Optimization of Time-Varying Electricity Rates

Jacob Mays and Diego Klabjan***

ABSTRACT

Current consensus holds that 1) passing through wholesale electricity clearing prices to end-use consumers will produce maximal efficiency gains and 2) simpler forms of time-varying retail rates will capture only a small portion of potential benefits. We show that neither holds in the presence of capacity costs typical in U.S. wholesale markets. Using an optimization model describing the short-term problem faced by an electricity retailer, we find hourly prices that optimally pass through capacity costs. We estimate benefits for a retailer using these prices as well as optimal configurations of a number of time-varying rate structures. Testing a range of realistic assumptions, we find that in the absence of a well-designed demand charge, passing through clearing prices may miss up to three quarters of the benefits possible from optimal hourly prices. By contrast, a simpler critical peak pricing structure enables retailers to achieve approximately two-thirds of the total possible benefits.

Keywords: Rate design, time-of-use rates, critical peak pricing, real-time pricing

https://doi.org/10.5547/01956574.38.5.jmay

1. INTRODUCTION

The cost of generating electricity can vary tremendously over the course of the day and the year, but end-use customers have traditionally seen prices that are flat or close to it. Residential customers are particularly insulated from cost movements, with only 4 percent of U.S. households facing time-varying rates (U.S. Energy Information Agency (2015)). The mismatch between the cost of supplying electricity and the price of using it leads to inefficiency: customers use too much electricity when costs are high and too little when costs are low. The effect of this inefficiency is compounded by the capital-intensive nature of electricity generation. Capacity is built to meet the highest load of the year, leading to the construction of generating stations that operate only rarely. For instance, the average capacity factor for natural gas-fired combustion turbines, used primarily for peak loads in the United States, is under 5 percent (U.S. Energy Information Agency (2015)).

Recognizing this problem, many economists and policy makers over the past several decades have advocated a shift from fixed to time-varying electricity prices. Allowing prices to change can both reduce the deadweight loss arising from differences between wholesale and retail prices and lead to a reduction in required capacity. This proposal has gained popularity in recent years, facilitated by the spread of advanced metering infrastructure (AMI). Smart meters represented 41

* Corresponding author. PhD Candidate in Industrial Engineering & Management Sciences, Northwestern University. Send correspondence to 2145 Sheridan Road, Room C210, Evanston, IL 60208. E-mail: jacobmays@u.northwestern.edu.

** Professor of Industrial Engineering & Management Sciences and Director of Master of Science in Analytics Program, Northwestern University.

percent of all meters in the U.S. in 2014, with an additional 33 percent enabling automated meter reading (AMR).[1] Over 500 utilities across the United States had at least some customers enrolled in time-varying rates in 2014 (U.S. Energy Information Agency (2015)). The states of Massachusetts (Massachusetts Department of Public Utilities (2014)) and California (Public Utilities Commission of the State of California (2015)) have taken a step further, moving toward establishment of default time-varying rates.

While momentum for time-varying rates has been gaining, there is little agreement on the form such rate structures should take. A majority of the economic literature encourages a move to real-time pricing (RTP), generally understood to be hourly retail prices that pass through the clearing cost in either the day-ahead or real-time wholesale market. Prices that change by the hour can track the volatility of wholesale markets and send accurate price signals. However, the resulting increase in complexity could be undesirable if customers are unable to respond to them efficiently or are unhappy with potential increases in bill volatility. While this trade-off is well-understood intuitively, quantitative comparisons of potential rate structures have been limited. This paper addresses the benefits side of this cost-benefit analysis.

Optimization of retail electricity prices has two components: accurately reflecting energy costs that vary over time, and choosing the best method by which to pass through capacity costs. For participants in wholesale markets, one solution to this problem comes in the form of an Energy Only market design. Retailers[2] in this setting have no explicit capacity charge, paying only for energy consumed. This leads to high prices in hours of peak demand, during which generators recover their capacity costs. The resulting wholesale prices reflect both marginal costs and capacity constraints. In the U.S., however, only ERCOT operates with an Energy Only design. All other markets have opted for an Installed Capacity (ICAP) market design, in which retailers must either self-supply or purchase sufficient capacity to cover their customers' coincident peak load plus a reserve margin. Accordingly, retailers in these markets have a large cost corresponding to their maximum demand that is not reflected in wholesale prices. In this setting, passing through wholesale energy prices focuses on the first goal of time-varying rates, eliminating deadweight loss, and misses a substantial portion of the benefits possible from capacity reductions.

There are at least four strategies retailers could adopt to recover these capacity costs and promote capacity savings: fixed charges, non-coincident peak demand charges, coincident peak demand charges, and combining energy and capacity in a single volumetric price. Fixed charges represent a worst case scenario, as they do not allow for any customer response. Non-coincident peak demand charges encourage end users to limit their individual maximum load; however, given the diversity of electricity consumers, most of these individual peaks are unlikely to align with the system peak that actually drives cost. Coincident peak demand charges may be difficult for customers to manage, since they require each consumer to predict when system peaks are likely to occur. This paper focuses on the fourth strategy, determining retail prices that bundle energy and capacity into a single signal. However, unlike typical residential rates that recover capacity costs by adding a small amount to prices in all hours of the year, our goal is inject these capacity costs into the best possible hours. To distinguish them from the wholesale clearing prices utilized in RTP, we designate these optimal hourly prices RTP$^+$.

1. AMR allows one-way transmission of data from the customer to the utility, whereas AMI allows two-way communication. Accordingly, AMR is sufficient for time-varying rates, but unlikely to produce the same level of customer response.

2. For convenience, throughout the paper we refer to buyers of electricity as retailers instead of Load Serving Entities or utilities. That said, retail choice is not a prerequisite for the applicability of the results.

With a focus on the ICAP market design, we construct a model that determines optimal retail prices for a variety of rate structures. We use this model to compare different rate structures under a range of realistic assumptions. Previous comparisons have either utilized an Energy Only market design (Borenstein (2005)) or set prices without regard to capacity (Ata et al. (2015); Hogan (2014); Holland and Mansur (2006); Spees and Lave (2008)). In addition, all have focused on a limited set of rate configurations, covering only select two- and three-tier Time of Use (TOU) rates. Accordingly, the applicability of these estimates to markets with significant capacity charges is uncertain. We solve for a wider range of TOU configurations, as well as extending to Critical Peak Pricing (CPP) and RTP$^+$. Through the use of a model that finds optimal prices for several rate structures, this paper both broadens the range of structures tested and enhances the validity of the resulting comparisons.

Two modeling choices facilitate these more complete comparisons. First, we consider only the incentives of electricity retailers and their customers. Crucially, the welfare of generators is not included. This choice is realistic in a deregulated setting, in which retailers divest all ownership in generation. The primary effect of this decision is that retailers in the model do not need to consider ongoing expenses from stranded generation assets. One consequence is that a long-term model is not needed to estimate welfare gains; the effect of capacity reductions can be captured as early as the next capacity auction.[3] Without this assumption, the cost of capacity is effectively zero up to the currently installed amount, with savings only possible from deferral of future investments. Nevertheless, the fundamental problem of passing through capacity costs in the optimal hours does not change in the regulated setting.

The second choice is in the population moving to time-varying rates. An actual deregulated market contains many retailers, each offering several rate structures to a wide variety of customers, each of which have distinct demand characteristics. Previous models have assumed that fixed percentage of baseline demand moves to time-varying rates. Consistent with a focus on retailers, we instead assume that entire geographical areas move to time-varying rates. We test our model on several such zones, allowing us to confirm that the performance of time-varying rates is not contingent on geographical differences in demand characteristics. Further, testing combinations of zones provides a natural way to test how the effectiveness of these rates changes as when applied to a larger proportion of customers. To make comparisons as clear as possible, each scenario of the model splits customers into two groups: one zone or group or zones that all shifts to the same time-varying rate structure, and a second that does not change its demand from the baseline, fixed-rate scenario.

In addition to these modeling decisions, a third departure relates to the presentation of numerical results. We place an emphasis on the relative rather than absolute performance of the tested rate structures. Estimates of the benefits of time-varying rates are highly dependent on uncertain input values, key among them the demand elasticity of electricity users. In the ICAP setting, a second factor is wide geographic differences in the price of capacity. In practice, retailers will need to revisit decisions on time-varying rates frequently as values for these inputs are updated. Given these uncertainties, we focus on evaluating the conditions under which a particular rate structure is likely to be favorable.

3. A secondary effect is that retailers have an incentive to set prices slightly higher than they would otherwise; this reduces demand and leads to a lower clearing price. For small retailers this effect is small, since most of the benefit from this price suppression goes to other buyers. Nevertheless, we mitigate this possibility by enforcing a profit constraint in our model.

One simplification deserves special attention. Like the retailers we model, many large consumers of electricity already have a component of their bill determined by coincident peak demand. Accordingly, many approaches to predict the timing of coincident peaks and reduce demand in those hours already exist in practice. In this paper, the RTP rate structure passes through wholesale clearing prices, and a separate demand charge is required for the retailer to recover capacity costs. All other rate structures tested recover capacity costs by adding to rates throughout the year in the optimal hours, obviating the need for a demand charge. We assume that customers do not respond to the demand charge in RTP, except to the extent it is already present in historical data. Accordingly, all results for RTP can be seen as a lower bound on potential benefits. Conversely, results for the remaining rate structures can be seen as upper bounds, since any response present in the historical data might disappear.

Our objective is to find retail rates that maximize surplus for a retailer and its customers subject to the constraints of the chosen rate structure. Required inputs include models for customer demand, wholesale supply, and the cost of capacity. Our chosen representation for these three inputs leads to a nonlinear convex optimization problem, guaranteeing optimality of the resulting prices. We test the model using data from PJM, a regional transmission organization (RTO) covering an area of approximately 61 million people in the eastern half of the United States (Monitoring Analytics, LLC (2015)).

In a typical setting, RTP may achieve only 25 percent of the surplus of RTP$^+$. By contrast, a simple three-tier TOU structure can achieve 26 percent, whereas CPP can reach 65 percent. These results represent a significant departure from previous estimates, demonstrating the importance of tailoring to the ICAP setting. Two observations explain the high-level results. First, the most effective time-varying rates derive benefits primarily from curtailing peak load. Reductions in deadweight loss throughout the year, often taken to be the main goal of time-varying rates, make a much smaller contribution to welfare. Second, passing through wholesale clearing prices will not accomplish the goal of curtailing load in the ICAP setting. The optimal prices proposed in this paper, or separate incentives like a demand charge or direct load control, are required to achieve all the possible benefits of time-varying rates.

Our main contributions are as follows. First, we propose a model able to find optimal retail prices for a variety of rate structures in an ICAP setting. Second, we propose a new rate structure, RTP$^+$, denoting hourly prices that optimally combine energy and capacity costs. Third, we compare the performance of several rate structures currently in use as well as RTP$^+$.

We provide an overview of the relevant literature in Section 2. The energy and capacity markets are described in Section 3, along with the effect of time-varying retail prices on each. We present the optimization model as well as modeling considerations in Section 4. Data for calibration to PJM is described in Section 5. Comparisons of the performance of different time-varying rate structures under a range of conditions are shown in Section 6, and we conclude in Section 7.

2. LITERATURE REVIEW

The concept of tying retail electricity rates to underlying generation costs goes back several decades. A survey of much of the theory related to peak-load pricing is given in Crew et al. (1995). This theory is extended to competitive markets with a fraction of customers constrained to fixed rates in Borenstein and Holland (2005). A basic result of this literature is that the lack of real-time pricing leads to over-building of capacity, as electricity customers have no incentive to curtail usage in the highest-load hours of the year.

Despite long-standing theoretical support, the deployment of time-varying and dynamic retail electricity rates in the real world involves several complications.[4] An overview of these considerations is provided in Faruqui et al. (2012). The authors describe the advantages and disadvantages of several rate structure choices, identifying a risk-reward tradeoff between the variance of retail prices and the potential benefits. Prices that change by the hour can entail risk for customers, but also provide the most savings. Rate structures with less volatility can protect customers, but also fail to capture all the potential savings. This relationship forms the motivation for the comparisons in our paper: without an estimate of the potential benefits, evaluating the tradeoff is impossible. As described by Faruqui et al. (2012), volatility is likely to affect both the acceptance of time-varying rates and the ability of customers to respond optimally; this could manifest, for example, as weaker elasticity of demand for more complex rate structures. Consistent with other authors, our model makes no adjustment for this possibility. Additional discussion of these issues, with a particular focus on the effect of distributed energy resources, is provided by Glick et al. (2014).

Several authors have recognized the need for an estimate of the comparative benefits of RTP and simpler time-varying rate structures. Building on the theoretical model in Borenstein and Holland (2005), Borenstein (2005) develops a two-stage entry model that projects long-run efficiency gains from RTP in an Energy Only market. Generators set capacity for three different production technologies in the first stage based on how much revenue they will generate from electricity sales in the second stage. Accordingly, capacity savings result from reduced entry in the first stage. Because clearing prices are optimal in the Energy Only setting, no distinction between RTP and RTP$^+$ is needed. Three different TOU rates are compared to RTP across a wide range of scenarios for demand elasticity and share of customers moving to time-varying rates; the best performing TOU rate achieves between 13 and 29 percent of the benefits of RTP, depending on the scenario. As a consequence of the two-stage design, none of the tested TOU rates are guaranteed to be optimal. By contrast, our approach adapts the theoretical model from Borenstein and Holland (2005) to the ICAP setting and solves the problem in a single-stage optimization, enabling the determination of optimal prices.

Three estimates for the relative effectiveness of TOU come from short-run models without capacity costs. The model used by Holland and Mansur (2006) is distinguished by a detailed description of generators in PJM, allowing a more complete accounting of generator profits and the environmental effects of time-varying rates. The authors test a TOU rate in use by a PJM utility at the time, finding that only 15 percent of deadweight loss is eliminated by TOU pricing. Changing flat rates on a monthly basis, however, eliminates 30 percent of this loss. The model of Spees and Lave (2008) also uses PJM data, fitting a parametric model to describe changes in supply cost. As in Holland and Mansur (2006), the authors test a TOU structure then in use, projecting 20 to 22 percent of the benefits of RTP. Lastly, the model developed in Ata et al. (2015) uses demand characteristics from household-level data collected through a field experiment in Ireland along with constant wholesale electricity prices. The TOU rates tested in the model, chosen to match the structures used in the experiment, achieve roughly 28 percent of the surplus of RTP in the competitive setting. In each of these three models, prices in each TOU period are the unique prices that result in zero profit within each period; reductions in peak load are reported only as a consequence of these prices. A fourth short-term estimate, also without capacity costs, comes from a simple

4. "Time-varying" refers to any rate that changes over the course of the day, whereas "dynamic" implies that rates react to current market conditions. In this paper, the dynamic rates are CPP, RTP, and RTP$^+$.

variability index proposed in Hogan (2014). The author estimates that even a complex TOU rate can only achieve 23 percent of the benefits of RTP.

The key feature distinguishing our model from these four comparisons is the explicit inclusion of capacity costs. Over the past decade, several wholesale markets have developed capacity markets as a solution to the "missing money" problem described in Cramton and Stoft (2006) and Joskow (2008). The recent history of the PJM capacity market is covered in Bowring (2013). In the current paper, we do not address the necessity or desirability of a capacity market, instead taking its presence as a given. Further, we do not directly model the capacity market, instead assuming an exogenous price of capacity and testing a range of potential inputs.

The strategy of combining capacity and energy costs into a single rate can be found in the model of Newell et al. (2009). The authors construct a rate that add capacity costs to roughly the top one percent of consumption hours, estimating the effect if all consumers in New York switched to a rate equal to hourly wholesale prices plus this capacity adder. This heuristic leads to a rate structure that looks like a hybrid between RTP$^+$ (because prices change hourly with wholesale market energy prices) and CPP (because a uniform capacity cost is added to a select number of hours). Our approach replaces this heuristic with an optimization model and extends the analysis to a wider set of potential rate structures.

The long-term model of Borenstein (2005) is applied to the ICAP setting in Alcott (2013). The author compares long-term results from RTP in the ICAP and Energy Only designs, finding that the benefits of RTP in an ICAP setting are under half those in an Energy Only setting. Further, the author recognizes that this gap could be bridged with optimal pass-through of prices in an ICAP setting. This observation leads to one goal of our model, namely, determining these optimal prices. Accordingly, our results for RTP and RTP$^+$ roughly correspond to results for RTP in the ICAP and Energy Only settings.

Estimates of the effectiveness of time-varying rates depend on how end users of electricity respond to changing prices. Models for consumer behavior are described in Cappers et al. (2013). These models generally include two types of elasticity: own-price elasticity reflecting the overall change in consumption when moving from fixed to time-varying rates, and substitution elasticity measuring shifts in consumption from more expensive to less expensive time periods. Many estimates for these have been generated from pilots described in Faruqui and Sergici (2009) and several programs established through the U.S. Department of Energy's Smart Grid Investment Grant program U.S. Department of Energy (2015). Following the lead of previous authors (Alcott (2013); Borenstein (2005); Holland and Mansur (2006); Spees and Lave (2008)), we opt for a demand model that includes only own-price elasticity. This simplification is justified both by the uncertainty in presently available elasticity estimates and the increased tractability of the resultant model.

3. MARKET DESCRIPTION

Electricity sales are made at two levels, between which electricity retailers stand in the center. Generators offer electricity to retailers on the wholesale market; retailers purchase this electricity, re-price it, and distribute it to end-use customers. Traditional, regulated utilities consolidate this process into a single level, not requiring a wholesale market since they are both generator and retailer. Optimization of retail pricing applies to either setting. However, we focus on the competitive market due to greater data on capacity costs as well as hourly changes in the cost of electricity.

Optimization of retail pricing relies on the assumption that customers will change their consumption in response to different prices. These changes in consumption in turn have an effect

on the wholesale market: lower demand means that the highest cost generators will no longer be needed, so the clearing price of electricity will fall. Conversely, lower retail prices are likely to lead to higher consumption and therefore higher wholesale prices. The problem faced by the retailer is then to select optimal prices given predicted customer response and its effects on the wholesale market. Many suitable objectives are possible, e.g., maximizing profit, minimizing total consumption, minimizing peak demand, minimizing pollution, or some combination. We elect to optimize total (consumer plus retailer) surplus. Assuming they have the ability to extract surplus through a fixed customer charge that does not affect demand, this aligns with the objectives of a profit-maximizing retailer.

With this framework, we need to model customer demand, the effect of demand on the wholesale clearing price, and the workings of the capacity market. These in combination define the consumer and retailer surplus resulting from given retail rates.

3.1 End-use Demand

Estimating the customer response to time-varying rates is a complex task, considerations for which are discussed in Section 2. We begin by choosing a portion of demand that will move to a given time-varying rate structure. This could be a region, customer segment, or any other subset of participants in the market. Given that pricing decisions are made by retailers who typically have a concentrated geographic footprint, a natural choice for this subset is a geographical zone. We assume that retail customers have a constant elasticity of demand $\varepsilon \leq 0$ within each hour. The value of ε could be set based on pilot studies specific to the retailer and could vary over the course of the day; we assume a single value for all hours that represents the aggregate short-term response of all customer segments. Let $\mathbf{p}$ represent the vector of retail prices by hour, indexed by h. Then, load in an hour in this zone $Q_h^{var}(\mathbf{p})$ can be found as

$$Q_h^{var}(\mathbf{p}) = A_h p_h^{\varepsilon}, \tag{1}$$

with A_h a demand coefficient specific to the hour. To calculate these coefficients, we use historical usage and assume that all customers were on a flat rate $\bar{p}$ through the year. This flat rate is calculated by summing the weighted average of all electricity sales within the zone, the total capacity cost divided by total consumption, and a fixed distribution cost per MWh.

This demand model includes two assumptions worth highlighting. First, there is no cross-price elasticity between hours. While an incomplete picture of individual customer response, this simplification is consistent with the literature and is assumed to provide sufficient accuracy for the aggregate population. Second, as previously described, end-use demand is determined only by the rate in that hour. Under any rate design, sophisticated customers facing a separate demand charge can and do attempt to predict the timing of coincident peak hours and moderate their consumption accordingly, allowing them to reduce future demand charges. The model assumes that no such response occurs.

Our approach, which allows any subset of demand to move to time-varying rates, offers a contrast to the previous literature, which has typically assumed that the moving subset is a fixed percentage of baseline demand. This decision is in keeping with our focus on retailers, whose load profiles may not match that of the overall system. For example, retailers whose underlying demand is already low during system peaks may not need to be as concerned with reducing capacity obligations. The model also offers a natural extension to situations in which zones are further subdivided

into residential, commercial, and industrial customers with their own demand profiles, elasticities, and rate structures.

Moving an entire zone mimics the "all or nothing" decision faced by a typical retailer: trials described in Faruqui (2015) suggest that enrollment in time-varying rates jumps from 20 percent when customers opt in to 84 percent when they opt out. Moreover, it allows us to avoid problems in the estimation of elasticity. It is likely that early adopters of time-varying rates will be customers with flatter than typical load profiles or particular capabilities enabling greater responsiveness. Further, some load was already on time-varying rates in 2014, meaning that some of the gains from shifting to time-varying rates have already been realized. Instead of producing estimates for these effects, we test a range of population-wide elasticities.

3.2 Wholesale Supply

Changes in end-use demand will result in changes in the wholesale price of electricity. We assume that all energy is procured in the day-ahead market. The clearing price of energy on the day-ahead market is the result of a unit commitment problem taking customer demand and generator bids as inputs. Previous authors have modeled the wholesale market using a stylized unit commitment problem (e.g., Borenstein (2005); Alcott (2013)), a unit commitment with a more detailed description of generators in PJM (e.g., Holland and Mansur (2006)), or a parametric model (e.g., Spees and Lave (2008)). To guarantee a unique clearing price, it suffices to represent the marginal cost of supply as a non-negative, non-decreasing function of load within each hour. We model the clearing price of energy by a constant elasticity of supply $\eta \geq 0$. Price is dependent on the total load in the RTO, $Q_h(\mathbf{p}) = Q_h^{var}(\mathbf{p}) + Q_h^{fixed}$, where the two components reflect load in the zones on time-varying rates and the fixed load in the remaining zones.[5] Then the clearing price in an hour c_h can be found from the load $Q_h(\mathbf{p})$ and an hourly supply coefficient $B_h \geq 0$ as

$$c_h(\mathbf{p}) = B_h(Q_h(\mathbf{p}))^{\eta}. \tag{2}$$

This model has several advantages. First, through the coefficients B_h we can guarantee that resulting clearing prices can exactly match those seen in historical data, allowing a fair comparison against the baseline fixed rate. Second, clearing prices are guaranteed to be non-negative, which is the case for the day-ahead market. Third, price is monotonically increasing in load and, with high elasticity, mimics the "hockey-stick" shape characteristic to the wholesale electricity market. Fourth, in comparison to solving a unit commitment problem, a functional form for the clearing price is more straightforward as an input into a retail price optimization model.

We assume that the clearing price of electricity is a not a function of installed capacity. This cannot hold in an Energy Only market design: the marginal unit of generation would not enter the market if it did not decrease the marginal cost of generating electricity. It can, however, be the case with a high enough capacity price, since costs can be recovered without selling any electricity. Regardless, this assumption is likely valid for the short term, since in most cases generators will not exit the market immediately.

5. Load in remaining zones could change as a result of changes in the weighted average clearing price of electricity in the wholesale market. However, these effects are very small, generally leading to consumption differences of under 0.01 percent between rate structures.

Given a supply elasticity η, inputting historical load and clearing price into Eq. 2 determines the supply coefficient for each hour. Combining this supply equation with the end-use demand given by Eq. 1 and the fixed demand contributed by other zones, we can calculate a clearing price in the wholesale market given any retail price for electricity. With the assumed monotonically decreasing demand and monotonically increasing supply, a unique clearing price is guaranteed to exist in each hour.

3.3 Capacity Market

Most electricity markets in the U.S. require participating retailers to purchase or self-supply sufficient capacity to meet their customers' peak load. The details of this calculation vary by market. Consider the PJM market: PJM forecasts a system-wide peak three years in advance and procures capacity to cover this demand plus a reserve margin. Retailers then pay PJM a cost proportional to their share of peak demand. This proportion is determined by finding their average load in the summer coincident peaks and applying a weather normalization. The summer coincident peaks are the 5 highest-load hours in the RTO occurring from June 1 to September 30, with the further restriction that they must occur on 5 different days.

Given this calculation, there are three avenues by which a retailer could achieve lower capacity costs: 1) a reduction in load in the coincident peak hours, 2) a reduction in the clearing price for capacity, or 3) a change in the coincident peaks to hours that are more advantageous for the retailer. We focus on the first. The clearing price of capacity is assumed to be fixed. The results of Alcott (2013) project a small decrease in this cost with more customers on time-varying rates, but finds equilibrium costs substantially higher than values seen in the most recent PJM Auction. This result is consistent with the observation that recent clearing prices have been far lower than the Net Cost of New Entry calculated by PJM (PJM (2014a)). Given these opposing forces, we elect to keep this parameter constant and test a range of possible input values.

We avoid the third tactic both because it would be difficult for a retailer to implement and because advantageous hours are largely advantageous due to weather, a factor that is controlled for in the capacity obligation calculation. To prevent the retailer from taking advantage of this third possibility, when calculating capacity cost we start with the baseline capacity cost for the zones, find the new RTO coincident peaks under time-varying rates (maintaining the same reserve factor), and give the zones with time-varying rates 100 percent of the credit for the difference. That is, with $Z_{rto}(\mathbf{p})$ as the average load in the RTO coincident peaks multiplied by the reserve factor, $\bar{Z}_{rto}$ the same value in the baseline case, and $\bar{Z}_{zone}$ the capacity procured by the zones in the baseline case, the new capacity obligation is

$$Z_{zone}(\mathbf{p}) = \bar{Z}_{zone} + (Z_{rto}(\mathbf{p}) - \bar{Z}_{rto}). \tag{3}$$

The capacity obligation is a function of the entire vector of prices $\mathbf{p}$, since the coincident peak hours can shift with changing customer demand. While an explicit mathematical form for $Z_{rto}(\mathbf{p})$ that identifies the 5 coincident peaks in PJM is possible, we omit it for brevity.

3.4 Consumer and Retailer Surplus

The demand model implies a willingness to pay for each unit of electricity, from which we compute consumer surplus in units of dollars. Since the chosen model results in infinite consumer

surplus, we instead focus on the change in consumer surplus from the baseline scenario. For the entire year, this can be calculated as

$$\Delta CS(\mathbf{p}) = \sum_h \int_{p_h}^{\bar{p}} A_h p^{\varepsilon} dp = \sum_h \frac{A_h}{(1+\varepsilon)} (\bar{p}^{(1+\varepsilon)} - p_h^{(1+\varepsilon)}). \tag{4}$$

Typical estimates for ε are small and negative, so a price higher than the baseline will result in a decrease in consumer surplus in a given hour while a lower price will result in an increase.

Retailer surplus is the profit of the retailer. We have assumed that the fixed rate in the baseline case results in zero profit from electricity sales, so the change in surplus is identical to the profit under the new rates. The retailer has revenue from electricity sales and costs from purchasing energy and capacity. Accordingly, assuming an exogenous hourly fixed load Q_h^{fixed} and a capacity cost k per MW-day we can calculate retailer surplus as

$$\Delta RS(\mathbf{p}) = \sum_h A_h p_h^{\varepsilon} (p_h - B_h (Q_h^{fixed} + A_h p_h^{\varepsilon})^{\eta}) - 365 \cdot k Z_{zone}(\mathbf{p}). \tag{5}$$

In all cases, the retail price of electricity is assumed to include a constant distribution cost per MWh. Other bill items, e.g., a metering charge, are neglected since they would have no effect on the modeled electricity consumption. Such charges result in a transfer of surplus from the customer to the retailer, but do not change the total surplus.

3.5 Rate Structures

In recent years, retailers have experimented with a number of potential time-varying rate structures. We test several of these structures on historical data, allowing us to estimate the effectiveness of each structure had retail rates been optimally set for the chosen zones and input parameters. Additionally, we test a rate structure in which prices are not set optimally, but rather wholesale clearing prices are passed through to retail customers.

Of the rate structures covered in Faruqui et al. (2012), we do not show results for three. The first of these, inclining block rates, are judged by Faruqui et al. (2012) to have the smallest potential reward and do not fit directly into our model formulation. The second, Peak-time Rebate programs, are guaranteed to be less successful than CPP in our formulation. Initial tests showed that the third, Variable Peak Pricing, tracks closely with CPP. We also add one rate structure. Consistent with the broader literature, the authors use RTP to refer to a rate structure that passes through wholesale clearing prices on an hourly basis. Unaware of any term already in use, we designate the optimal hourly prices found by our model RTP$^+$.

The following rate structures, listed in order of complexity, are included in the results shown in Section 6. Several pilot programs and experiments utilizing these rate structures are given in Faruqui and Sergici (2009) and U.S. Department of Energy (2015). In addition, for all rate structures except RTP$^+$, specific examples of programs active in 2015 are listed below.

- **Fixed**: Retail customers are charged a single rate in all hours of the year. When retailer profit is zero this corresponds to the baseline case, and thus has zero change in surplus. Accordingly, we do not show results for this case.
- **Monthly**: Retail customers are charged a rate that changes on a monthly basis. This rate structure is popular among competitive energy supply companies, such as NRG Home (2015),

and has the attractive quality that no advanced metering infrastructure is required for its implementation.

- **Two-Tier Time of Use (TOU)**: Customers are charged a higher rate during peak hours, i.e., during the day on all non-holiday weekdays, and a lower rate at all other times. See, for example, the residential program of We Energies (2015) or National Grid (2015). Unless otherwise noted, results for this rate use the same prices and time windows for the entire year.
- **Three-Tier Time of Use (TOU)**: On top of two-tier TOU prices, a higher rate is charged during a shorter peak on summer afternoons. This strategy is employed by, e.g., Pacific Gas & Electric (2015) and Southern California Edison (2015) for their commercial and industrial rates. More complicated three-tier rate structures can vary by season, changing time windows or extending a higher rate to winter mornings and evenings (e.g., the programs of Baltimore Gas & Electric (2015) or Ontario Hydro (2015)). In the results shown, the lower two tiers utilize the same prices and time windows year round; initial tests resulted in only mild improvements when these were allowed to change.
- **Critical Peak Pricing (CPP)**: On top of two-tier TOU prices, the retailer may call critical peak events, each lasting four hours with a shared higher price, on a limited number of days over the course of the summer. See, for example, the Peak Day Pricing of Pacific Gas & Electric (2015).
- **Real Time Pricing (RTP)**: The customer sees the clearing price in the wholesale market. This corresponds, e.g., to the ComEd Hourly Pricing Program (ComEd, 2015). This rate assumes that capacity costs are recovered through fixed charges that do not provoke any response from customers.
- **Real Time Pricing with Optimal Pass-Through (RTP$^+$)**: The retailer chooses a price that changes hourly. While we are not aware of any retailer employing this strategy, it would roughly correspond to passing through the wholesale price with an Energy Only market design (e.g., a retailer participating in ERCOT). In contrast to RTP, capacity costs are recovered by adding to retail prices in coincident peak hours.

4. RETAIL PRICE OPTIMIZATION

We model the problem faced by a retailer serving a subset of customers within a competitive market. The retailer chooses a vector of prices $\mathbf{p}$, which determines customer demand in those zones for each hour of the year, leading to new hourly clearing prices in the wholesale market and a new capacity obligation for the retailer. The model finds the optimal price vector given demand and supply parameters inferred from historical data. Accordingly, the optimization is backwards-looking, using observed load and wholesale prices to determine optimal retail prices. A forward-looking model would instead rely on forecasts of these quantities, with the effectiveness of the resultant solution depending on the quality of the forecast. Developing these forecasts and setting forward-looking or real-time rates represents a promising avenue for future research.

4.1 Optimization Model

The objective of the retailer is to maximize total surplus for itself and its customers. Notably absent from this calculation is the surplus of the generators. Retailers may have a long-term interest in generator profitability in order to maintain a competitive market, and would certainly incorporate this profit in situations where they themselves own generation assets. These consider-

ations are set aside, with the model instead focusing on the short-term priorities of end-use customers of electricity. The primary implication of leaving out generator surplus is that reductions in capacity can be realized as a short-term benefit.

There are three types of constraints. First, we assume zero profit for the retailer, consistent with a competitive market. This is represented by retailer surplus being less than or equal to zero. In a solutions with negative profit, the retailer could add fixed charges to customer bills without affecting customer behavior.[6] Second, the vector of prices must fit a pre-defined form depending on the rate structure under examination. At one extreme, the price is constrained to be equal in all hours; this corresponds to the baseline, fixed-rate case. At the other extreme, there is no constraint apart from non-negativity; this corresponds to prices that change hourly. The third set of constraints arises from the capacity obligation. We introduce the dummy variable z to take the place of $Z_{rto}(\mathbf{p})$ in the retailer surplus calculation. Applying the PJM rule for calculating capacity obligation, we include the constraints that z must be greater than or equal to the average load in any eligible combination of 5 coincident peak hours times a multiplier r for the reserve margin.

Assume that the chosen rate structure partitions the set of hours H into T subsets $H_1,\ldots,H_T$, within which the retail price must be equal. For example, a two-tier TOU structure might separate H into two subsets, the first including all hours from 9 AM to 9 PM on non-holiday weekdays and the second including nights, holidays, and weekends. Further, let CP indicate the set of all possible sets of the 5 coincident peak hours. Then we can write the optimization model as

$$\underset{\mathbf{p},z}{\text{maximize}} \quad \Delta CS(\mathbf{p}) + \Delta RS(\mathbf{p},z) \tag{6}$$

$$\text{subject to} \quad \Delta RS(\mathbf{p},z) \leq 0 \tag{7}$$

$$p_{\tilde{h}} = p_h \qquad \forall t,\tilde{h} \in H_t, h \in H_t \tag{8}$$

$$z \geq \frac{r}{5} \sum_{h \in C} Q_h(\mathbf{p}) \qquad \forall C \in CP. \tag{9}$$

Different market rules would result in a different form for Eq. 9. While this complication does not significantly impact the results for most rate structures, it does have implications for the optimal number of CPP events.

It is clear that for optimality at least one of the capacity constraints must be active when $k>0$; otherwise, a superior solution could be constructed by reducing z. Under the loose assumption that demand and cost fall with an increase in price, it is also the case that the constraint governing retailer profit, i.e., Eq. 7, is active. The following Proposition, with proof given in the Appendix, presents this result.

Proposition. *If* $\varepsilon<0$ *and for every h,* $\dfrac{\partial c_h}{\partial p_h}<0$*, then every optimal solution to Eqs. 6–9 satisfies* $\Delta RS(\mathbf{p},z)=0$.

6. Most importantly, this applies to RTP, which has negative profit whenever capacity costs are positive

4.2 Computation

We solve the model using Ipopt, described in Wächter and Biegler (2006). Numerical experiments confirm that the chosen representation results in a concave objective function for a wide range of input assumptions, guaranteeing that the model identifies a unique global maximizer. To identify the optimal start and end times for TOU pricing, we test all combinations of start times between 6 AM and 3 PM and end times between 6 PM and 11 PM, solving the model for each. For CPP, we identify 4 hour periods surrounding the highest load hours in the data. We then test the same combinations of start and end times as well as any number of CPP events between 0 and 15.

A particular complexity in the model is the combinatorial nature of the set of all possible sets of coincident peak hours. Since each set can be composed of any hour chosen from the 122 days of summer, with at most one from any given day, there are $\binom{122}{5} \cdot 24^5$ possible sets of hours. We note, however, that since the constraints take identical form it is simple to identify which if any is violated by a potential solution. Thus the strategy we employ is to start with only one constraint, corresponding to the historical coincident peak hours. Given the optimal prices under this relaxed model, we find the new coincident peak hours, check if the associated constraint is violated, and if so add it to the model. Iterating this procedure results in a solution that is feasible and optimal for the original problem without generating all of the possible constraints.

Source code and data for the model and numerical results, as well as for several additional tests performed are in the online Appendix (http://sites.northwestern.edu/jacobmays/research).

5. CALIBRATION TO PJM

We model the short-term problem faced by a retailer participating in PJM's wholesale power markets. PJM manages markets for energy, capacity, transmission, and several ancillary services. We focus on the two biggest sources of cost to retailers, which result from the energy and capacity markets. From 2010 to 2014, these two markets represented an average of 76 percent of the billings administered by PJM PJM (2013, 2015b). PJM is split into 20 zones; the wholesale energy price is shared by all zones, but transmission and capacity costs can vary between them. As discussed below, in our model we neglect transmission and assume a single capacity cost, so wholesale prices are identical for all zones.

Transmission-related cost, representing an average of 18 percent of PJM billings over the same period, comprises several different components.[7] Some of these costs are related to local (i.e., non-coincident) peak load; the model could be extended to allocate these costs into the optimal hours in a manner analogous to capacity costs without sacrificing concavity. However, the effect of end-use demand on a majority of transmission costs is likely to be small in the short term. The largest element of transmission cost in PJM, network integration transmission service, is subject to an annual revenue requirement. Efforts to avoid this cost could shift responsibility to another retailer or customer class within a zone, but not eliminate it. Congestion and loss charges are available on an hourly basis, but are only loosely correlated with load at the zone or RTO level. This suggests that the mechanism of time-varying rates applied at the zone level may be too blunt to reduce these costs. While there has been little appetite thus far to include a finer-grained locational component

7. In order of cost, this estimate includes Network Transmission Service, Congestion, Loss, FTR Auction Revenues, and Transmission Enhancement.

in retail prices, this possibility is discussed in Glick et al. (2014). Furthermore, congestion costs are often hedged with Financial Transmission Rights (FTRs), dulling the short-term financial impact of a reduction in congestion. Accordingly, transmission costs are not included in the model.

Distribution cost can vary substantially by geography and customer class. We assume a fixed cost of $15/MWh. As with transmission, some distribution costs can scale with non-coincident peak load, and could accordingly be incorporated into a time-varying rate. Including these costs would likely increase the projected benefits of time-varying rates. A higher per-MWh distribution cost, on the other hand, leads to weaker customer response in our demand model and lower projected benefits from time-varying rates. The same holds for inclusion of the remaining 6 percent of PJM costs, the majority of which come from ancillary services that are charged on a per-MWh basis. Since this affects all time-varying rates, however, increasing the assumed distribution cost has little impact on the relative performance of the rate structures.

Data on historical loads by zone and energy prices in the day-ahead market are drawn from PJM data covering the one year period from February 2014 through January 2015, with the period chosen to avoid the anomalous high energy costs seen in January 2014. Capacity costs ranged from $120 to $215/MW-day in PJM's 2017/2018 Reliability Pricing Model Base Residual Auction; the modal result, $120/MW-day, is used as the default value. Weather normalized capacity obligations for the baseline case are found in PJM (2014b). The reserve margin is calculated as the reported capacity obligations divided by the observed coincident peak demand in Summer 2014.

Straightforward application of the clearing price of capacity misses many possible complications. Capacity cost can vary significantly by region and by year. In addition to the previously mentioned variation within PJM, other geographies may have even lower cost: for example, an area with excess capacity and no consumption growth may be unconstrained and therefore effectively have no capacity cost until generators are retired. We assume that the retailer owns no generation assets; any potential loss in income from the capacity market could be interpreted as a reduction in effective capacity cost. One difficulty in this regard is the participation of demand resources in the capacity market: time-varying rates could lower the baseline for these resources, reducing their potential to bid. Lastly, no attempt is made to discount based on the timing of capacity-related cash flows. Despite these complications, any individual retailer should be able to identify its own capacity cost for use in the model, justifying the assumption of a single fixed price. Given the wide range of possible values, we test a range of capacity cost inputs from $0 to $250/MW-day. The effect of k is discussed in Section 6.4.

A major determinant of the effectiveness of time-varying prices is the demand elasticity. As discussed in Section 2, estimates of this elasticity vary. Because the data we use includes a certain segment of demand already on time-varying rates, we opt for a default elasticity of -0.05, at the low end of published estimates. We test a range from -0.01 to -0.25 to assess the impact of customer responsiveness. In general, the effect of stronger elasticity is similar to that of a greater share of customers moving to time-varying rates. Since there is a much wider range of plausible customer shares (anywhere from 0 to 100 percent) than of elasticities, we focus on the first when showing results in Section 6.5.

The supply cost curve can change over the course of the year due to several factors, such as generator maintenance taking units offline or changes in fuel cost. These underlying changes are small within a given day, so we estimate the supply elasticity η by fitting a version of the model in Eq. 2 in which the coefficients B are fixed for each day. This strategy attempts to isolate the effect of load on price. This model has $R^2 = 0.9065$ and results in a supply elasticity of $\eta =$ 2.7424. Hourly supply coefficients are calculated by inserting this value for η into Eq. 2 with

Table 1: Comparative Performance of Time-Varying Rate Structures

Rate Structure	Max Price ($/MWh)	Change in Capacity (%)	Capacity Savings ($M)	Change in Surplus ($M)	Relative Performance (%)
Monthly	78	−1.1	9	12	8
Two-tier TOU	129	−3.7	31	21	13
Three-tier TOU	248	−6.6	56	42	26
CPP	1,741	−15.3	128	103	65
RTP	412	−2.2	18	40	25
RTP$^+$	4,890	−19.3	162	160	100

Notes: Dominion zone with default assumptions for ε, η, and k.

historical loads and clearing prices. As with other inputs, we test a range of input values (from 2.0 to 4.5) to estimate the impact of the supply curve. Changes in the supply elasticity have a very weak impact on both the absolute and relative performance of the rate structures, and accordingly no results are shown below. These tests suggest that our simplified model of supply cost is adequate.

6. NUMERICAL RESULTS

In this section we compare the performance of the above rate structures, computing optimal solutions of the model described in Section 4.1 for each. To ensure broad validity of the numerical results, we test each of the rate structures on each zone individually, as well as select larger combinations of zones. Unless otherwise stated, results below are for the Dominion zone, chosen for its balance between energy and capacity costs.[8] Comparative performance of the various rate structures is broadly similar for each individual zone, with Dominion representing close to average results.

As previously described, the results for RTP assume that customers will be unable to predict coincident peaks and respond to a demand charge, and accordingly represent a low estimate. The results for the other rate structures assume that retailers are able to achieve the true optimum found by the backwards-looking model, and accordingly represents a high estimate. Nevertheless, the large differences in performance do highlight the need for retailers to carefully consider their choice of rate structure. Furthermore, tests across a wide range of input values show that several high-level insights are likely to hold in different geographies and market contexts.

6.1 Rate Structure Comparison

Table 1 shows the key results for each rate structure for the Dominion zone using the default assumptions for demand elasticity, supply elasticity, and capacity cost described above. Since RTP$^+$ relaxes constraints present for the other rate structures, it is guaranteed to achieve the highest objective function value. Here and in subsequent results, relative performance is thus calculated as the change in surplus for the given rate structure divided by the change in surplus for RTP$^+$. The maximum price reflects the highest retail price in any hour of the year excluding the fixed distribution cost, and change in capacity is given in relation to the baseline, fixed-rate scenario.

8. For the period under study, this zone was responsible for 12.25 percent of energy and 12.23 percent of capacity payments included in the model. Since the PJM region includes around 61M people, we can infer that the zone includes approximately 7.5M people.

Several features are worth highlighting. With a total surplus of $40M, passing through wholesale clearing prices (RTP) potentially misses a full 3/4 of the $160M in benefits possible using optimal hourly prices (RTP$^+$). Since the results for RTP assume no customer response to separate demand charges, the performance gap between these two structures is overstated. However, even with generous assumptions on customers' ability to predict the timing of coincident peak events a substantial gap is likely to remain.

Benefits of time-varying rates come from two sources: reducing deadweight loss in each hour of the year and reducing capacity costs for the entire year. Comparing the capacity savings to the change in surplus shows that for most of the rate structures, the main driver of surplus is reduction in capacity costs. For two-tier TOU, for example, capacity charges fall by $31M while total surplus improves by only $21M, implying that the optimal prices under this structure are actually a worse approximation of wholesale clearing prices than the baseline fixed rate. This observation is in stark opposition to the priority given to deadweight loss in previous evaluations of time-of-use rates. It is well-aligned, however, with the results of Newell et al. (2009), which found in its base case that capacity savings amounted to $153.6M out of a total change in consumer surplus of $162.2M. As explained in Hogan (2014), deadweight loss is more important when energy prices have higher variance; as described in Alcott (2013), capacity markets have the effect of reducing volatility in energy prices. Taken together, these observations can help explain the primacy of capacity and the poor relative performance of RTP.

As expected, increasing complexity in the rate structure leads to greater change in surplus. Moving from Two- to Three-tier TOU doubles the benefit from $21M to $42M. The addition of a different sort of complexity, allowing the prices and time windows in Two-tier TOU to change between the summer and non-summer months, can lead to similar results. In this case, the optimal time window and price for the summer peak nearly matches the highest tier in Three-tier TOU. Accordingly, this "Two-tier TOU with Seasonality" option achieves nearly the same capacity reductions and overall benefits as Three-tier TOU. Due to the similarity of the results, they are not shown separately. The increase in complexity from Three-tier TOU to CPP leads to an even larger jump in benefits, from $42M to $103M. Whereas Three-tier TOU results in a moderately higher price ($248/MWh) on all summer afternoons, CPP gives the retailer the ability to charge significantly higher prices ($1,741/MWh) on the most important afternoons of the summer.[9] The outcomes show just how valuable this flexibility is.

6.2 Start and End Timing

A central decision in the design of TOU and CPP rates is the timing of the price tiers. Figure 1 shows the effect that these decisions have in the Two-tier TOU structure.[10] Most utilities experimenting with this rate structure have opted for peak periods that last most of the day. The model results, however, suggest that a short peak period in the afternoon from 3 PM to 6 PM is nearly twice as effective as a period that lasts the whole day (e.g., 9 AM to 9 PM). This observation matches the primacy of capacity costs seen in the previous section; a shorter period allows the retailer to concentrate higher prices on the hours in which peak demand occurs.

9. Dominion uses the higher CPP price on 6 occasions in the optimal solution.

10. The optimal Three-tier TOU configuration for the Dominion zone retains the 3 PM-6 PM peak as the highest tier in the summer, with the second tier lasting from 6 AM to 10 PM year-round.

Figure 1: Two-tier TOU Surplus by Start and End Time ($M)

		End					
		6 PM	**7 PM**	**8 PM**	**9 PM**	**10 PM**	**11 PM**
	6 AM	11.0	11.4	12.0	12.4	12.3	11.4
	7 AM	11.1	11.4	11.9	12.2	12.0	11.0
	8 AM	10.5	10.7	11.1	11.3	11.0	10.0
	9 AM	10.7	10.9	11.1	11.2	10.8	9.8
	10 AM	11.3	11.3	11.4	11.4	10.9	9.8
Start	**11 AM**	12.0	11.8	11.8	11.7	11.1	9.9
	12 PM	13.1	12.7	12.5	12.2	11.4	10.1
	1 PM	14.9	14.1	13.6	13.1	12.1	10.6
	2 PM	17.4	15.9	15.0	14.2	13.0	11.3
	3 PM	21.1	18.4	16.8	15.6	14.0	12.0
	4 PM	18.9	15.3	13.4	12.1	10.6	8.7

Notes: Dominion zone with default assumptions for ε, η, and k.

Optimal start and end times depend on demand characteristics. With larger populations on time-varying rates and/or stronger demand elasticity, a longer peak period becomes optimal. Intuitively, increased responsiveness of demand means that a wider range of hours have the potential to be in the set of 5 coincident peak hours. A short window with high prices could bring down demand substantially, leading to coincident peak hours that occur in the off-peak period.

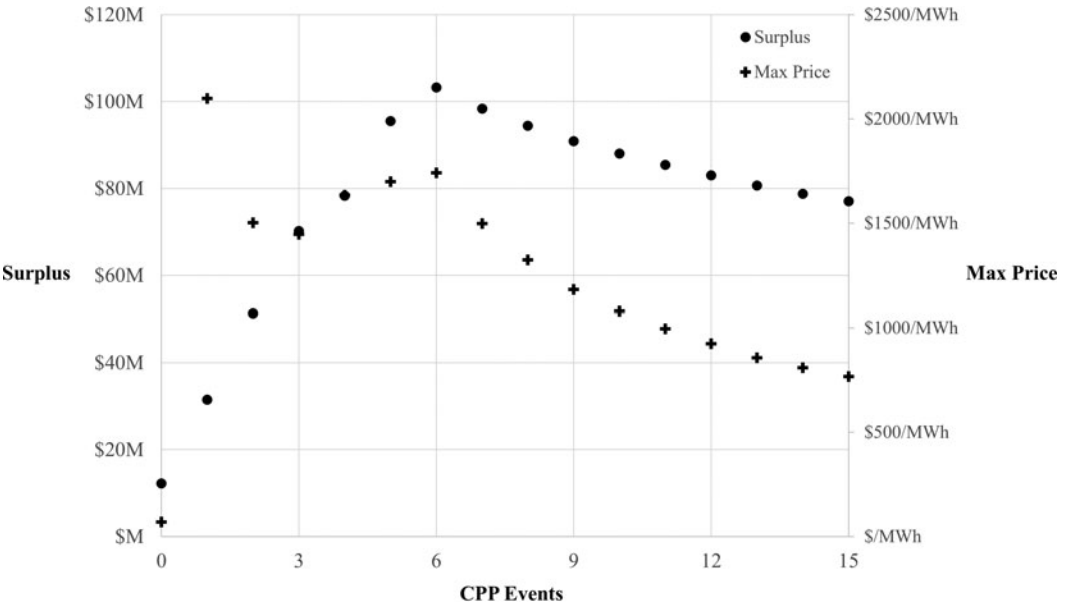
6.3 CPP Events

While the increased flexibility of CPP brings large benefits, retailers need to exercise caution in determining how often to call critical events. In currently active programs, utilities generally have the option to call CPP events approximately a dozen times throughout a summer. Figure 2 exhibits the importance of choosing the appropriate number of CPP events.[11] For the Dominion zone, the first 6 CPP days add a substantial amount to total surplus, reaching the globally optimal value of $103M previously seen in Table 1. However, adding subsequent CPP days detracts from surplus. These additional, lower-priority days militate toward a lower peak price, weakening the power of the CPP events to bring down peak demand.

The result of 6 days is specific to the particulars of PJM's calculation for capacity obligation as well as the weather seen in the summer under study. Even in the same geography, the optimal number of CPP events has the potential to change year to year. Additionally, as with TOU start and end times, the optimal number of CPP events is higher with more zones on time-varying rates or stronger demand elasticity. Accordingly, the important result from this analysis is not the specific number of days but instead the observation that results are quite sensitive to choosing the right number of events. Understanding this sensitivity and its implications for calling CPP events in real time is an area for future research.

11. This figure holds the timing of the two base pricing tiers constant, with the higher tier lasting from 6 AM to 10 PM on non-holiday weekdays.

Figure 2: CPP Surplus by Number of Events



Notes: Dominion zone with default assumptions for ε, η, and k.

6.4 Capacity Cost

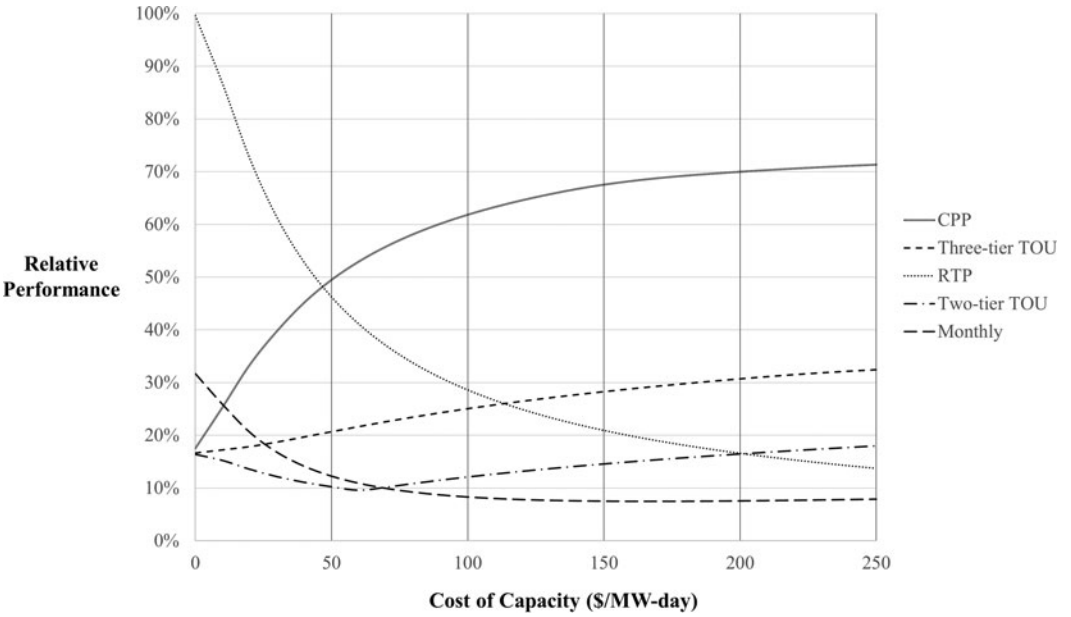
Reductions in capacity obligation is the primary source of benefits from time-varying rate structures, but capacity costs can vary significantly by geography and year. Accordingly, it is important to understand both the absolute and relative performance of the rate structures at different capacity costs. We test this by finding the optimal solution for each rate structure given values of k from \$0 to \$250/MW-day. Higher values of k lead both to greater overall costs and an increase in the proportion of total costs coming from capacity. In absolute terms, as might be anticipated, the potential benefits from every rate structure grow with increasing capacity prices; in general, this growth is weakly superlinear.[12]

Figure 3 shows the relative performance of the rate structures at varying levels of capacity cost. We draw attention first to the results when there is no capacity cost, since this most closely corresponds to the previous literature. The relative performance of Two-tier TOU in this case is 16.4 percent, in line with previous estimates with PJM data that range from 15-23 percent (see Hogan (2014); Holland and Mansur (2006); Spees and Lave (2008)). The Monthly rate is nearly twice as effective, at 31.7 percent, echoing the findings of Holland and Mansur (2006).

The largest changes in effectiveness arise for CPP and RTP. This exhibits the fundamental difference between the two approaches. Conceptually, the optimal solution in CPP prioritizes the reduction of capacity obligations. Capacity reductions in the RTP structure, on the other hand, are merely a side benefit of passing through the higher wholesale clearing prices seen in hours of high

12. For example, for the Dominion zone an increase from $k = \$120/\text{MW}-\text{day}$ to $k = \$130/\text{MW}-\text{day}$ results in an increase of \$13.5M in the potential benefit from RTP$^+$, while an increase from $k = \$240/\text{MW}-\text{day}$ to $k = \$250/\text{MW}-\text{day}$ results in an increase of \$15.2M.

Figure 3: Performance of Rate Structures by Capacity Cost



Notes: Dominion zone with default assumptions for ε and η.

demand. In the range of capacity costs currently experienced across PJM, from $120 to $215/MW-day, CPP exceeds the efficiency of RTP by a substantial margin. At sufficient levels of k, both Two- and Three-tier TOU also surpass RTP in effectiveness.

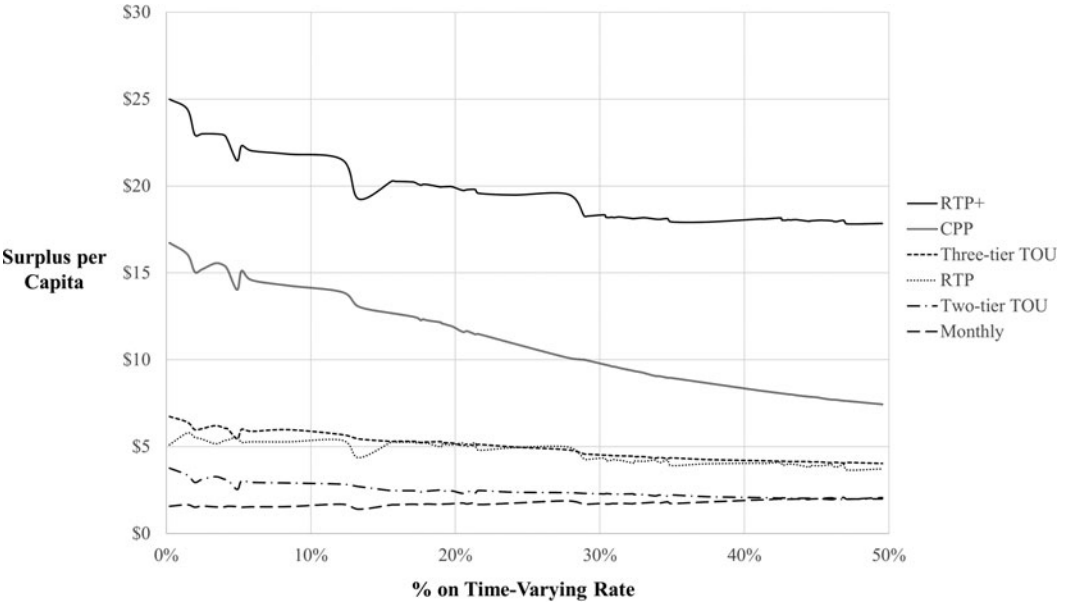
A final interesting feature in Figure 3 is the non-monotonicity of Two-tier TOU. Despite the relative decline, benefits are monotonically increasing in absolute terms. The shift reflects a change in optimal start and end times: at low capacity costs, the optimal peak period lasts from 6 AM to 10 PM, but at higher k a peak from 3 PM to 6 PM is preferable. Relative benefits are decreasing in k for the former window and increasing in k for the latter. This conforms with earlier observations, since the longer window is better at matching clearing prices while the shorter window more effectively reduces capacity obligation.

6.5 Share of Customers on Time-Varying Rates

While time-varying rates are expected to become more common, the pace at which they will be adopted is uncertain. Experts surveyed in Faruqui and Mitarotonda (2011) project that between 7.5 to 20 percent of residential and 10 to 30 percent of commercial and industrial customers will be on dynamic pricing by 2020.[13] Accordingly, it is important to understand how the potential benefits of each of the rate structures could change with larger numbers of customers moving to time-varying rates. Importantly, both the relative performance of the rate structures and the maximum prices in the optimal solutions can see significant changes with increasing uptake.

13. These estimates exclude TOU rates, which are time-varying but not dynamic.

Figure 4: Surplus per Capita by Share of Customers



Notes: Default assumptions for ε, η, and k.

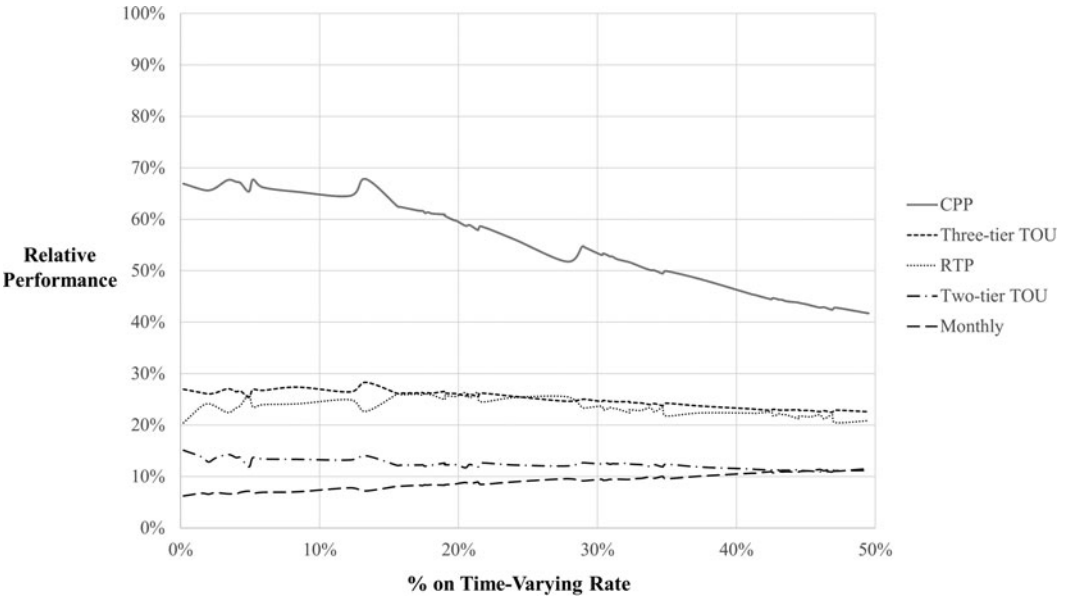
To test the effect of an increasing share of customers moving to time-varying rates, we solve for the model for each rate structure for 74 configurations of zones. At smaller percentages, this represents single zones switching to time-varying rates; for larger percentages, we model combinations of zones jointly optimizing and setting identical prices. All 20 zones were modeled individually, and a further 54 combinations of zones were chosen to achieve coverage over the range from 0.2 to 49.5 percent. The percentage of demand on time-varying rates is estimated as the average of two percentages: the percentage of total energy consumption and the percentage of total capacity obligation in the switching zones.

Figure 4 shows the decline in per capita benefits for all rate structures except Monthly as time-varying rates spread to a larger share of total customers. Surplus per capita is estimated by dividing total surplus by the approximate number of people switching to time-varying rates based on a PJM-wide population of 61M people. The results are not completely smooth, reflecting quirks of usage in the zones chosen for the time period under study.[14] For example, the AEP zone was responsible for 14.6 percent of capacity and 16.6 percent of energy costs in the period under study, while the ComEd zone was responsible for 14.1 percent of capacity and only 12.4 percent of energy. As demonstrated by Section 6.4, the relative importance of the two cost components can produce changes in the relative effectiveness of the rate structures. Despite these sometimes large differences in demand characteristics, the rank order of the rate structures is stable over the range examined.[15] As consolation for these declining marginal benefits, maximum prices fall with increasing penetration for all rate structures except Monthly. Whereas the maximum prices for CPP and RTP$^+$ were

14. For clarity, results for zones within 0.2% share of each other have been averaged for inclusion in Figures 4 and 5.

15. Two exceptions to this stability are that RTP outperforms Three-tier TOU in 7 out of the 74 cases modeled, and Monthly starts to outperform Two-tier TOU at the high end of the range.

Figure 5: Performance of Rate Structures by Share of Customers



Notes: Default assumptions for ε, η, and k.

1,741 and 4,890 $/MWh in the case of the Dominion zone, they fall to 269 and 1,688 $/MWh for the largest combination of zones tested. By comparison, many currently active CPP programs have a peak price halfway between these two cases (e.g., $849/MWh for Pacific Gas & Electric (2015)). These reductions are particularly important because a key objection to time-varying rates is increased bill volatility for customers: lower peak prices may lead to greater customer acceptance of these rate structures.

While the rank order is relatively stable, the relative performance of each of the time-varying rates is not. The magnitude of this effect is shown in Figure 5. The performance of all rate structures except Monthly falls relative to RTP$^+$ with increasing coverage. The drop is particularly significant for CPP: whereas the relative performance of CPP was 65 percent for the Dominion zone, covering 12.2 percent of the zone, it falls to around 42 percent of RTP$^+$ when half of customers move to time-varying rates. Intuitively, an increased level of customer response smooths hour-to-hour variations in demand. This compression is particularly strong in the highest demand hours of the year. Accordingly, replacing one of the coincident peak hours with the next highest in the capacity obligation calculation has a smaller effect. Two-tier TOU, Three-tier TOU, and RTP see drops in relative performance of 2, 3, and 4 percent over the same range, whereas Monthly improves by 4 percent.

7. CONCLUSION

Time-varying retail electricity rates have been gaining in popularity, resulting in a wide array of rate structures and prices currently in use. To date, however, efforts to compare the effectiveness of these rate structures have been limited. The above results show that the choice of rate structure can have a significant impact on welfare: moving from a two-tier to a three-tier TOU scheme, for example, could double the resulting benefits. Within a certain rate structure, making

optimal choices has an equally significant impact: identifying the right TOU windows, choosing the right number of CPP days, and setting the right hourly prices all strongly impact the effectiveness of these rates.

Several opportunities to enhance the realism of the model are readily apparent. We rely on a simple estimate of own-price elasticity; a more accurate value, potentially one that changes by hour, day, or season, would be straightforward to incorporate. Replacing our stylized model of the supply curve with a more detailed bid curve would be similarly straightforward. While we tested several of the most popular time-varying rate structures, the possible configurations are endless. Other cost components could be incorporated into the optimization. Some transmission and distribution costs could be recovered through a time-varying rate, promoting greater customer response and further increasing benefits. Models of customer response are sure to be refined as experience with time-varying rates grows, allowing a more precise demand model. Lastly, more work is required to produce a forward-looking model that can capture the benefits of our backward-looking model.

This paper takes the position that passing through capacity costs optimally will provide the best possible signal to end-use consumers and produce an optimal response. Other mechanisms to curtail capacity are available, including peak-time rebates, direct load control, and demand charges. The best strategy might include a combination of these. For example, demand charges may be effective for larger industrial customers, but residential users are unlikely to be able to predict the timing of coincident peaks. Understanding the comparative effectiveness of these and other strategies is of central importance to future decision making. The wide gap in effectiveness between RTP and RTP$^+$ shows what is at stake in designing these programs.

The above numerical results are not the final word, but do allow us to draw several conclusions. Many of the outcomes hold under a wide range of input assumptions and across many geographical areas, leading to several clear implications for retailers and regulators:

- Most of the benefits of time-varying rates come in the form of reduced capacity requirements.
- In markets with an ICAP design, retailers can benefit from capacity reductions in the short term and should prioritize this goal when designing time-varying rates.
- Passing through wholesale clearing prices (RTP) is an ineffective way to reduce capacity requirements in the ICAP setting, missing most of the potential benefits of optimal hourly pricing (RTP$^+$).
- Among rate structures currently in use, CPP is a much more promising route than RTP for approximating the efficiencies of RTP$^+$.

APPENDIX

Proof of Proposition. Starting with the optimization problem in Eqs. 6–9, we introduce vector $\mathbf{y}$, with y_t representing the retail rate applied to all hours in subset H_t. Making this substitution means that any solution automatically satisfies the constraints in Eq. 8. For simplicity, let $c_h(y_t)$ indicate the supply cost when using price y_t in hour h in Eq. 2. Then, expressing consumer surplus from Eq. 4 and retailer surplus from Eq. 5 as sums over all price tiers, we can rewrite the problem as

$$\underset{\mathbf{y},z}{\text{maximize}} \qquad \sum_{t=1}^{T} \sum_{h \in H_t} \frac{A_h}{(1+\varepsilon)} (\bar{p}^{(1+\varepsilon)} - y_t^{(1+\varepsilon)}) + \sum_{t=1}^{T} \sum_{h \in H_t} A_h y_t^{\varepsilon} (y_t - c_h(y_t)) - 365 \cdot k \cdot z$$

$$\text{subject to} \quad \sum_{t=1}^{T} \sum_{h \in H_t} A_h y_t^{\varepsilon}(y_t - c_h(y_t)) - 365 \cdot k \cdot z$$

$$\leq 0 \sum_{t=1}^{T} \sum_{h \in C} 1_{\{h \in H_t\}} \cdot \frac{r}{5} \cdot A_h y_t^{\varepsilon} - z \leq 0 \qquad \forall C \in CP.$$

Let us assume there exists an optimal solution at $\hat{\mathbf{y}}, \hat{z}$ for which the retailer profit is strictly negative, i.e., the profit constraint is not active. We introduce KKT multipliers $\mu \geq 0$ for the profit constraint and $\lambda_C \geq 0$ for the capacity constraints. Differentiating with respect to z, stationarity requires that

$$\sum_{C \in CP} \lambda_C = 365 \cdot k \cdot (1 - \mu). \tag{10}$$

Since we have assumed that the profit constraint is not active, complementary slackness implies that $\mu = 0$. Differentiating with respect to any y_t, stationarity requires that

$$\sum_{h \in H_t} \left[\varepsilon A_h \hat{y}_t^{\varepsilon} - \varepsilon A_h \hat{y}_t^{\varepsilon - 1} c_h(y_t) - A_h \hat{y}_t^{\varepsilon} \cdot \frac{\partial c_h}{\partial y_t} \Big|_{y_t = \hat{y}_t} \right] = \sum_{C \in CP} \sum_{h \in C} 1_{\{h \in H_t\}} \cdot \lambda_C \cdot \frac{r}{5} \cdot \varepsilon A_h \hat{y}_t^{\varepsilon - 1}. \tag{11}$$

Multiplying both sides of this equation by $\hat{y}_t / \varepsilon$ and rearranging terms leads to

$$\sum_{h \in H_t} \left[A_h \hat{y}_t^{\varepsilon} (\hat{y}_t - c_h(y_t)) - \frac{\hat{y}_t}{\varepsilon} \cdot A_h \hat{y}_t^{\varepsilon} \cdot \frac{\partial c_h}{\partial y_t} \Big|_{y_t = \hat{y}_t} \right] = \sum_{C \in CP} \sum_{h \in C} 1_{\{h \in H_t\}} \cdot \lambda_C \cdot \frac{r}{5} \cdot A_h \hat{y}_t^{\varepsilon}. \tag{12}$$

Summing the equations over all T price tiers and focusing on the right-hand side, we note that the multiplier λ_C associated with any inactive capacity constraint equals 0. We can then substitute z for the average demand in all the active constraints multiplied by the reserve factor r and utilize Eq. 10, leading to

$$\sum_{t=1}^{T} \sum_{h \in H_t} \left[A_h \hat{y}_t^{\varepsilon} (\hat{y}_t - c_h(y_t)) - \frac{\hat{y}_t}{\varepsilon} \cdot A_h \hat{y}_t^{\varepsilon} \cdot \frac{\partial c_h}{\partial y_t} \Big|_{y_t = \hat{y}_t} \right] = 365 \cdot k \cdot z. \tag{13}$$

Rearranging,

$$\sum_{t=1}^{T} \sum_{h \in H_t} A_h \hat{y}_t^{\varepsilon} (\hat{y}_t - c_h(y_t)) - 365 \cdot k \cdot z = \sum_{t=1}^{T} \sum_{h \in H_t} \frac{\hat{y}_t}{\varepsilon} \cdot A_h \hat{y}_t^{\varepsilon} \cdot \frac{\partial c_h}{\partial y_t} \Big|_{y_t = \hat{y}_t}. \tag{14}$$

The left-hand side of Eq. 14 precisely equals retailer profit. We note than the partial derivative on the right-hand side is negative by assumption. Since $\varepsilon < 0$, this implies that the right-hand side of Eq. 14 is positive, which is a contradiction. We can therefore conclude that the constraint governing retail profit is active at any optimal solution.

ACKNOWLEDGMENTS

We would like to thank Andreas Wächter and audience participants at the CORS/INFORMS 2015 Joint International Meeting for helpful discussions, as well as three referees for valuable comments.

REFERENCES

Alcott, H. (2013). Real-time pricing and electricity market design. Working Paper.

Ata, B., A. S. Duran, and O. Islegen (2015). An analysis of time-based pricing in electricity supply chains. Working Paper.

Baltimore Gas & Electric (2015). Time of use pricing. Available at https://www.bge.com/waystosave/manageyourusage/Pages/Time-of-Use-Pricing.aspx.

Borenstein, S. (2005). The long-run efficiency of real-time electricity pricing. *The Energy Journal* 26(3): 93–116. http://dx.doi.org/10.5547/ISSN0195-6574-EJ-Vol26-No3-5.

Borenstein, S. and S. Holland (2005). On the efficiency of competitive electricity markets with time-invariant retail prices. *RAND Journal of Economics* 36(3): 469–493.

Bowring, J. (2013). Capacity markets in PJM. *Economics of Energy & Environmental Policy* 2(2). http://dx.doi.org/10.5547/2160-5890.2.2.3.

Cappers, P., A. Todd, M. Perry, B. Neenan, and R. Boisvert (2013). Quantifying the impacts of time-based rates, enabling technology, and other treatments in consumer behavior studies: Protocols and guidelines. Available at https://www.smartgrid.gov/files/LBNL EPRI AnalysisProtocols FINAL-20130716.pdf.

ComEd (2015). Hourly pricing program. Available at https://hourlypricing.comed.com/.

Cramton, P. and S. Stoft (2006). The convergence of market designs for adequate generating capacity. Report for California Electricity Oversight Board.

Crew, M. A., C. S. Fernando, and P. R. Kleindorfer (1995). The theory of peak-load pricing: A survey. *Journal of Regulatory Economics* 8(3): 215–248. http://dx.doi.org/10.1007/BF01070807.

Faruqui, A. (2015). A global perspective on time-varying rates. Available at http://www.brattle.com/system/publications/pdfs/000/005/183/original/A_global_perspective_on_time-varying_rates_Faruqui_061915.pdf?1436207012.

Faruqui, A., R. Hledik, and J. Palmer (2012). Time-varying and dynamic rate design. Available at http://www.raponline.org/document/download/id/5131.

Faruqui, A. and D. Mitarotonda (2011). Energy efficiency and demand response in 2020-a survey of expert opinion. Available at http://www.brattle.com/system/publications/pdfs/000/004/697/original/Energy_Efficiency_and_Demand_Response_in_2020_Faruqui_Mitarotonda_Nov_2011.pdf.

Faruqui, A. and S. Sergici (2009). Household response to dynamic pricing of electricity-a survey of the experimental evidence. Available at http://www.hks.harvard.edu/hepg/Papers/2009/The

Glick, D., M. Lehrman, and O. Smith (2014). Rate design for the distribution edge. Available at http://www.rmi.org/elab_rate_design.

Hogan, W. W. (2014). Time-of-use rates and real-time prices. Working Paper.

Holland, S. and E. Mansur (2006). The short-run effects of time-varying prices in competitive electricity markets. *The Energy Journal* 27(4): 127–155. http://dx.doi.org/10.5547/ISSN0195-6574-EJ-Vol27-No4-6.

Joskow, P. (2008). Capacity payments in imperfect electricity markets: Need and design. *Utilities Policy* 16(3): 159–170. http://dx.doi.org/10.1016/j.jup.2007.10.003.

Massachusetts Department of Public Utilities (2014, June). Anticipated policy framework for time varying rates. Available at http://www.mass.gov/eea/docs/dpu/orders/d-p-u-14-04-b-order-6-12-14.pdf.

Monitoring Analytics, LLC (2015). 2014 state of the market report for PJM. Available at http://www.monitoringanalytics.com/reports/PJM_State_of_the_Market/2014.shtml.

National Grid (2015). Time-of-use. Available at https://www.nationalgridus.com/masselectric/home/rates/4_tou.asp.

Newell, S., A. Faruqui, M. Swider, C. Brown, D. Pratt, A. Jaggi, and R. Bowers (2009). Dynamic Pricing: Potential Wholesale Market Benefits in New York State. Available at http://www.nyiso.com/public/webdocs/media_room/publications_presentations/White_Papers/White_Papers/Dynamic_Pricing_NYISO_White_Paper_102709.pdf.

NRG Home (2015). NRG home electricity plans. Available at http://www.nrghomepower.com/plans/.

Ontario Hydro (2015). Ontario hydro rates. Available at http://www.ontario-hydro.com/index.php?page = current_rates.

Pacific Gas & Electric (2015). Time-of-use. Available at http://www.pge.com/en/mybusiness/rates/tvp/toupricing.page.

Pacific Gas & Electric (2015). Peak Day Pricing. Available at https://www.pge.com/en_US/business/rate-plans/rate-plans/peak-day-pricing/peak-day-pricing.page.

PJM (2015). PJM data miner. Available at http://www.pjm.com/markets-and-operations/etools/data-miner.aspx.

PJM (2013). PJM financial report 2012. Available at http://www.pjm.com/ /media/about-pjm/newsroom/annual-reports/2012-financial-report.ashx.

PJM (2014a). 2017/2018 RPM base residual auction planning period parameters. Available at https://www.pjm.com/ /media/markets-ops/rpm/rpm-auction-info/2017-2018-rpm-bra-planning-parameters-report.ashx.

PJM (2014b). Summer 2014 weather normalized RTO coincident peaks. Available at https://www.pjm.com/ /media/planning/res-adeq/load-forecast/summer-2014-pjm-5cps-and-w-n-zonal-peaks.ashx.

PJM (2015a). Annual transmission revenue requirements and rates. Available at http://www.pjm.com/ /media/markets-ops/settlements/network-integration-trans-service-january-2015.ashx.

PJM (2015b). PJM financial report 2014. Available at http://www.pjm.com/ /media/about-pjm/newsroom/annual-reports/2014-financial-report.ashx.

Public Utilities Commission of the State of California (2015). Decision on residential rate reform for Pacific Gas and Electric company, Southern California Edison company, and San Diego Gas & Electric company and transition to time-of-use rates. Available at http://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M153/K024/153024891.PDF.

Southern California Edison (2015). Time-of-use rates FAQ. Available at https://www.sce.com/wps/portal/home/business/tools/time-of-use/Time-of-Use-Rates-FAQs/.

Spees, K. and L. Lave (2008). Impacts of responsive load in PJM: Load shifting and real time pricing. *The Energy Journal* 29(2): 101–121. http://dx.doi.org/10.5547/ISSN0195-6574-EJ-Vol29-No2-6.

U.S. Department of Energy (2015). Interim report on customer acceptance, retention, and response to time-based rates from the consumer behavior studies. Available at http://energy.gov/oe/downloads/interim-report-customer-acceptance-retention-and-response-time-based-rates-consumer.

U.S. Energy Information Agency (2015). Electric power sales, revenue, and energy efficiency Form EIA-861 detailed data files. Available at https://www.eia.gov/electricity/data/eia861/.

U.S. Energy Information Agency (2015). Electric power monthly. Available at http://www.eia.gov/electricity/monthly/pdf/epm.pdf.

Wächter, A. and L. T. Biegler (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* 106(1): 25–57. http://dx.doi.org/10.1007/s10107-004-0559-y.

We Energies (2015). Time-of-use. Available at https://www.we-energies.com/residential/acctoptions/time-of-use.htm.

Connect with
IAEE
on facebook