p_t) - D_t(\bar{p})]. \qquad (A5)$$

Note that the summands in (A5) are always positive. For example, if $p_t > \bar{p}$, the difference $U_t(D(p_t)) - U_t(D(\bar{p}))$ is negative but it is smaller in absolute value than $-p_t[D_t(p_t) - D_t(\bar{p})] > 0$. Conversely, if $p_t < \bar{p}$, the difference $U_t(D(p_t)) - U_t(D(\bar{p}))$ is positive and larger (in absolute value) than $-p_t[D_t(p_t) - D_t(\bar{p})] < 0$. Note, however, that these summands depend on the shape of the demand curve between $D_t(p_t)$ and $D_t(\bar{p})$. This implies that the summands can be made arbitrarily small by making the demands more concave (convex) for p_t above (below) $\bar{p}$ while holding constant the $D_t(p_t)$, $D_t'(p_t)$, $D_t(\bar{p})$, $D_t'(\bar{p})$.[45]

[45] In the case of $p_t > \bar{p}$, for instance, the welfare gain from switchers would be arbitrarily small—without changing slopes or demands at p_t and $\bar{p}$—if demand were a concave right angle between p_t and $\bar{p}$,

Finally, consider any equilibrium where $\frac{\partial W}{\partial \bar{p}} > 0$. By perturbing the demand curves between $D_t(p_t)$ and $D_t(\bar{p})$ without changing $D_t(p_t)$, $D'_t(p_t)$, $D_t(\bar{p})$, or $D'_t(\bar{p})$, the term $\frac{\partial W}{\partial \alpha}$ can be made arbitrarily small without changing $\frac{\partial W}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha}$.

This example is obviously an extreme case since it relies on making the gains to switchers arbitrarily small by making peak demand curves concave and off-peak demand curves convex. Our simulations and empirical work have failed to generate this situation, but further work is required to understand the policy relevance of this example.

i.e., if demand were identical to $D_t(p)$ for $p > p_t - \epsilon$ and for $p < \bar{p} + \epsilon$ but were constant at $D_t(\bar{p} + \epsilon)$ for $p \in [\bar{p} + \epsilon, p_t - \epsilon]$. Although this demand curve would be discontinuous at $D(p_t - \epsilon)$, continuous examples could be similarly constructed.

REFERENCES

Bergstrom, Ted and Jeffrey K. MacKie-Mason. "Some Simple Analytics of Peak-Load Pricing" *RAND Journal of Economics*, Summer 1991, **22**, 2, 241-49.

Boiteaux, Marcel. "La tarification des demandes en point: application de la théorie de la vente au coût marginal." *Revue Général de l'Electricité*, August 1949, **58**, 321-40, translated as "Peak Load Pricing." *Journal of Business*, April 1960, **33**, 157-179.

Borenstein, Severin. "The Economics of Costly Risk Sorting in Competitive Insurance Markets," *International Review of Law and Economics*, **9**(June 1989).

Borenstein, Severin. "Time-Varying Retail Electricity Prices: Theory and Practice," in Griffin and Puller, eds., *Electricity Deregulation*, Chicago: University of Chicago Press, forthcoming.

Borenstein, Severin. "Estimating the Long-Run Impact of Real-time Pricing," Center for the Study of Energy Markets Working Paper, University of California Energy Institute, August 2003.

Borenstein, Severin and Stephen P. Holland. "Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices," CSEM Working Paper CSEMWP-106, University of California Energy Institute, revised July 2003. Available at http://www.ucei.org/PDF/csemwp106r.pdf.

Carlton, Dennis. "The Rigidity of Prices." *American Economic Review*, 1986 **76**, 4, 637-658.

Carlton, Dennis. "Peak Load Pricing with Stochastic Demand." *American Economic Review*, 1977 **67**, 5, 1006-1010.

Chao, Hung-po. "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty." *Bell Journal of Economics*, 1983, **14**(1), 179-190.

Crew, Michael, Chitru S. Fernando, Paul R. Kleindorfer. "The Theory of Peak-Load Pricing: A Survey." *Journal of Regulatory Economics*, November 1995, **8** 3, 215-248.

Dana, James. "Using Yield Management to Shift Demand When the Peak Time Is Unknown." *RAND Journal of Economics*, Autumn 1999, **30**, 3, 456-474.

Doucet, Joseph A. and Andrew Kleit, "Metering in Electricity Markets: When is More Better?" *Markets, Pricing, and Deregulation of Utilities* (Michael A. Crew and Joseph C. Schuh, editors), Kluwer, 2003.

Jaske, Michael. "Practical Implications of Dynamic Pricing," in Severin Borenstein, Michael Jaske and Arthur Rosenfeld, *Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets*, October 2002. Available at http://www.ucei.org/PDF/csemwp105.pdf.

Panzar, John C. "A Neoclassical Approach to Peak-Load Pricing." *Bell Journal of Economics*, Autumn 1976, **7**(2), 521-530.

Panzar John C. and David S. Sibley. "Public Utility Pricing under Risk: The Case of Self-Rationing." *American Economic Review*, 1978, **68** (5), 888-895.

Samuelson, Paul R. "The Consumer Does Benefit From Feasible Price Stability." *Quarterly Journal of Economics*, August 1972, **86** (3), 476-493.

Steiner, Peter O. "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics*, November 1957, **72**(1), 585-610.

Wenders, John T. "Peak Load Pricing in the Utility Industry." *Bell Journal of Economics*, Spring 1976, **7**(1), 232-241.

Williamson, Oliver E. "Peak Load Pricing and Optimal Capacity Under Indivisibility Constraints." *American Economic Review*, September 1966, **56**(4), 810-827.

Williamson, Oliver E. "Peak Load Pricing: Some Further Remarks." *Bell Journal of Economics and Management Science*, Spring 1974, **5**(1), 223-228.

Figure 1

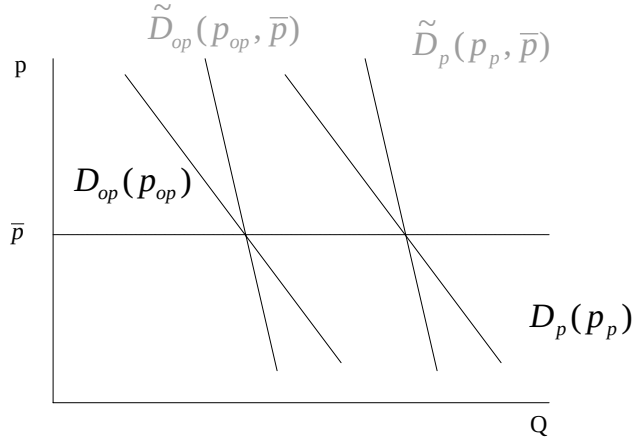

Figure 2

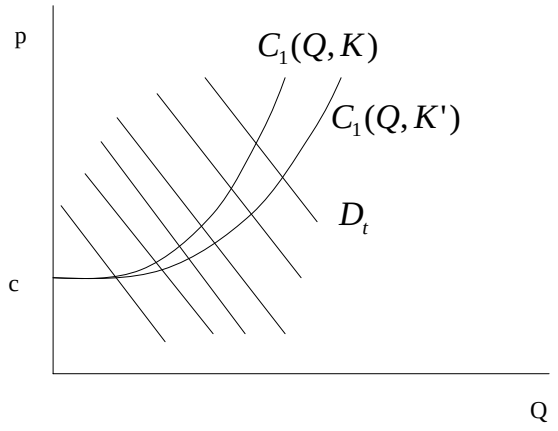

Figure 3

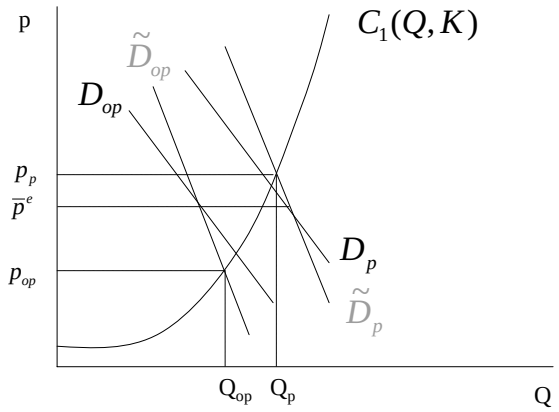

Figure 4