The Long-Run Efficiency of Real-Time Electricity Pricing

*Severin Borenstein**

Retail real-time pricing (RTP) of electricity – retail pricing that changes hourly to reflect the changing supply/demand balance – is very appealing to economists because it "sends the right price signals." Economic efficiency gains from RTP, however, are often confused with the short-term wealth transfers from producers to consumers that RTP can create. Abstracting from transfers, I focus on the long-run efficiency gains from adopting RTP in a competitive electricity market. Using simple simulations with realistic parameters, I demonstrate that the magnitude of efficiency gains from RTP is likely to be significant even if demand shows very little elasticity. I also show that "time-of-use" pricing, a simple peak and off-peak pricing system, is likely to capture a very small share of the efficiency gains that RTP offers.

1. INTRODUCTION

Over the last few years, a great deal has been written about time-varying retail pricing of electricity. Many authors, myself included, have argued that real-time retail electricity pricing (RTP) – retail prices that change very frequently, e.g., hourly, to reflect changes in the market's supply/demand balance – is a critical component of an efficient restructured electricity market. During the California electricity crisis in 2000-2001, RTP boosters pointed out its value in reducing

The Energy Journal, Vol. 26, No. 3.

I would like to thank Carl Blumstein, Jim Bushnell, Ali Hortacsu, Ed Kahn, Erin Mansur, Karen Notsund, Celeste Saravia, Ralph Turvey, Bert Willems, Frank Wolak , Catherine Wolfram, participants in seminars at U.C. Berkeley, U.C. Santa Barbara, U.C. Irvine, Wharton, and University of Florida, and two anonymous referees for valuable comments. Meredith Fowlie and Amol Phadke provided excellent research assistance. This work grew directly from related research with Stephen Holland. Many hours of valuable discussion with Stephen have shaped my thinking on RTP issues, though he bears no responsibility for any errors in this paper.

* 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.

the ability of sellers to exercise market power. While nearly all economists have supported RTP conceptually, Ruff (2002) among others has argued that it is important to distinguish between RTP's long-run societal benefits and the short-run wealth transfers it might bring about. In particular, the reductions in market power primarily prevent a short-run wealth transfer from customers to generators, though the transfers can still be quite large.

In this paper, I estimate the magnitude of the potential long-run societal gains from RTP, abstracting from market power issues and short-run wealth transfers in general. I do this by formulating a model of competitive electricity generation with demand and production costs based on actual data from U.S. markets. I solve computationally for the model's long-run competitive equilibrium, with the results indicating the amount of each possible type of capacity that would be built, the prices that would be charged to customers on RTP and on flat-rate service, and the total social surplus that would be generated by the system. The model also allows estimation of the transfers that would occur among customers if customers on RTP had demands that were (absent RTP) peakier or flatter than customers not on RTP.

The estimates indicate that RTP would substantially reduce peak electricity production and thereby reduce the use of low-capital-cost/high-variable-cost peaker generation. The social gains from RTP for at least the largest customers in the system are estimated to far outweigh reasonable estimates of the metering cost. The magnitudes of the social gain are sensitive to the demand elasticity that is assumed, but the results indicate that even with quite small elasticities, the benefits are substantial.

Section 2 presents the economic model that is the basis for simulations. Section 3 explains the data used in the simulations and the process used to compute long-run equilibria. The results of the simulations are presented and their implications discussed in Section 4. In section 5, I carry out a similar analysis on a much simpler pricing system, time-of-use (TOU) pricing, in which there are simple peak and off-peak periods, with the prices differing between periods, but being held constant for months or even years at a time. Section 6 discusses a number of factors that are omitted from the simulations and suggests how those factors are likely to affect the results. I conclude in Section 7.

2. LONG-RUN COMPETITION IN ELECTRICITY MARKETS

The model that is the basis for the simulations is adapted from Borenstein and Holland (revised 2003, hereafter BH).[1] It assumes a simple competitive wholesale and retail market structure. The retail structure is identified only by the way in which it charges end-use customers for electricity, using a flat rate

1. A slightly different version of this model with continuous marginal cost functions is in Borenstein and Holland, forthcoming. Holland and Mansur (2005) analyze the *short-run* efficiency, distributional, and environmental effects of RTP in a very similar model.

or RTP. The price(s) charged to each group allow the retailer to exactly break even on service to that group. As in BH, this reflects the outcome of competition among many retail providers, but it also could be interpreted as a single regulated retail provider that is required to exactly cover its costs and required not to cross-subsidize between flat-rate and RTP customers. Following BH, I assume for simplicity that retailers have no other transaction costs.

I assume free-entry of generators of three different types. Generation exhibits no scale economies, with each generation unit having a capacity of one megawatt. The types of generation differ in their fixed and variable costs, higher fixed costs being associated with lower marginal cost of production. For generator type *j*, annual generator costs are modeled as a fixed cost plus variable costs that are linear in the number of megawatt-hours produced during the year, $TC_j = F_j + m_j \cdot MWh_j$. Startup costs and restrictions on ramping are not considered, an issue discussed in section 6. Parameters used for this and all other aspects of the simulations are discussed in the next section.

Demand is modeled as constant elasticity, using a range of possible elasticities. Within any one simulation, demand is first assumed to have the same elasticity in all hours. I then consider the effect of demand elasticity varying positively or negatively with the level of demand. The level of demand in each hour is taken from the distribution based on the actual levels of demand in various US electricity regions, as explained in the following section. Cross-elasticities across hours are assumed to be zero, another issue discussed in section 6.

Some proportion of customers, α, are on real-time pricing, and the remainder are on flat-rate service. I assume that all customers have identical demand up to a scale parameter. Thus, following BH, if the total demand in hour h is $D_h(p_h)$ and the flat-rate service customers are charged $\overline{p}$ in every hour, the wholesale demand is

$$\widetilde{D}_h(p_h, \overline{p}) = \alpha \bullet D_h(p_h) + (1 - \alpha) \bullet D_h(\overline{p}). \tag{1}$$

In this case, demand is modeled as constant elasticity, $D_h(p_h) = A_h \bullet p_h^{\varepsilon}$.

Under these assumptions, for any set of installed baseload, mid-merit, and peaker capacity, K_b, K_m, K_p, there is a unique market-clearing wholesale price in each hour, provided that total installed capacity exceeds demand from flat-rate customers in every hour, $K_b + K_m + K_p > (1 - \alpha) \bullet D_h(\overline{p})\ \forall h$. In the following section, I discuss the algorithm for finding the short-run equilibrium for any set of installed capacity and the long-run equilibrium allowing capacity to vary. In presenting the algorithm, I demonstrate that there is a unique long-run equilibrium.

In addition to establishing long-run equilibria for any $0 \leq \alpha < 1$, it will be important, as a baseline, to determine an equilibrium with no customers on RTP. The model above is not applicable to a market with no RTP customers, because without RTP there is no short-run demand elasticity, so in order to meet demand in all hours, sufficient capacity must be built so that the market always clears "on the supply side," i.e., at a price no greater than the marginal generation cost of the

technology with the highest marginal cost. Such an organization requires some sort of additional wholesale payment to generation in order to assure that demand does not exceed supply in any period and, at the same time, that generators' revenues exceed their variable costs over a year by an amount sufficient to cover their fixed costs.

It is straightforward to show that the annual capacity payment that assures sufficient generation and the optimal mix of generation is equal to the annual fixed costs of a unit of peaker capacity. To avoid distorting the mix of capacity, this payment is made to all units of capacity, regardless of type.[2] The payment is financed by increasing the price of the flat-rate electricity service until it generates sufficient revenue to cover the capacity payments. That is how simulation of the baseline flat-rate service is implemented in the following section. In contrast, in the RTP simulations no capacity payment is made; generators earn all revenues through energy sales.

3. DATA, MODEL DETAILS AND SOLUTION ALGORITHM

The value of the simulation results depends on the realism of the underlying assumptions. In this section, I describe in detail the modeling of demand and supply, and then the algorithm for finding the long-run competitive equilibrium. I first present the details of the model, and then discuss the data used to parameterize the model.

3.1 Demand, Supply and Equilibrium Modeling

Within each hour, each customer's demand is modeled as constant elasticity. Each customer i is assumed to have a demand that is simply a fixed proportion, γ_i, of total demand. In the base simulations, I assume that total demand has the same elasticity in all hours, but this is later relaxed to allow elasticity to vary positively or negatively with the overall demand level.

The aggregate demand function for hour h can be specified as $D_h(p_h) = A_h \bullet p_h^{\varepsilon_h}$, where elasticity may or may not vary by hour depending on the simulation run. For any share of demand on RTP, α, the demand from customers on RTP is then $D_h(p_h) = \alpha \bullet A_h \bullet p_h^{\varepsilon_h}$ and the demand function for customers on flat rate service is $D_h(\overline{p}) = (1 - \alpha) \bullet A_h \bullet \overline{p}^{\varepsilon_h}$. The aggregate demand in the wholesale power market is then $\widetilde{D}_h(p_h, \overline{p}) = \alpha \bullet A_h \bullet p_h^{\varepsilon_h} + (1 - \alpha) \bullet A_h \bullet \overline{p}^{\varepsilon_h}$.

Given an elasticity for a certain hour, ε_h, and the assumption of a constant-elasticity functional form, demand is fully specified by A_h, the scale parameter. A_h is determined by any one price/quantity point on the demand curve, which I refer to as the demand "anchor point" for the hour. I assume that at a

2. This would also be the outcome if the wholesale price exceeded the marginal cost of the peaking generation only in the highest demand hour of the year, and the price in that hour was equal to the marginal cost of the peaker plus its annual fixed cost.

given constant price (discussed next), the anchor quantity demanded takes on a distribution equal to the actual distribution of quantities demanded from a certain electricity control region.

The constant price used to specify the anchor points is chosen to be the price that would allow producers to break even if it were charged as a flat retail price to all customers. This is not the actual flat rate (or time-of-use rate) that was charged to customers during the observed period from which the demand distribution data are taken. The difference, however, will not substantially change the results for two reasons. First, at the low elasticities I consider in the simulations, a change of 10%-20% in the base flat rate that I assume (which is the magnitude of the potential difference between the rate assumed and the actual flat rate in use) will change quantity demanded very little. Second, and more important, the overall level of base demand is just a scale factor in the simulations. The value of using an actual distribution comes from accurately representing the *shape* of the distribution; that changes negligibly with the assumption made about the level of the flat retail rate.

Once the wholesale demand function has been specified each hour, that can be combined with the production technologies to calculate the long-run equilibrium capacity of each technology type. Note that from any given baseload, mid-merit, and peaker capacities, K_b, K_m, K_p, one can determine a short-run industry supply function and therefore wholesale prices for each hour. From those prices, one can calculate the profits of owners of each technology type. In the long-run each technology type is built to the point that one more unit of that capacity would cause profits of all owners of the capacity to be negative. So, the goal is to identify the mix of capacity that causes this condition to hold for all three technologies simultaneously.

At first, this might seem difficult, and it might seem that there could be multiple long-run equilibria or none, but in fact there is a unique technology mix that satisfies this condition. To see this, begin with the peaker technology which, if it is used at all, will be used in the highest demand hour. It is straightforward to find a unique long-run equilibrium if supply is restricted to use only the peaker technology. One simply expands the quantity of peaker capacity, recalculating the associated short-run equilibria with each increment in capacity, until expansion of capacity by one more unit causes profits to go negative. Call the capacity level that satisfies this condition K_{tot} since that will generally turn out to be the equilibrium total amount of capacity.

In this peaker-only equilibrium, all rents to generators are earned when production quantity is equal to K_{tot}. In hours with lower equilibrium quantity, price must be equal to peaker marginal cost. Now, begin substituting mid-merit capacity for peaker capacity. Once built, the mid-merit capacity will all be used in any given hour before any of the peaker capacity is used; it is lower on the supply function than the peaker capacity. The key is to recognize that substituting mid-merit for peakers units, holding total capacity constant, does not change the rents earned by the remaining peaker units. In fact, so long as one peaker unit

remains, the rents it earns are unchanged by substituting lower-MC technologies for the other units.[3]

Continuing to substitute mid-merit for peaker units will drive down the equilibrium profits of mid-merit units until one more unit would drive the profits of all mid-merit units to be negative. Call the largest capacity of mid-merit units that still earns positive profits, K_{bm} because this will generally turn out to be the total of the baseload and mid-merit capacity. Next, begin substituting baseload capacity for mid-merit units. Note that this does not change the rents to mid-merit units. Continue this substitution until one more baseload unit would drive baseload profits negative. This is K_b. Then, $K_m = K_{bm} - K_b$ and $K_p = K_{tot} - K_m - K_b$. These are the unique long-run competitive equilibrium capacity levels for a given set of available technologies, share of customers on RTP (α), and flat rate ($\overline{p}$).[4]

This equilibrium, however, may not satisfy the retailer breakeven condition, so one must calculate the profits retailers earn on flat rate customers in this equilibrium. If it is not zero, then one adjusts $\overline{p}$ up or down and resimulates capacity. When the resulting equilibrium yields zero profits for retailers as well as generators, this is the unique long-run competitive equilibrium in the generator and retailer markets given the set of available technologies and share of customers on RTP (α). Using this supply function, one can then calculate the equilibrium distribution of prices, loads (quantities), and the consumer surplus for each group.[5]

3.2 Data Inputs for Simulation

The critical inputs for the simulation are a load profile, demand elasticities, and cost characteristics of the production technologies.

The load profile determines the distribution of quantity demand and the flat rate when all customers are on flat-rate service, as described in the previous section. For the simulations presented in here, I use five years of hourly demand data from the California Independent System Operator, 1999 through 2003.[6] This

3. This description assumes that equilibrium capacity investment includes at least one unit of each type of capacity. If peaker capacity is dominated by mid-merit or baseload for even the least utilized peaker unit, or if mid-merit is dominated by baseload for the least utilized mid-merit unit, then the same process is followed omitting the dominated technology.

4. These searches were done inefficiently from a computing standpoint, as grid searches with a 1 MW grid over a very wide range of possible capacity quantities. They still converged quite quickly on a desktop PC.

5. The updating algorithm for $\overline{p}$ was to always reset it to the level that would have broken even given the prior iteration's quantities demanded by flat-rate customers and the wholesale prices from the current iteration. This usually converged in two to four iterations on $\overline{p}$.

6. I adjust the baseline hourly demand data for the fact that about half of all demand is on time-of-use rates (TOU). I do this by assuming that the elasticity of demand with respect to TOU price variation is -0.1 and that the price ratios among TOU periods are equal to the average ratios in the TOU rate schedules offered by Pacific Gas & Electric and Southern California Edison. This adjustment has only a slight effect on the results.

Table 1: Generation Costs Assumed in Long-Run RTP Simulations

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

period includes both relatively cool summers and quite hot summers.[7] As pointed out earlier, the importance of the load distribution used is in the shape of the load duration curve, not the overall size of the loads. It appears that load duration curves don't differ that much in shape from one control area to another.

Electricity demand elasticities are a subject of nearly endless contention. The relevant elasticity would be a short-run elasticity, but still recognizing that customers would know well in advance that prices would be volatile. The actual elasticity will depend in great part on technology, as automated response to price changes will surely become easier over time. I simulate for a fairly wide range of elasticities from -0.025 to -0.500. The range -0.025 to -0.150 illustrates that likely impact of RTP in the short run and under current available technologies for demand response. Probably the two most current and relevant sources for elasticity estimates, Patrick and Wolak (1997) and Braithwait and O'Sheasy (2002), derive estimates that span this range. In the longer run, however, real-time demand response will become easier to automate and larger elasticities might be expected, so I include results using -0.3 and -0.5 as well. All demand levels are calculated based on the full retail price, which is assumed to be the cost of power plus $40/MWh for transmission and distribution (T&D).[8]

The assumptions about production technology are presented in Table 1. They are intended to represent typical capital and variable costs of baseload, mid-merit, and peaker technologies, corresponding roughly to coal, combined-cycle gas turbine, and combustion turbine generation. The numbers were derived from conversations with industry analysts. The variable costs depend on fuel prices, and are meant to include variable O&M.[9] The annual fixed costs are more difficult to determine precisely, in part because they depend on the cost of capital and on the rate of economic depreciation of the plant. These figures appear to be in what most industry analysts would consider to be a reasonable range.

Two further comments on plant costs are warranted. First, the results are not particularly sensitive to the exact cost assumptions on the baseload and mid-merit technology. The different effects of RTP under varying assumptions on

7. I've carried out the same analysis using datasets from the ECAR (upper midwest) and NPCC (New England) regions with very similar results.

8. I assume that the T&D charge is not time-varying. T&D could also be subject to real-time pricing if capacity constraints become binding at some times.

9. For these costs, the price of natural gas is assumed to be $4.25/MMBtu and variable O&M is assumed to be $1/MWh.

elasticity and the share of customers on RTP are driven mostly from changes in the amount of peaker capacity that is built. In future analysis, I will include a range of cost assumptions. Second, this paper presents an easily-replicated algorithm for analyzing the long-run effect of introducing demand elasticity. For whatever cost assumptions the policy analyst believes are appropriate, this technique can be used to analyze the long-run implications.

4. SIMULATION RESULTS AND IMPLICATIONS

The first line of Table 2 presents the equilibrium flat rate ($79.68/MWh, which includes $40/MWh for transmission and distribution), as well as the capacity that is utilized in efficiently providing the demand under the flat rate, and the total energy consumed and cost of that energy. The remainder of the table presents the equilibrium capacities and information about equilibrium price distributions under scenarios with varying proportions of customers on RTP and with those customers exhibiting various demand elasticities. Within each simulation, demand has the same elasticity in all hours.

It is apparent from Table 2 that with even moderate demand elasticity, RTP will significantly change the composition of generation, as indicated in columns F, G, H and I. The greatest effect will be a large decline in the amount of installed peaker capacity (column H). Mid-merit capacity (column G) would likely also decline and baseload capacity (column F) would increase, though these changes would be small in comparison to the potential for drastic reductions in peaker capacity. Figure 1 shows the load duration curves for simulations with varying elasticities and one-third of customers on RTP.[10] The highest curve, representing all customers on a flat-rate tariff, has one hour in the upper left corner in which quantity hits 46928, which is the highest quantity demanded when all customers are on flat rates. Note that the other curves, representing differing demand elasticities for the one-third of demand on RTP, flatten out at different load levels, with lower peak demand levels associated with greater demand elasticity. For demands in these regions, the market clears "on the demand side," i.e., on the vertical portion of the supply curve (constant quantity, varying price). This illustrates the effect shown in column I in table 2: RTP has a very significant effect on the total capacity needed because for the highest demand periods, the market equilibrates by raising price rather than building additional generation capacity that is used for only a few hours per year.

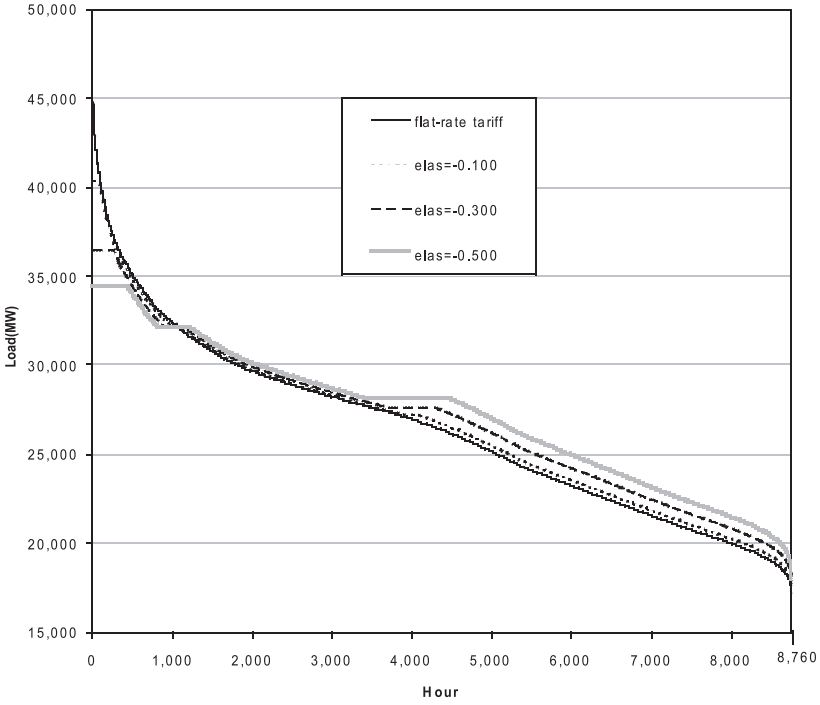
A question that frequently arises with RTP is how high prices could get and whether "bill shock" during a high-price month would undermine the program. This concern, of course, is greatly mitigated by forward contracts and other financial instruments, as explained in Borenstein (forthcoming). Customers that hold fixed-quantity forward contracts can eliminate most price risk without reducing the strong price incentives on marginal purchases.

10. A load duration curve shows the number of hours (horizontal axis) in which the quantity demanded will be at least a certain level (vertical axis).

Table 2: Capacity, Price and Quantity Effects of RTP

A	B	C	D	E	F	G	H	I	J	K	L	M
Elas-ticity	Share on RTP	Total Annual Energy Consumed (MWh million)	Total Annual Energy Bill ($ million)	Flat Rate ($/MWh)	EQUILIBRIUM CAPACITY (MW) Base-Load	Mid-Merit	Peaker	Total	PRICE DURATION CURVE Peak Price ($/MWh)	Avg Hrs per year at Peak Quantity (of 8760)	Pctg of annual bill from top 10 hours in sample	Pctg of annual bill from top hour in sample
All On Flat Rate												
---	0.000	231,095,835	9,170,521,267	79.68	26984	5384	14560	46928				
Some On RTP												
-0.025	0.333	231,405,274	9,048,736,469	79.65	27028	5341	12038	44407	90772	4	60.8%	22.0%
-0.025	0.666	231,691,153	8,945,358,991	79.47	27074	5258	10014	42346	45292	30	44.0%	10.5%
-0.025	0.999	231,933,022	8,871,844,654	79.22	27118	5184	8603	40905	19505	67	23.5%	4.4%
-0.050	0.333	231,711,476	8,958,308,425	79.52	27075	5258	10251	42584	54052	25	48.3%	12.6%
-0.050	0.666	232,212,872	8,826,608,560	79.08	27169	5113	7732	40014	11890	97	15.5%	2.6%
-0.050	0.999	232,625,430	8,739,797,966	78.84	27256	4974	6176	38406	4405	157	6.6%	0.9%
-0.100	0.333	232,326,272	8,848,470,458	79.18	27178	5116	8074	40368	18834	84	21.6%	4.2%
-0.100	0.666	233,214,051	8,689,157,551	78.73	27361	4837	5211	37409	3038	206	4.7%	0.6%
-0.100	0.999	233,932,035	8,572,879,071	78.48	27531	4556	3364	35451	1321	348	2.2%	0.3%
-0.150	0.333	232,953,671	8,780,950,177	78.97	27284	4978	6733	38995	9302	132	11.8%	2.0%
-0.150	0.666	234,209,237	8,594,016,415	78.53	27554	4558	3568	35680	1577	328	2.5%	0.3%
-0.150	0.999	235,202,138	8,455,153,100	78.24	27799	4154	1573	33526	752	556	1.3%	0.1%
-0.300	0.333	234,955,611	8,659,285,409	78.68	27612	4564	4266	36442	3505	264	4.8%	0.7%
-0.300	0.666	237,327,726	8,409,265,790	78.12	28133	3759	547	32439	647	682	1.0%	0.1%
-0.300	0.999	238,825,409	8,238,485,575	77.59	28606	1786	0	30392	340	1891	0.6%	0.1%
-0.500	0.333	237,926,466	8,576,661,386	78.47	28062	4026	2361	34449	2302	438	3.1%	0.4%
-0.500	0.666	241,571,384	8,289,753,486	77.58	28942	1445	0	30387	370	2006	0.6%	0.1%
-0.500	0.999	243,110,229	8,139,900,836	76.73	28986	0	0	28986	209	5498	0.3%	0.0%

Figure 1: Load Duration Curve with Varying Demand Elasticities of RTP Customers (1/3 of total demand on RTP, 2/3 on flat-rate tariff)



Setting aside hedging instruments, however, it is apparent from Table 2, columns J, K, L and M, that an RTP program could yield very high prices for a few hours. With very inelastic demand, the prices would be extremely high in some hours. The reason for these high prices is shown in column K, which shows the average number of hours per year in which all capacity was used and, thus, the price was above the marginal cost of a peaker plant. With extremely inelastic demand, the peaker plants must recover all of their fixed costs over just a few hours per year, so spectacular price spikes are dictated. But taken in the context of the annual bill, even the very high prices seem more manageable. With a demand elasticity of -0.1, column L shows that the highest price hour would amount to 4.2% of the annual bill. Column M indicates that the 10 most expensive hours of the 5-year period, if they all occurred in a single month, would account for about 22% of the annual bill. Although these amounts would be substantial in monthly bills, the suggestion that a customer would find that half or more of its bill occurs in two or three hours is not consistent with my findings.[11]

11. Note that unlike the surplus comparisons I make below, this comparison is to the total bill including non-energy (T&D) components of the bill. This seems appropriate given that the concern is bill shock. Roughly half of the total bill is energy and the remainder is T&D.

Before leaving table 2, it is worth pointing out that RTP is not an energy conservation program. In these simulations, the aggregate energy consumed actually increases slightly (0%-2%), though that this could be due to the constant-elasticity demand function; in theory, total quantity consumed could increase or decrease. By lowering off-peak prices and lowering overall average prices, there is a real possibility that RTP would stimulate increased aggregate consumption of electricity.

The overall effect of RTP on social welfare is presented in Table 3. Because I use constant-elasticity demand curves, for which total consumer surplus is undefined, I evaluate the effects by calculating the *change* in consumer surplus from the flat-rate tariff consumer's faced before RTP was introduced. Thus, the equation for aggregate change in consumer surplus over the H hours simulated is:

$$\Delta CS = (1-\alpha)\sum_{h=1}^{H}\frac{A_h}{\varepsilon+1}\bullet(\hat{P}^{\varepsilon+1}-\bar{P}^{\varepsilon+1})+\alpha\sum_{h=1}^{H}\frac{A_h}{\varepsilon+1}\bullet(\hat{P}^{\varepsilon+1}-P_h^{\varepsilon+1}) \qquad (2)$$

where $\hat{P}$ is the flat rate prior to introduction of RTP and $\bar{P}$ is the flat rate in equilibrium after α share of demand is on RTP. The A_h for each hour are set so as to include the actual quantity demanded at a price of $\hat{P}$, as described earlier.

The annual average ΔCS is shown in column C of table 3. Columns E and G break out that number into the two terms in equation (2), which represent the change in surplus, still compared to having everyone on flat rate, for customers who stay on flat rate (column E) and for customers who move to RTP (column G).

It is immediately clear that the surplus gains from real-time pricing are substantial, even if demand of customers on RTP is quite inelastic. With an elasticity of only -0.025, the surplus gain from putting one-third of demand on RTP, shown in column C, is over $100 million per year. To give these figures some context, in 2001 the state of California appropriated $35 million as a *one time* cost of installing real-time meters for the largest customers in the state, representing slightly under one-third of total demand. That isn't the only cost of switching these customers to RTP, since billing systems must be changed as well, but there are also other benefits to the meters, including remote meter reading that can yield big labor savings. Nonetheless, as shown in column D, the savings are still a fairly modest share of the total energy cost for the system, less than 10% for all but the most optimistic case, and quite possibly less than 5%. Still, as discussed in section 6, the long-run energy market impact analyzed here is only one part of the value of RTP.

It is also clear that the total surplus gains from RTP are highly non-linear in both the elasticity of demand and the share of demand that is on RTP. There are diminishing returns to both greater elasticity and a greater share of demand on RTP. For most elasticities, putting one-third of demand on RTP achieves more than one-half the benefits of putting all demand on RTP. For any given $\alpha > 0$, a demand elasticity of -0.05 generates more than half the benefits of a demand elasticity of -0.15.

Table 3: Welfare Effects of RTP

A	B	C	D	E	F	G	H	I	J
Elas-ticity	Share on RTP	Annual Total Surplus Change from All on Flat ($)	Annual TS Change as percentage of original energy bill	Annual CS Change of Customers on Flat Rate ($)	Annual CS change "per customer" on Flat Rate ($)	Annual CS Change of Customers on RTP ($)	Annual CS change "per customer" on RTP ($)	Annual Incremental Surplus to Switchers ($)	Annual Incremental Externality ($)
-0.025	0.333	112,060,365	1.2%	4,602,394	69	107,457,971	3,227	107,457,971	4,602,394
-0.025	0.666	205,800,109	2.2%	16,195,248	485	189,604,862	2,847	92,504,684	1,235,061
-0.025	0.999	271,333,946	3.0%	107,052	1,071	271,226,894	2,715	74,262,205	-8,728,369
-0.050	0.333	196,836,537	2.1%	24,879,553	373	171,956,984	5,164	171,956,984	24,879,553
-0.050	0.666	314,219,558	3.4%	46,572,214	1,394	267,647,344	4,019	121,402,546	-4,019,525
-0.050	0.999	388,316,857	4.2%	194,639	1,946	388,122,219	3,885	82,941,297	-8,843,997
-0.100	0.333	302,262,176	3.3%	77,399,306	1,160	224,862,870	6,753	224,862,870	77,399,306
-0.100	0.666	439,987,363	4.8%	73,366,291	2,197	366,621,072	5,505	144,668,903	-6,943,716
-0.100	0.999	537,284,137	5.9%	276,546	2,765	537,007,592	5,375	105,855,899	-8,559,124
-0.150	0.333	370,238,483	4.0%	108,757,099	1,631	261,481,384	7,852	261,481,384	108,757,099
-0.150	0.666	530,960,593	5.8%	89,145,379	2,669	441,815,214	6,634	166,610,585	-5,888,475
-0.150	0.999	647,620,518	7.1%	333,189	3,332	647,287,329	6,479	126,883,966	-10,224,041
-0.300	0.333	509,388,631	5.6%	154,467,302	2,316	354,921,329	10,658	354,921,329	154,467,302
-0.300	0.666	730,577,275	8.0%	120,644,221	3,612	609,933,053	9,158	227,848,668	-6,660,025
-0.300	0.999	888,877,347	9.7%	484,978	4,850	888,392,369	8,893	175,847,779	-17,547,706
-0.500	0.333	641,472,723	7.0%	187,262,169	2,808	454,210,554	13,640	454,210,554	187,262,169
-0.500	0.666	922,328,312	10.1%	162,892,786	4,877	759,435,525	11,403	286,227,054	-5,371,466
-0.500	0.999	1,098,811,460	12.0%	687,144	6,871	1,098,124,316	10,992	203,636,356	-27,153,207

Decomposing the change in total surplus reveals two effects that BH demonstrate theoretically. First, column E shows that flat-rate customers are made better off by other customers moving to RTP. Column F calculates the "per capita" benefit for a hypothetical customer who makes up 0.001% of the total demand ($D_h(p_h)$) in any given hour.[12] This customer on flat rate billing benefits as an increasing share of other customers moves to RTP. This effect is frequently argued by parties who advocate subsidizing RTP participants.

A second effect, however, suggests that policy is not always wise: as demonstrated theoretically by BH, customers moving to RTP harm other customers who are already on RTP. This is shown numerically in column H, which presents the "per capita'" benefit of a customer (again representing 0.001% of total demand) on RTP when the total share of customers on RTP is the α in column B. We see that the benefits to a customer on RTP decline as more customers switch to RTP. In fact, the overall externality from a group of customers moving to RTP can be positive or negative, as shown in column J.[13]

4.1 Elasticity Varying with Demand Level

In the simulations presented thus far, the elasticity of demand has been the same in all periods, the case in which BH show that the equilibrium flat rate will be equal to the optimal flat rate. BH also show that if demand elasticity is greater in high-demand periods than in low-demand periods, the equilibrium flat rate will be below its optimal level. BH demonstrate that in that case it is theoretically possible that moving more customers on to RTP could lower long-run equilibrium total surplus.

I simulate this case by allowing elasticity of demand to vary with the level of demand, where the level is indicated by the quantity demanded if all customers were charged the flat rate.[14] The elasticity of demand varies linearly with demand level, in this case from 50% of the original demand elasticity for the lowest demand level to 192% of the original demand elasticity for the highest demand level. These boundaries were chosen so that the demand-weighted average elasticity is equal to the original demand elasticity in order to allow some comparability to the previous simulations.

Omitting a few of the columns, table 4 presents results comparable to tables 2 and 3, but for simulations in which demand is more elastic at higher demand levels. In fact, the introduction of RTP yields greater benefits in this case

12. This would be a customer with a peak demand of about 450kW. In California, there were approximately 8,000 customers of at least this size during the sample period.

13. BH show that the net externality from a *marginal* change in α is zero when demand in all periods has the same elasticity. There is a non-zero net externality in the cases shown here because the change is not incremental: Some of the externality of any one customer switching to RTP is captured by other customers in the switching group, so is not an externality from the group as a whole.

14. As explained above, this is by assumption the actual CAISO load during each hour.

Table 4: Larger Elasticity with Higher Demand

A	B	C	D	E	F	G	H	I	J	K	L	M
Elas-ticity	Share on RTP	Total Surplus Change from All on Flat	CS Change of Customers on Flat Rate	CS Change of Customers on RTP	Total Energy Consumed (MWh)	Total Energy Bill ($)	TS Chg as pctg of orig energy bill	Flat Rate ($/MWh)	CAPACITY Base-Load	Mid-Merit	Peaker	Total
All On Flat Rate												
---	0.000				231,095,835	9,170,521,267		79.68	26984	5384	14560	46928
Some On RTP												
-0.025	0.333	185,012,684	24,753,508	160,259,175	231,299,627	8,960,830,174	2.0%	79.52	27028	5320	10396	42744
-0.025	0.666	290,445,187	45,902,053	244,543,133	231,410,802	8,834,577,684	3.2%	79.09	27073	5246	8031	40350
-0.025	0.999	354,501,597	190,899	354,310,698	231,466,667	8,752,105,413	3.9%	78.85	27116	5167	6611	38894
-0.050	0.333	279,563,402	77,004,082	202,559,320	231,471,033	8,851,986,763	3.0%	79.18	27077	5248	8324	40649
-0.050	0.666	397,074,357	71,576,380	325,497,977	231,606,958	8,700,275,606	4.3%	78.75	27167	5093	5720	37980
-0.050	0.999	476,254,443	266,271	475,988,172	231,655,045	8,591,255,016	5.2%	78.53	27251	4946	4036	36233
-0.100	0.333	381,877,312	126,459,616	255,417,696	231,802,179	8,731,108,942	4.2%	78.86	27181	5105	6141	38427
-0.100	0.666	528,647,468	97,383,322	431,264,146	231,966,565	8,530,420,782	5.8%	78.42	27357	4798	3087	35242
-0.100	0.999	631,606,056	359,894	631,246,162	231,970,742	8,382,906,188	6.9%	78.12	27520	4497	1241	33258
-0.150	0.333	450,449,991	151,066,725	299,383,266	232,154,887	8,651,179,570	4.9%	78.70	27288	4958	4768	37014
-0.150	0.666	621,625,868	115,027,545	506,598,324	232,347,230	8,412,227,233	6.8%	78.19	27548	4510	1472	33530
-0.150	0.999	742,093,305	430,021	741,663,284	232,300,973	8,239,591,317	8.1%	77.82	27782	3716	0	31498
-0.300	0.333	588,538,248	191,696,925	396,841,323	233,366,127	8,503,667,355	6.4%	78.44	27616	4534	2329	34479
-0.300	0.666	812,804,128	157,464,303	655,339,825	233,634,657	8,196,068,488	8.9%	77.65	28123	2673	0	30796
-0.300	0.999	951,596,870	603,932	950,992,938	233,361,253	8,008,016,774	10.4%	77.08	28560	603	0	29163
-0.500	0.333	713,407,144	215,805,278	497,601,865	235,355,289	8,401,517,798	7.8%	78.29	28058	3994	577	32629
-0.500	0.666	971,233,473	198,634,806	772,598,667	235,752,041	8,063,581,334	10.6%	77.13	28911	343	0	29254
-0.500	0.999	1,112,675,961	788,031	1,111,887,930	235,191,765	7,915,433,819	12.1%	76.31	28304	0	0	28304

than the base case in which elasticity is the same in all periods. The reason is clear from looking at the equilibrium capacities. Elasticity in the peak periods is what drives the reduction in peaker capacity when customers move to RTP. This effect is larger when demand elasticity is greater in the peaks. So, having greater elasticity in peak periods means both greater demand response when there is more demand and a larger change in the equilibrium level of capacity, both of which contribute to a greater surplus gain from moving to RTP.

Table 5 presents the opposite case, in which demand is more elastic in low-demand periods than in high demand periods. The elasticity of demand varies linearly with demand level, in this case from 127% of the original demand elasticity for the lowest demand level to 50% of the original demand elasticity for the highest demand level. These boundaries were again chosen so that the demand-weighted average elasticity is equal to the original demand elasticity.

BH demonstrate that when elasticity is greater in low demand periods, the equilibrium flat rate will be above optimal and increasing the share of customers on RTP must necessarily increase total surplus. Nonetheless, the surplus gains in this case are smaller than in the base case, and much smaller than in the case in which demand is more elastic at peak times. The result follows intuitively after recognizing that inelastic demand during peak times means that RTP has less effect of reducing the amount of peaker capacity necessary to meet demand.

4.2 The Efficiency of RTP with Heterogeneous Customers

Throughout this analysis, I have assumed that all customers have identical demand patterns. Technically, this means that each customer's demand function is a fixed proportion of the aggregate demand function $D_{hi}(P_h) = \gamma_h \bullet A_h \bullet p_h^{\varepsilon_h}$.[15] One might ask how the results would change if customers differed in their demand patterns.

I do not carry out a complete exploration of this complex topic, but a few observations are useful. First, if the customers switching to RTP are chosen randomly from the population as a whole, and each customer is small relative to the aggregate demand, then the results presented here will apply. The aggregate wholesale demand will still be approximately $\tilde{D}_h(p_h, \bar{p}) = \alpha \bullet A_h \bullet p_h^{\varepsilon_h} + (1 - \alpha) \bullet A_h \bullet \bar{p}^{\varepsilon_h}$.

More interesting, however, is the recognition that the RTP adopters are likely to differ from the population on average in two important ways. First, they are likely to have demand profiles that, even absent any adjustment to RTP prices, are less peaky at high-demand times than the aggregate demand. These are the customers who cross-subsidize the peaky-demand customers when all are under a common flat-rate tariff. RTP gives them an opportunity to reduce or end this

15. Note that this means that the demand *function* is a fixed proportion of the aggregate demand *function*. Because different customers face different prices in a given hour – depending on whether they are on a fixed-rate tariff or RTP – this does not mean that a given customer will consume the same share of total system quantity in all hours.

Table 5: Smaller Elasticity with Higher Demand

A	B	C	D	E	F	G	H	I	J	K	L	M
Elas-ticity	Share on RTP	Total Surplus Change from All on Flat	CS Change of Customers on Flat Rate	CS Change of Customers on RTP	Total Energy Consumed (MWh)	Total Energy Bill ($)	TS Chg as pctg of orig energy bill	Flat Rate ($/MWh)	CAPACITY Base-Load	Mid-Merit	Peaker	Total
All On Flat Rate												
---	0.000				231,095,835	9,170,521,267		79.68	26984	5384	14560	46928
Some On RTP												
-0.025	0.333	60,973,382	0	60,973,382	231,456,637	9,109,076,008	0.7%	79.68	27028	5359	13169	45556
-0.025	0.666	121,784,247	2,460,451	119,323,796	231,815,267	9,047,779,004	1.3%	79.65	27074	5268	11952	44294
-0.025	0.999	177,024,608	22,150	177,002,458	232,164,082	8,992,105,924	1.9%	79.58	27119	5192	10819	43130
-0.050	0.333	118,437,413	4,494,537	113,942,876	231,821,793	9,052,836,335	1.3%	79.65	27074	5268	12051	44393
-0.050	0.666	220,414,820	14,073,842	206,340,978	232,522,183	8,950,840,222	2.4%	79.50	27168	5124	9953	42245
-0.050	0.999	297,190,391	98,205	297,092,186	233,161,482	8,873,523,930	3.2%	79.26	27259	4993	8381	40633
-0.100	0.333	210,667,653	21,371,969	189,295,684	232,564,028	8,965,285,145	2.3%	79.54	27171	5121	10221	42513
-0.100	0.666	349,596,306	43,941,433	305,654,873	233,890,389	8,828,532,614	3.8%	79.11	27360	4859	7407	39626
-0.100	0.999	444,180,558	193,792	443,986,766	235,055,299	8,734,565,308	4.8%	78.84	27536	4592	5607	37735
-0.150	0.333	280,447,527	47,059,933	233,387,593	233,332,159	8,903,748,804	3.1%	79.37	27273	4989	8874	41136
-0.150	0.666	439,799,926	62,288,190	377,511,736	235,253,128	8,749,697,717	4.8%	78.87	27554	4587	5783	37924
-0.150	0.999	552,454,609	246,286	552,208,323	236,918,971	8,639,981,640	6.0%	78.62	27808	4200	3771	35779
-0.300	0.333	423,048,000	103,852,046	319,195,955	235,774,058	8,796,788,098	4.6%	79.01	27596	4583	6365	38544
-0.300	0.666	638,364,596	92,432,780	545,931,816	239,438,461	8,602,426,632	7.0%	78.49	28135	3806	2652	34593
-0.300	0.999	801,790,577	352,184	801,438,393	242,528,799	8,454,748,414	8.7%	78.16	28629	3056	274	31959
-0.500	0.333	557,558,695	141,926,277	415,632,419	239,327,549	8,728,896,973	6.1%	78.76	28041	4044	4376	36461
-0.500	0.666	842,778,837	118,889,088	723,889,749	245,404,704	8,497,863,423	9.2%	78.15	28938	2778	101	31817
-0.500	0.999	1,049,404,926	527,912	1,048,877,014	249,300,982	8,341,841,585	11.4%	77.41	29690	0	0	29690

cross-subsidy. Second, the RTP adopters are likely to be more able to respond to high peak prices by reducing consumption, i.e., to have demand that exhibits more price-elasticity in response to peak prices.

While this heterogeneity has obvious and important implications for the wealth transfers that RTP would effect, it also has potential implications for the efficiency of RTP. To the extent that the RTP adopters exhibit less peaky demand (but still the same demand *elasticity* in each hour as all other customers), this selection of customers moving to RTP would reduce the efficiency gains from the change. This is because the RTP adopters would in aggregate be a smaller proportion of total demand at peak times than at other times. The primary efficiency gains come from price-responsive demand reduction at peak times, so the potential for gains from such response is reduced if RTP adopters have relatively less demand at those times.

The fact that RTP adopters are likely to be more able to respond to high prices, however, will tend to improve the efficiency gains from RTP. If RTP adopters have the same peakiness in their demands as the system aggregate, analyzed for instance at the original flat-rate tariff, but have greater elasticity, then the gains from RTP would be greater than suggested by the previous calculations. The RTP adopters would simply have higher demand elasticity, so one would want to use a different row of the tables than if RTP adopters were representative of the overall demand elasticity of all customers.

5. IS TIME-OF-USE PRICING A GOOD SUBSTITUTE FOR RTP?

Though RTP has not been implemented in many electricity systems, the alternative assumption I've made thus far - that all customers are on flat rates - is also not accurate. In fact, in nearly all systems, prices for some customers vary over time, but in a pre-set manner. These "time-of-use" (TOU) pricing systems generally include peak/shoulder/off-peak prices that are set months in advance and are in effect for fixed hours of each week. For example, Pacific Gas & Electric's basic TOU rate for small commercial customers in summer 2004 was 30.0¢/kWh during peak hours (noon-6pm on non-holiday weekdays), 13.9¢/kWh during shoulder hours (8am-noon and 6pm-11pm on non-holiday weekdays) and 8.7¢/kWh during off-peak (all other) hours. Thus, a worthwhile question to ask is how much of the welfare gains I've identified from RTP are captured with simple TOU pricing.

Unfortunately, while the allocation of costs in a flat-rate or an RTP system is straightforward, this is not necessarily the case in "middle" cases such as TOU or seasonally varying prices. To illustrate, consider a simple example with one L-shaped production technology and two TOU pricing periods. Assume that the utility must break even overall, exactly covering its fixed plus variable costs, and that it must build enough capacity to meet the highest quantity demanded. These seem like minimal constraints, but this problem in many cases still has no solution.

One approach is to allocate all of the fixed costs of capacity to the "peak" period, when the full capacity is used in at least one hour, and set the price during the off-peak period equal to the variable production cost. This approach is appealing because it mimics the outcome that would obtain if the prices were equal to the weighted-average (competitive) wholesale price during a TOU period (assuming that the wholesale demand exhibited just the slightest bit of elasticity so prices were not indeterminate in the peak hour). Even in this case, however, things do not work out simply. If the "off-peak" period has even one high-demand hour, then with a sufficiently high retail demand elasticity and peak-period price, the highest quantity demanded hour for the system could occur during the off-peak period, making it effectively the peak period.[16] One solution is to constrain the prices so that the maximum quantity always occurs during the designated peak period, but this is just artificially constraining the peak/off-peak price difference to the level that (nearly) equalizes demand in the highest demand peak-period hour and the highest-demand off-peak period hour.

Another approach is to allocate the fixed cost of a unit of capacity equally over all periods in which that unit is used. For example, the cost of capacity used at the minimum quantity time would be spread over all hours, because that capacity is used in all hours, while the cost of the last unit of capacity, used only at the maximum demand time, would be borne entirely by the consumers in that period. This greatly reduces (though does not eliminate) the peak-switching problem, but it also greatly dampens the price swing across TOU periods. It does have the popular appeal that only, and all, those who use a given unit of capacity pay for it.

I have tried three different approaches to constructing a TOU pricing scenario that could then be compared to the RTP and the flat rate scenarios. The first, which I call the "quasi wholesale" scenario attempts to mimic the weighted average competitive wholesale price with all capacity costs of the peaker capacity allocated to the hour in which quantity demanded is highest.[17] This has a solution for low demand elasticities, but does not have a solution if demand elasticity is too large. With the five years of California data that I am using, "too large" is an elasticity greater in absolute value than 0.05.

The second approach is the "cost-share" scenario in which the allocation of capacity costs is determined by the number of hours in which a given unit of

16. This is obviously related to the "shifting peaks" problem which was identified in the early peak-load pricing literature of the 1950s and then, with the help of a Lagrangian multiplier, solved. See, for instance, Steiner (1957). The present problem, however, does not disappear so easily. A first-best solution cannot be implemented, because there is not a complete set of prices for every demand state.

17. Baseload, mid-merit, and peaker capacities are set to minimize total production costs for a given peak, shoulder, and off-peak price, which determine quantity demanded in each hour. The competitive wholesale price for each hour is then calculated using those demand quantities (without elasticity). Then the TOU prices during each period are reset to be weighted-average wholesale prices during the period. This iteration continues until a fixed point is found.

Table 6: Welfare Effects of RTP versus TOU Pricing

A	B	C	D	E	F
Elas-ticity	Share on RTP/TOU	ANNUAL TOTAL SURPLUS CHANGE VS FLAT RATE			
		RTP	"Quasi-wholesale" TOU	Actual TOU price ratios	"Cost-share" TOU
-0.025	0.333	112,060,365	16,269,127	10,657,394	6,928,165
-0.025	0.666	205,800,109	32,538,254	21,314,789	13,856,330
-0.025	0.999	271,333,946	48,807,381	31,972,183	20,784,495
-0.050	0.333	196,836,537	32,226,253	21,322,177	13,683,652
-0.050	0.666	314,219,558	64,452,506	42,644,355	27,367,305
-0.050	0.999	388,316,857	96,678,759	63,966,532	41,050,957
-0.100	0.333	302,262,176	N/A	42,006,103	26,159,344
-0.100	0.666	439,987,363	N/A	84,012,206	52,318,689
-0.100	0.999	537,284,137	N/A	126,018,309	78,478,033
-0.150	0.333	370,238,483	N/A	61,775,434	37,387,646
-0.150	0.666	530,960,593	N/A	123,550,868	74,775,291
-0.150	0.999	647,620,518	N/A	185,326,302	112,162,937
-0.300	0.333	509,388,631	N/A	N/A	65,167,555
-0.300	0.666	730,577,275	N/A	N/A	130,335,110
-0.300	0.999	888,877,347	N/A	N/A	195,502,666
-0.500	0.333	641,472,723	N/A	N/A	92,710,676
-0.500	0.666	922,328,312	N/A	N/A	185,421,352
-0.500	0.999	1,098,811,460	N/A	N/A	278,132,028

capacity is used during each of the TOU periods.[18] With the data I am using, this produces a solution for all elasticities up to 0.5 in absolute value.

The third approach is a "fixed-ratio" scenario in which the ratios of peak to shoulder and off-peak prices are set exogenously and then the prices and capacity are set in much the same way as in the flat-rate simulation described in the previous section. The price ratios were set to a level that reflects the average of the (fairly similar) pricing structures used by Pacific Gas & Electric and Southern California Edison, the two major utilities in California. This yielded a solution for elasticities up to 0.15 in absolute value.

For all three scenarios, all prices were allowed to vary between winter and summer as well. In particular, similar to the utilities' actual TOU rate structures, there were two prices in the winter: a peak price that was in effect 8am-9pm on non-holiday weekdays, and an off-peak price that was in effect on at all other times in the winter. In the summer, there were three TOU periods: Peak

18. Baseload, mid-merit, and peaker capacities are set to minimize total production costs for a given peak, shoulder, and off-peak price, which determine quantity demanded in each hour. The allocation of the fixed capacity costs is then determined by the quantities demanded in each period and the levels of each type of capacity. Based on this allocation, the TOU prices during each period are reset to cover each period's variable costs plus share of capacity costs. This iteration continues until a fixed point is found.

was noon-6pm on non-holiday weekdays; Shoulder was 8am-noon and 6pm-11pm on non-holiday weekdays; Off-peak was in effect at all other times.[19] Summer was defined as June-September and winter was defined as October-May.

In the simulations I present, the prices change by season, but not year-to-year. The summer peak price, for instance, is the same in all years. This is meant to reflect the fact that the year-to-year variation during this period is mostly not predictable growth, but idiosyncratic weather variation that would not be predictable at the time that the TOU prices were set for each time period.

The welfare results of these simulations are presented in table 6, with the figures for RTP also presented for comparison. The conclusion is clear: TOU rates capture a small share of the benefits that would be obtained from RTP. Even the most efficient form of TOU ("quasi-wholesale"), which generates peak to off-peak price ratios well above those observed in actual TOU programs, captures only one-quarter or less of the RTP gains for those elasticities for which it is feasible. Using actual fixed-ratios of prices, the gains also seem to get up to about one-quarter of RTP before those price ratios become infeasible at higher elasticities.

I should note, however, that there is a critical assumption in these calculations, that elasticity of demand in responding to long-run TOU prices is the same as the elasticity in response to RTP prices. Put differently, one can think of RTP prices as decomposable into different averages for TOU-like periods and deviations from those averages in any given hour. The underlying assumption is that customers would be equally responsive to the variations in averages as to the deviations from those averages in a particular hour.

In reality, elasticity with respect to short-term fluctuations could be lower or higher than with respect to longer-term predictable average price differences. One could argue that the short-term less-predictable deviations are more difficult to respond to because of the lack of advanced notice. For instance, companies could not reschedule work shifts based on a price spike that becomes apparent only hours before it actually occurs. In the extreme, if the only electricity-consumption modifications that a customer could make would be the result of months-ahead planning, then RTP offers a much smaller advantage over TOU.[20] The elasticity of demand with respect to deviations from months-ahead expected price for a given hour would be virtually zero.

On the other hand, there may be short-duration adjustments that a firm could make to respond to a price spike that they could not maintain for a longer period. For instance, if a company knew that a heat spell is driving prices to very high levels today, but will likely break by tomorrow, it could possibly shift some electricity-intensive activity to tomorrow. The potential for these sort of

19. For the fixed-ratio scenario, all prices were fixed as a proportion of the summer peak price. The proportions were summer/shoulder 57.4%, summer/off-peak 45.0%, winter/peak 61.9%, and winter/off-peak 47.7%.

20. RTP would still offer better granularity of prices, as the 3-4pm expected price for a day six months hence would differ at least slightly from the 2-3pm expected price for that same day.

short-term adjustments suggests that the elasticity could be greater for short-term deviations than for long-term average price differences.

The relatively small efficiency gains from TOU pricing are quite intuitive when one recalls that the inefficiency from non-optimal pricing in a given hour goes up in proportion to the square of the deviation of price from marginal cost. Thus, the most costly "mistakes" occur during the times when prices deviate most from the mean during a given TOU period. Intuitively, then it would be more effective to attain a given average price within a certain TOU period by having very high retail prices during the few hours with highest wholesale prices and slightly lower retail prices at all other times, thereby substantially mitigating the largest pricing mistakes rather than addressing slightly a larger number of small price mistakes. A program that roughly takes this approach, called "Critical Peak Pricing" is currently being tested in California and elsewhere. My preliminary analysis suggests CPP could capture a much greater share of the RTP efficiency gains than could TOU.

Still, the TOU results do make clear that the gains I've claimed from RTP in the previous section were slightly overstated. The baseline from which most systems begin is with 50% or more of total demand on TOU, including most customers that would be initially put on RTP if only a share of customers were moved to RTP. Thus, the gains from moving these customers to RTP should be scaled down by between 15% and 25% (using the assumption that elasticity of demand is the same for longer-term changes as for shorter-term price variations).

6. LIMITATIONS OF THE RTP SIMULATION MODEL

Though these simulations are useful in giving an idea of the potential gains from RTP, they don't take into account all aspects of electricity markets. Incorporating many of these characteristics will be challenging, but it is clear even without that additional analysis that these simulations are likely to understate the benefits of RTP.

The most important area of omission is the stochastic elements of supply and demand. The model does not incorporate the unpredictability of demand or the probabilistic outages of generation supply. Currently, responses to these stochastic elements of the supply/demand balance are addressed almost entirely with supply adjustment. Unless, short-run demand adjustment is impossible, which there is increasing evidence is not the case, responding entirely on the supply side is clearly not the most efficient way to address such outcomes.

Including RTP in system balancing will further enhance system efficiency. It seems almost certain that RTP would decrease system peak loads, so using standard proportional reserve rules, it would reduce the amount of reserve capacity needed and the payments for that capacity. More importantly, RTP would increase the responsiveness of demand to system stress and thus would reduce the level of reserves needed for any given level of demand. In economic terms, RTP would not just shift demand to the left at peak times, it would make demand

more price elastic, so more balancing could be accomplished with less supply-side adjustment. Likewise, incorporating generator outages raises the benefits of demand responsiveness by reducing the need to compensate for a generator outage completely on the supply side.

Assuming competitive supply, an upper bound on the "reserves cost" savings from RTP is the total cost of reserve payments. In most systems, operating reserves average 5-10% of energy costs. Planning reserves costs may be covered by energy and operating reserve payments, or they may require additional payments, which would also be subject to reduction through use of RTP. RTP is likely to reduce these costs by a significant amount, but much of these costs will remain for a long time. Nonetheless, the benefits from RTP are likely to be underestimated from the simulations presented, because they do not incorporate the benefits from reduced need for reserves.

Closely related to reserves costs are the effect of non-convexities in operation of plants and lumpiness in the size of plants. As discussed in detail by Mansur (2003), generation units do not costlessly or instantly switch from off to full production. There are start-up costs and "ramping" constraints (on the speed with which output can be adjusted). These constraints make it more costly to adjust supply to meet demand fluctuations. As with reserves, RTP would allow some of this adjustment to occur on the demand side in a way that would enhance efficiency. Similarly, I have assumed the plants can be scaled to any size at the same long-run average cost. If this were not the case, then there would be greater mismatches between demand and the capital stock. In conventional electricity systems, these mismatches have been handled by over-building and then either selling excess production on the wholesale market or leaving excess capacity idle. Having the additional option of demand-side adjustment could only lower long-run costs.

The simulations also have ignored market power issues, instead assuming that free entry would bring a completely competitive market over the longer run. As has been discussed elsewhere, e.g., Borenstein and Bushnell (1999) and Bushnell (forthcoming), demand elasticity introduced by implementing RTP reduces the incentive of sellers to exercise market power. However, it is unclear how much incremental inefficiency the exercise of market power itself introduces in a flat-rate system, since it simply changes the flat retail rate that is charged in all time periods. In fact, Borenstein and Holland's analysis (forthcoming) suggests that if the equilibrium flat rate is less than the surplus-maximizing flat rate, $\overline{p}^e < \overline{p}^*$, seller market power could increase efficiency. In a full RTP system, market power could not increase efficiency. Thus, it is difficult to analyze the bias from excluding seller market power.

The demand system I've analyzed departs from reality by assuming all cross-elasticities are zero. Simulation with a complete matrix of own- and cross-elasticities would increase the complexity substantially. Still, if demands are generally substitutes across hour, it seems very likely that incorporation of cross-elasticities would increase the gains from RTP. Essentially, RTP increases efficiency by reducing the volatility of quantity consumed and increasing the

utilization rate of installed capacity. Holding constant own-price elasticities, increasing cross-price elasticities from zero to positive (substitutes) will tend to further reduce quantity volatility by increasing off-peak quantity when peak prices rise and reducing peak quantity when off-peak prices fall.

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

7. CONCLUSIONS

Real-time electricity pricing has tremendous appeal to economists on a theoretical level, because it has the potential to improve welfare by giving customers efficient consumption incentives. The theoretical analysis, however, does not indicate how large the gains from RTP are likely to be. With a simple simulation exercise, I have tried to generate some numbers to go with the theory. This is obviously just a first cut, but the results suggest a number of likely findings:

- The benefits of RTP are likely to far outweigh the costs for the largest customers.

- The incremental benefits of putting more customers on RTP are likely to decline as the share of demand on RTP grows. At the same time, the cost of increasing the share of demand on RTP is likely to increase as the size of each customer declines. Thus, while there seems to be clear net social value from putting larger customers on RTP, the additional gains from putting smaller customers on RTP may not justify the cost. A factor weighing against this conclusion is that small customers are thought by many electricity analysts to be the most price responsive. If that is true, then the argument for RTP metering of them is, of course, strengthened. Further analysis of both the costs and benefits is needed.

- Time-of-use rates are a very poor substitute for RTP. Roughly speaking, TOU rates capture only 20% of the efficiencies of RTP, though this finding has the caveat that it assumes as high an elasticity for response to short-run price variation as long-run differences in average prices.

The findings of this study should be viewed as a first step. A number of factors have not been addressed in the analysis thus far, though incorporating them seems likely to lead to larger estimated gains from RTP. Incorporation of these factors into the analysis is not particularly complex. A larger barrier is likely to be the data necessary to permit reliable estimates of demand elasticities and supply flexibility, which I've shown have very large impacts on the efficiency gains.

Finally, it is worth pointing out that RTP is being adopted in a number of places in the U.S. and elsewhere. The programs are relatively young – the oldest began in the early 1990s – but there are already a number of examples of programs with which both the utilities and the customers are quite happy, and that have documented both peak-demand reductions and reduced need for peaking capacity. For a very thorough description of voluntary dynamic pricing programs in the U.S., see Barbose, Goldman and Neenan (2004). The large RTP programs operated by Georgia Power, Gulf Power and Niagra Mohawk should be of great interest to those evaluating the efficacy of RTP.

REFERENCES

Barbose, Galen, Charles Goldman and Bernie Neenan (2004). "A Survey of Utility Experience with Real-Time Pricing," Lawrence Berkeley National Laboratory Working Paper No. LBNL-54238, December 2004. Available at http://eetd.lbl.gov/ea/ems/reports/54238.pdf.

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

Borenstein, Severin and James B. Bushnell (1999), "An Empirical Analysis of Market Power in a Deregulated California Electricity Market," *Journal of Industrial Economics* 47: 285-323.

Borenstein, Severin and Stephen P. Holland (2003). "Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices," Center for the Study of Energy Markets Working Paper No. 106, University of California Energy Institute, revised July. Available at http://www.ucei.org/PDF/csemwp106.pdf.

Borenstein, Severin and Stephen P. Holland (forthcoming). "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices," *RAND Journal of Economics*.

Bushnell, James B. (forthcoming) "Looking for Trouble: Competition Policy in the U.S. Electricity Industry," in Griffin and Puller, eds., *Electricity Deregulation: Choices and Challenges*, Chicago: University of Chicago Press.

Braithwait, Steven D. and Michael O'Sheasy (2002). "RTP Customer Demand Response – Empirical Evidence on How Much Can You Expect," in Faruqui and Eakin eds. *Electricity Pricing in Transition*, Boston, MA: Kluwer Academic Publishers.

Holland, Stephen P. and Erin T. Mansur (2005) "The Distributional and Environmental Effects of Time-Varying Prices in Competitive Electricity Markets," Center for the Study of Energy Markets Working Paper No. 140, University of California Energy Institute, October. Available at http://www.ucei.org/PDF/csemwp140.pdf.

Mansur, Erin T. (2003). "Vertical Integration in Restructured Electricity Markets: Measuring Market Efficiency and Firm Conduct," Center for the Study of Energy Markets Working Paper No. 117, University of California Energy Institute, October. Available at http://www.ucei.org/PDF/csemwp117.pdf.

Patrick, Robert H. and Frank A. Wolak (1997). "Estimating the Customer-Level Demand for Electricity Under Real-Time Market Prices," Stanford University Working Paper. Available at ftp://zia.stanford.edu/pub/papers/rtppap.pdf.

Ruff, Larry E. (2002). "Demand Response: Reality versus 'Resource'," *The Electricity Journal*, 15(10):10-23.

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