CSEM
Center for the Study of Energy Markets

CSEM WP 106R

Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices

Severin Borenstein and Stephen P. Holland

Revised July 2003

UCEI
University of California Energy Institute

2510 Channing Way, # 5180
Berkeley, California 94720-5180
www.ucei.org

Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices

Severin Borenstein and Stephen P. Holland[1]

July, 2003

Abstract: The standard economic model of efficient competitive markets relies on the ability of sellers to charge prices that vary as their costs change. Yet, there is no restructured electricity market in which most retail customers can be charged realtime prices (RTP), prices that can change as frequently as wholesale costs. We analyze the impact of having some share of customers on time-invariant pricing in competitive electricity markets. Not only does time-invariant pricing in competitive markets lead to outcomes (prices and investment) that are not first-best, it even fails to achieve the second-best optimum given the constraint of time-invariant pricing. We then study a number of policy interventions that have been proposed to address the perceived inadequacy of capacity investment. We show that attempts to correct the level of investment through taxes or subsidies on electricity or capacity are unlikely to succeed, because these interventions create new inefficiencies. We demonstrate that the most common proposal, a subsidy to capacity ownership financed by a tax on retail electricity, is particularly problematic. An alternative approach to improving efficiency, increasing the share of customers on RTP, has some surprising effects. We show that such a change lowers the equilibrium price to flat rate customers and makes them better off, but it makes incumbent RTP customers worse off. Also, an increase in RTP customers, does not necessarily reduce equilibrium capacity investment. If the equilibrium flat rate is higher than optimal, then an increase in RTP customers improves welfare. However, if the equilibrium flat rate is lower than optimal, such an increase does not necessarily improve welfare. We present an example in which welfare decreases, but the construction of the example suggests it is not likely to be policy relevant. Finally, we demonstrate that the analysis is robust to inclusion of a simple form of reserve capacity.

[1] *Borenstein:* Director of the University of California Energy Institute (www.ucei.org) and E.T. Grether Professor of Business Administration and Public Policy at the Haas School of Business, U.C. Berkeley (www.haas.berkeley.edu). Email: borenste@haas.berkeley.edu. *Holland:* Visiting Researcher, University of California Energy Institute. Email: sholland@uclink.berkeley.edu. Our thinking on this topic has benefitted a great deal from discussions with Jim Bushnell and Erin Mansur.

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

Perhaps nowhere is the disconnect between retail pricing and wholesale costs so great as in restructured electricity markets. In the last decade, it has become apparent that wholesale electricity price fluctuations can be extreme, but retail prices have in nearly all cases 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 demand, it is not uncommon for wholesale prices within one day to vary from five cents to twenty-five cents or more per kilowatt-hour while retail prices do not adjust at all.

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

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

Closely tied to time-invariant retail pricing is the issue of investment adequacy. Many participants in the electricity industry have argued, generally without an economic explanation, that deregulated electricity markets will result in inadequate investment in

[2] See Steiner (1957), Boiteaux (1960), Wenders (1976) and Panzar (1976).

production capacity. While this clearly is not the case with peak-load pricing under regulation – as explored by Steiner (1957), Boiteaux (1960), Wenders (1976) and Panzar (1976) – 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 when some customers are on flat retail rates. Not only is the level of investment not the first-best level that results when all customers are on RTP, it is not even the second-best optimal level of capacity investment *given the constraint that some customers cannot be charged realtime prices.*

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

We also analyze an alternative approach to improving the efficiency of competitive markets: increasing the share of customers on RTP. While this approach has been advocated by economists, it has not been studied extensively, and we show that it can have some surprising effects on investment and efficiency.

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

We begin in section I by presenting a model of competitive wholesale and retail electricity markets in which some share of customers is able to be charged realtime electricity prices. We demonstrate the short-run pricing and long-run investment inefficiency that results from the inability to charge all customers realtime prices. In section II, we explore the possible use of subsidies or taxes to overcome the inefficiency from such "inaccurate" retail pricing. In section III, we examine the effect of changing the proportion of customers on RTP and derive a somewhat surprising result that increasing this proportion does not necessarily reduce the equilibrium investment in capacity. The model we use for the analysis thus far does not incorporate a demand for reserve capacity, capacity paid to stand by for use in case it is needed to equilibrate supply and demand. This increases the complexity of the analysis, but we present a simplified model with reserve capacity in section IV. We conclude in section V.

I. Competition in wholesale and retail electricity markets

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

Since electricity cannot be stored economically, demand must equal supply at all times. Assume there are T periods per day with retail demand in period t given by $D_t(p)$ where $D'_t < 0$.[3] A fraction, α, of the customers pay realtime prices, *i.e.,* retail prices that vary hour to hour. The remaining fraction of customers, $1 - \alpha$, pay a flat retail price $\bar{p}$. We assume that $\alpha \in (0, 1]$ is exogenous and that customers on realtime pricing do not differ systematically from those on flat-rate pricing.[4] Aggregate demand from the customers is then $\tilde{D}_t(p, \bar{p}) = \alpha D_t(p) + (1 - \alpha)D_t(\bar{p})$ which implies that $\tilde{D}_t$ is decreasing in $\bar{p}$ and p. Note that $\tilde{D}_t(\bar{p}, \bar{p}) = D_t(\bar{p})$. For $p > \bar{p}$, the flat-rate customers do not decrease consumption in response to the higher realtime price so $\tilde{D}_t(p, \bar{p}) > D_t(p)$, and for $p < \bar{p}$, the flat-rate customers do not increase consumption in response to the lower realtime price so $\tilde{D}_t(p, \bar{p}) < D_t(p)$. Finally, $\tilde{D}_t(p, \bar{p})$ is decreasing in α for $p > \bar{p}$, and $\tilde{D}_t(p, \bar{p})$ is increasing in α for $p < \bar{p}$. That is, increasing alpha increases the elasticity of wholesale demand by rotating $\tilde{D}_t$ around the point $(D_t(\bar{p}), \bar{p})$.

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

Generators install capacity and sell electricity in the wholesale market. Assume that each identical generator is small relative to the market and can produce up to capacity at constant marginal cost c. Further assume that generators can install incremental capacity at constant cost r per unit of capacity per day. Let K be the total installed generation

[3] Following the literature on peak-load pricing, we also assume that cross price elasticities between demands in different periods are zero.

[4] Incorporating cross-elasticities among periods and endogenizing the choice of pricing system for a customer are both areas for further research, but we believe that the basic insights from our analysis carry over to those cases.

capacity. Thus, the short-run supply curve for electricity is L-shaped. In the long run, generators can add or retire capacity. Profits for the generators are then $\sum_{t=1}^{T}(w_t - c)\tilde{D}_t(p_t, \bar{p}) - rK$ per day where w_t is the wholesale price in period t.

Figure 2 illustrates the L-shaped, short-run supply curve with capacity K and shows demand curves for six different time periods. In the four low-demand periods, capacity is not fully utilized so a market-clearing price would equal marginal cost. At capacity, a market-clearing price can be determined by the intersection of the supply and demand curves. Capacity would be fully utilized in the two high-demand periods.

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. The retail firms choose realtime retail prices, p_t, and the flat retail rate, $\bar{p}$, engaging in Bertrand competition over these prices. Bertrand competition represents quite accurately the competition among retail electricity providers, because they would be price takers in the wholesale market, would be selling a nearly homogeneous product in the retail market, and would face no real capacity constraints. Profit of the retail sector is given by $\sum_{t=1}^{T}(\bar{p} - w_t)(1 - \alpha)D_t(\bar{p}) + (p_t - w_t)\alpha D_t(p_t)$ per day. Since electricity cannot be stored economically, demand greater than capacity in any period would require non-price rationing. The flat retail price, $\bar{p}$, is *feasible* if there exists some p such that $\tilde{D}_t(p, \bar{p}) \leq K$ for all t, *i.e.,* if rationing is not necessary. In other words, $\bar{p}$ is feasible if enough customers are on RTP to allow the wholesale market to clear at some price.

A. Competitive equilibrium in wholesale and retail markets

Equilibrium prices in the retail sector are determined by competition among retailers. First, consider the customers on RTP. If a realtime price, p_t, were greater than the wholesale price, a competitor 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 wholesale and retail realtime price is $p^e_{t_{SR}} = w_t$ for every t. In other words, competition among retailers drives retail prices for RTP customers to be equal to wholesale prices in each period.

Similarly, competition forces the flat retail rate to be set to cover exactly the cost of providing electricity to the flat-rate customers. Since this implies zero profits for the retail sector, the condition $\sum_{t=1}^{T}(\bar{p}^e_{SR} - w_t)(1 - \alpha)D_t(\bar{p}^e_{SR}) = 0$ determines the short-run equilibrium flat retail price $\bar{p}^e_{SR}$. Note that this zero profit condition can be written $\bar{p}^e_{SR} = \sum_{t=1}^{T} w_t D_t(\bar{p}^e_{SR}) / \sum_{t=1}^{T} D_t(\bar{p}^e_{SR})$. In other words, the equilibrium flat retail price is a weighted average of the realtime wholesale (and retail) prices where the weights are the relative quantities demanded by the customers facing a flat retail price. Thus, competition among retailers drives $\bar{p}^e_{SR}$ to be equal to the demand-weighted average wholesale price.[5]

In the short run, equilibrium prices in the wholesale market are determined by the intersection of the demand curve and the L-shaped supply curve in each period. If $\tilde{D}_t(c, \bar{p}) \leq K$, then the equilibrium wholesale price, w_t, is c. If $\tilde{D}_t(c, \bar{p}) > K$, then the equilibrium price is such that $\tilde{D}_t(w_t, \bar{p}) = K$. As in standard peak-load pricing models, the generators make positive short-run profits (scarcity rents) when capacity is fully utilized but make no short-run profits when capacity is not fully utilized. Thus, we have:

Result 1: Short-run Competitive Equilibrium — *For a given capacity, K, and share of customers on RTP, α, the short-run competitive equilibrium is characterized by realtime retail prices $p^e_t = w^e_t$ and flat 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 prices are determined by $(w^e_t - c) \cdot (\tilde{D}_t(p^e_t, \bar{p}^e) - K) = 0$.*

The equilibrium characterized in Result 1 is illustrated in Figure 3 for two demand periods: peak and off peak. Since capacity is not fully utilized off peak, the equilibrium wholesale price p_{op} is c. On peak, capacity is fully utilized, so the equilibrium wholesale price p_p is determined by the intersection of the aggregate demand $\tilde{D}_p$ and the L-shaped supply curve. The equilibrium flat rate $\bar{p}^e$ is the demand-weighted average of p_p and p_{op}. The demand-weighted average $\bar{p}^e$ is closer to p_p than to p_{op} since the flat-rate customers demand more in the peak than off peak.

In the long-run, generation capacity will enter (exit) the wholesale market as long as profits are positive (negative). Thus, competition drives long-run profits to zero. The zero profit condition on the wholesale sector is $\sum_{t=1}^T (w_t - c)\tilde{D}_t(p_t, \bar{p}) = rK$. Since competitive margins are positive only when capacity is fully utilized, the long-run equilibrium condition can be written $\sum_{t=1}^T (w_t - c) = r$. Thus,

Result 2: 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 in Result 1 and $\sum_{t=1}^T (w_t - c) = r$.*

Figure 2 also illustrates the long-run competitive equilibrium for the case in which there are two periods and only one period has positive scarcity rents, so the long-run peak price is $p_p = c + r$.

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

A question remains about the feasibility of this competitive equilibrium in the short and long run. To see that the equilibrium flat price is always feasible in the short run, define $\bar{p}^{min}(K)$ as the smallest feasible flat retail price for installed capacity K. Assume each D_t has a choke price and $\hat{T}$ is the period with highest demand.[6] Thus $\bar{p}^{min}(K)$ is defined implicitly by $K = (1-\alpha)D_{\hat{T}}(\bar{p}^{min}(K))$. If any flat retail price greater than $\bar{p}^{min}(K)$ has zero profit for the retail sector, then that price is a short-run equilibrium price and the equilibrium price is feasible. If every price greater than $\bar{p}^{min}(K)$ has positive profit, then $\bar{p}^{min}(K)$ is the equilibrium price. To see this, note that the equilibrium realtime price in $\hat{T}$ is not uniquely defined by the short-run supply and demand curves, both of which are vertical at K, *i.e.,* any price exceeding the choke price could be the equilibrium realtime price in period $\hat{T}$. Therefore, competition between retailers will bid up the equilibrium wholesale price in period $\hat{T}$ until there are no retail profits in the short-run. In other words, if there could be excess retail profits by charging $\bar{p}^{min}(K)$, retail competition would force up the wholesale price in period $\hat{T}$ and transfer the excess profit to the generators.[7] Therefore $\bar{p}^{min}(K)$ is the equilibrium flat rate and the short-run equilibrium price is always feasible. Feasibility of the long-run equilibrium price is implied by feasibility of the short-run equilibrium price.[8]

B. (In)efficiency of competitive equilibrium

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

Result 3: Efficiency When All Customers Are on RTP ($\alpha = 1$) — *If all customers are on realtime prices, i.e.,* $\alpha = 1$*, then the competitive equilibrium is Pareto efficient, i.e., attains the first-best electricity allocation and capacity investment.*

Result 3 follows immediately once $\alpha = 1$, because there is no missing market and all

[6] If each D_t does not have a choke price, then a similar argument holds where $\bar{p}^{min}(K)$ is the infimum of the feasible flat retail prices.

[7] If retail competition forces the flat price to $\bar{p}^{min}(K)$, the realtime prices can be higher than the choke price. This implies that realtime prices may not be an accurate measure of willingness to pay for electricity in periods of binding capacity.

[8] If demand does not have a choke price, equilibrium realtime prices can still serve to transfer potential excess profits from the retail sector to the generators by being bid up arbitrarily high.

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. For each period t if $D_t(c) \leq K$, then the competitive price for period t is $w_t = p_t = c$, and if $D_t(c) > K$, then the competitive price will clear the market, so it is defined implicitly by $D_t(p_t) = K$. These competitive outcomes are also the welfare maximizing prices for each period.

For the long run, define γ_t as the marginal social value of capacity in period t. The marginal social value of capacity, γ_t, is equal to the marginal value of output minus the marginal cost of variable inputs. The social optimum would dictate building capacity so long as $\sum_{t=1}^{T} \gamma_t > r$ and stopping when $\sum_{t=1}^{T} \gamma_t = r$. This, however, mirrors the private incentive to construct capacity if all customers are on realtime pricing. When all customers face the realtime prices, utility maximization implies that $p_t(= w_t)$ is the marginal value of output in period t. Since the marginal cost is c, the profit margin reflects the marginal social value of capacity in each period, *i.e.,* $w_t - c = \gamma_t$. Thus, a price-taking firm will find it profitable to build one more unit of capacity, *i.e.,* if $\sum_{t=1}^{T}(w_t - c) > r$, precisely when it is socially optimal to build one more unit of capacity, *i.e.,* if $\sum_{t=1}^{T} \gamma_t > r$. Thus, private incentives for investment accurately reflect social incentives when all customers are on realtime pricing.

If $\alpha < 1$, however, the efficient outcome is lost.

Result 4: Inefficiency with Some Flat-Rate Customers ($\alpha < 1$) — *If some customers do not face the realtime prices, i.e., $\alpha < 1$, the competitive equilibrium is not Pareto efficient, i.e., does not attain the first-best electricity allocation and capacity investment.*

Consider the short run in which K is fixed, and the short-run supply function is L-shaped. 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 wholesale price. If $\tilde{D}_t(p_t, \bar{p}) = K$, then the inefficiency on the margin manifests as a misallocation of the fixed quantity between RTP and flat-rate customers, who face different prices and therefore have different marginal benefits of consumption. In periods in which production is less than capacity, the marginal inefficiency is simply from underconsumption by the flat-rate customers because they face a price greater than marginal cost.

While it is clear that first-best resource allocation will not occur when some customers are charged a flat rate, regardless of the level of that rate, there is still a question of what flat rate minimizes the deadweight loss that results. In particular, does the competitive equilibrium flat rate, $\bar{p}^e_{SR}$, minimize the deadweight loss associated with having flat-rate customers and, if not, could government policy improve resource allocation? To answer this question, consider the flat retail rate, $\bar{p}^*_{SR}$, and realtime prices $p^*_{t_{SR}}$ that minimize deadweight loss in the short run. $\bar{p}^*_{SR}$ and $p^*_{t_{SR}}$ can be found from the optimization:[9]

$$\max_{p_t,\bar{p}} \sum_{t=1}^{T} [\tilde{U}_t(p_t,\bar{p}) - c\tilde{D}_t(p_t,\bar{p})] - rK \qquad s.t. \qquad \tilde{D}_t(p_t,\bar{p}) \leq K \;\; \forall \; t \tag{1}$$

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.[10,11] Let λ_t be the Lagrange multiplier of the capacity constraint in period t. The optimization yields two first-order conditions.

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

$$\alpha\{U_t'(D_t(p_t)) \cdot D_t'(p_t) - c \cdot D_t'(p_t) - \lambda_t \cdot D_t'(p_t)\} = 0, \tag{2}$$

which, since $U_t'(D_t(p_t)) = p_t$, implies that $p_t - c = \lambda_t$. The complementary-slackness conditions, $(p_t - c)(K - \tilde{D}_t(p_t,\bar{p})) = 0$ for every t, determine the optimal realtime prices, and show that λ_t is non-zero only if the capacity constraint is binding. Note that the shadow value of capacity, λ_t, reflects the marginal value of output to the *realtime* customers since they would benefit directly from additional capacity.

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

$$\sum_{t=1}^{T} [\bar{p}^*_{SR} - c - \lambda_t](1-\alpha)D_t'(\bar{p}^*_{SR}) = 0. \tag{3}$$

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

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

[11] 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, $\sum (w_t - c)\tilde{D}_t(p_t,\bar{p}) - rK$. Note that w_t is simply a transfer and does not affect deadweight loss.

Substituting $w^*_{t_{SR}} - c$ for λ_t for all t in [3] yields

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

which implies

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

We refer to the result of this optimization as the *second-best optimal allocation*. Thus, the flat retail price that minimizes the deadweight loss is a weighted average of the realtime prices where the weights are the relative slopes of the demand curves.[12] Recall that the equilibrium flat rate is a weighted average of the realtime prices where the weights are the relative quantities demanded by the flat-rate customers, *i.e.,* $\bar{p}^e_{SR} = \sum_{t=1}^{T} w^e_{t_{SR}} D_t(\bar{p}^e_{SR}) / \sum_{t=1}^{T} D_t(\bar{p}^e_{SR})$. Since both $\bar{p}^e_{SR}$ and $\bar{p}^*_{SR}$ are weighted averages of the realtime prices but their weights are not necessarily equal, comparison of the two weighted averages implies that $\bar{p}^e_{SR}$ does not necessarily equal $\bar{p}^*_{SR}$. Thus,

Result 5: Non-attainment of the Second Best in the Short Run — *The short-run competitive equilibrium does not attain the second-best optimal electricity allocation. Furthermore, the equilibrium flat rate, $\bar{p}^e_{SR}$, can be either higher or lower than optimal.*

To see that the equilibrium flat retail price may be either too high or too low, consider a simple example with two time periods, peak and off-peak, where realtime prices, w_p and w_{op}, are such that $w_p > w_{op} = c$. If the off-peak demand is steeper than the peak demand, the second-best flat retail price is closer to w_p than to w_{op}.[13] In fact, for a steep off-peak demand, $\bar{p}^*_{SR}$ can be arbitrarily close to w_p.[14] Since the equilibrium flat retail price is between the peak and off-peak prices, $\bar{p}^*_{SR} > \bar{p}^e_{SR}$ when off-peak demand is sufficiently steep relative to peak demand. Conversely, if peak demand is relatively steep, then $\bar{p}^*_{SR} < \bar{p}^e_{SR}$. Thus the equilibrium flat retail rate can be either too high or too low.

Figure 4 illustrates the case where off-peak demand is perfectly inelastic and $\bar{p}^*_{SR} > \bar{p}^e_{SR}$. The equilibrium flat price, $\bar{p}^e_{SR}$, is a demand-weighted average of the peak wholesale

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

[13] This follows since the consumption distortion is greater when demand is flatter. Thus the consumption distortion (and deadweight loss) is minimized by a flat retail price close to the realtime price in the time period with flatter demand. Mathematically, if demand is flatter, D'_t is large so the weighted average puts more weight on the price with flatter demand.

[14] If off-peak demand were perfectly inelastic, there would be no consumption distortion off peak, and the second-best flat price would equal w_p.

price p_p^e and off-peak wholesale price $p_{op}^e = c$. Since off-peak demand is perfectly inelastic, there is no inefficiency off-peak since the prices $p_{op}^e = c$ and $\bar{p}_{SR}^e$ induce customers to consume the same amounts. Because customers are not price sensitive off-peak, increasing or lowering the flat rate does not affect consumption (or efficiency) off-peak. Note, however, that increasing the flat rate in this case improves efficiency since it does not affect consumption off-peak but improves the peak-period misallocation between flat-rate and RTP customers. Clearly, setting $\bar{p}_{SR}^* = p_p^*$ eliminates the peak-period misallocation since flat-rate and RTP customers both face the same prices.[15] Note also that increasing the flat rate from $\bar{p}_{SR}^e$ to $\bar{p}_{SR}^*$ decreases the peak demand from $\tilde{D}_p$ to $\tilde{D}_p'$ and lowers the peak realtime price.

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

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

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

Although competition distorts the consumption of the flat-rate customers relative to the second best, competition does not introduce additional distortions into the realtime market for a given flat rate. For a given $\bar{p}$, the optimal realtime prices are determined by the first-order and complementary-slackness conditions from the planner's problem, which imply (since $w_t = p_t$) that $(w_t - c)(K - \tilde{D}_t(w_t, \bar{p})) = 0$ for every t. Note that these optimal prices are exactly the realtime prices that would result from competition, given a $\bar{p}$, namely, $w_t = c$ if $\tilde{D}_t(c, \bar{p}) < K$ and, otherwise, w_t is defined implicitly by $\tilde{D}_t(w_t, \bar{p}) = K$. Thus, if a regulator were to force the retail sector to charge $\bar{p}_{SR}^*$ to flat-rate customers, the realtime prices resulting from retail competition would be second-best optimal. In this manner, the second-best optimal allocation could be achieved in the short run.

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

C. Inefficiency in the long run

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

To determine the second best in the long run, consider the flat retail rate, $\bar{p}^*_{LR}$, realtime prices, $w^*_{t_{LR}}$, 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.[16] The first-order conditions for w_t and $\bar{p}$ are given by [2] and [3] and the first-order condition for K is

$$\sum_{t=1}^{T} \lambda_t = r \qquad [6]$$

Since [2] implies $\lambda_t = w_t - c$, the first order conditions can be solved in terms of the realtime prices. As in the short run, the second-best price, $\bar{p}^*_{LR}$, is a weighted average of the realtime prices where the weights are the relative slopes of the demand curves. The complementary slackness condition, which determines optimal realtime prices, is $(w^*_{t_{LR}} - c)(K^*_{LR} - \tilde{D}_t(w^*_{t_{LR}}, \bar{p})) = 0$ for every t. Since λ_t is equal to the margin that competitive generators would earn in each period, equation [6] implies that the second-best optimal capacity yields daily operating profits on the margin that are just equal to the daily cost of capital, or zero profits net of capital costs.

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

Result 6: Non-attainment of the Second Best in the Long Run — *The long-run competitive equilibrium does not attain the second-best optimal electricity allocation and capacity investment. Furthermore, the equilibrium flat rate, $\bar{p}^e_{LR}$, can be either higher or lower than optimal, and the equilibrium capacity investment, K^e_{LR}, can be either larger or smaller than optimal.*

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

To see that K^e_{LR}, can be either larger or smaller than K^*_{LR}, suppose that demand elasticities 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 regulator tries to improve efficiency by increasing the flat retail price to $\bar{p}^*_{LR}$. In the short run, this will simply shift consumption from the flat-rate customers to the realtime customers in periods in which capacity is binding. Since consumption of the realtime customers has increased, their marginal benefit of consumption has decreased in these periods. This implies that the shadow value of capacity has strictly decreased in all periods for which it was positive. Since the shadow values of capacity in the long-run equilibrium, $w_t - c$, were such that $\sum_{t=1}^{T} w_t - c = r$, the sum of the shadow values now must be less than r. Therefore, in the long run, the regulator would increase surplus by decreasing capacity. This implies that the equilibrium long-run capacity was too large relative to the second-best optimal long-run capacity given that $\bar{p}^*_{LR} > \bar{p}^e_{LR}$. A symmetric argument shows that $K^*_{LR} > K^e_{LR}$ if $\bar{p}^*_{LR} < \bar{p}^e_{LR}$.

As in the short run, the distortion in the competitive equilibrium stems from the flat retail price. In particular, if a regulator were to impose the optimal flat rate, $\bar{p}^*_{LR}$, then competition would lead to the second-best optimal realtime prices and capacity investment, K^*_{LR}. As above, the complementary slackness conditions insure that the competitive realtime prices are optimal given $\bar{p}^*_{LR}$. In addition, the condition $\sum_{t=1}^{T} w_t - c = r = \sum_{t=1}^{T} \lambda_t$ insures that competitive investment is optimal provided that the regulator imposes the optimal flat retail price.

II. Subsidies/Taxes on Capacity or Electricity

In restructured wholesale electricity markets, many parties have suggested that in order to assure sufficient investment in generation, “capacity payments” to producers are necessary. These payments directly subsidize the holding of capacity, generally without a commitment on the producer’s part to offer any certain quantity of energy or any certain price.[17] Such payments can be seen as part of a general category of market interventions designed to move the equilibrium outcome closer to the (constrained) social optimum. In this section, we consider such policies.

Among such interventions, there are two characteristics that are central to the economic analysis of the policy. First, the subsidy/tax can be directed at the retail price of electricity or it can be directed at capacity. Second, the revenues from a subsidy/tax can

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

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, generally the flat-rate group because the RTP group begins from a second-best optimum.

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

A. Subsidies or taxes on retail electricity with external financing

The very simplest policy intervention to analyze is a tax or subsidy on flat-rate retail electricity prices. Such a tax would drive up the retail price paid by the flat-rate customers thereby decreasing wholesale demand during all periods. The decrease in wholesale quantity demanded would cause wholesale prices to decrease, generators to exit in the long run, and industry generation capacity to decrease. A subsidy to the flat-rate retail price would have the opposite effect.

We can characterize the long-run competitive equilibrium with a retail tax τ on the flat-rate customers. As in Result 2, realtime customers pay the wholesale prices, *i.e.,* $p_t = w_t$; wholesale demand equals supply, *i.e.,* $(w_t - c)[K - \tilde{D}_t(p_t, \bar{p})] = 0$; and wholesale profits cover capacity costs, *i.e.,* $\sum_{t=1}^{T}(w_t - c) = r$. In the flat-rate retail market, however, there is now a tax wedge between the flat-rate price paid by the customers $\bar{p}$ and the flat rate received by the retail sector, $\bar{p} - \tau$. Thus, the equilibrium flat rate is determined by $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$.

Given this characterization of the equilibrium, it is straightforward to show that the optimal tax or subsidy will be $\tau^* = \bar{p}^*_{LR} - \sum_{t=1}^{T} w^*_t D_t(\bar{p}^*_{LR}) / \sum_{t=1}^{T} D_t(\bar{p}^*_{LR})$ charged to all customers paying a flat retail rate. The second term is the quantity-weighted average price of buying wholesale power for flat-rate customers when the flat rate is $\bar{p}^*_{LR}$. Thus, τ^* is the tax or subsidy that allows the retailer to break even while charging $\bar{p}^*_{LR}$ and, therefore, $\bar{p}^*_{LR}$ will be the competitive equilibrium flat rate when a tax of τ^* is imposed on all electricity consumption by flat-rate customers. τ^* may be positive or negative depending on whether $\bar{p}^*_{LR}$ is greater or less than $\bar{p}^e_{LR}$.[18] As was the case earlier, if the flat-rate is optimal (now including the tax/subsidy), competition does not introduce any additional distortions in consumption of the realtime customers or in investment. Thus the second-best optimum

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

can be attained with a retail tax or subsidy τ^* charged to the flat-rate customers.[19]

Result 7: Optimality of Retail Tax/Subsidy on Flat-Rate Retail Customers — *With external financing, a tax/subsidy $\tau^* = \bar{p}^*_{LR} - \sum_{t=1}^{T} w^*_t D_t(\bar{p}^*_{LR}) / \sum_{t=1}^{T} D_t(\bar{p}^*_{LR})$ on the flat-rate customers achieves the second-best optimal allocation and capacity investment. The optimal policy may be a tax or a subsidy.*

Result 7 can be illustrated with Figure 4 in the short-run. If the retail sector were to charge the flat rate $\bar{p}^*_{SR}$, profits would be positive since its margin on the flat-rate customers is positive in the off-peak, but its margin (loss) is zero in the peak. Taxing the flat-rate customers forces the equilibrium flat rate up which improves efficiency. Note that in the short run, this would decrease wholesale demand so capacity will exit in this example.

While a tax/subsidy on flat-rate customers can achieve the second-best optimal price, a tax/subsidy on all retail customers (flat-rate and RTP) cannot. If all retail customers are taxed, there are tax wedges in both the realtime and flat-rate markets. The equilibrium is then characterized by the equality of wholesale demand and supply, *i.e.,* $(w_t - c)[K - \tilde{D}_t(p_t, \bar{p})] = 0$; and wholesale profits covering capacity costs, *i.e.,* $\sum_{t=1}^{T}(w_t - c) = r$; plus the two conditions on the distorted markets: $p_t - \tau = w_t$ and $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$.

A tax/subsidy on all retail customers cannot achieve the second best because the RTP customers are served optimally absent the tax/subsidy, as was shown in the previous section. Setting τ to achieve the optimal second-best optimal price for flat-rate customers distorts the prices for RTP customers away from the second-best optimal level for them that is achieved if no tax/subsidy is applied to RTP customers.[20]

Result 8: Non-optimality of Retail Tax/Subsidy on All Retail Customers — *With external financing, a tax/subsidy on all flat-rate customers cannot achieve the second-best optimal allocation and capacity investment.*

Result 8 is illustrated in Figure 5. Without the retail tax, the equilibrium flat rate is $\bar{p}^e_{LR}$, the peak price is $c + r$, and the off-peak price is c. A tax on all electricity can increase the equilibrium flat rate to $\bar{p}^*_{LR}$ which decreases wholesale demand and causes

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

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

capacity to exit. If the tax were applied only to flat-rate retail customers, as in Result 7, the long-run peak price would be $c + r$ and the off-peak price would be c and capacity investment would be at the second-best optimal level K^*_{LR}. However, taxing the realtime customers τ implies that in the long run the realtime prices must rise to $c+r+\tau$ and $c+\tau$ so that the wholesale prices are $c + r$ and c. The tax τ introduces deadweight loss (shown by the shaded triangles) into the realtime market so the second-best optimal electricity allocation is not attained. Note also that the reduction in realtime consumption from the tax implies that additional capacity exits, *i.e.,* K' is less than K^*_{LR}.

B. Capacity subsidies/taxes with external financing

Though retail taxes/subsidies may seem the natural policy instrument to address the efficiency problem caused by flat retail pricing, the public policy debate has focused on taxes or subsidies (actually, just subsidies) for capacity. A policy of subsidizing generation capacity can affect efficiency, by lowering the capital cost of new generation and inducing new entry, thereby driving down wholesale prices. This would increase profits of the retailers which would drive down the equilibrium flat retail rate. As above, this could improve efficiency if the equilibrium flat retail price were too high, *i.e.,* $\bar{p}^e > \bar{p}^*$.

To see if the second-best allocation can be attained in competitive equilibrium, consider a capacity subsidy (or tax, if negative), σ, from the general fund that changes the producer's cost of capital from r to $r-\sigma$. The long-run competitive equilibrium can be characterized by: $p_t = w_t$; $\bar{p} = \sum_{t=1}^T w_t D_t(\bar{p}) / \sum_{t=1}^T D_t(\bar{p})$; and $(w_t - c)[K - \tilde{D}_t(p_t, \bar{p})] = 0$; plus the tax-distorted condition in the generation capacity market, *i.e.,* $\sum_{t=1}^T (w_t - c) = r - \sigma$.

Since both $\bar{p}^*_{LR}$ and $\bar{p}^e_{LR}$ are weighted averages of realtime prices (with non-negative weights), it is straightforward to show that there is a σ that would yield an equilibrium $\bar{p}^e_{LR} = \bar{p}^*_{LR}$.[21] At that equilibrium, $\sigma = Tc + r - \sum_{t=1}^T p^*_t$ where $\sum_{t=1}^T p^*_t$ is such that $\bar{p}^*_{LR} = \sum_{t=1}^T p^*_t D_t(\bar{p}^*_{LR}) / \sum_{t=1}^T D_t(\bar{p}^*_{LR})$. That is, the σ that yields $\bar{p}^*_{LR}$ is equal to the operating plus capital costs less the sum of the wholesale prices where the wholesale prices are such that $\bar{p}^*_{LR}$ is the average wholesale price weighted by the quantities demanded at $\bar{p}^*_{LR}$. Thus, the second-best optimal flat price can be attained with a capacity subsidy (or tax) from the general fund.

Note, however, that a capacity subsidy σ will lower some wholesale prices below the second-best optimal levels since $\sum_{t=1}^T w_t - c = r - \sigma < r$. This implies that customers paying the realtime prices are consuming more than their efficient amounts in some periods.

[21] From the previous section, $\bar{p}^e_{LR}$ is continuous in σ, is equal to c if σ is sufficiently positive and increases as σ grows more negative over the relevant range of prices.

In periods in which capacity is not fully utilized, the price is c with or without the capacity subsidy. In these periods, the capacity subsidy does not distort realtime consumption. However, if capacity is fully utilized, the capacity subsidy decreases the wholesale price, and realtime consumption is distorted above the second-best optimum. Thus, the capacity subsidy σ from the general fund can reduce the deadweight loss from the customers on the flat rate to zero but it will create deadweight loss in the consumption of the realtime customers in some periods.[22] This implies that

Result 9: Non-optimality of Capacity Tax/Subsidy — *With external financing, a capacity tax/subsidy cannot achieve the second-best optimal allocation and investment.*

Result 9 is illustrated in Figure 6. Let s be the capacity tax which results in $\bar{p}^*_{LR}$ as an equilibrium flat rate. Without the capacity tax s (σ is negative here), the long-run equilibrium flat rate is $\bar{p}^e_{LR}$, the peak price is $c + r$, and the off-peak price is c. The capacity tax raises the cost of capital to $c+r+s$. This implies that the peak realtime (and wholesale) price must rise to $c+r+s$. This induces deadweight loss (the shaded triangle) in the realtime peak consumption so the capacity tax s does not achieve the second-best optimal consumption or capacity investment. Note, however, that the off-peak realtime price remains c so there is no additional inefficiency in the off-peak realtime market from the capacity tax.

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

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

The retail electricity tax used to fund the capacity payments can be administered in a number of ways. First, we analyze the simpler case where the tax is levied only on the flat-rate customers. Combining the analyses above, the long-run competitive equilibrium can be characterized by: $p_t = w_t$; $\bar{p} - \tau = \sum_{t=1}^T w_t D_t(\bar{p}) / \sum_{t=1}^T D_t(\bar{p})$; $(w_t - c)[K - \tilde{D}_t(p_t, \bar{p})] = 0$; and $\sum_{t=1}^T (w_t - c) = r - \sigma$. The balanced-budget condition is $\tau \sum_{t=1}^T (1-\alpha) D_t(\bar{p}) = \sigma K$. The balanced-budget condition ensures that the tax revenue collected by the retail sector

[22] An optimal capacity subsidy would not equate the flat rate with $\bar{p}^*_{LR}$ but would allow some distortion in the flat-rate market in order to lessen the distortion in the peak realtime market. Clearly, the optimal capacity subsidy does not attain the second best.

exactly funds the capacity payments made to the wholesale sector.[23]

Such a capacity payment has two off-setting effects. The scheme includes a tax on the retail sector, which increases the equilibrium flat retail price, and a capacity subsidy to the wholesale sector, which decreases wholesale prices and, thereby, decreases the equilibrium flat retail price. Though at first it may seem that these effects would be offsetting, that isn't generally true. It is true, however, in the special case where there are no customers on RTP.

If there are no customers on RTP, then the realtime prices are irrelevant, and capacity is determined by the highest demand period. Let $\hat{T}$ be the highest demand period. The equilibrium with capacity payments is then characterized by $D_{\hat{T}}(\bar{p}) = K$; $w_{\hat{T}} = c + r - \sigma$ and $w_t = c$ for all other periods; $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$; and the balanced-budget condition $\tau \sum_{t=1}^{T} D_t(\bar{p}) = \sigma K$. Note, however, that this implies that

$$\begin{aligned} \bar{p} &= \tau + \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p}) \\ &= \tau + c + (r - \sigma) D_{\hat{T}}(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p}) = c + r D_{\hat{T}}(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p}). \end{aligned} \qquad [7]$$

[7] implies that the flat rate that results from competition under any level of capacity payments is exactly the flat rate that would result with no capacity payments, *i.e.,* the flat rate is a weighted average of wholesale prices $c + r$ in the highest peak period and c in every other period. Thus the capacity payments have no effect on the competitive equilibrium.

Result 10: Neutrality of Capacity Payments With No RTP Customers — *If all customers are on flat-rate pricing, then a capacity payment σ funded by an excise tax on electricity has no effect on the equilibrium prices, allocation or capacity investment.*[24]

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

[24] A capacity payment, σ, which is funded by the flat-rate customers, also has no effect if α is small and $D_t(c + r - \sigma) = 0 \;\; \forall t$, *i.e.,* demand has a choke price. If α is sufficiently small then capacity only binds in one peak period, defined as period $\hat{T}$. This implies that $w_{\hat{T}} = c + r - \sigma$ and that $w_t = c$ for all other periods. Since realtime demand is choked off in period $\hat{T}$, capacity is determined by $(1 - \alpha) D_{\hat{T}}(\bar{p}) = K$. The equilibrium is then fully characterized by the additional conditions $p_t = w_t$; $(w_t - c)[K - \tilde{D}_t(p_t, \bar{p})] = 0$; $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$; and the balanced-budget condition $\tau \sum_{t=1}^{T} (1 - \alpha) D_t(\bar{p}) = \sigma K$. As in [7], it is easy to show from this characterization that $\bar{p} = c + r D_{\hat{T}}(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$. This implies that the equilibrium flat rate with capacity payment σ is exactly the flat rate that would result if there were no capacity payment.

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

With the above characterization of the equilibrium, it can be shown that the second-best optimal price $\bar{p}^*_{LR}$ is the equilibrium outcome from a capacity payment of $\sigma^* = (1-\alpha)/K \cdot [\bar{p}^*_{LR} \sum_{t=1}^{T} D_t(\bar{p}^*_{LR}) - \sum_{t=1}^{T} p^*_t D_t(\bar{p}^*_{LR})]$ where $\sum_{t=1}^{T} p^*_t = Tc + r - \sigma^*$ and $(p^*_t - c)[K - \tilde{D}_t(p^*_t, \bar{p})] = 0$.[25] Note that this implies that σ^* should be positive if $\bar{p}^e_{LR} < \bar{p}^*_{LR}$, *i.e.,* if the equilibrium flat price is too low. This is equivalent to the effect of a tax on retail electricity. However, note also that although σ^* minimizes the deadweight loss in the flat-rate market, it leads to deadweight loss in the realtime market since $\sum_{t=1}^{T}(p^*_t - c) = r - \sigma^* < r$ implies that some of the realtime prices are too low.[26] As in the analysis of a capacity subsidy, this implies a distortion in the realtime market in periods in which capacity is fully utilized. This distortion implies,

Result 11: Non-optimality of Capacity Payments from Flat-rate Customers — *Capacity payments, funded by an excise tax on electricity sold to the flat-rate retail customers, cannot achieve the second-best optimal allocation and investment.*

Policy makers have generally proposed capacity payments to be funded by payments from all retail customers and not just the flat-rate customers. Combining the analyses of a retail tax on all customers from above with a capacity subsidy, the long-run competitive equilibrium can be characterized by: $p_t - \tau = w_t$; $\bar{p} - \tau = \sum_{t=1}^{T} w_t D_t(\bar{p}) / \sum_{t=1}^{T} D_t(\bar{p})$; $(w_t - c)[K - \tilde{D}_t(p_t, \bar{p})] = 0$; and $\sum_{t=1}^{T}(w_t - c) = r - \sigma$. The balanced-budget condition is now $\tau \sum_{t=1}^{T} \tilde{D}_t(p_t, \bar{p}) = \sigma K$. As above, the balanced-budget condition ensures that the revenue collected exactly covers the capacity subsidy where now revenue is collected from all customers.

Capacity payments funded by all retail customers have a number of effects. Consider a positive capacity payment. The capacity payment has two components: a tax on all

[25] Since σ^* is defined implicitly by highly non-linear equations, it is difficult to prove that a general solution exists to the system of equations. Examples can be shown where σ^* can be easily derived. For example, if α is small such that capacity is only fully utilized in one period and the choke price is such that $D_T(c+r)$ is positive, σ^* can be derived.

[26] Alternatively some of the realtime prices would be too high if σ were positive.

retail electricity customers and a subsidy to capacity. If the equilibrium flat rate was too low, the tax on the flat-rate customers may improve efficiency. However, it distorts RTP consumption in all periods by driving up the realtime prices. The capacity subsidy decreases the cost of capital which drives down the wholesale prices in periods in which capacity is fully utilized. This lessens the inefficiency of the realtime tax during peak periods but does nothing to lessen the inefficiency when capacity is not fully utilized. It follows that

Result 12: Non-optimality of Capacity Payments from All Customers — *Capacity payments, funded by an excise tax on electricity sold to all retail customers, cannot achieve the second-best optimal allocation and capacity investment.*

III. Changing Proportion of Customers on Realtime Pricing

This section examines the effect of changing the proportion of customers that are on RTP. There are three ways in which such a change might occur. First, policy makers could simply change the proportion of customers on RTP without regard to characteristics of the customers. Second, policy makers could treat heterogenous customers differently, *e.g.,* by putting the customers with the largest or most elastic peak demand on RTP. Third, policy makers could allow customers to choose whether to purchase electricity at realtime prices or at a flat rate. We address only the first of these policies by analyzing an exogenous change in α.[27]

A. The Effect of Increasing RTP Customers on Prices

Increasing the proportion of customers on RTP increases the elasticity of demand by rotating $\tilde{D}_t$ around $\bar{p}^e$. This has two effects on wholesale demand. For periods in which the wholesale price is above the flat rate, increasing α decreases demand since more customers face the higher realtime price. Because capacity is fully utilized in these periods, decreased demand drives down the wholesale price. Conversely, for periods in which the wholesale price is below the flat rate, demand *increases* with α since more customers face the lower realtime price. When the wholesale price is low, capacity may or may not be fully utilized. If capacity is not fully utilized, the increased demand does not affect the wholesale price. However, if capacity is fully utilized, the increased demand increases the wholesale price in

[27] In reality, because metering and billing costs are virtually independent of demand level, RTP has been considered primarily for large industrial and commercial customers. Our analysis goes through without change if RTP for large customers is mandatory and these customers have approximately the same *relative* demand across hours as do smaller customers. Borenstein (2002) discusses the possible impact of an RTP program open to voluntary participation.

these periods. Thus, in the short run, increasing α decreases some wholesale prices (*i.e.,* in peak periods) while increasing or not changing others (*i.e.,* in off-peak periods). We show in the next subsection that an increase in α can cause capacity to increase, decrease or stay the same in the long run.

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

Result 13: Effect of Increasing RTP Customers on Flat Retail Rate — *In the long run, an increase in the proportion of customers on RTP reduces $\bar{p}^e_{LR}$.*

The key to this result, the proof of which is presented in the appendix, is recognizing that $\bar{p}^e = \sum_{t=1}^{T} w_t^e D_t(\bar{p}^e) / \sum_{t=1}^{T} D_t(\bar{p}^e)$, the equilibrium $\bar{p}$ is a weighted average of wholesale prices with greater weight on high-demand periods, *i.e.,* those in which customers purchase the highest quantities if faced with $\bar{p}^e$ in all periods. The long run equilibrium in the wholesale market requires that $\sum_{t=1}^{T}(w_t - c) = r$, so the increase in customers on RTP, since it leaves c and r unchanged, must leave the unweighted average of w_t unchanged, $\sum_{t=1}^{T} \Delta w_t = 0$. One can then show that since putting more customers on RTP lowers the highest prices (associated with the highest quantities at a given $\bar{p}$) and raises the lowest price (associated with the lowest quantities at a given $\bar{p}$), if the unweighted average of prices is constant, this weighted average of prices must decline. 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.[28]

B. The Effect of Increasing RTP Customers on Capacity

One of the frequently-touted attractions of RTP is the belief that it will cut peak demand and therefore reduce the need for peaking capacity. Surprisingly, however, an increase in α does not necessarily decrease the equilibrium level of capacity; the equilibrium effect on capacity of changing α is indeterminate.

To illustrate a capacity decrease, the more intuitive case, return to the simple two-period model in which $p_p > p_{op} = c$ and assume that $\tilde{D}_{op}(c, \bar{p}) < K$. Recall that the long-run equilibrium requires that $p_p = c + r$. Marginally increasing α has two effects on peak demand. First, since more customers face the higher realtime price, peak demand decreases. Second, since, as argued above, the short-run equilibrium flat rate must decrease, peak demand increases. However, the first effect must dominate—otherwise $\bar{p}^e_{SR}$ would

[28] Here and throughout this section, we do not state results as being *weakly* true, though there are special cases where changes are zero. We note here that zero-change cases can be constructed, but they appear to have no generality, so we do not emphasize them.

increase—so that p_p falls.[29] Since the peak price falls, and the off-peak price remains constant, increasing α decreases wholesale profits. Therefore wholesale profits are negative in the short run, and equilibrium capacity decreases in the long run.

To show that increasing the proportion of customers on RTP can lead to increased investment, consider a two-period linear model in which capacity is fully utilized in both periods and $D_t(p_t) = A_t - B_t p_t$. Since $K = \tilde{D}_t(p_t, \bar{p})$ implies that $p_t = \frac{1}{\alpha}[\frac{A_t - K}{B_t} - (1-\alpha)\bar{p}]$ the retail profits can be written:

$$\pi^R = (1-\alpha)\sum_{t=1}^{2}[\bar{p} - p_t]D_t(\bar{p}) = \frac{1-\alpha}{\alpha}\sum_{t=1}^{2}[\bar{p} - \frac{A_t - K}{B_t}]D_t(\bar{p}). \qquad [8]$$

This implies that for a given flat rate the profits are proportional in $(1-\alpha)/\alpha$. Therefore, if retail profits are zero for a given flat rate, profits remain zero when more customers are put on RTP. Thus, the short-run equilibrium flat rate does not change with α.

However, increasing α does decrease the peak wholesale price and increase the off-peak wholesale price. Since the quantity-weighted average of the peak and off-peak prices does not change (*i.e.,* the short-run equilibrium flat rate does not change) and the average puts more weight on the peak price than the off-peak price, the off-peak price must change more than the peak price. This implies that the price increase off peak must be greater than the price decrease on peak, *i.e.,* the sum of the two prices has increased.[30] Since investment incentives depend on the wholesale profits, which depend on the sum of the wholesale prices, the increase in α leads to an increase in investment.

Result 14: Indeterminant Effect of Increasing RTP Customers on Capacity — *An increase in the proportion of customers on RTP can increase or decreases long-run equilibrium capacity K^e_{LR}.*

Result 14 is illustrated in Figure 7. As shown above, the short-run equilibrium flat rate does not change when the proportion of customers on RTP increases for linear demands. With a small number of customers on RTP, the aggregate demands are given by $\tilde{D}_t$, and the peak price is $c + r$ while the off-peak price is c. Starting from an initial capacity of

[29] A simple proof by contradiction shows that p_p falls. Starting from a long-run equilibrium, *i.e.,* $p_p = c+r$, increase α. Let $\tilde{p}_p$ be the short-run peak price after increasing α. Suppose that $\tilde{p}_p \geq c+r$. This implies that retail losses from serving the peak flat-rate customers would increase which implies that $\bar{p}$ must increase in the short run. But this contradicts the result from above that $\bar{p}^e$ must strictly decrease in the short run from increasing α. Therefore, $\tilde{p}_p < c + r$.

[30] The simple sum is relevant here because the peak and off-peak periods are assumed to be of equal length.

K_1, charging more customers the realtime prices decreases peak demand and decreases the peak wholesale price. Since capacity is not fully utilized off-peak, the price does not increase so wholesale profits decrease, and capacity decreases from K_1 due to exit. When sufficient customers are on RTP such that demands are $\tilde{D}_t'$, equilibrium capacity is K_2. Now putting more customers on RTP drives down the peak wholesale price but increases the off-peak wholesale price. These two offsetting effects on wholesale profits may imply that profits are positive. As argued above, wholesale profits will increase for linear demands so capacity will increase above K_2. As drawn in Figure 7, the first-best capacity, attained in equilibrium if all customers are on RTP, lies between K_1 and K_2.

C. The Effect of Increasing RTP Customers on Efficiency

As shown above, if all customers are on RTP, allocation and investment are efficient. When some customers are not on RTP, electricity is allocated inefficiently between the flat-rate and RTP markets. The question remains about the welfare effects of a marginal increase in the proportion of customers on RTP when $\alpha < 1$. This question is more subtle than it may appear at first glance since the welfare theorems are not applicable.[31]

To analyze the long-run welfare effects of increasing the proportion of customers on RTP, we analyze the surplus accruing to different groups: the generators, the retail service providers, the customers on RTP, the customers on flat-rate pricing, and the customers who switch from flat rates to RTP. First, the generators and retail service providers receive no surplus in the long run, so their surplus is unaffected by increasing α. Second, Result 13 shows that $\bar{p}^e_{LR}$ decreases in α. Therefore, the customers on flat-rate pricing consume more at a lower price. Thus, the flat-rate customers are better off with an increase in α.

Third, the customers who switch from the flat rate to RTP receive higher surplus. This can be shown by a revealed preferences argument. Since $\sum_{t=1}^{T} p_t D_t(\bar{p}) = \sum_{t=1}^{T} \bar{p} D_t(\bar{p})$, the switchers could consume exactly the same electricity quantities as the flat rate customers choose at the exact same total bill. Since they choose to consume different quantities, they must be strictly better off.

Finally, the surplus to the customers on RTP decreases in α. To see this, first note that the envelope theorem implies that the change in consumer surplus in period t is given

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

by $-\frac{dp_t}{d\alpha}D_t(p_t)$. Thus, the change in surplus to RTP customers is

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

$$= \frac{1-\alpha}{\alpha}\sum_{t=1}^{T}\frac{dp_t}{d\alpha}D_t(\bar{p}) \qquad [9]$$

The first equation follows since $\alpha D_t(p_t) + (1-\alpha)D_t(\bar{p}) = K$ whenever $\frac{dp_t}{d\alpha} \neq 0$, and the second equation follows since $\sum_{t=1}^{T}\frac{dp_t}{d\alpha} = 0$. We can show that [9] is negative by differentiating the zero profit retail condition $\sum_{t=1}^{T}\bar{p}D(\bar{p}) = \sum_{t=1}^{T}p_tD(\bar{p})$. Differentiation implies that

$$\frac{d\bar{p}}{d\alpha}\sum_{t=1}^{T}D_t(\bar{p}) + \bar{p}\sum_{t=1}^{T}D_t'(\bar{p})\frac{d\bar{p}}{d\alpha} = \sum_{t=1}^{T}[\frac{dp_t}{d\alpha}D_t(\bar{p}) + p_tD_t'(\bar{p})\frac{d\bar{p}}{d\alpha}]$$

which implies

$$\frac{d\bar{p}}{d\alpha}\sum_{t=1}^{T}[D_t(\bar{p}) + D_t'(\bar{p})(\bar{p}-p_t)] = \sum_{t=1}^{T}\frac{dp_t}{d\alpha}D_t(\bar{p}). \qquad [10]$$

Note that $(1-\alpha)\sum_{t=1}^{T}[D_t(p_t) + D_t'(\bar{p})(\bar{p}-p_t)]$ is the derivative of the retail profit function with respect to $\bar{p}$ taking the wholesale prices as given. Since the competitive equilibrium $\bar{p}$ results from Bertrand competition over the flat rates, the derivative must be greater than or equal to zero, i.e., $\sum_{t=1}^{T}D_t(p_t) + D_t'(\bar{p})(\bar{p}-p_t) \geq 0$. Since $\frac{d\bar{p}}{d\alpha} \leq 0$ by Result 13, the righthand side of [10] must be less than zero. Combining this with [9] shows that the consumer surplus to the RTP customers is decreasing in α.

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

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

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

We have shown that in the competitive equilibrium, $\frac{\partial W}{\partial K} = 0$, *i.e.,* capacity is set efficiently given the choices of prices. Likewise, $\frac{\partial W}{\partial p_t} = 0$ for all t, since we have explained earlier that realtime prices are set efficiently given the equilibrium $\bar{p}$. Similarly, $\frac{\partial W}{\partial \lambda_t}\frac{d\lambda_t}{d\alpha} = 0$ since either $p_t = c$ or capacity is fully utilized. Thus [11] reduces to: $\frac{dW}{\mathrm{d}\alpha} = \frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} + \frac{\partial W}{\partial \alpha}$.

Result 13 shows that $\frac{d\bar{p}}{d\alpha} \leq 0$, and in section I, we showed that $\frac{\partial W}{\partial \bar{p}}$ can be positive or negative depending on whether $\bar{p}^e$ is greater or less than $\bar{p}^*$.[32] The last term, $\frac{\partial W}{\partial \alpha}$, is the direct welfare gain from customers switching from flat-rate to RTP and can be written as

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

which is positive by the revealed preference argument made above. Thus, we have shown that if $\bar{p}^e > \bar{p}^*$, so that a decrease in $\bar{p}$ improves welfare, then increasing α increases total welfare. If $\bar{p}^e < \bar{p}^*$ so that $\frac{\partial W}{\partial \bar{p}} > 0$, then $\frac{\partial W}{\partial \bar{p}}\frac{d\bar{p}}{d\alpha} < 0$ and $\frac{\partial W}{\partial \alpha} > 0$ so the net impact on welfare is ambiguous.[33]

So, to summarize,

Result 15: Welfare Effects of Increasing RTP Customers — *In the long run, an increase in the proportion of customers on RTP (i) increases consumer surplus of customers remaining on flat-rate service, (ii) increases consumer surplus of customers switching from flat-rate to RTP, and (iii) decreases consumer surplus of incumbent RTP customers, and (iv) has no effect on generator or retailer profits, . Total welfare increases with an increase in the proportion of customers on RTP if $\bar{p}^e > \bar{p}^*$, but may decrease welfare 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.*

We show in the appendix that an example in which increasing α lowers welfare can be constructed, but the construction suggests that this may be a fairly unusual case. More generally, we know from section I that increasing α to 1 from any lower value increases welfare. So, the example in the appendix demonstrates the increase in welfare need not always be monotonic as it moves to the maximum welfare at $\alpha = 1$.

IV. Reserves in Peak-Load Pricing Analysis

Thus far, we have assumed that the market can be relied upon for instantaneous price adjustment that matches supply and demand in each period. This is an important issue,

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

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

because instantaneous supply/demand balancing is necessary to prevent destabilizing the grid, which would cause costly service disruptions. This concern arises because of the stochastic natures of demand and available supply, which are not modeled in our analysis, and the non-storability of electricity. In response to these realities, independent system operators (ISOs) carry operating reserves, capacity that is paid to standby to be available for production on short notice. A complete model of a deregulated market with operating reserves under stochastic supply and demand is beyond the scope of the current analysis, but a simple model of reserves demonstrates why we believe that the basic insights of our analysis would apply to such a model.

A first step towards incorporating reserves is as an augmentation to demand. Assume that the ISO decides to hold reserves equal to a fraction v of quantity demanded, so effectively demand for production capacity is $(1+v)\tilde{D}_t(p,\bar{p})$. With a minor modification, the analysis then goes through as in the previous sections. The modification is necessary because generation owners must be paid for some capacity that is not actually producing power. In effect, the reserves requirement means that a unit of production requires $1+v$ units of capacity.[34]

If each capacity owner "self-provided" reserves, by simply using only $\frac{1}{1+v}$ of its capacity, leaving the remaining capacity as reserve, then the analysis of the previous sections goes through without further modification. In periods in which $\tilde{D}_t(c,\bar{p}) < \frac{K}{1+v}$, the wholesale and retail realtime price are $w_t = p_t = c$. If $\tilde{D}_t(c,\bar{p}) > \frac{K}{1+v}$, the wholesale and retail realtime price are defined by $\tilde{D}_t(w_t,\bar{p}) = \frac{K}{1+v}$. In the long run, capacity adjusts until $\sum_{t=1}^{T} w_t - c = r \cdot (1+v)$. The additional scarcity rent for each unit of capacity that gets used would be exactly sufficient to cover the cost of the extra capacity that is never used.

In reality, reserves are not provided by operating every generating unit at less than full capacity. Because of startup costs and differential fuel costs, some generators operate at full capacity while others sit idle but ready to operate as backup if necessary. Owners of the units that provide reserves need not be the same as owners of the units that generate power. In this situation, a single price for power is not sufficient to compensate all generation owners. Instead, generators receive one payment for generating power and a different payment for being available as standby capacity. We'll call the payment for providing one unit of reserve capacity θ_t.

Applying this structure to our simple model, which has homogeneous capacity and

[34] Alternatively, every unit of capacity that previously cost r per day could now be considered to cost $r \cdot (1+v)$ per day.

no startup costs, one can derive the short-run equilibrium by recognizing that (1) retailers must cover their costs of power and reserve capacity, (2) generators must always be indifferent between generating power and providing reserve capacity, and (3) demand must never exceed $\frac{K}{1+v}$. Thus, for a given K, if $\tilde{D}_t(c, \bar{p}) < \frac{K}{1+v}$ in a given period, then $p_t = w_t = c$ and $\theta = 0$.

If $\tilde{D}_t(c, \bar{p}) > \frac{K}{1+v}$, the capacity constraint is binding and the retail realtime price is defined by $\tilde{D}_t(p_t, \bar{p}) = \frac{K}{1+v}$. If it were the case that $w_t = p_t$ and $\theta = 0$, then all generators would want to provide power and none would provide reserves. This would drive down the wholesale price of power and drive up the price of reserve capacity (which the retailer is required to have in proportion v of the power it is providing) until $w_t^e = c + \frac{p_t^e - c}{1+v}$ and $\theta_t^e = \frac{p_t^e - c}{1+v}$. This would satisfy generator indifference between providing power or reserves, because either way they earn $\frac{p_t^e - c}{1+v}$. It would also allow retailers to exactly cover their costs, because their cost of providing one unit of power at retail would be $w_t^e + v\theta_t^e = c + \frac{p_t^e - c}{1+v} + v\frac{p_t^e - c}{1+v} = c + (p_t^e - c) = p_t^e$. The equilibrium retail price to flat-rate customers would be the quantity-weighted average of the cost of providing the power, $w_t^e + v\theta_t^e = p_t^e$, so $\bar{p}^e$ is defined implicitly by $\bar{p}^e = \sum_{t=1}^T p_t^e D_t(\bar{p}^e) / \sum_{t=1}^T D_t(\bar{p}^e)$.

In the long run, capacity adjusts until $\sum_{t=1}^T w_t - c = r$ which, by construction, occurs at the same K that results in $\sum_{t=1}^T \theta_t = r$. Generators are indifferent between providing power or reserve capacity and exactly cover their costs of capital regardless of which they do. So we have,

Result 16: Competitive Equilibrium with Reserves — *If reserves are modeled as unused capacity that must be held in proportion v of power production, then the competitive equilibrium is characterized by (1) if $\tilde{D}_t(c, \bar{p}^e) \leq \frac{K}{1+v}$ in a period, realtime retail and wholesale price is $p_t^e = w_t^e = c$ and reserve capacity payment is $\theta_t^e = 0$, (2) if $\tilde{D}_t(c, \bar{p}^e) > \frac{K}{1+v}$, realtime retail price is determined by $\tilde{D}_t(p_t^e, \bar{p}^e) = \frac{K}{1+v}$, realtime wholesale price is $w_t^e = c + \frac{p_t^e - c}{1+v}$, and realtime reserve capacity payment is $\theta_t^e = \frac{p_t - c}{1+v}$, (3) a flat-rate retail price defined implicitly by $\bar{p}^e = \sum_{t=1}^T p_t^e D_t(\bar{p}^e) / \sum_{t=1}^T D_t(\bar{p}^e)$, and (4) long-run adjustment of capacity so that $\sum_{t=1}^T w_t^e - c = \sum_{t=1}^T \theta_t^e = r$.*

Given the characterization of the equilibrium in Result 16, all the results of the preceding analysis can be applied to the equilibrium with reserves. In particular, if all customers are on RTP, then investment, electricity allocation, and provision of reserves are all efficient. However, if some customers do not face the realtime prices, then neither the efficient allocation nor the second-best optimum is attained.

V. Conclusion

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

In order to assure adequate capacity investment, many market participants and advisors have argued for "capacity payments," which are effectively subsidies that reduce the cost of owning capacity and, thus, increase equilibrium investment. We have demonstrated that capacity subsidies (or taxes) cannot achieve the second-best optimum, because they create other distortions as they address the distortion caused by flat-rate customers. Furthermore, capacity investment distortion under flat-rate pricing can lead to either excessive or insufficient investment. We also examine taxes or subsidies on retail electricity as a policy response to the inefficiency caused by flat-rate pricing. A tax or subsidy on the flat-rate customers alone can indeed achieve the second-best optimal flat-rate price and capacity investment, but a tax or subsidy that applies to all customers—flat-rate and those on RTP—will distort the RTP market, so it will not achieve the second-best optimum.

Many economists and some industry participants have argued strongly for increasing the proportion of customers on RTP. We have shown that putting all customers on RTP increases market efficiency and attains the first best. While that is not surprising, the effect of such increases on capacity investment may be. Though increasing the use of RTP can, as many have suspected, decrease equilibrium investment in capacity, it can also increase equilibrium capacity. Which outcome obtains depends on the specific parameters of the market, in particular whether demand at peak times is more or less elastic than demand during hours of lower demand.

Incrementally increasing the proportion of customers on RTP can also have surprising effects on the efficiency of the competitive equilibrium. For the long-run equilibrium, we have shown that putting an additional customer on RTP benefits that customer and the customers who remain on the flat rate but harms the existing RTP customers. If the equilibrium flat rate is higher than the optimal flat rate, we have shown that the losses of the RTP customers can indeed be compensated by the other customers. In this case, increasing the proportion of customers on RTP improves welfare. If the equilibrium flat rate is lower than optimal, then we cannot prove that welfare increases monotonically in the number of customers on RTP. In fact, we present an example in which welfare decreases. However, the construction of the example suggests that it is not likely to be policy relevant.

The argument in favor of capacity payments is often supported by pointing out the need for excess capacity to provide operating reserves in case prices cannot instantaneously balance supply and demand. We have demonstrated in a simple model that the need for reserve capacity does not alter the basic insights of our analysis. The need for a given proportional capacity reserve changes optimal prices and capacity, but the problems raised by flat-rate pricing remain and the inability of capacity subsidies to correct these problems is unchanged.

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

Like much of the peak-load pricing literature, we have made certain restrictive assumptions to simplify this analysis. We have assumed that there is no cross-elasticity of demand across periods, that all generation technology is identical, that all customers have identical distributions of demands across periods, and that demand has no stochastic component. In future work, we intend to relax these restrictive assumptions, but at this point we see no reason that such changes to the model will alter the basic insights of the analysis.

Appendix

Result 13: Effect of Increasing RTP Customers on Flat Retail Rate — *In the long run, an increase in the proportion of customers on RTP (α) reduces $\bar{p}^e_{LR}$.*

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

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

Note that long run equilibrium in the wholesale market requires that $\sum_{t=1}^{T}(w_t - c) = r$, so increasing α, since it leaves c and r unchanged, must leave the unweighted average of w_t unchanged, $\sum_t \frac{dw_t}{d\alpha} = 0$. We want to know the sign of $\sum_t -D_t(\bar{p})\frac{dw_t}{d\alpha}$, which is a weighted average of the $\frac{dw_t}{d\alpha}$. If we can show, that the weighting is systematically greater for negative values of $\frac{dw_t}{d\alpha}$ than for positive values, then the weighted average is negative.

We show this by demonstrating that there is a $\hat{Q}$ such that $\frac{dw_t}{d\alpha} < 0 \quad iff \quad D_t(\bar{p}) > \hat{Q}$, that is, all of the negative $\frac{dw_t}{d\alpha}$ are associated with greater weights than any of the positive $\frac{dw_t}{d\alpha}$.

Note that

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

(A1) applies for all periods in which capacity is fully utilized. For periods in which capacity is not fully utilized, $\frac{dw_t}{d\alpha} = 0$, so these periods do not affect the *change* in the weighted average wholesale price of power consumed by flat-rate customers.

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

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

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

$$-\alpha D_t'(w_t)\frac{dw_t}{d\alpha} = \frac{K - D_t(\bar{p})}{\alpha} - \frac{dK}{d\alpha} \qquad (A3)$$

or

$$-\alpha^2 D_t'(w_t)\frac{dw_t}{d\alpha} = K - \alpha\frac{dK}{d\alpha} - D_t(\bar{p}) \qquad (A4)$$

Note that $-\alpha^2 D_t'(w_t)$ is always positive and define $\hat{Q} \equiv K - \alpha\frac{dK}{d\alpha}$, then $\frac{dw_t}{d\alpha}$ has the same sign as $\hat{Q} - D_t(\bar{p})$, that is $\frac{dw_t}{d\alpha} < 0 \quad iff \quad D_t(\bar{p}) > \hat{Q}$. ▮

Example of an Increase in RTP Customers that Decreases Welfare

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

First, recall that the competitive equilibrium is characterized completely by p_t, $\bar{p}$, α, c, r, K, and the demand functions D_t. Note, however, that the equilibrium does not depend on the entire demand functions, but rather only on two points, $D_t(p_t)$ and $D_t(\bar{p})$, of each demand function. Thus, any system of demand equations which does not change these $2T$ points (or α, c, or r) will have an equilibrium with the same prices and capacity.

Next, consider $\frac{d\bar{p}}{d\alpha}$, $\frac{dp_t}{d\alpha}$, and $\frac{dK}{d\alpha}$. By the Implicit Function Theorem, these derivatives can be found by totally differentiating the system of equations that characterize the competitive equilibrium. This implies that $\frac{d\bar{p}}{d\alpha}$ can be written as a function of the $5T + 5$ parameters: p_t, $D_t(p_t)$, $D'_t(p_t)$, $D_t(\bar{p})$, $D'_t(\bar{p})$, $\bar{p}$, α, c, r, and K. Since $\frac{\partial W}{\partial \bar{p}}$ can also be written in terms of these $5T + 5$ parameters, the product $\frac{\partial W}{\partial \bar{p}} \frac{d\bar{p}}{d\alpha}$ does not change if we were to perturb the demand curves such that the demands and slopes at p_t and $\bar{p}$ were unchanged.

Now consider $\frac{\partial W}{\partial \alpha}$. Since $\sum_t p_t D_t(\bar{p}) = \sum_t \bar{p} D_t(\bar{p})$, [12] can be written

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

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

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

REFERENCES

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

Borenstein, Severin. "Frequently Asked Questions about Implementing Real-Time Electricity Pricing in California for Summer 2001." mimeo, University of California Energy Institute, March 2001. Available at http://www.ucei.org/PDF/faq.pdf.

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

Doucet, Joseph and Andrew Kleit. "Metering in Electricity Markets: Should It Be Encouraged?" mimeo, University of Alberta, 2002.

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

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

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

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

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

Figure 1

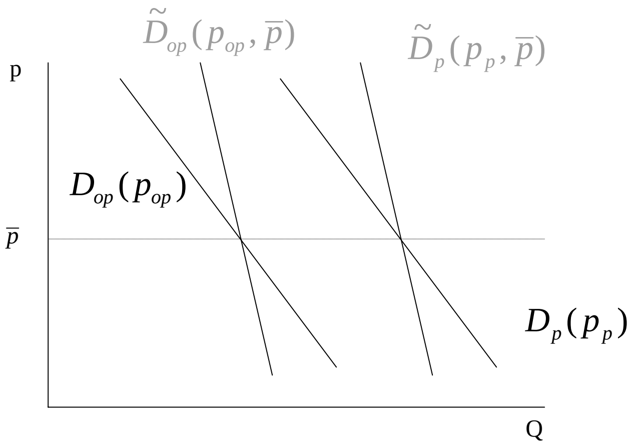

$\tilde{D}_{op}(p_{op},\overline{p})$
$\tilde{D}_{p}(p_{p},\overline{p})$
p
$D_{op}(p_{op})$
$\overline{p}$
$D_{p}(p_{p})$
Q

Figure 2

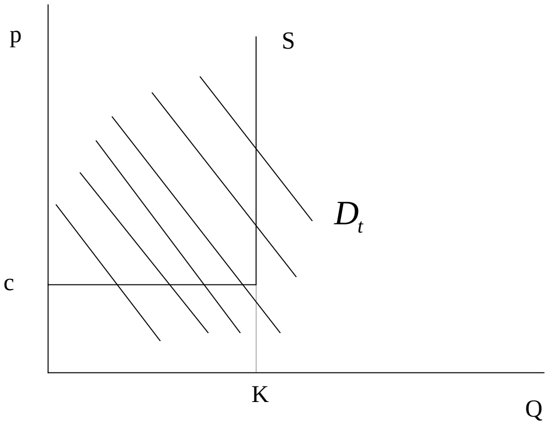


Figure 3

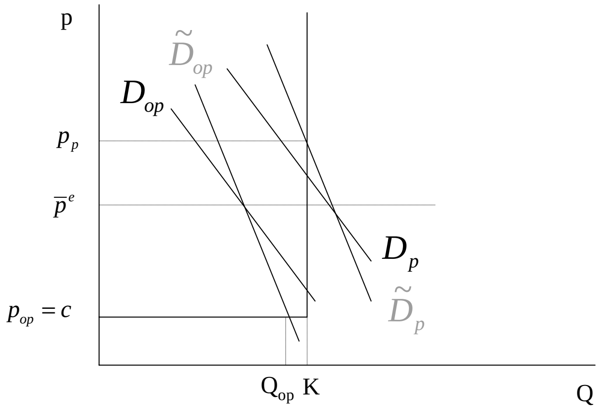


Figure 4

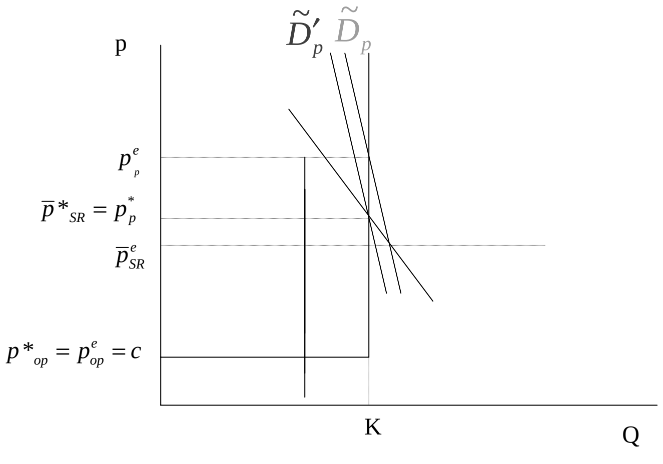


Figure 5

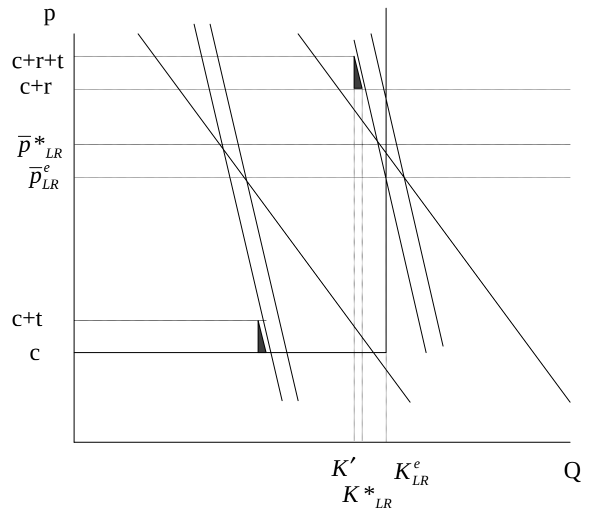


Figure 6

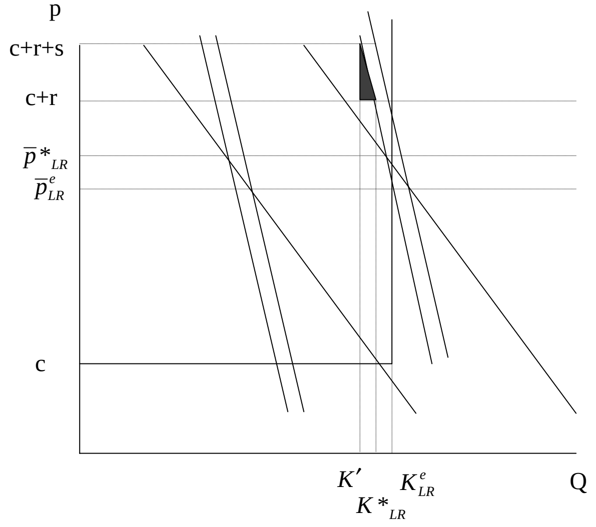
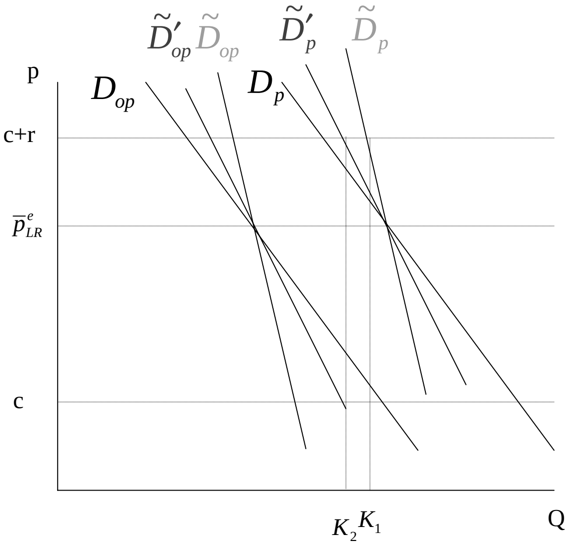


Figure 7


$\tilde{D}'_{op}$
$\tilde{D}_{op}$
$\tilde{D}'_{p}$
$\tilde{D}_{p}$
p
D_{op}
D_{p}
c+r
$\bar{p}^{e}_{LR}$
c
K_2
K_1
Q