Journal of Economic Perspectives—Volume 16, Number 1—Winter 2002—Pages 191–211

The Trouble With Electricity Markets: Understanding California's Restructuring Disaster

Severin Borenstein

Starting in June 2000, California's wholesale electricity prices increased to unprecedented levels. The June 2000 average of $143 per megawatt-hour (MWh) was more than twice as high as in any previous month since the market opened in April 1998. These high prices produced enormous profits for generating companies and financial crises for the regulated utilities that were required to buy power in the wholesale markets and sell at much lower regulated prices in the retail markets. The state's largest utility, Pacific Gas & Electric, declared bankruptcy in March 2001. The state of California took over wholesale electricity purchases and spent more than $1 billion per month buying power in the spring of 2001, with average prices more than ten times higher than they had been a year earlier. Accusations of price gouging and collusion among the sellers were widespread. Some observers blamed the problems on the format of the wholesale auctions in California, while others focused on the way that transmission capacity is priced and how prices varied by location. A number of economists, myself included, did studies that concluded that sellers exercised significant market power.

While some of these issues played a role in the difficulties that electricity markets encountered in California and elsewhere, the policy discussion thus far has not focused on the fundamental problem with electricity markets: In nearly all electricity markets, demand is difficult to forecast and is almost completely insensitive to price fluctuations, while supply faces binding constraints at peak times, and storage is prohibitively costly. Combined with the fact that unregulated prices for homogeneous goods clear at a uniform, or near-uniform, price for all sellers—regardless of their costs of production—these attributes necessarily imply that

■ *Severin Borenstein is Director, University of California Energy Institute, and E. T. Grether Professor of Business Administration and Public Policy, Haas School of Business, University of California, Berkeley, California. His website is <http://haas.berkeley.edu/~borenste>.*

short-term prices for electricity will be extremely volatile. Problems with market power and imperfect locational pricing can exacerbate the fundamental trouble with electricity markets.

Two market design adjustments would greatly mitigate the fundamental trouble: long-term contracts between wholesale buyers and sellers; and real-time retail pricing of electricity, which indicates to the final customer on an hourly basis when electricity is more or less costly to consume. Historically, long-term contracts have been a standard feature of electricity markets, with cost-of-service regulation being the most detailed and extreme form of long-term contracting. Long-term contracts allow buyers to hedge against price booms and sellers to hedge against price busts.

While long-term contracts alone could be used to avoid situations like the California crisis, a much more efficient approach to the problem combines long-term contracting with real-time retail pricing. Variable retail prices can reflect real-time variation in the cost of procuring electricity, while monthly electricity bills can remain quite stable through the use of long-term contracts. Implementing real-time retail pricing would lower the total production capacity needed to meet peaks in demand and would substantially reduce the prices that buyers would need to offer to procure power on long-term contracts. Together, these two policy responses would help to produce an electricity market that operates in a smoother, more cost-effective and more environmentally responsible manner.

California's Road to Electricity Deregulation

California began serious consideration of restructuring its electricity market in 1994, motivated in part by the high electricity prices the state's customers faced at the time and in part by the example of electricity deregulation in the United Kingdom. In 1993, California's average retail electricity price was 9.7 cents per kilowatt-hour, compared to the national average of 6.9 cents. The state's high electricity prices were primarily the result of investment and procurement decisions that were made by the investor-owned utilities, with the oversight of the California Public Utilities Commission (CPUC), during the previous two decades. The utilities had built nuclear power plants that turned out to be far more expensive than originally forecast, and they had, under pressure from the CPUC, signed long-term contracts with small generators that committed them to very high wholesale purchase prices.

These mistakes were, for the most part, sunk costs, so restructuring couldn't eliminate them. Some of the customers supporting the change hoped that restructuring could be used to shift those sunk costs from ratepayers to the shareholders of the investor-owned utilities—Pacific Gas & Electric (PG&E), Southern California Edison (SCE) and San Diego Gas & Electric (SDG&E). As potent political forces in the state, however, the utilities made sure that any restructuring bill would allow for full recovery of their sunk investments, just as would have occurred if no changes in regulation had taken place.[1] Thus, the 1996 restructuring bill that was passed by the California state legislature and signed by Governor Pete Wilson contained a scheme that most observers believed would permit the utilities full recovery of their bad investments, which were often referred to as "stranded costs."

The restructuring plan treated each of the components of the electricity industry quite differently. Electricity generation was to be deregulated, and the investor-owned utilities were to reduce substantially their ownership of generation facilities. Long-distance electricity transmission was to remain a regulated function, with the utilities owning the lines and receiving compensation for their use. Local distribution of electricity also remained a regulated utility function, but the financial aspects of retail markets were opened to competition among "energy service providers." These energy service providers could contract to sell electricity to end users, while the utilities would be compensated for carrying the power to these customers' locations.

The scheme implemented for stranded cost recovery was a "Competition Transition Charge." Instead of a simple fixed surcharge on electricity consumption, the Competition Transition Charge fixed the retail price for electricity at about 6 cents per kilowatt-hour.[2] It then required customers to pay for the wholesale price of electricity and, in addition, to pay to the investor-owned utilities the difference between 6 cents and the actual wholesale price of electricity, which was expected to be much lower than 6 cents.[3] The effect was to freeze retail rates for consumers and allow the recovery of stranded costs to vary inversely with the wholesale price of electricity. The Competition Transition Charge was to end for a utility at the point that it had recovered all of its stranded costs or in March 2002, whichever came first. When the charge ended for a given utility, the utility would then switch to simply passing through the (assumed lower) wholesale price of electricity.

San Diego Gas & Electric did, in fact, end its stranded cost recovery in 1999, so when wholesale prices jumped in June 2000, SDG&E passed them through to San Diego customers. These increases raised howls of protest, and the California legislature quickly reimposed the frozen retail price on SDG&E, though with the understanding that SDG&E would be reimbursed eventually for the additional costs. The other two utilities were still under the Competition Transition Charge in June 2000, when they found themselves buying power at prices averaging more than 10 cents per KWh and reselling to customers at the frozen rate of about 6 cents per KWh.

[1] Borenstein and Bushnell (2000) discuss at greater length the reasonable and the unsupported promises that have been made in support of electricity deregulation.

[2] This is the retail price just for electricity before adding in fees for transmission and distribution. Retail prices are usually expressed in cents per kilowatt-hour (KWh). Wholesale prices are usually expressed in dollars per megawatt-hour (MWh). One MWh is equal to 1000 KWh. One cent per KWh is equal to $10 per MWh.

[3] The customer was required to make this Competition Transition Charge payment to the utilities regardless of whether the customer switched to a retail provider other than the utilities.

Besides the scheme for covering stranded costs, the most controversial aspect of the restructuring was the design of the wholesale electricity market. Essentially, there were two models of how the market could operate, an electricity pool or a market based on bilateral trades. Joskow (2000) discusses the pros and cons of these organizational structures in detail. In an electricity pool, all producers sell their power into a centrally operated electricity pool, and all customers (or their retail providers) purchase from the pool. The pool market is run by an independent system operator that also controls the physical structure of the electricity grid and thus moves power to where it is demanded and adjusts prices to reflect the supply/demand balance at each point on the grid. Parties are still free to make financial arrangements to hedge price risk associated with the market. For instance, if a producer and customer wished to contract on price, they would still be required to sell to and buy from the pool at the pool spot price, but they could sign a contract that offset any variations in that pool price and thus locked in a buy and sell price in advance.

The alternative plan was for a bilateral market, with buyers and sellers striking one-on-one deals and then notifying the independent system operator where they intended to produce and consume power. The system operator would step in only if the transactions that were planned for a given time period would overload some part of the transmission grid. In that case, the system operator would set grid usage charges that would induce changes in transaction plans so that the grid would not become overloaded. Such transmission charges would determine the price difference between locations and would reflect the shadow value of capacity to carry power between those locations. The independent system operator would also run a real-time "imbalance market," which market participants would have to use to make real-time (more precisely, after the fact) transactions, since both production and consumption usually deviate at least slightly from the advance plan. Proponents argued that this was a more free-market approach to restructuring and that if a centralized pool was so valuable, the market would create one. In addition, if such a pool were created, it would be under constant pressure to operate efficiently to keep traders using the pool rather than trading bilaterally.

What came out of the 1996–1998 market design process was a hybrid of the two visions. The independent system operator was set up to operate with approximately the vision of those proposing the bilateral model. But the California Power Exchange was also created to run a day-ahead market as a pool. For the first four years, all three California utilities, who together had most of the retail customers and a large share of the production capacity, were required to transact all their business in the Power Exchange (or the independent system operator's imbalance market). The Power Exchange ran a day-ahead trading market with both demand and supply bids. Beginning in 1999, the Power Exchange also started to run a forward market in which power could be traded for delivery many months in advance. This forward market never achieved sufficient volume to be considered a reliable market.[4] The utilities purchased nearly all of their power in the Power Exchange day-ahead market.

On April 1, 1998, California's deregulated wholesale electricity market began operation. At that time, the three utilities owned most of the electricity generation capacity in the state, which included nuclear, hydroelectric, coal, natural gas and geothermal units. Under pressure from the state, the utilities sold off nearly all of their natural gas powered generation over the following year, capacity that at the time produced 30 to 40 percent of the state's power. Five companies purchased most of this capacity, with each ending up with between 6 and 8 percent of the state's generation capacity.

For the first two years, prices fluctuated substantially within a month and even within a day. On a few days, the market registered severe shortages, and the independent system operator's real-time market price shot up to its price cap, which was $250/MWh, until October 1, 1999, when it was raised to $750/MWh. Still, the average wholesale price was never greater than $50/MWh in any month. Then, in June 2000, the precarious balance that the market had maintained fell apart. Wholesale prices increased dramatically, the independent system operator found itself unable to purchase as much power as it needed through its real-time market, and the utilities were paying wholesale prices that vastly exceeded the retail prices they were allowed to charge. Many people were surprised by the market disruption, but in retrospect, the surprise should have been that the market, as it was designed, took two years to self-destruct.

Why are Electricity Prices so Volatile?

Because storage of electricity is extremely costly and capacity constraints on generation facilities cannot be breached for significant periods without risk of costly damage, there are fairly hard constraints on the amount of electricity that can be delivered at any point in time. Yet, because of the properties of electricity transmission, an imbalance of supply and demand at any one location on an electricity grid can threaten the stability of the entire grid and can disrupt delivery of the product for *all* suppliers and consumers on the grid.

Given these unusual characteristics on the supply side of the electricity market, it is all the more remarkable how little flexibility has been built in to the demand side of the market. Metering technology to record consumption on an hourly basis

[4] Attempts by other trading forums, including the New York Mercantile Exchange, to create futures markets for electricity have also met with little success. It is hard to see how futures markets in electricity could achieve the depth and liquidity of markets that exist for other commodities, such as oil, natural gas or gold. Because electricity is not storable and transmission can become congested, prices can fluctuate dramatically over time and location. Thus, trades for any given location and time will not be very useful in hedging the price of power at another place or time.

is widely available and has even been installed at many industrial and commercial customers. Thus far, however, the meters have seldom been used to charge time-varying retail prices that reflect the time-varying wholesale cost of procuring electricity. Nearly all customers in California, and the rest of the United States, receive either a constant price or a simple fixed peak/off-peak price that captures very little of day-to-day variation in the cost of procuring electricity.

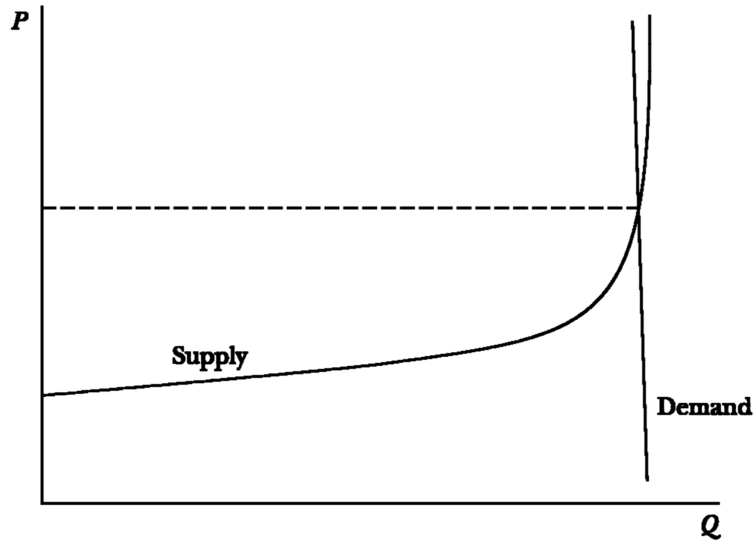
The price volatility resulting from inelastic demand and inelastic supply (when output nears capacity) is further exacerbated by the high capital intensity of electricity generation. Because a significant part of generation costs are fixed, the marginal cost of production will be well below the average cost for a plant operating at below its capacity. So long as the market price is above a plant's marginal operating cost, a competitive firm is better off generating than not. As a result, excess capacity in a competitive market will cause prices to fall to a level well below the average cost of producing electricity. This occurred in the capital-intensive memory chip industry in the early 1990s, when excess capacity caused prices of memory chips to collapse and producers to lose billions of dollars.

Figure 1 illustrates these characteristics of the electricity market graphically. Assume that the price at which the very inelastic supply and demand intersect allows the firm just to cover its fixed and variable costs. It is easy to see, however, that if capacity cannot adjust quickly and demand is difficult to forecast precisely, Figure 1 is an unlikely outcome. Even small changes will lead to a price boom or bust.

For example, a slight rightward shift of demand will cause price to skyrocket. Unlike, for instance, in the airline industry, where capacity on a route can adjust quickly and demand is responsive to price changes, there is no elasticity on the supply or demand side that allows the electricity market to adjust to such a mismatch. Extremely high prices may elicit a bit more output as generators run their plants harder—risking heavier maintenance costs—due to the tremendous profit opportunity. In nearly all current restructured markets, the demand response from high prices is primarily limited to actions by the independent system operator, which can reduce reserve margins (standby capacity it pays some generators to have ready on short notice) and can exercise contract rights it has to interrupt power to certain customers, an extreme measure that causes significant disruption to the affected customers.

The tight supply situation is exacerbated if markets are not fully competitive. Tight supply conditions in electricity markets put even a fairly small seller in a very strong position to exercise market power unilaterally, because there is very little demand elasticity and other suppliers are unable to increase their output appreciably (Borenstein, 2000). Because market power is easier to exercise in electricity markets when the competitive price would have been high anyway, it exacerbates the volatility of prices and further reduces the chance that prices will remain in a reasonable range.

Figure 1
Supply and Demand in the Electricity Market



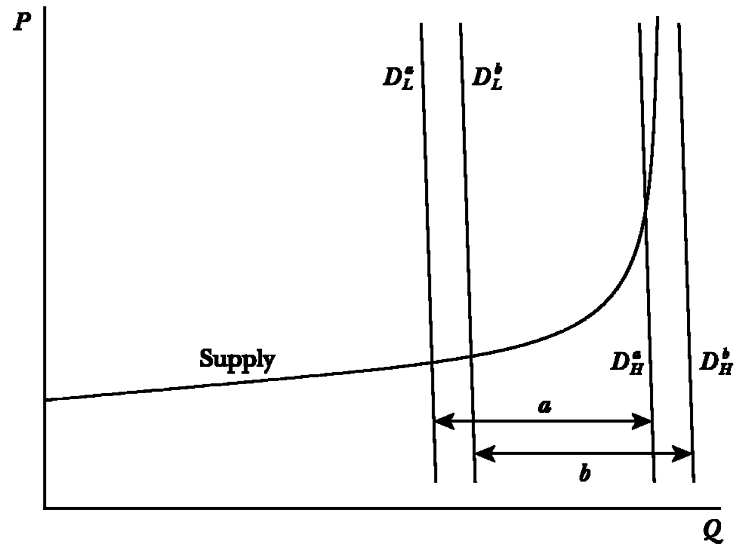
Many observers of deregulation have said that the root of the problem in California is that the state's expected surplus of capacity disappeared due to strong economic growth throughout the western U.S. electricity grid. If the surplus had remained, however, the result would have been a crisis of a different sort. A slight leftward shift of demand in Figure 1 causes price to collapse to the low marginal running costs of the marginal unit. These prices would almost certainly fail to cover the average costs of operating the plants, a situation similar to the 1990s memory chip market. In the newly deregulated electricity market, this outcome would surely have led to calls for subsidies to producers.

While Figure 1 and the discussion thus far has focused on one supply/demand interaction, the concept applies equally to a market in which demand varies by hour. In Figure 2, assume that demand in a month is distributed uniformly between D_L^a and D_H^a. Now, consider a relatively small rightward shift of the demand distribution to between D_L^b and D_H^b. This small shift replaces hours that were at very low prices on the left of the distribution with hours that are at extremely high prices at the right side of the distribution, causing the average price to increase drastically.

The discussion so far has assumed that all sellers in a short-term market for electricity receive the same price for delivery of power at the same time. In the policy debate, there was a great deal of discussion about the fact that sellers who have low production costs are paid a much higher market-clearing price. Some policymakers, and even a few economists, blamed this on the uniform-price auctions that were used by the Power Exchange and the independent system operator. This is, however, the way that all commodity markets work. Producers sell their output at the market price regardless of whether they are producing from low-cost or high-cost sources.

This demonstration of the law of one price is not a function of the auction

Figure 2
A Shift in Variable Demand in the Electricity Market



format or some design flaw in the electricity market.[5] It is true in all commodity markets, whether or not firms are able to exercise market power. Nonetheless, this outcome means that when a supply/demand mismatch causes extreme price volatility, it changes the price for all power being sold in the market at that time. This one-price outcome is in sharp contrast to the outcome under regulation, in which each production facility is compensated at its own average cost of production, and the price that consumers pay is set to cover the average of all of these production costs.

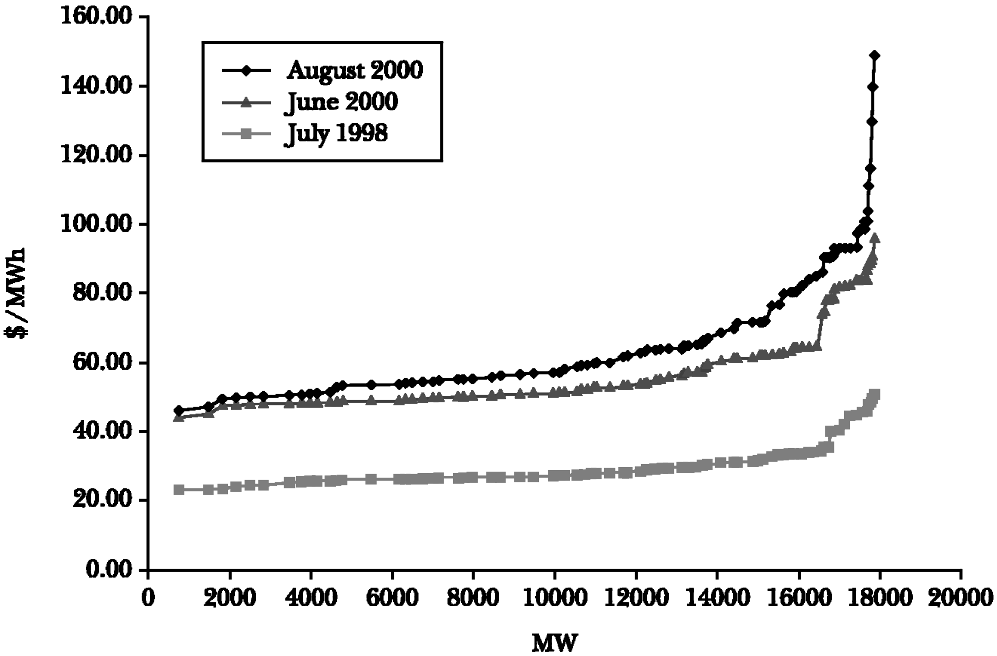
If production were just as efficient under regulation as in a competitive market, average-cost regulatory pricing would yield lower prices when supply is tight, because the marginal cost of production would be above the average cost. The difference would be even greater if the unregulated market were not completely competitive and unregulated prices were above marginal cost. California faced that situation in summer 2000. But in a situation of surplus capacity, marginal cost will be below average cost. In that case, the price from a market process may be below the price that regulation would produce, which is the situation that in 1996 many people believed California would face during the early years of restructuring.

The Upheaval in California's Electricity Market

California's summer 2000 electricity market illustrates the inherent volatility discussed in the previous section. A dryer-than-normal year, which reduced hydroelectric production, combined with a hotter-than-normal summer and continued

[5] Kahn, Cramton, Porter and Tabors (2001) analyze uniform-price versus pay-as-bid auctions in the California electricity market. Wolfram (1999) discusses the same issue in the U.K. electricity market.

Figure 3
California Thermal-Generation Supply Curve
(various months)



economic growth throughout the western United States shifted the supply/demand balance and caused the market to tighten up suddenly. Although the investor-owned utilities had by 2000 received permission to buy a limited amount of power under long-term contracts, they were doing very little of it. They were still procuring about 90 percent of their "net short" position—the power that they were not producing with their own generation and did not have under contracts that predated the restructuring—in the Power Exchange's day-ahead or the system operator's real-time market.

In addition, cost increases for thermal generating plants (in California, nearly all of which are natural-gas fueled) raised production costs and, importantly, did so much more for the marginal production units. Figure 3 shows the marginal cost curve from all thermal plant capacity in California. This omits production from nuclear and hydroelectric production, which are inframarginal in nearly all hours, and renewable sources (wind, solar and geothermal), which have less reliable production patterns. Thermal plant production is nearly always the marginal power source in California.

The lowest line is the marginal cost curve during July 1998, when gas prices were low and the costs of pollution permits for emitting nitrogen oxide were negligible. The next highest line shows costs during June 2000, when natural gas prices were almost double their 1998 levels. Not only has the curve shifted up, it has rotated, with the costs of the most expensive units increasing more, because the most expensive units convert natural gas to electricity at about half the efficiency rate of the least expensive generators. By August 2000, shown in the highest line, the problem was further exacerbated as the price of nitrogen oxide pollution permits increased from about \$1 per pound to over \$30 per pound (and gas prices increased further). The least efficient generators were also the biggest emitters of nitrogen oxide, so the rotation was even more pronounced.

Thus, even absent any exercise of market power, the cost and demand changes that took place during summer 2000 would have greatly increased market prices. The rotation of the supply curve meant that the increased price of natural gas and nitrogen oxide pollution permits not only raised electricity prices to cover increased costs, they also greatly increased the inframarginal rents that suppliers were able to earn. In July 1998, the most expensive gas-fired generators had costs \$20/MWh greater than the least expensive plants. By August 2000, the difference was more than \$100/MWh. Thus, when the high-cost plants needed to run, it created enormous inframarginal rents for low-cost producers.

Market Power in California's Wholesale Market

A number of empirical studies have concluded that sellers have exercised significant market power in California's wholesale electricity market (Borenstein, Bushnell and Wolak, 2001; Wolak, Nordhaus and Shapiro, 2000; Puller, 2001; Joskow and Kahn, 2001; Hildebrandt, 2001; Sheffrin, 2001). Harvey and Hogan (2000, 2001) have disputed these conclusions by suggesting that the studies did not appropriately control for costs and scarcity, but their work does not offer an alternative empirical analysis. This debate over market power has differed from those in many other industries because it has focused on unilateral exercise of market power by firms that have a comparatively small share of total production in the market. The unregulated generation owners that have been accused of exercising market power own between 6 and 8 percent of the production capacity in the independent system operator control area. The Federal Energy Regulatory Commission (FERC) has the power to monitor and to mitigate market power, but until 2001, it was committed to the view that firms with a market share below 20 percent could not exercise significant market power.

This focus on market share analysis ignores the reality that in a market with no demand elasticity and strict production constraints, a firm with even a small percentage of the market could exercise extreme market power when demand is high. On a hot summer afternoon, when the system operator needs 97 percent of all generators running to meet demand, a firm that owns 6 percent of capacity can exercise a great deal of market power. In fact, a seller will find it profitable to exercise market power any time the elasticity of residual demand the firm faces is sufficiently small. That elasticity is determined by the elasticity of market demand and the elasticity of supply from other producers.

Figure 3 shows that in summer 2000, beyond about 14,000 MW of thermal generation, the marginal cost curve becomes increasingly steep, implying a less elastic residual demand curve faced by any single producer. Restricting output becomes more profitable when the cost of the next highest cost generation unit exceeds the market price by a greater amount, that is, when the industry supply function is steeper.

Thus, while the exact degree of market power is an empirical question, a reasonable first-cut analysis leads one to ask why a seller with 3,000–4,000 MW of capacity *wouldn't* exercise market power. Borenstein and Bushnell (1999) simulated the market using a Cournot quantity-setting model. Even with an assumed demand elasticity of −0.1—larger than any plausible estimate under the California transition plan—we found the potential for very significant markups without any collusion among sellers.[6]

The Role of Long-Term Contracting

In unregulated markets that exhibit a great deal of spot-price volatility, buyers and sellers commonly smooth their transaction prices by signing long-term contracts. Nearly all electricity markets outside of California have taken this approach. In many cases, the sale of utility generation facilities to other firms has been accompanied by "vesting contracts" that require a certain amount of power sales back to the utility at a predetermined price. Also, the regulated utilities have in many cases retained some of their generation facilities. The price customers end up paying for the power from those facilities is then based on their costs of operation, not the market price. While California had virtually no vesting contracts, the California utilities did retain generation facilities, and they had some long-term contracts that predated restructuring. Together, these sources accounted for more than 60 percent of the power the California utilities delivered to customers.

Some participants in the debate have suggested that utilities in California and elsewhere will get systematically lower prices buying power on long-term contracts than they will get in the spot market. Spot prices, however, are very unlikely to exceed forward prices for power to be delivered on the same day in a systematic way, because such a situation would set up a profitable arbitrage opportunity. In summer 2000 in California, power contracted in advance was cheaper than spot power for the same delivery hour, but the reason sellers were willing to contract at those lower prices in advance—in late 1999 or early 2000—was that their best guess of summer 2000 prices was below the spot prices that actually resulted. In contrast, the forward prices for power to be delivered during 2001–2002 in California shot up in early 2001, and contracts signed at that time turned out to be well above the

[6] Some observers have argued that any capital-intensive industry will always be imperfectly competitive, so measuring margins above short-run marginal cost is meaningless. This is incorrect on both counts. First, even if markets are imperfectly competitive, measuring price-cost margins is the appropriate way to see how imperfect that competition is and to monitor changes in the degree of imperfection. Second, many capital-intensive industries are populated by price-taking firms. Gold mining, for instance, is a highly capital-intensive industry in which all sellers are price takers. In fact, the same is true for most of the goods listed on the commodities page of the *Wall Street Journal,* such as oil, natural gas, corn, oats, silver and coffee. This is also the page on which the *Journal* lists California electricity prices.

spot price for summer 2001, when spot prices collapsed. *On average, a purchaser buying power in forward markets (or through long-term bilateral contracts) will not receive lower power costs than a purchaser buying in the spot market.*[7]

The buyer's concern with long-term forward contracting, of course, is that it might lock in a higher price than it would have had to pay if it had purchased in nearer-term markets. This fear is especially large for regulated utilities acting as energy service providers in a restructured market. They are concerned that in such a situation the state regulatory agency might decide that the contract purchase price was "imprudent" and not allow the utility to pass through the costs to customers. Credible commitment by regulators is difficult. Nonetheless, it is clear that the correct standard for judging the prudence of these contracts is based on the information available at the time the contract is signed, not looking backward after the actual spot prices have become available. Such opportunistic behavior by regulatory agencies simply discourages prudent long-term contracting.

Long-Term Contracts and Market Power

While forward prices won't systematically beat spot prices, there is a potential price-lowering effect in *both* forward and spot markets if, in aggregate, buyers purchase more power through long-term contracts. Locking in some sales in advance reduces the incentives of multiple firms to behave less competitively among themselves (Allaz and Vila, 1993).

The idea is that if firms are maintaining high prices by foregoing aggressive price cutting, then the existence of many forums for trading, especially over time, makes it more difficult to maintain such mutual forbearance. The forbearance could take the form of implicit or explicit collusion, or it could be the result of unilateral decisions that result in a less competitive outcome, such as under Cournot competition. The possibility of selling in advance makes it more difficult for firms to restrain competition. Once a firm has sold some output in advance, it has less incentive to restrict its output in the spot market in an attempt to push up prices in that market, since it does not receive the higher spot price on the output it has already sold through a forward contract. Thus, in anticipation of more aggressive competition in the spot market—because some firms have presold a significant quantity in a forward market—firms are likely to price more aggressively in the forward market.

More generally, the incentive of a generating company to exercise market power will depend on its *net* purchasing position in the market at a given point in time. If a firm were a large net seller, it would likely have an incentive to restrict output to raise price. If it had sold much of its output under forward contracts, then it would have much less incentive to restrict its output to increase the spot price.

[7] The study by Borenstein, Bushnell, Knittel and Wolfram (2001) looks at the relationship between the California Power Exchange's day-ahead price and the California independent system operator's balancing market price. We note that prices in the forward market could be lower on average if sellers are systematically more risk averse than buyers, but we argue that this is unlikely.

The actual equilibrium impact of forward contracting on both spot and forward market prices is uncertain. It can do no more than eliminate the portion of price premia that are due to market power, and it might have a substantially smaller effect. Forward contracting cannot lower the average price a buyer pays to below the level that a buyer will obtain in a competitive market.

Long-Term Contracting is Only Part of the Solution

Long-term contracting is an important part of the solution to the fundamental problem of electricity markets, but it does not "solve" the mismatches between supply and demand. It just prevents large fluctuations in electric bills when those mismatches occur. It can, however, be used to pay for excess or standby capacity by assuring that the generating companies receive payments sufficient to cover their capital costs even if demand turns out to be low and some of the capacity does not get used.

In fact, this is what the old regulatory system did. Utilities were assured of revenues to cover their costs and in return built sufficient capacity to make sure that all contingencies could be covered. Supply always exceeded demand by a significant amount, and the cost of all that idle capacity was rolled into the price that customers paid for the power that they did use.

Many players in the California market now advocate a return to this type of system in a quasi-deregulated electricity market. Utilities could sign long-term contracts for power and capacity that assured generators they could recover their costs even if the capacity were not actually used. A number of state and federal policymakers have argued that the state should always make sure that capacity exceeds expected demand by at least 15 percent.

A policy of holding excess capacity would assure that spot prices were always very low (assuming that no generator held a large market share) and that many new "peaker" plants were built to assure excess capacity but virtually never used. This outcome would be unfortunate, since it does not make sense to hold such capacity if the customer's value of consuming the additional power when it is used is less than the full cost of making the power available. Real-time retail prices that reflect the cost imposed by additional consumption in each hour are the ideal mechanism for making that tradeoff.

Thus far, California and other states have attempted to make electricity markets work almost entirely on the supply side of the market. This approach has worked relatively well in some markets, but the California crisis has demonstrated the variety of constraints that exist on the supply side. Deregulating only the supply side of the market seems to be the equivalent of making an electricity market operate with one arm tied behind its back. Combining long-term contracts with real-time pricing can provide the right economic incentives to reduce demand at peak times when the system is strained, while still assuring customers of relatively stable monthly bills.

Real-Time Retail Price Signals and Stable Monthly Bills

Although the marginal cost of producing electricity varies tremendously over time and producers face hard capacity constraints, in very few electricity markets do retail prices reflect these cost variations. Peak/off-peak pricing is fairly common for commercial and industrial customers, but it is virtually always implemented as "time-of-use" pricing, a two- or three-price system with, for instance, one price for daytime usage and a lower price for nighttime usage. Real-time retail pricing, in contrast, allows prices to change with each given time interval, such as ten minutes or one hour, and prices need not be the same at a given time from one day to the next.[8] The effect of customers facing a single constant price for electricity is that they have no more incentive to conserve during peak consumption times, like a hot summer afternoon, than during low consumption times, like a cool afternoon or the middle of the night. They also have no incentive to shift consumption away from times when the production capacity of the grid is strained and production costs are highest. As a result, more capacity needs to be built to accommodate all of the demand at the highest peak times than would otherwise be the case. Real-time pricing would reduce the need to build new plants that would run for only a few days of peak demand each year.

While many people have advocated greater price responsiveness in demand through real-time retail electricity pricing, at the same time, there have been calls for greater protection of customers from price spikes. These goals may seem to conflict, but it is possible to expose customers to hourly price fluctuations, so that price-responsive demand will be meaningful, and still assure them of relative stability in their monthly bills. The key to meeting both of these goals is to recognize that the *average level* of prices can be stabilized without damping the variation in prices. For an energy service provider to offer both real-time retail price variation and monthly bill stability, without risking substantial losses, it needs to hedge a significant portion of its energy cost through long-term contracts.

To be concrete, assume that the energy service provider begins by engaging in no hedging. It charges customers a fixed per-kilowatt-hour transmission and distribution charge plus the spot price of energy in each hour. This approach satisfies the real-time pricing goal, but the monthly bills would be as variable as the month-to-month variation in the weighted-average spot energy prices. To attain the goal of monthly bill stability, the energy service provider would sign a long-term contract to buy some amount of power at a fixed price. Such a contract is likely to be at about the average spot price of the electricity that the parties anticipate over the life of the contract, but in any given month, the contract price could be greater or less than the average spot price.[9]

This contract can be considered a financial investment that is completely

[8] Borenstein (2001) discusses the large advantages of real-time pricing over time-of-use pricing.
[9] Alternatively, if the energy service provider owns generation capacity, it need only to hedge the price of fuel to run the generation. Capacity ownership itself hedges much of the electricity price risk.

independent of the retailing function. The critical point is that the energy service provider's return on this financial investment varies directly with the average spot price of energy, and that return can be applied to change the average level of customer bills. When viewed this way, it becomes clear that the long-term contract can affect the average price level without damping the price variation. The gains from the long-term contract (when the average spot price is higher than the contract price) or losses (when the average spot price is lower than the contract price) could be distributed to customers to stabilize bills. The distribution could be done with a constant (over the month) surcharge or discount on each kilowatt-hour sold during that month or—even more attractive to economists—as a lump-sum transfer based, perhaps, on the customer's past usage levels.

The most important impact of this approach would be that it would lower quantities demanded at peak times, and by doing so, it would lower the market prices at those times. Hearkening back to Figure 1, the demand curves would become much flatter, since customers would be able to see and to respond to high prices. This would prevent extreme price spikes. It would also reduce the financial incentive of sellers to exercise market power, since one firm's reduction of output would have a smaller effect on price than it does when demand is completely price-inelastic. Thus, real-time pricing would lower the overall average wholesale cost of power.[10]

The effect of real-time pricing also has very important implications for the negotiation of long-term contracts. If sellers, at the time of negotiation, believe that real-time pricing is likely, then they will reduce their forecasts of the average spot prices they would be able to earn if they did not sell through a long-term contract. As a result, the sellers will be willing to accept a lower long-term contract price than they otherwise would. Unfortunately, California did not make such a commitment to real-time pricing before it negotiated many long-term contracts in the spring of 2001.

Though real-time pricing has not been widely used in the United States, the technology is well established. Most large commercial and industrial customers in California have real-time meters already, and communication of the day-ahead or imbalance market price to those customers can easily take place through the Internet. In the near future, it may not be practical or necessary to include residential customers in a real-time pricing program, but as the cost of real-time meters declines, including residential customers can be straightforward. It is critical to understand that the *variation* in prices can be separated from the *average level* of prices. For any given level of flat retail price that is contemplated, the same systemwide average price level can be attained each month with real-time retail

[10] It is also worth noting that setting retail prices below the sum of the wholesale price and the transmission and distribution charge can move prices closer to the actual marginal cost, even if there is no market power present. Transmission and distribution is charged on a marginal basis, but these costs are largely fixed. Therefore, reducing price by up to the transmission and distribution fee that would otherwise be in the retail price has the effect of moving price closer to marginal cost.

pricing. Doing it with real-time pricing will reduce the cost of procuring the power and reduce the need to build more power plants, ultimately allowing lower retail prices.

While real-time pricing would increase total welfare, those customers who now consume disproportionately at times when the system demand is highest could be made worse off. Under the current flat pricing of electricity, these customers are subsidized by those that consume a smaller share of the system load at peak times than at off-peak times. In the case of California, this cross-subsidy roughly runs geographically from coastal communities that use less air conditioning to central valley communities that use more.

However, even with a moderate amount of price responsiveness, the wholesale electricity price at peak times would be reduced as demand at those times declines, so the increase in the retail price at peak times relative to flat retail pricing would not be nearly as great as one would infer from looking at price patterns during 2000. To the extent that policymakers wish to cross-subsidize areas that consume more power at peak times, this could be done through an explicit subsidy of power use in those areas, preferably one that does not continue to subsidize consumption at peak times most heavily. In the end, however, the only way absolutely to assure that *no one* will be made worse off by ending this cross-subsidy is to continue with flat pricing, which gives no incentive to reduce peak-time consumption

The Difference Between Real-Time Retail Pricing and Paying for "Negawatts"

Many alternative programs have been proposed that mimic, to some extent, the effect of real-time pricing. These programs generally are based on the idea of paying customers to reduce consumption at certain times. Paying for demand reduction at peak demand times may seem, at first, more attractive than real-time pricing, because it "rewards" those who conserve at peak times rather than "punishing" those who consume when the system is strained. The distinction is, of course, misleading, since the rewards are paid for through either electricity rate increases that are spread across all consumption or other taxes that are unrelated to the cost of electricity consumed.

While these programs can, in theory, offer many of the benefits of real-time pricing, in practice they offer much less benefit and about the same cost. On the cost side, implementing any sort of demand-reduction market requires the same real-time metering equipment and about as much information on price, quantity demanded or reserve margins as real-time pricing. You can't reward demand reduction unless you know when and how much reduction occurred.

The more difficult problem with paying for demand reduction is the baseline from which the payment is made. Unless the program is mandatory and the baseline is set based on information that is completely out of the control of the customer (such as demand information from a number of years earlier), the program will be subject to extensive manipulation and self-selection problems. The manipulation occurs if the baseline is set based on any consumption information that can be affected after the program is announced or anticipated. For instance, one recent suggestion in California would pay a customer on superpeak hot summer days to reduce demand from its average level over the previous *x* days. This plan would greatly diminish any incentive to reduce demand in other days, since such actions would lower the baseline the customer started from on the superpeak days.

The self-selection problem exists even if the baseline is set from truly exogenous information. The entities that would opt to sign up for these programs will disproportionately be the ones who know already that they will be reducing their demands, such as companies that are reducing their operations or that have already changed their production process to use less power.[11] Likewise, those entities whose baseline has been set inordinately high, due to some unusual activity during the period used for determining the baseline, would also be more likely to join the program.

The Role of Price Caps

In California and other wholesale electricity markets, price spikes have led to a debate about imposing price controls. In reality, price caps are, and will continue to be, a critical element of virtually all wholesale electricity markets. The extreme inelasticity of both supply and demand means that supply shortages, whether real or due to market power, can potentially drive prices many thousands of times higher than their normal level. Such outcomes would destroy the market. Therefore, the debate should be about the level of price caps and mechanisms for their adjustment.

Price cap opponents have said that such controls reduce investment in production facilities and reduce production from facilities that already exist. Both statements are potentially true. *If price caps are set too low, they will have detrimental effects.* The question is at what level these effects will occur.

In the short run, a price cap will deter production from an existing facility if the cap is below the short-run marginal cost of production. Until summer 2000 in California, suggestions that a \$250/Mwh price cap would deter production were hard to credit. During that summer, the additional cost of air pollution permits in the south coast may have pushed the incremental cost for the least efficient plants in that area above the cap and thus deterred them from producing. The problem became very salient in November and December 2000, when a spike in the price of natural gas—rising from \$4–\$6 per million BTU (British thermal units) to over \$30—put the incremental cost of nearly all natural gas plants above the price cap.

Price caps, however, can also deter the exercise of market power. A cap set at or above the competitive price, but below the price that would have resulted without the cap, will lower prices and *increase* aggregate output from the firms in the

[11] A very similar self-selection problem occurs if real-time pricing is implemented on a voluntary basis. Those entities that know they consume disproportionately at the peak times will not opt for the program and will thus continue to have no incentive to conserve when the system is strained.

market.[12] The intuition is that with a price cap in place, firms with market power do not have an incentive to restrict output any more than would be necessary to raise price to the cap. Thus, the appropriate level for price caps trades off the risk of setting them too low and deterring production with the risk of setting them too high and permitting the exercise of excessive market power.

The long-run impact of price caps is straightforward to analyze conceptually, but more difficult to study empirically. A price cap will deter investment in new capacity if it is set, or if investors believe it will be set, at a level that does not allow a return on investment that exceeds the investors' cost of capital. The data available on costs of building a power plant are necessarily rougher than the data on variable costs of production, because the costs of building a power plant are subject to many idiosyncratic factors related to location, siting restrictions and other attributes. Furthermore, the beliefs of investors play a critical role, because the return is calculated over the life of the plant. Just as under cost-of-service regulation, uncertainty about future regulatory intervention is likely to deter investment. Thus, price caps should be used with great caution.

Still, it is a well-established result that absent significant scale economies, a price cap that is set at or above the competitive price level in every hour will not deter efficient investment. In a fully restructured electricity market with price-responsive demand and long-term contracts, price caps should exist only as a backstop measure. The debate over price caps in California took place in a setting with no price-responsive demand and very limited use of long-term contracts.

Finally, economic analysis of price caps has generally assumed that an announced price cap is credible and is never breached. That was not the case in California during 2000; the independent system operator frequently violated the cap, both during the summer, when the competitive price was probably below the cap nearly all of the time, and in November and December 2000, when, for a few weeks, the competitive price almost certainly exceeded the cap due to soaring natural gas prices. In the latter situation, violation of the cap was the only reasonable action, since generators were better off shutting down than selling power at $250/MWh.

During summer 2000, however, the breaches of the cap made it very difficult to convince sellers that attempts to raise the price above the cap through exercise of market power would fail. Absent such credibility, the price cap creates a game of "chicken" between sellers and buyers. In the case of California, the independent system operator's unwillingness to curtail demand, and its inability to elicit demand-side response with real-time retail prices, put it in a very weak position in these showdowns.

[12] See Carlton and Perloff (1994, pp. 864–870) for an example. Also see Viscusi, Vernon and Harrington (2000, pp. 500–503).

The Aftermath of the California Electricity Crisis

Because retail electricity rates remained frozen through 2000, the California utilities lost millions of dollars per day buying power at high wholesale prices and selling at lower retail prices. In early 2001, with the utilities teetering on the edge of bankruptcy and no longer creditworthy, the state of California stepped in to become the wholesale power buyer for the utilities.

At the same time, the state and the California utilities pleaded with the Federal Energy Regulatory Commission to impose price caps on the wholesale market. The FERC had imposed "soft caps" in December 2000, which were largely ineffective due to half-hearted enforcement. Throughout spring 2001, the federal and state government were at loggerheads over the price cap issue, until May 2001, when FERC quite suddenly reversed its position and imposed price caps that were lower and more likely to be enforced.

During spring 2001, the state of California also signed long-term power contracts, ranging from one to 20 years, with nearly all of the major generators selling power in the California market. The contract prices are difficult to characterize easily due to the varying lengths and contract conditions, but they were clearly at prices that most observers would have considered astoundingly high a year earlier. In part, the high prices spread over many years were a way for the state to hide astronomical prices it was implicitly going to pay for power during summer 2001 and 2002.

In early June 2001, just as the new price cap policy was taking effect and the state was completing negotiation of the long-term contracts that covered most of the utilities' net short position for at least the next few years, the price of natural gas suddenly collapsed in California, falling from around $10 per million BTU to the $3 range, a level comparable to the eastern United States. Some observers argued that the sudden collapse, which coincided with a change in one company's transmission rights on the main pipeline into southern California, indicated that the price had been artificially inflated.

The collapse of gas prices occurred at nearly the same time that the California Public Utilities Commission finally raised retail electricity prices. The price increase was 40 percent to 50 percent for most industrial and commercial customers, but less than half that for most residential customers. In addition, the state instituted a summer 2001 rebate plan that rewarded customers who reduced consumption by at least 20 percent from summer 2000 levels.

Mild weather and aggressive conservation (which reduced weather-adjusted demand by 5 to 10 percent) combined with price caps, long-term contracts and, most importantly, the collapse of natural gas prices sent spot electricity prices tumbling in June 2001. By mid-summer 2001, spot electricity prices were back to pre-crisis levels, and the state was committed to over $40 billion worth of long-term electricity contracts at prices that are likely more than 50 percent above the expected future spot prices. These are the new stranded costs of the California electricity industry. Of course, many large customers then attempted to avoid paying for these costs by switching from the utility providers to other energy service providers, who could once again offer prices well below the utility prices that have the stranded cost recovery bundled in. Rather than imposing a non-bypassable charge to cover contract costs, the California Public Utilities Commission responded by canceling retail competition. In many ways, California has returned to 1996, albeit with customers many billions of dollars poorer.

Finally, during summer 2001, the state also established the California Public Power Authority. The CPPA has set about building state-owned "peaker" plants to assure that there won't be another shortage. Its goal is a 15 percent reserve capacity margin, which it argues is necessary to ensure a competitive wholesale market.

In spring 2001, California launched a $35 million program to install real-time meters at all large industrial and commercial customers. Despite that expenditure, real-time electricity pricing has stalled at the California Public Utilities Commission and at this writing seems unlikely to be adopted on even a widespread voluntary basis in the near future.

Conclusion

The movement toward restructuring of electricity markets was born from a history of well-supported dissatisfaction with outcomes under cost-of-service regulation. Nonetheless, electricity markets have proven to be more difficult to restructure than many other markets that served as models for deregulation—including airlines, trucking, natural gas and oil—due to the unusual combination of extremely inelastic supply and extremely inelastic demand. Real-time retail pricing and long-term contracting can help to control the soaring wholesale prices recently seen in California and can buy time to address other important structural problems that need to be solved to create a stable, well-functioning electricity market. These problems include creating a workable structure for retail competition, determining the most efficient way to set locational prices and transmission charges, implementing a coherent framework for investing in new transmission capacity and optimizing the procurement of reserve capacity.

Those states and countries that have not yet started down the road of electricity deregulation would be wise to wait to learn from the experiments that are now occurring in California, New York, Pennsylvania, New England, England and Wales, Norway, Australia and elsewhere. The difficulties with the outcomes so far, however, should not be interpreted as a failure of restructuring, but as part of the lurching process toward an electric power industry that is still likely to serve customers better than the approaches of the past.

■ *I have benefited a great deal from discussions with Carl Blumstein, James Bushnell, Erin Mansur, Paul Joskow, Steve Puller, Steve Stoft, Frank Wolak, Catherine Wolfram and Hal Varian, but the opinions in this paper do not necessarily reflect their views. Erin Mansur provided excellent research assistance.*

References

Allaz, Blaise and Jean-Luc Vila. 1993. "Cournot Competition, Forward Markets, and Efficiency." *Journal of Economic Theory.* February, 59:1, pp. 1–16.

Borenstein, Severin. 2000. "Understanding Competitive Pricing and Market Power in Wholesale Electricity Markets." *Electricity Journal.* July, 13:6, pp. 49–57.

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

Borenstein, Severin and James Bushnell. 1999. "An Empirical Analysis of the Potential for Market Power in California's Electricity Industry." *Journal of Industrial Economics.* September, 47:3, pp. 285–323.

Borenstein, Severin and James Bushnell. 2000. "Electricity Restructuring: Deregulation or Reregulation?" *Regulation.* 23:2, pp. 46–52.

Borenstein, Severin, James Bushnell and Frank Wolak. 2001. "Diagnosing Market Power in California's Deregulated Wholesale Electricity Market." POWER Working Paper PWP-064, University of California Energy Institute, revised December.

Borenstein, Severin, James Bushell, Christopher Knittel and Catherine Wolfram. 2001. "Trading Inefficiencies in California's Electricity Markets." POWER Working Paper PWP-086, University of California Energy Institute, October.

Carlton, Dennis W. and Jeffrey M. Perloff. 1994. *Modern Industrial Organization, Second Edition*. New York: Harper Collins.

Joskow, Paul L. 2000. "Deregulation and Regulatory Reform in the U.S. Electric Power Sector," in *Deregulation of Network Industries: What's Next?* Sam Peltzman and Clifford Winston, eds. Washington, D.C.: AEI-Brookings Joint Center for Regulatory Studies, pp. 113–54.

Joskow, Paul L. and Edward Kahn. 2001. "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market During Summer 2000." National Bureau of Economic Research Working Paper #8157, March.

Harvey, Scott M. and William W. Hogan. 2000. "Issues in the Analysis of Market Power in California." Mimeo, October 27.

Harvey, Scott M. and William W. Hogan. 2001. "On the Exercise of Market Power Through Strategic Withholding in California." Mimeo, April 24.

Hildebrandt, Eric. 2001. "Further Analysis of the Exercise and Cost Impacts of Market Power in California's Wholesale Electricity Market." Report from the California Independent System Operator's Department of Market Analysis, March.

Kahn, Alfred E., Peter C. Cramton, Robert H. Porter and Richard D. Tabors. 2001. "Pricing in the California Power Exchange Electricity Market: Should California Switch from Uniform Pricing to Pay-as-Bid Pricing?" Commissioned by the California Power Exchange, January 23.

Puller, Steve L. 2001. "Pricing and Firm Conduct in California's Deregulated Electricity Market." POWER Working Paper PWP-080, University of California Energy Institute, January.

Sheffrin, Anjali. 2001. "Empirical Evidence of Strategic Bidding in California ISO Real Time Market." Report from the California Independent System Operator's Department of Market Analysis, March.

Viscusi, W. Kip, John M. Vernon and Joseph E. Harrington, Jr. 2000. *Economics of Regulation and Antitrust.* Cambridge, Mass.: MIT Press.

Wolak, Frank A., Robert Nordhaus and Carl Shapiro. 2000. "An Analysis of the June 2000 Price Spikes in the California ISO's Energy and Ancillary Services Markets." Report of the California Independent System Operator's Market Surveillance Committee, September.

Wolfram, Catherine. 1999. "Electricity Markets: Should the Rest of the World Adopt the United Kingdom's Reforms?" *Regulation.* 22:4, pp. 48–53.