RAND Journal of Economics
Vol. 36, No. 3, Autumn 2005
pp. 469–493

On the efficiency of competitive electricity markets with time-invariant retail prices

Severin Borenstein*

and

Stephen Holland**

Most customers in electricity markets do not face prices that change frequently to reflect changes in wholesale costs, known as real-time pricing (RTP). We show that not only does time-invariant pricing in competitive markets lead to prices and investment that are not first best, it even fails to achieve the constrained second-best optimum. 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. Simulations demonstrate that the efficiency gains from RTP are potentially quite significant.

1. Introduction

■ In many industries, retail prices do not adjust quickly to changes in costs or market conditions. Restaurants keep stable menu prices on most dishes even when ingredient prices fluctuate.[1] Service providers, from barbers to veterinarians, regulate fluctuating demand with nonprice 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

* University of California Energy Institute and University of California at Berkeley; borenste@haas.berkeley. edu.

** University of North Carolina at Greensboro; sphollan@uncg.edu.

For helpful comments and discussions, we thank James Bushnell, Joseph Farrell, Joseph Harrington (Editor), Morten Hviid, Paul Joskow, Erin Mansur, Michael Riordan, Lawrence White, two anonymous referees, and seminar participants at UC Berkeley (2002), the UC Energy Institute (2002), Columbia University/NYU (2003), the 2003 Econometric Society Summer Meetings in Evanston, IL, the 2003 International IO Conference in Boston, and the 2003 Center for Research in Regulated Industries Western Conference in San Diego, CA.

[1] For items with highly volatile ingredient costs, however, some restaurants list only "market price" on the menu, indicating that the benefits of time-invariant pricing are outweighed in those cases by the ability to quickly change price. We return to this endogeneity of variable pricing later in the article.

production or wholesale acquisition costs, have been among the most vocal proponents of real-time 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 real-time prices and others see time-invariant prices, more commonly called flat-rate service. In this article 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.[3,4]

Closely tied to time-invariant retail pricing is the issue of investment adequacy.[5] 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 real-time 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 commonly proposed policies cannot overcome the inefficiency caused by suboptimal investment.

We then analyze the impact of expanding the use of RTP. We show that if customers have homogeneous 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. In the following section, we present preliminary estimates of the magnitude of the distortion resulting from flat-rate pricing in electricity. Using realistic parameters for costs and demand, we find that a lower-bound estimate of the inefficiency is probably 5–10% of wholesale energy costs.

We focus in this article 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

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

[3] Numerous studies have analyzed market power and market design issues in restructured electricity markets, such as Joskow and Kahn (2002) and Borenstein, Bushnell, and Wolak (2002). We analyze the efficiency of competitive markets in the absence of these other potential distortions.

[4] We analyze inefficiencies resulting from prices being fixed over time. Pricing distortions also occur because wholesale prices vary by location, due to transmission constraints, but retail prices frequently do not reflect that variation. Analysis of prices that are fixed across locations is beyond the scope of this article.

[5] Concerns about investment adequacy also stem from regulatory inefficiencies such as price caps, regulatory credibility, and restrictions in siting power plants. These regulatory inefficiencies would not be relevant in deregulated, competitive markets.

has significant implications for capital investment and long-run efficiency, particularly in service industries and other markets with little or no ability to carry inventories.[6]

We begin in Section 2 by presenting a model of competitive wholesale and retail electricity markets in which some share of customers is able to be charged real-time electricity prices. We demonstrate the short-run pricing and long-run investment inefficiency that results from the inability to charge all customers real-time prices. We also show that subsidies or taxes that have been suggested to overcome these inefficiencies cannot generally correct the problem. In Section 3 we examine the welfare effects of changing the proportion of customers on RTP and the customer's incentives to switch to RTP. In Section 4 we describe the basic simulations we have carried out to evaluate the magnitude of these inefficiencies and show that they are likely to be significant compared to the total costs of operating the system. We conclude in Section 5.

2. 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 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 $D_t' < 0$.[7] A fraction, α, of the customers pay real-time 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 real-time pricing do not differ systematically from those on flat-rate pricing.[8,9] 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 real-time 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 real-time 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 wholesale demand curves if everyone were on RTP, D_t (solid lines), and if $1 - \alpha$ share of customers were on flat-rate service, $\tilde{D}_t$ (dashed lines), where there are only two periods: the high-peak period, H, and the low off-peak period, L. Note that the less-elastic curves are the aggregate wholesale demand when some customers are on flat-rate service. For prices above $\bar{p}$, wholesale quantity demanded is greater than the quantity demanded if everyone were on real-time prices, since the flat-rate customers do not decrease consumption in response to the higher real-time price. Similarly, for prices below $\bar{p}$, wholesale quantity demanded is less than the quantity demanded if everyone were on real-time prices, since the flat-rate customers do not increase consumption in response to the lower real-time 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 that the

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

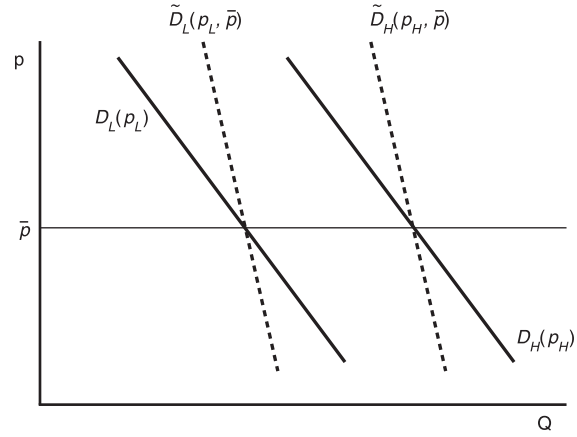
[7] Following the literature on peak-load pricing, we assume that cross-price elasticities between demands in different periods are zero but discuss extending the analysis to nonzero cross-elasticities. Bergstrom and MacKie-Mason (1991) allow nonzero cross-elasticities but assume homothetic preferences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

[8] Although some of the results, including Theorems 1a and 1b, hold for $\alpha = 0$, we assume $\alpha > 0$ for expositional ease. We do not consider peak-load pricing with stochastic demand (see Carlton, 1977; Panzar and Sibley, 1978; and Chao, 1983), but we note here that with $\alpha = 0$, demand may not be met at all times.

[9] Throughout, we assume that customers on RTP are risk neutral with respect to the price of electricity. We discuss hedging of price risk later.

FIGURE 1

WHOLESALE DEMAND CURVES WITH AND WITHOUT SOME CUSTOMERS ON FLAT RATES



marginal costs of each generator are continuous and increasing in output.[10] 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 variable 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

(i) $C_q > 0$, for a given K, increasing generation output increases variable costs;

(ii) $C_k < 0$, variable costs of generating a given quantity of electricity is lower with more installed capacity;

(iii) $C_{qq} > 0$, short-run marginal costs are increasing in quantity;

(iv) $C_{kk} > 0$, 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

(v) $C_{qk} < 0$, 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 so that, since all firms are identical, $w_t = C_q(q, K)$, where w_t is the wholesale price in period t. Thus, the short-run industry supply curve is upward sloping.[11] 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 that increases capacity from K to K' lowers the marginal cost of generation and lowers the market-clearing price in each period.

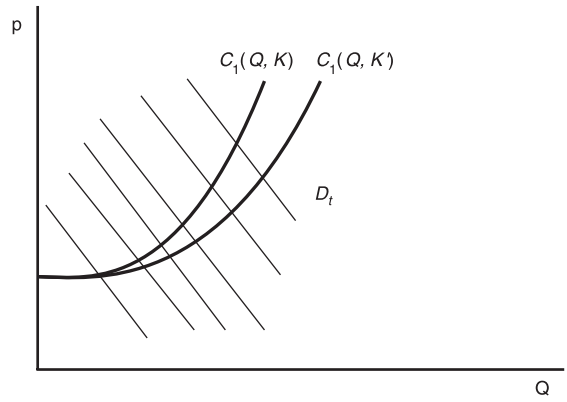
In the long run, generators can add or retire capacity. Assume that the cost per unit of capacity is r per day. If q_t megawatts of electricity are generated in period t, industry profits for the generators are $\sum_{t=1}^{T}[w_t q_t - C(q_t, K)] - rK$ per day. Since each firm has identical technology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

[10] We show similar results for reverse-L-shaped marginal cost curves (constant marginal cost up to a capacity constraint) in Borenstein and Holland (2003a). We assume continuity and nonzero derivatives here for ease of exposition.

[11] The assumption of identical technologies implies that the industry supply curve, found by inverting $w_t = C_q(q, K)$, 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.

FIGURE 2

SHORT-RUN INDUSTRY SUPPLY CURVES WITH DIFFERENT GENERATION CAPACITIES



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.[12] The retail firms choose real-time 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.[13] Since electricity cannot be stored economically, demand greater than capacity in any period would require nonprice 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.[14]

□ **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 real-time 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 real-time price is $p^e_{tSR} = w_t$ 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 real-time wholesale prices where the weights are the relative quantities demanded by the customers facing a flat retail

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

[13] We assume that the system operator can bill each retailer for its customers' real-time consumption. If the customer doesn't have a real-time meter, the operator must use an assumed pattern of demand, known as a load profile. Joskow and Tirole (2004) show that the equilibrium is not substantially affected by load profiling.

[14] We have assumed here that retailers charge linear tariffs. Joskow and Tirole (2004) analyze a similar model and show that the equilibrium two-part tariff under load profiling, i.e., without real-time meters, has no fixed charge.

price. Thus, competition among retailers drives $\bar{p}^e_{SR}$ to be equal to the demand-weighted average wholesale price.[15]

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 $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$.[16] We then have the following.

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

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. Due to symmetry, this can be written as a zero-profit condition on the wholesale sector, i.e., $\sum_{t=1}^T [w_t q_t - C(q_t, K)] - rK = 0$. Thus, we have the following.

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 $\sum_{t=1}^T [w^e_t \tilde{D}_t(p^e_t, \bar{p}) - C(\tilde{D}_t(p^e_t, \bar{p}), K)] = rK$.[17]

□ **(In)efficiency of competitive equilibrium.** The First Welfare Theorem ensures efficiency of the competitive equilibrium under certain conditions.[18] 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 real-time prices or with producers, because 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 real-time 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 $w^e_t = C_q(D_t(p^e_t), K)$ implies that the marginal cost of production is equal to the wholesale price in every period. Since all customers are on real-time 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 $\sum_{t=1}^T -C_k(D_t(p_t), K) = r$.[19] 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

[15] Existence of the equilibrium can be shown since (i) retail profits are continuous in $\bar{p}$, (ii) retail profits are negative for $\bar{p} = 0$, and (iii) retail profits are positive if $\bar{p}$ is equal to the highest wholesale price that occurs during the time period.

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

[17] In Borenstein and Holland (2003b) we show that the equilibrium flat rate is always feasible in the short and long run.

[18] These conditions include market completeness and the absence of externalities, market power, and asymmetric information.

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

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 customers are on real-time pricing.

If some customers do not face the real-time 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 the 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 real-time 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

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

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.

We refer to the result of this optimization as the *second-best optimal allocation*.[20] The optimization can be described by two first-order conditions.

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

$$\alpha\{U_t'(D_t(p_t)) \cdot D_t'(p_t) - C_q(\tilde{D}_t(p_t, \bar{p}), K) \cdot D_t'(p_t)\} = 0, \tag{2}$$

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

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

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

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

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

which implies

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

Thus, the flat retail price that minimizes the deadweight loss is a weighted average of the real-time prices where the weights are the relative slopes of the demand curves.[21] Since $\bar{p}^e_{SR}$ is also a weighted average of the real-time prices but with different weights, we have the following.

[20] 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.

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

Theorem 1a (nonattainment of the second best in the short run). The short-run competitive equilibrium does not in general 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 $D_t(\bar{p}) < KS(\bar{p})$. *Q.E.D.*

To illustrate that the equilibrium flat retail price may be either too high or too low, consider a simple example with two time periods: high-demand peak, H, and low-demand off-peak, L. Clearly, the competitive-equilibrium flat rate is less than the peak real-time price and greater than the off-peak price. If peak demand were perfectly inelastic, i.e., if $D'_H = 0$, and if off-peak demand exhibited some elasticity, 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. Conversely, if $D'_L = 0$ and $D'_H < 0$, 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.

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

$$\frac{\bar{p}}{D_i(\bar{p})}D'_i(\bar{p}) = \frac{\bar{p}}{D_j(\bar{p})}D'_j(\bar{p}) \quad \Leftrightarrow \quad \frac{D'_i(\bar{p})}{D_i(\bar{p})} = \frac{D'_j(\bar{p})}{D_j(\bar{p})} \quad \Leftrightarrow \quad \frac{D'_i(\bar{p})}{D'_j(\bar{p})} = \frac{D_i(\bar{p})}{D_j(\bar{p})}. \tag{6}$$

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 $D'_i(\bar{p})/D'_j(\bar{p}) > D_i(\bar{p})/D_j(\bar{p})$. Therefore, the weighted average, using 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.

□ **Inefficiency in the long run.** In the long run, supply and demand are equated by the real-time 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.

To determine the second-best optimum in the long run, consider the flat retail rate, $\bar{p}^*_{LR}$, real-time 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. The first-order conditions for p_t and $\bar{p}$ are given by (2) and (3), and the first-order condition for K is

$$\sum_{t=1}^{T} -C_k(\tilde{D}_t(p_t, \bar{p}), K) = r. \tag{7}$$

As in the short run, the second-best price, $\bar{p}^*_{LR}$, is a weighted average of the real-time prices where the weights are the relative slopes of the demand curves. The optimal real-time prices are determined by $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$ for every t. Note that equation (7) 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, so that each firm is investing to the point that it is earning zero profits net of capital costs. This implies that, given the second-best flat rate, competition in investment would lead to the second-best optimal capacity investment.

As in the short run, $\bar{p}^e_{LR}$ and $\bar{p}^*_{LR}$ are different weighted averages of the real-time 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, we have the following.

Theorem 1b (nonattainment of the second best in the long run). The long-run competitive equilibrium does not in general 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 the slopes of the demand curves are such that $\bar{p}^*_{LR} > \bar{p}^e_{LR}$, i.e., the equilibrium flat price is too 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 $K^*_{LR} > K^e_{LR}$ if and only if $\bar{p}^*_{LR} < \bar{p}^e_{LR}$ (see Borenstein and Holland, 2003b). *Q.E.D.*

Although competition distorts the consumption of the flat-rate customers relative to the second best, competition does not introduce additional distortions into the real-time market or investment for a given flat rate. For a given $\bar{p}$, the optimal real-time 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 real-time prices that would result from competition, given a $\bar{p}$, namely, the prices such that supply equals demand. In addition, 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.

Although we have assumed that demand in each period depends only on the price in that period, Theorems 1a and 1b can be extended to incorporate nonzero cross-price elasticities. To see this, first note that the characterization of the competitive equilibrium (namely as a quantity-weighted average of the wholesale prices) does not depend on the elasticities (slopes) of the demands, but rather on the quantities demanded at $\bar{p}^e$. This characterization would not change fundamentally if we assumed instead that demand in each period depended on the entire vector of (flat) prices. In other words, the equilibrium flat price would still depend only on the quantities demanded and not on own- or cross-price elasticities of demand.

The second-best optimal flat price, derived by maximizing (1), would, however, depend on the own- and cross-price elasticities of demand. The planner would recognize that raising the flat price would affect the quantity demanded in each period not only since the price in that period would be higher, but also because the price in all other periods would be higher. Thus $\bar{p}^*$ would depend on the own- and cross-elasticities of demand.[22] Since $\bar{p}^e$ does not depend on the own- or cross-elasticities of demand, the second-best optimum would still not be attained by competition even when cross-price elasticities are nonzero.

Finally, it is useful to distinguish Theorem 1 from the contestable-markets theory of the classic natural monopoly. First, that literature deals with the market failure due to nonexistence of a Walrasian competitive equilibrium. We analyze the market failure that results from flat-rate pricing when customers on flat-rate pricing cannot trade with customers on real-time pricing. Second, contestable equilibria, characterized by average cost pricing, do not attain the first best, but attain the second best where the planner is constrained to choose a price that achieves at least normal profit. If we were to define the profit-constrained second best as the flat price that maximized welfare subject to achieving normal profit, the competitive equilibrium still would not

[22] Clearly, the characterization of $\bar{p}^*$ in (5) would be less intuitive and much more complex with nonzero cross-elasticities.

generally attain the profit-constrained second best. In particular, if $\bar{p}^* > \bar{p}^e$, then $\bar{p}^*$ yields the retailers more than normal profit and competition still would not achieve the profit-constrained second best. On the other hand, if $\bar{p}^* < \bar{p}^e$, then $\bar{p}^*$ does not yield a normal profit. In this case, $\bar{p}^e$ would be the profit-constrained second best and would be attained by competition.

□ **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.[23] 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 subsection, we analyze such policies.

Among such interventions, there are three 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 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.

To see whether these policies can attain the second-best optimum, we characterize a long-run competitive equilibrium with a tax/subsidy on the RTP customers, τ_{rtp}; a tax/subsidy on the flat-rate customers, τ_{flat}; and a tax/subsidy on capacity, σ.[24] As above, the equilibrium is characterized by the conditions on equilibrium in the product market and in the wholesale and retail sectors. First, the tax on RTP customers implies a tax wedge between the wholesale prices received by the generators and the price paid by the customers, so $p_t = w_t + \tau_{rtp}$. Second, there is a tax wedge between the flat-rate price paid by the customers $\bar{p}$ and the flat rate received by the retail sector, $\bar{p} - \tau_{\text{flat}}$. Thus, the equilibrium flat rate is determined by $\bar{p} - \tau_{\text{flat}} = [\sum_{t=1}^{T} w_t D_t(\bar{p})]/[\sum_{t=1}^{T} D_t(\bar{p})]$. Third, the capacity tax raises the cost of capital to $r + \sigma$. The long-run equilibrium condition on generator profits is then $\sum_{t=1}^{T} w_t \tilde{D}_t(p_t, \bar{p}) - C(\tilde{D}_t(p_t, \bar{p}), K) = (r + \sigma)K$. The final condition equates supply and demand in every period: $w_t = C_q(\tilde{D}_t(p_t, \bar{p}), K)$.

Given this characterization of the equilibrium, it is straightforward to show that the second-best optimum will be attained in equilibrium by a policy with a tax/subsidy to the flat-rate customers of $\tau^*_{\text{flat}} = \bar{p}^*_{LR} - [\sum_{t=1}^{T} p_t^* D_t(\bar{p}^*_{LR})]/[\sum_{t=1}^{T} D_t(\bar{p}^*_{LR})]$ and no taxes or subsidies to the real-time customers or to capacity, i.e., $\tau_{rtp} = 0$ and $\sigma = 0$. The second term of τ^*_{flat} is the quantity-weighted average price of buying wholesale power for flat-rate customers when the flat rate is $\bar{p}^*_{LR}$. Thus, τ^*_{flat} is the tax or subsidy that allows the retailer to break even while charging $\bar{p}^*_{LR}$.[25] Therefore, we have the following.

Theorem 2 (achieving second-best optimality with taxes/subsidies). With external financing, a policy with a tax/subsidy on the flat-rate customers of

$$\tau^*_{\text{flat}} = \bar{p}^*_{LR} - \frac{\sum_{t=1}^{T} p_t^* D_t(\bar{p}^*_{LR})}{\sum_{t=1}^{T} D_t(\bar{p}^*_{LR})}$$

and no taxes or subsidies on the real-time customers or on capacity, i.e., $\tau_{rtp} = 0$ and $\sigma = 0$, achieves the second-best optimal allocation and capacity investment. The optimal policy, τ^*_{flat},

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

[24] Note that the description of these policy instruments is quite general and that each can be either positive, negative, or zero.

[25] The optimal flat-rate tax/subsidy is not, in general, equal to the difference between the second-best optimal flat rate and the equilibrium flat rate, $\bar{p}^e_{LR} - \bar{p}^*_{LR}$. The tax/subsidy, τ^*, is like a Pigouvian tax/subsidy on an externality, but it does not attain the first best.

may be a tax or a subsidy. Any policy that taxes or subsidizes real-time customers or capacity cannot attain the second-best optimum.

Proof. If $\tau_{\text{flat}} = \tau^*_{\text{flat}}$ and $\tau_{rtp} = \sigma = 0$, the retailers break even when charging $\bar{p}^*_{LR}$ and paying the retail tax. Therefore, $\bar{p}^*_{LR}$ is the equilibrium flat rate, and equilibrium 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 real-time customers or in investment, since real-time prices and investment costs are not distorted and thus the second-best optimum is attained.

Any policy with $\tau_{rtp} \neq 0$ or $\sigma \neq 0$ cannot attain the second-best optimum. First, if $\bar{p}^*_{LR}$ is not the equilibrium flat rate, then the consumption of the flat-rate customers is distorted. Alternately, if $\bar{p}^*_{LR}$ is the equilibrium flat rate, then the second best is attained when $\tau_{rtp} = \sigma = 0$, and any other policy would distort either consumption of real-time customers or investment. *Q.E.D.*

Most of the public policy debates regarding investment in electricity markets have not actually considered taxes or 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 time-invariant retail electricity tax that applies to all retail customers. It is a straightforward implication of Theorem 2, discussed in more depth in Borenstein and Holland (2003b), that such policies do not attain the second-best optimum consumption or investment.

3. Changing the proportion of customers on real-time pricing

■ While it is clear that, absent metering costs, charging real-time prices to all customers would be Pareto efficient, in reality any changes toward 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.

□ **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 quantity demanded, since more customers face the higher real-time 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, quantity demanded *increases* with α, since more customers face the lower real-time price. This drives up the wholesale prices in these periods. Thus, some wholesale prices increase and some decrease when more customers are put on real-time pricing.

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

Theorem 3 (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 the Appendix.

The key to this result is recognizing that retailers break even on the flat-rate customers by 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 periods. Because increasing α will rotate the wholesale demand, increasing the proportion of customers on RTP decreases the peak wholesale prices and increases off-peak wholesale 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 holding $\bar{p}$ constant, the decreased retail losses in the peak periods are greater than the decreased retail gains in off-peak periods 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.

□ **The effect on capacity of increasing RTP customers.** Under regulation, investment in the 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.[26]

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 real-time 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, $\partial \pi^w_{SR}/\partial \bar{p} = 0$ and $\partial \pi^w_{SR}/\partial p_t = \tilde{D}_t$, so

$$\frac{d\pi^w_{SR}}{d\alpha} = \frac{\partial \pi^w_{SR}}{\partial \alpha} + \frac{\partial \pi^w_{SR}}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha} + \sum_t \frac{\partial \pi^w_{SR}}{\partial p_t} \frac{dp_t}{d\alpha} = \sum_t \tilde{D}_t(p_t, \bar{p}) \frac{dp_t}{d\alpha}. \tag{8}$$

Since (8) 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.

Theorem 4 (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 the Appendix.

The proof of Theorem 4 highlights the importance of the convexity of the marginal costs across the relevant range. If the marginal cost curve is relatively flat at off-peak demand levels, then putting additional customers on RTP will not increase the off-peak prices very much. 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. In the simulations we have run with realistic parameters, presented in Section 4, increases in α cause capacity to decline in the long run in all cases.

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

[26] Bergstrom and MacKie-Mason (1991) argue against the conventional wisdom by showing that peak-load pricing could increase investment under regulation. We analyze competitive markets and do not assume homothetic preferences. In a related model of airline competition, Dana (1999) shows that stochastic peak-load pricing can lead to lower capacity costs.

more customers on RTP leads to increased investment. Note that in this example, the marginal cost curve is not concave.

□ **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 competitive equilibrium is not efficient, we cannot rely on comparative statics results from a constrained optimization problem.

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, Theorem 3 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 customer who switches from the flat rate to RTP receives 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 switcher could consume exactly the same electricity quantities that the flat-rate customers choose at the exact same total bill. Since the switcher chooses to consume different quantities, it must be better off.[27]

Finally, the surplus to the customers who are already 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

$$\alpha \frac{\mathrm{d}CS_{RTP}}{\mathrm{d}\alpha} = \sum_{t=1}^{T} -\frac{dp_t}{d\alpha} \alpha D_t(p_t) = \sum_{t=1}^{T} \frac{dp_t}{d\alpha} (1-\alpha) D_t(\bar{p}), \tag{9}$$

where the second equality follows from (8), recognizing $\frac{d\pi^w}{d\alpha} = 0$ in the long run (because $\pi^w = 0$ 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 (9) 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

$$0 = \frac{\partial \pi^r}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha} + \sum_{t=1}^{T} \frac{\partial \pi^r}{\partial p_t} \frac{dp_t}{d\alpha} = \frac{\partial \pi^r}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha} - \sum_{t=1}^{T} \frac{dp_t}{d\alpha} (1-\alpha) D_t(\bar{p}). \tag{10}$$

Since the competitive-equilibrium $\bar{p}$ results from Bertrand competition over the flat rates, the derivative $\partial \pi^r / \partial \bar{p}$ must be greater than or equal to zero. Since $d\bar{p}/d\alpha < 0$ by Theorem 3, $\sum_{t=1}^{T} (dp_t/d\alpha)(1-\alpha) D_t(\bar{p})$ must be less than zero. Combining this with (9) shows that the consumer surplus to the incumbent RTP customers is decreasing in α.

A bit of intuition for this result comes from thinking about a single small customer who is the only customer on RTP. That RTP customer is better off than a customer on a flat rate because it can reoptimize against the volatile real-time prices (the revealed-preferences argument above). As more customers move to RTP, that real-time price volatility is muted, reducing the benefits from responding to the volatility. On the other hand, as more customers move to RTP, the weighted average wholesale price is lower, benefiting all customers. Our result shows that the first effect outweighs the second for the existing RTP customers, making them worse off.

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

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

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

$$\frac{\mathrm{d}W(K, p_t, \bar{p}, \alpha)}{\mathrm{d}\alpha} = \frac{\partial W}{\partial K}\frac{dK}{d\alpha} + \sum_{t=1}^{T}\frac{\partial W}{\partial p_t}\frac{dp_t}{d\alpha} + \frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} + \frac{\partial W}{\partial \alpha}. \tag{11}$$

We have shown that in the competitive equilibrium, $\partial W/\partial K = 0$, i.e., capacity is set efficiently given the equilibrium prices. Likewise, $\partial W/\partial p_t = 0$ for all t, since we have explained earlier that real-time prices are set efficiently given the equilibrium $\bar{p}$. Thus (11) reduces to $dW/\mathrm{d}\alpha = (\partial W/\partial \bar{p})(d\bar{p}/d\alpha) + (\partial W/\partial \alpha)$.

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

$$\frac{\partial W}{\partial \alpha} = \sum_{t=1}^{T}[U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - p_t D_t(\bar{p})], \tag{12}$$

which is positive by the revealed-preferences argument made above. Theorem 3 shows that $d\bar{p}/d\alpha < 0$, and in Section 2 we showed that $\partial W/\partial \bar{p}$ can be positive or negative depending on whether $\bar{p}^e$ is greater or less than $\bar{p}^*$.[28] 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 $\bar{p}$. To summarize, we have the following.

Theorem 5 (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, (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 increase or 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. Parts (i)–(iv) are proved in the text. Since $dW/\mathrm{d}\alpha = (\partial W/\partial \bar{p})(d\bar{p}/d\alpha) + (\partial W/\partial \alpha)$ and $\partial W/\partial \alpha > 0$, if $\bar{p}^e > \bar{p}^*$, so that $\partial W/\partial \bar{p} < 0$, then $(\partial W/\partial \bar{p})(d\bar{p}/d\alpha) > 0$ and increasing α increases total welfare. If $\bar{p}^e < \bar{p}^*$, then $(\partial W/\partial \bar{p})(d\bar{p}/d\alpha) < 0$. Since $\partial W \partial \alpha > 0$, the net impact on welfare in this case is ambiguous. In the Appendix we demonstrate how an example in which increasing α lowers welfare can be constructed. *Q.E.D.*

We know, however, from Section 2 that increasing α to one 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 that the increase in welfare need not always be monotonic as it moves to the maximum welfare at $\alpha = 1$.[29]

□ **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. We now consider the incentives of customers to adopt RTP under competition.

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

[29] Simulations with linear demands (in which $\frac{\partial W}{\partial \bar{p}} \gg 0$) and simulations presented in Borenstein (2005b) that use actual California system load profiles and assume constant elasticity demand with higher elasticities during peak periods showed no cases in which welfare declined with an increase in α.

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.[30] Let M be the additional daily cost (variable plus amortized fixed cost) of metering and billing one customer when that customer adopts RTP.

Assume, for now, that each customer constitutes a share γ of the total demand, where γ is very small. Since customers can avoid this additional metering cost by choosing the flat-rate service, customers will adopt RTP until in equilibrium we have

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

Equation (13) determines the equilibrium share of customers on RTP, α. The benefit to adopting RTP is positive by the revealed-preferences argument and bounded. If M is large enough, then no customers will adopt RTP. Conversely, if M is small but still positive, then all customers adopt RTP. Even the last customer left on a flat-rate tariff when all others have switched will have strictly positive benefits from switching that will outweigh a sufficiently small M. Since the left-hand side of (13) is decreasing in α—an implication of Theorem 5—if M is positive but not too large, there will be a solution with $0 < \alpha \leq 1$.[31] The long-run competitive equilibrium with customer choice over rate structure is then fully described by (13) plus the conditions described in the characterization of the long-run equilibrium above.

First note that, as in Section 2, the competitive equilibrium with competitive RTP selection does not attain the second-best optimum, since the equilibrium flat rate will be suboptimal. To analyze the efficiency of competitive RTP adoption, we define W as the welfare attained in the competitive equilibrium where now W incorporates the metering cost $(\alpha/\gamma)M$, i.e., the costs of metering the RTP customers. From a long-run competitive equilibrium, differentiating W as in (11) yields the following result.

Theorem 6 (nonoptimality of competitive RTP selection). If metering costs are positive and customers choose between flat rates or real-time prices, the long-run competitive equilibrium does not in general attain the second-best optimal electricity allocation, capacity investment, and RTP metering. Competition leads to excessive adoption of RTP if $\bar{p}^e < \bar{p}^*$. If $\bar{p}^e > \bar{p}^*$, RTP is adopted less than is optimal.

Proof. Nonoptimality of the electricity allocation and capacity investment follow directly from Theorem 1b. Next, consider $dW/d\alpha$ evaluated at the competitive equilibrium. The partial derivative of W with respect to α is now

$$\begin{aligned}
\frac{\partial W}{\partial \alpha} &= \sum_{t=1}^{T} \{[U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - p_t D_t(\bar{p})]\} - \frac{M}{\gamma} \\
&= \sum_{t=1}^{T} \{[U_t(D(p_t)) - p_t D_t(p_t)] - [U_t(D(\bar{p})) - \bar{p} D_t(\bar{p})]\} - \sum_{t=1}^{T} (\bar{p} - p_t) D_t(\bar{p}) - \frac{M}{\gamma} \\
&= 0.
\end{aligned} \tag{14}$$

The first equality follows from differentiation of W as in (12), the second equality is algebra, and the third equality follows from (13) and the condition on retail profit. Since $\partial W/\partial \alpha = 0$, it follows from (11) that $dW/\mathrm{d}\alpha = (\partial W/\partial \bar{p})(d\bar{p}/d\alpha)$. Since $d\bar{p}/d\alpha < 0$ from Theorem 3, increasing metering beyond the competitive level increases welfare if and only if $\partial W/\partial \bar{p} < 0$. *Q.E.D.*

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

[31] In what follows, we ignore the corner solutions.

Theorem 6 obtains because customers do not recognize that by adopting 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 adopt RTP. On the other hand, if the flat rate is too low, then the externality is harmful, and too many customers adopt RTP.[32]

If customers differ in size but 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, (13) implies that the customers with the largest γ_i would be the first to adopt RTP. However, the marginal customer still would not consider the effect on $\bar{p}$ of its decision to adopt, and adoption 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 adopt is further complicated in two ways. First, an *elasticity effect* will cause customers with more elastic demand to be more inclined to adopt RTP. For instance, if two customers on flat-rate service demand 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 adopt RTP. This welfare effect, however, seems to be fully captured by the adopter. 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 adopt. 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 does not by itself change total surplus, gives some customers inefficiently large incentives to pay M in order to adopt RTP.[33] The adverse-selection effect will tend to cause $\bar{p}$ to rise as these customers adopt RTP, which may or may not outweigh the tendency for $\bar{p}$ to fall with adoption when customers are identical. If $\bar{p}$ were to rise, this most likely could raise or lower welfare, depending on whether $\bar{p}^e$ is greater or less than $\bar{p}^*$.

Finally, we again address the extension of the model to incorporate nonzero cross-price elasticities. With nonzero cross-partial derivatives, the proof of Theorem 3 holds under the additional assumption on the cross-partials: $\sum_{j=1}^{T}(\partial D_t/\partial p_j)[(dp_j/d\alpha)/(dp_t/d\alpha)] < 0$ for every t. This condition clearly holds if all cross-partials are zero, as we've assumed above, since demand is downward sloping. Note that it also holds, for example, if electricity consumption in each peak period is complementary to consumption in all other peak periods but a substitute for consumption in off-peak periods and vice versa.

Under nonzero cross-price elasticities, Theorem 4 is still valid since it describes the ambiguity of the effect on capacity. Parts (i) and (iii) of Theorem 5 relied on Theorem 3 and thus obtain under the above condition on the cross-partials. Part (ii) remains valid by the revealed-preferences argument, and part (iv) and the remainder of the theorem remain valid by the arguments presented above. Theorem 6 also remains valid. Thus the main theoretical results of this section obtain if the model allows nonzero price elasticities.

Throughout the analysis, we have ignored the potential for price volatility to lower the welfare of RTP customers due to risk aversion. In reality, these risks can be nearly entirely eliminated through forward contracting for fixed quantities of power for each hour of the year. Such fixed-quantity contracts, similar to standard futures contracts that are used for hedging risk, can eliminate most of the wealth risk caused by fluctuating electricity prices while still giving the customer the full incentive effect of time-varying prices on the margin. They would have no effect on our analysis.[34] Such contracts are quite different from the "requirements contracts" that nearly all retail customers currently face, in which the customer has a right to buy any quantity it chooses at a price specified months in advance through a regulatory process or long-term contract. Customers under requirements contracts correspond to flat-rate customers in our analysis.

[32] Brennan (2002) and Doucet and Kleit (2003) do not recognize this externality in their analyses of competitive adoption of real-time pricing.

[33] 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.

[34] Borenstein (2005a) explains in more detail how forward contracts can be used to mitigate these risks.

4. How much difference would RTP make?

■ While the analysis thus far demonstrates that RTP is likely to improve welfare, such a change would also involve costs, so it is important to evaluate the magnitude of the effect that RTP would have. Full estimation of the expected effect is beyond the scope of this article, but in this section we offer some preliminary analysis that suggests that the welfare gain is likely to be significant and to vastly outweigh the costs of implementing RTP for at least the largest customers.

To accomplish this, we simulate long-run competitive equilibria under a set of realistic assumptions about the system demand profile and the production technologies available. We simulate the system first using a flat-rate retail pricing scheme and then putting some share of the customers on RTP. No such simulation can include all factors that would affect the welfare change due to RTP, but we believe that we capture the primary effects. We conclude this section with a discussion of omitted factors and how they would be likely to alter the analysis.

Since we have assumed competitive generation markets and no cross-subsidy between RTP and flat-rate customers, the simulations could also be interpreted as applying to multiple retail providers that participate in the same wholesale market. For instance, the customers of one utility, constituting perhaps two-thirds of the total wholesale market, might remain on flat rates, while the customers of a competing retail provider that participates in the same wholesale market and comprises the remainder of demand might be on RTP.[35] The analysis thus far shows that the RTP customers would produce positive spillovers for the utility customers. The simulations suggest how large that effect might be.

□ **Assumptions.** We depart somewhat from the simple theoretical model in the previous sections by assuming that there are multiple technologies of production that range from high-capital/low-variable cost to low-capital/high-variable cost. Each technology has a simple cost function: a unit of capacity has an annual capital cost (independent of usage) and a constant per-megawatt-hour marginal cost up to the unit's capacity of one megawatt.[36] There are no economies of scale in any of the technologies, or equivalently, over the relevant range of usage, each technology can be scaled up or down at constant long-run average cost.

To be concrete, we assume that there are three technologies: a baseload technology (highest capital, lowest marginal cost) for which we set the parameters to roughly reflect coal-fired generation, a mid-merit technology, which is set to reflect combined-cycle gas turbine generation, and a peaker technology, which is meant to reflect combustion turbine generation. The cost assumptions are shown in Table 1. We assume that in the long run there is free entry and exit of any of these technologies. We also assume that no company owns a sufficient amount of production to be able to exercise market power. The retail price charged to customers reflects production costs plus $40/MWh to cover transmission and distribution, a figure that we assume is not subject to temporal variation.

On the demand side, we start from a basic distribution of demand levels that we have taken from the actual reported demand levels in the California Independent System Operator's control area over the two-year period November 1998 through October 2000.[37] This period includes periods of record high consumption and very low consumption, a mild summer in 1999 and an unusually hot summer in 2000. To get from these demand quantities to a set of demand curves, we assume that each quantity occurred at the same flat-rate retail price.[38] Further, we assume that the

[35] Such a system would not require real-time meters for the flat-rate customers, since the demand of the flat-rate customers in aggregate could be calculated by the grid operator as system demand minus the real-time demand of RTP customers.

[36] Borenstein and Holland (2003a) prove the equivalent of Theorems 1 through 5 for production with such inverted L-shaped cost curves.

[37] The CAISO (www.caiso.com) is the quasi-governmental organization that controls the electricity grid for most of California.

[38] Incorporating the fact that some customers were charged somewhat higher prices during preset peak periods than during off-peak periods ("time-of-use" rates) would have little effect on the results, due in part to the inelasticity of demand.

TABLE 1 Generation Costs Assumed in Long-Run RTP Simulations

Generation Type	Annual Capital Cost ($/MW)	Variable Cost ($/MWh)
Baseload	155,000	15
Mid-merit	75,000	35
Peaker	50,000	60

flat retail rate that resulted in these demand quantities is the one that we calculate below to be the break-even rate.[39] We assume that retail customer demand goes through the actual quantity at the equilibrium flat retail rate, and that it takes a constant-elasticity form with the assumed elasticity, which we discuss next.[40]

For the switch to RTP to have any effect, demand must be price elastic. Estimates of price elasticity of electricity demand vary widely, but the very short-run price elasticity is commonly argued to be in the range of $-.1$. In the longer run, we would expect greater elasticities, as customers adapt to price variability and response technologies—e.g., thermostats that change settings in response to price—improve and come down in cost. Therefore, we also evaluate the effect for assumed demand elasticities of $-.3$ and $-.5$.[41] We assume that all customers have the same demand elasticity and the same "load profile," i.e., all customers are identical up to a scale factor.

These cost and demand specifications, along with assumptions of price-taking behavior by all market participants and free entry/exit of producers, are sufficient to determine a unique long-run competitive equilibrium for any $\alpha > 0$.[42] The algorithm for determining the competitive equilibrium and its uniqueness are discussed in greater detail in Borenstein (2005b).

To determine the base case in which all customers are on flat-rate service requires a somewhat different approach. Any given flat rate, $\bar{p}$, determines the quantity demanded in each hour. Given the quantity to be produced in each hour, straightforward algebra determines the cost minimizing amount of each type of capacity that should be installed and the number of hours each would be used. This determines the total cost of production. The $\bar{p}$ is then adjusted iteratively until the total revenue generated with that flat rate is equal to the total (minimized) cost of production, which is the competitive-equilibrium flat rate.

□ **Results.** The results of these simulations are presented in Tables 2 and 3. Table 2 indicates the capacity, price, and quantity outcomes, while Table 3 shows the changes in consumer surplus, which approximately equal total surplus because all sellers earn zero profit in equilibrium (up to the integer constraint on capacity). In Table 2, the results with all customers on flat-rate service are shown at the top. The rows below show the outcomes under varying assumptions about the demand elasticity and shares of customers on RTP.

Table 2 is consistent with our theoretical finding that as more customers switch to RTP, the equilibrium flat rate falls. It also indicates that RTP would have large effects on the equilibrium installed capacity. Even with an elasticity of only $-.1$, putting just a third of customers on RTP would cut the number of peakers by about 44% and the total installed capacity by more than 10%. Though we showed that theoretically RTP would not necessarily lower the total installed

[39] Also due to inelastic demand, adjusting for the fact that the prevailing flat rate wasn't exactly equal to the level that we calculate would break even given the production technologies makes very little difference in the analysis.

[40] During the observed period, the utilities had real-time pricing for only very few customers that were part of pilot programs. Many customers—constituting about 15% of total demand—had switched to nonutility retail providers, but virtually none of these customers was on RTP either.

[41] We do not consider cross-elasticities between hours, though we do discuss the effect of substitutability across hours, i.e., "load shifting," below.

[42] Throughout the simulations, we continue assuming that the customers on RTP in aggregate do not differ in their demand profile from those that remain on flat-rate service.

TABLE 2 Capacity, Price, and Quantity Effects of RTP

Elasticity	Share on RTP	Total Annual Energy Consumed (MWh)	Total Annual Energy Bill ($)	Flat Rate ($/MWh)	Equilibrium Capacity (MW)				Price Duration Curve	
					Base-load	Mid-Merit	Peaker	Total	Peak Price ($/MWh)	Hours per Year at Peak Quantity (of 8,760)
All on Flat Rate										
—	0	236,796,579	9,265,381,746	79.13	27,491	5,472	12,912	45,875		
Some on RTP										
−.100	.333	237,969,466	8,982,619,333	78.68	27,690	5,185	7,294	40,169	8,654	113
−.100	.666	238,809,721	8,840,599,109	78.35	27,876	4,891	4,846	37,613	1,864	252
−.100	.999	239,517,938	8,731,658,462	78.17	28,052	4,612	3,162	35,826	940	382
−.300	.333	240,435,717	8,805,179,930	78.30	28,124	4,616	3,897	36,637	1,843	312
−.300	.666	242,774,258	8,565,317,811	77.87	28,664	3,767	437	32,868	510	702
−.300	.999	244,269,124	8,396,112,525	77.38	29,157	1,713	0	30,870	297	1,914
−.500	.333	243,251,232	8,719,880,253	78.12	28,578	4,045	2,094	34,717	1,201	476
−.500	.666	246,834,756	8,440,185,760	77.37	29,478	1,353	0	30,831	314	2,023
−.500	.999	248,276,486	8,295,147,559	76.54	29,488	0	0	29,488	192	5,697

capacity, that is the effect we find in all simulations using realistic parameters. With enough elasticity, or most customers on RTP, baseload (high capital-cost, low variable-cost) capacity becomes relatively more cost effective, eventually eliminating peakers entirely as the quantity volatility is greatly reduced due to retail price volatility.

Retail price volatility would be a significant feature of the RTP market, though much less so if there were more elasticity or a large share of customers on RTP. With an elasticity of −.1 and only a third of customers on RTP, the peak price would be more than 100 times the average price. In only 113 hours of the year, about 1.3% of the time, would all of the capacity be used; in all other hours price would be no higher than the marginal production cost of the peaker units (plus transmission and distribution (T&D)). With greater elasticity, the capacity is used more intensively, with a great number of hours in which all capacity is used. An elasticity of −.3 yields a peak price that is 79% lower than results with an elasticity of −.1, with a third of customers on RTP. Our theoretical model does not predict the direction of change in either total consumption or the total energy bill, but these simulations indicate that total energy consumption could increase, while total energy bills could fall.

Table 3 presents the welfare effects of introducing RTP. All of the "change" columns are in comparison to the equilibrium with all customers on flat rate. The "Total Surplus Change" column indicates that the gains would be in the hundreds of millions of dollars per year for the California system, which would be between 3% and 11% of the total energy bill that obtained under flat rates.

In comparison to the cost of implementing RTP, these gains are likely to be large for at least some customers. In 2001, in the midst of the California electricity crisis, the state legislature passed a bill mandating real-time meters for all large customers, those with peak demand above 200 kilowatts. The cost of installing and operating meters for these 20,000 customers, which constituted about one-third of California demand, was about $35 million. Though other administrative costs would also accompany RTP, these other costs are unlikely to be larger than the original meter installation costs. Thus, it appears that the gains would almost certainly outweigh the costs.[43]

[43] Installation of these meters was not completed until 2004. As of this writing, there is no generally available RTP tariff in California. The meters are being used for RTP in only a few small pilot programs.

TABLE 3 Welfare Effects of RTP

Elasticity	Share on RTP	Annual Total Surplus Change from All on Flat ($)	Annual TS Change as Percentage of Original Energy Bill	Annual CS Change of Customers on Flat Rate ($)	Annual CS change "Per Customer" on Flat Rate ($)[a]	Annual CS Change of Customers on RTP ($)	Annual CS change "Per Customer" on RTP ($)[a]	Annual Incremental Surplus to Switchers ($)[b]	Annual Incremental Externality ($)[c]
−.100	.333	262,356,807	2.8%	71,409,037	1,071	190,947,770	5,734	190,947,770	71,409,037
−.100	.666	383,058,841	4.1%	61,428,039	1,839	321,630,802	4,829	125,164,413	−4,462,379
−.100	.999	472,596,878	5.1%	228,452	2,285	472,368,426	4,728	96,212,019	−6,673,982
−.300	.333	450,098,624	4.9%	131,733,121	1,975	318,365,503	9,561	318,365,503	131,733,121
−.300	.666	659,443,204	7.1%	99,888,884	2,991	559,554,320	8,402	214,009,350	−4,664,770
−.300	.999	815,479,611	8.8%	415,158	4,152	815,064,452	8,159	172,098,335	−16,061,929
−.500	.333	576,256,903	6.2%	159,412,061	2,390	416,844,842	12,518	416,844,842	159,412,061
−.500	.666	848,963,827	9.2%	140,337,732	4,202	708,626,095	10,640	274,726,516	−2,019,592
−.500	.999	1,023,011,668	11.0%	619,521	6,195	1,022,392,146	10,234	200,879,823	−26,831,983

[a] For a hypothetical customer who constitutes .001% of total demand at all times and remains on the same tariff (RTP or flat rate) in all simulations.

[b] The total effect on consumer surplus of customers who switch to RTP tariff as the share on RTP changes from the row above (from zero for .333 row).

[c] The total effect on consumer surplus of customers who do not switch tariffs as the share on RTP changes from the row above (from zero for .333 row).

Implementing RTP for the remaining customers might not be such a clear net gain. Table 3 indicates that the marginal surplus gains to RTP are declining in the share of customers on RTP. At all the elasticities in Table 3, putting the first one-third of the customers on RTP produces more than half the gains that result from putting all on RTP. Furthermore, the cost of switching the remaining customers would be greater due to the small size of individual customers: switching the first one-third of customers in California required replacing about 20,000 meters, while switching the remaining two-thirds would require replacing over 10 million meters.[44]

The aggregate measures of surplus are affected by the change in the share of customers on RTP and so may obscure the effect on individual customers. Comparing the rows in Table 3 with different shares of customers on RTP combines the effect on "switchers" with the effect on incumbent RTP and flat-rate customers who do not switch. To illustrate the effect on individual customers who do not switch, Table 3 also presents consumer surplus changes for a hypothetical customer that constitutes .001% of the total demand curve in any hour and remains on either RTP or a flat-rate tariff in all simulations.[45] The columns showing the surplus of these hypothetical customers on flat rate and RTP are consistent with the theoretical finding that switching customers to RTP benefits remaining flat-rate customers (which follows immediately from the decline in the flat rate shown in Table 2) and harms incumbent RTP customers.

The two right-hand columns of Table 3 decompose the surplus change that results from increasing the share of customers on RTP into the change captured by the switchers and the external effect on nonswitchers (both remaining flat-rate customers and incumbent RTP customers). The changes in each row are the result of increasing the share on RTP from the level in the previous row (from zero in the case of a .333 share on RTP). As expected, putting the first third of demand on RTP not only benefits the switchers, it also benefits the nonswitchers in aggregate, because all nonswitchers are flat-rate customers who benefit from the decline in the flat rate. Moving the next

[44] Some vendors, however, have suggested that the use of cellular technology for communicating with the meters creates large density economies, and that rollout in densely populated areas can actually pay for itself by reducing labor costs in meter reading.

[45] This .001% would represent a customer with a peak demand of about 450kW. In California, there are approximately 8,000 customers of at least this size.

third of customers to RTP, however, creates both positive (on flat-rate customers) and negative (on incumbent RTP customers) externalities. The net effect, found by properly weighting these positive and negative changes in consumer surplus, is a small negative externality. Putting nearly all of the remaining customers on RTP creates a larger net negative externality, because it harms incumbent RTP customers, while there are almost no flat-rate customers left to benefit from the declining flat rate.

Finally, it is worth pointing out that while demand elasticity is necessary for RTP to create social benefits, it may not take very much elasticity. There seems to be a declining marginal gain from increased elasticity. In the simulations, the gain from putting a given share of customers on RTP when the elasticity is $-.1$ is about 45% of the gain when the elasticity is $-.5$, five times greater.

□ **Omitted factors.** As we stated at the beginning of this section, these simulations omit a number of factors that one would want to address in a complete simulation of RTP. In this subsection, we discuss some of these factors and their likely impact.

Reserves. Demand and supply must balance exactly at all times in an electricity grid, so grid operators keep capacity on standby to respond as demand fluctuates stochastically. Holding such supply reserves is costly. The need for reserves would almost certainly be reduced with greater use of RTP, because price variation would substitute to some extent for quantity fluctuations. Thus, RTP would also reduce reserve costs. The potential for savings is bounded above by the size of reserve costs, which are 7–10% of total operating costs in a typical system. A significant proportion of this cost would remain if few customers would be willing to let the grid operator control their second-to-second consumption in order to balance the system.

Stochastic supply outages. The simulations assume that all generators are perfectly reliable. In fact, generators go out of service stochastically. If all units were very small and outages were uncorrelated, then the law of large numbers would imply that this requires simply a rescaling of effective capacity per unit. In fact, some generation units are large, and outages are not completely uncorrelated, so the grid operator must hold reserves also to respond to these outages. As above, however, RTP would reduce the need to hold reserves to respond to unforeseen supply/demand imbalances.

Nonconvexities in production. As discussed in detail by Mansur (forthcoming), generation units do not costlessly or instantly switch from off to full production. There are startup costs and "ramping" constraints (on the speed with which output can be adjusted). These constraints make it more costly to adjust supply to meet demand fluctuations. As with reserves, RTP would allow some of this adjustment to occur on the demand side in a way that would enhance efficiency.

Pricing of transmission and distribution. The simulations take a constant \$40/MWh charge for transmission and distribution. This is based on the historical recovery of the costs of these services, which are provided by a regulated monopoly. To the extent that minimum efficient capacity scale for T&D implies that they are never capacity constrained, introducing time-varying prices of these services would not improve efficiency. That may be the case with most local distribution, but transmission lines frequently face capacity constraints. By ignoring these constraints and holding the T&D cost per MWh constant, the simulations understate the potential gains for RTP that could also reflect time-varying (opportunity) cost of transmission.

Market power. In the simulations, we assume that sellers never exercise market power. As has been discussed elsewhere (see Borenstein and Bushnell, 1999, and Bushnell, 2005), demand elasticity introduced by implementing RTP reduces the incentive of sellers to exercise market power. However, it is unclear how much incremental inefficiency the exercise of market power itself introduces in a flat-rate system. In fact, for $\bar{p}^e < \bar{p}^*$, seller market power could increase efficiency by increasing the flat retail rate, providing that any production inefficiencies (due to misallocation of production across plants) were small. In a full RTP system, market power could

not reduce deadweight loss. Thus, it is difficult to analyze the bias from excluding seller market power.

Nonzero cross-elasticities of demand across hours. In the demand structure that we have analyzed, own-price elasticities are nonzero and all cross-elasticities are zero. Simulation with a complete matrix of own- and cross-elasticities would increase the complexity substantially. Still, if demands are generally substitutes across hour, it seems very likely that incorporation of cross-elasticities would increase the gains from RTP. Essentially, RTP increases efficiency by reducing the volatility of quantity consumed and increasing the utilization rate of installed capacity. Holding constant own-price elasticities, increasing cross-price elasticities from zero to positive (substitutes) will tend to further reduce quantity volatility by increasing off-peak quantity when peak prices rise and reducing peak quantity when off-peak prices fall.

5. 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.

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.[46]

To analyze these effects and assess their relevance for policy analysis, we developed a simulation model using three types of generation technology and realistic load profiles information. The simulation indicated that increasing the share of customers on RTP would decrease capacity and monotonically increase welfare. The effects on peaking capacity were particularly notable; we estimate that putting a third of customers on RTP would reduce peaking capacity requirements by 44%. The welfare gains were also substantial, with gains from 3% to 11% of the total energy bill. We found that the incremental gains from putting additional customers on RTP declined with the share that were already on RTP.

We've modelled the flat-rate retail price problem in the context and institutions of deregulated electricity markets, but the application is much broader.[47] In many markets, retail prices cannot,

[46] Like much of the peak-load pricing literature, we have made simplifying assumptions. We have extended the model to nonzero cross-elasticities above. Relaxing other assumptions, as we intend to do in future work, is unlikely to alter the basic insights of this analysis.

[47] 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 that requirements contracts may have greater adverse-efficiency effects than is generally recognized.

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 view of how markets may operate that is different from those presented by Carlton (1986) and others who examine nonprice rationing. In those models, all prices are sticky and therefore nonprice 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.

Appendix

■ Proofs of Theorems 3 and 4, along with examples demonstrating that increasing the share of customers on RTP can reduce equilibrium capacity and reduce equilibrium welfare, follow.

Proof of Theorem 3. We demonstrate this theorem 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, i.e., become positive, if $\bar{p}$ did not change. Since any higher flat rate would also have positive profit, the new equilibrium flat rate must be lower, i.e., competition in the retail sector reduces $\bar{p}$.

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

First note that competitive investment implies that in the long run,

$$0 = \frac{d\pi^w}{d\alpha} = \sum_t \tilde{D}(p_t, \bar{p})\frac{dw_t}{d\alpha} = K\sum_t S(w_t)\frac{dw_t}{d\alpha} \quad \Rightarrow \quad \sum_t S(w_t)\frac{dw_t}{d\alpha} = 0, \tag{A1}$$

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

Next note that

$$\alpha D_t(w_t) + (1-\alpha)D_t(\bar{p}) = KS(w_t) \quad \Leftrightarrow \quad D_t(w_t) - D_t(\bar{p}) = \frac{KS(w_t) - D_t(\bar{p})}{\alpha}. \tag{A2}$$

Differentiating the left-hand equation in (A2) with respect to α gives

$$D_t(w_t) - D_t(\bar{p}) + \alpha D_t'(w_t)\frac{dw_t}{d\alpha} + (1-\alpha)D_t'(\bar{p})\frac{d\bar{p}}{d\alpha} = KS'(w_t)\frac{dw_t}{d\alpha} + S(w_t)\frac{dK}{d\alpha}. \tag{A3}$$

Recognizing that $d\bar{p}/d\alpha = 0$ by assumption and substituting using the right-hand equation in (A2), (A3) can be rearranged as

$$\alpha[KS'(w_t) - \alpha D_t'(w_t)]\frac{dw_t}{d\alpha} = \left[K - \alpha\frac{dK}{d\alpha}\right]S(w_t) - D_t(\bar{p}). \tag{A4}$$

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

$$0 < \sum_t \left\{\left[K - \alpha\frac{dK}{d\alpha}\right]S(w_t) - D_t(\bar{p})\right\}\frac{dw_t}{d\alpha} = \sum_t -D_t(\bar{p})\frac{dw_t}{d\alpha}, \tag{A5}$$

where the equality holds because $[K - \alpha(dK/d\alpha)]\sum_t S(w_t)(dw_t/d\alpha) = 0$ from (A1). *Q.E.D.*

Proof of Theorem 4. Consider a long-run competitive equilibrium with two time periods: peak, H, and off-peak, L. 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 $C_q(\tilde{D}_L(p_L, \bar{p}), K)$, $C_q(\tilde{D}_H(p_H, \bar{p}), K)$, and $C(\tilde{D}_L(p_L, \bar{p}), K) + C(\tilde{D}_H(p_H, \bar{p}), K)$ did not change, then the long-run equilibrium would not change.

Note that $dp_t/d\alpha = C_{qq}(d\tilde{D}_t/d\alpha)$ in the short run, which implies that $dp_L/d\alpha > 0$ and $dp_H/d\alpha < 0$. Starting from a long-run equilibrium, we can increase the convexity of C_q without changing the long-run equilibrium if $C_{qq}(\tilde{D}_L(p_L, \bar{p}), K) > 0$. By increasing the convexity of C_q without changing the long-run equilibrium, we can make $C_{qq}(\tilde{D}_L(p_L, \bar{p}), K)$ smaller and $C_{qq}(\tilde{D}_H(p_H, \bar{p}), K)$ larger. But this implies that $dp_L/d\alpha$ is less positive and that $dp_H/d\alpha$ is more negative. Thus (8) can be negative. Similarly, an example can be constructed where (8) is positive by increasing the concavity of C_q. An example of a capacity increase follows. *Q.E.D.*

□ **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

$$\bar{p} - p_t = [(K + B)\bar{p} - A_t]/(K + B\alpha) = Y_t/(K + B\alpha), \tag{A6}$$

where $Y_t \equiv (K + B)\bar{p} - A_t$. This implies that retail profits can be written $\pi^r = f(\alpha)\sum Y_t D_t(\bar{p})$, where $f(\alpha) = (1-\alpha)/(K + B\alpha)$. Since $f(\alpha) \neq 0$, in short-run equilibrium, $\sum Y_t D_t(\bar{p})$ must equal zero. But since $\sum Y_t D_t(\bar{p})$ 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 if and only if $\bar{p} > p_t$, which occurs if and only if $\tilde{D}_t(p_t, \bar{p}) > D_t(\bar{p})$. Therefore $\sum Y_t\tilde{D}_t > \sum Y_t D_t(\bar{p})$, 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 $dW/\mathrm{d}\alpha = (\partial W/\partial\bar{p})(d\bar{p}/d\alpha) + (\partial W/\partial\alpha)$. We construct an example in which $dW/d\alpha$ 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 $p_t, \bar{p}, \alpha, r, K$, 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 that does not change these $2T$ points (nor α, S, or r) will have an equilibrium with the same prices and capacity.

Next, consider $d\bar{p}/d\alpha$, $dp_t/d\alpha$, and $dK/d\alpha$. 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 $d\bar{p}/d\alpha$ can be written as a function of the $7T + 4$ parameters: $p_t, D_t(p_t), D_t'(p_t), D_t(\bar{p}), D_t'(\bar{p}), \bar{p}, \alpha, S(p_t), S'(p_t), r$, and K. Since $\partial W/\partial\bar{p}$ can also be written in terms of these $7T + 4$ parameters, the product $(\partial W/\partial\bar{p})(d\bar{p}/d\alpha)$ 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 $\partial W/\partial\alpha$. [12] can be written

$$\frac{\partial W}{\partial\alpha} = \sum_{t=1}^{T}[U_t(D(p_t)) - U_t(D(\bar{p}))] - p_t[D_t(p_t) - D_t(\bar{p})]. \tag{A7}$$

Note that the summands in (A7) 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 $D_t(p_t), D_t'(p_t), D_t(\bar{p}), D_t'(\bar{p})$. For example, in the case of $p_t > \bar{p}$, 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 $\bar{p}$, i.e., if demand were identical to $D_t(p)$ for $p > p_t - \varepsilon$ and for $p < \bar{p} + \varepsilon$ but were constant at $D_t(\bar{p} + \varepsilon)$ for $p \in [\bar{p} + \varepsilon, p_t - \varepsilon]$. Although this demand curve would be discontinuous at $D(p_t - \varepsilon)$, continuous examples could be similarly constructed.

Finally, consider any equilibrium where $\partial W/\partial\bar{p} > 0$. 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 $\partial W/\partial\alpha$ can be made arbitrarily small without changing $(\partial W/\partial\bar{p})(d\bar{p}/d\alpha)$.

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.

References

BERGSTROM, T. AND MACKIE-MASON, J.K. "Some Simple Analytics of Peak-Load Pricing." *RAND Journal of Economics*, Vol. 22 (1991), pp. 241–49.

BOITEUX, M. "La tarification des demandes en point: application de la théorie de la vente au coût marginal." *Revue Général de l'Electricité*, Vol. 58 (1949), pp. 321–340 (translated as "Peak Load Pricing." *Journal of Business*, Vol. 33 (1960), pp. 157–179).

BORENSTEIN, S. "The Economics of Costly Risk Sorting in Competitive Insurance Markets." *International Review of Law and Economics*, Vol. 9 (1989), pp. 25–39.

———. "Time-Varying Retail Electricity Prices: Theory and Practice." In J. Griffin and S. Puller, eds., *Electricity Deregulation: Choices and Challenges*. Chicago: University of Chicago Press, 2005a.

———. "The Long-Run Efficiency of Real-Time Electricity Pricing." *Energy Journal*, Vol. 26 (2005b), pp. 93-116.

——— AND BUSHNELL, J.B. "An Empirical Analysis of the Potential for Market Power in California's Electricity Industry." *Journal of Industrial Economics*, Vol. 47 (1999), pp. 285–323.

——— AND HOLLAND, S.P. "Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices." Center for the Study of Energy Markets Working Paper no. 106, University of California Energy Institute, revised July 2003(a). Available at www.ucei.org/PDF/csemwp106.pdf.

——— AND ———. "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices." Center for the Study of Energy Markets Working Paper no. 116, University of California Energy Institute, August 2003(b). Available at www.ucei.org/PDF/csemwp116.pdf.

———, BUSHNELL, J.B., AND WOLAK, F.A. "Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market." *American Economic Review*, Vol. 92 (2002), pp. 1376–1405.

BRENNAN, T. "Market Failures in Real-Time Metering: A Theoretical Look." Resources for the Future Discussion Paper no. 02-53, October 2002.

BUSHNELL, J.B. "Looking for Trouble: Competition Policy in the U.S. Electricity Industry." In J. Griffin and S. Puller, eds., *Electricity Deregulation: Choices and Challenges*. Chicago: University of Chicago Press, 2005.

CARLTON, D.W. "Peak Load Pricing with Stochastic Demand." *American Economic Review*, Vol. 67 (1977), pp. 1006–1010.

———. "The Rigidity of Prices." *American Economic Review*, Vol. 76 (1986), pp. 637–658.

CHAO, H.-P. "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty." *Bell Journal of Economics*, Vol. 14 (1983), pp. 179–190.

CREW, M.A., FERNANDO, C.S., AND KLEINDORFER, P.R. "The Theory of Peak-Load Pricing: A Survey." *Journal of Regulatory Economics*, Vol. 8 (1995), pp. 215–248.

DANA, J.D. "Using Yield Management to Shift Demand When the Peak Time Is Unknown." *RAND Journal of Economics*, Vol. 30 (1999), pp. 456–474.

DOUCET, J.A. AND KLEIT, A. "Metering in Electricity Markets: When Is More Better?" In M.A. Crew and J.C. Schuh, eds., *Markets, Pricing, and Deregulation of Utilities*, Boston: Kluwer Academic Publishers, 2002.

JASKE, M. "Practical Implications of Dynamic Pricing." In S. Borenstein, M. Jaske, and A. Rosenfeld, *Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets*, Center for the Study of Energy Markets Working Paper no. 105, University of California Energy Institute, October 2002. Available at www.ucei.org/PDF/csemwp105.pdf.

JOSKOW, P.L. AND KAHN, E. "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market During Summer 2000." *Energy Journal*, Vol. 23 (2002), pp. 1–35.

——— AND TIROLE, J. "Retail Electricity Competition." Working Paper no. 10473, National Bureau of Economic Research, May 2004.

MANSUR, E.T. "Vertical Integration in Restructured Electricity Markets: Measuring Market Efficiency and Firm Conduct." *Journal of Laws and Economics*, forthcoming.

PANZAR, J.C. "A Neoclassical Approach to Peak-Load Pricing." *Bell Journal of Economics*, Vol. 7 (1976), pp. 521–530.

——— AND SIBLEY, D.S. "Public Utility Pricing Under Risk: The Case of Self-Rationing." *American Economic Review*, Vol. 68 (1978), pp. 888–895.

SAMUELSON, P.A.. "The Consumer Does Benefit from Feasible Price Stability." *Quarterly Journal of Economics*, Vol. 86 (1972), pp. 476–493.

STEINER, P.O. "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics*, Vol. 71 (1957), pp. 585–610.

WENDERS, J.T. "Peak Load Pricing in the Electric Utility Industry." *Bell Journal of Economics*, Vol. 7 (1976), pp. 232–241.

WILLIAMSON, O.E. "Peak-Load Pricing and Optimal Capacity Under Indivisibility Constraints." *American Economic Review*, Vol. 56 (1966), pp. 810–827.

———. "Peak-Load Pricing: Some Further Remarks." *Bell Journal of Economics and Management Science*, Vol. 5 (1974), pp. 223–228.