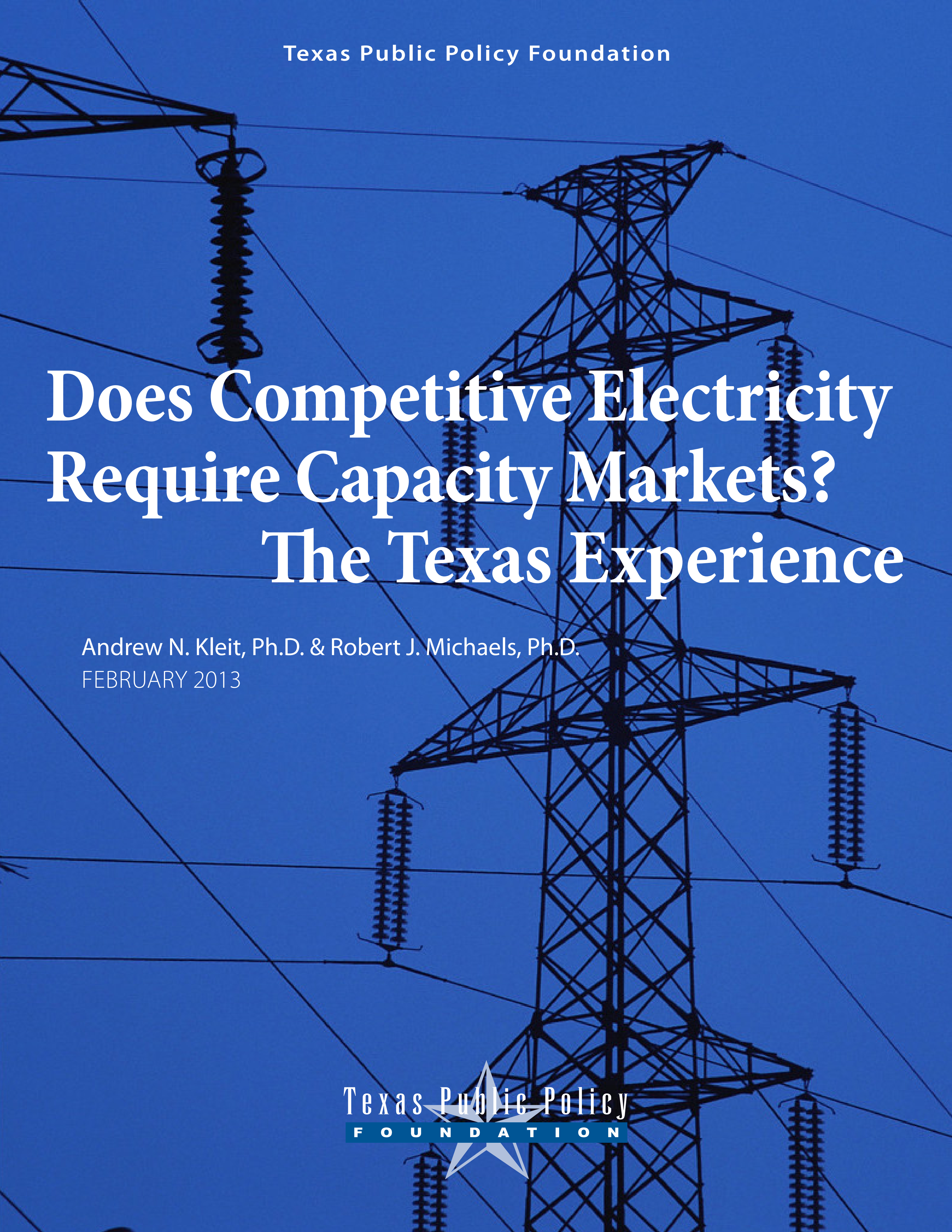

Texas Public Policy Foundation

Does Competitive Electricity Require Capacity Markets? The Texas Experience

Andrew N. Kleit, Ph.D. & Robert J. Michaels, Ph.D.

FEBRUARY 2013

Texas Public Policy
FOUNDATION

February 2013

Prepared for the Texas Public Policy Foundation
by Andrew N. Kleit and Robert J. Michaels

Table of Contents

Does Competitive Electricity Require Capacity Markets? The Texas Experience

Andrew N. Kleit, Ph.D. & Robert J. Michaels, Ph.D.

Executive Summary

The exception to the rule among U.S. power markets administered by Regional Transmission Operators is the Electricity Reliability Council of Texas (ERCOT). ERCOT's "energy-only" market relies on competitive market forces to meet the long-term electricity needs of the 23 million Texans in its service area. Shorter-term needs are also met through the competitive market, supplemented by markets for ancillary services.

Competition has worked remarkably well in ERCOT since its introduction about 15 years ago. Consumers can choose over a hundred different plans from dozens of providers. Billions of dollars invested in generation have provided Texas with a reliable supply of affordably-priced electricity.

However, recent concerns about the adequacy of generation investment have led to the consideration of imposing a "capacity market" in ERCOT. Proposals would make ERCOT more like other U.S. power markets, which require that sellers of power to end-users must own or have contractual access to generation capacity sufficient to cover their loads. In other words, in a capacity market the government rather than the market determines when supplies of electricity are adequate to meet long-term reliability needs.

This paper examines the potential value of a capacity market in Texas. We begin with a critique of the economic theory behind capacity markets, which we find deeply flawed. We then apply that theory to ERCOT. In the process, we reexamine research on investment adequacy in ERCOT and the value of energy prices as signals for generation investment. We conclude that investment in generation in ERCOT is likely to continue and, as it has in the past, provide sufficient reserves to maintain reliability. Shifting to a capacity market is unnecessary, and would in reality be a source of inefficiency and a barrier to competition that would likely increase the cost of electricity for consumers.

I. Introduction

In the U.S. and around the world, electricity restructuring is converting regulated monopolies into market regimes. The characteristics of those markets, however, are critical determinants of their performance and remain the subjects of active policy debate. One important issue is whether electricity markets can—without government intervention—provide adequate generation to reliably power society's needs.

In the U.S. and around the world, electricity restructuring is converting regulated monopolies into market regimes. The characteristics of those markets, however, are critical determinants of their performance and remain the subjects of active policy debate.

Advocates for government intervention believe that electricity's special characteristics require capacity markets for reliability and efficiency; others see capacity markets as little more than mechanisms to transfer wealth to owners of otherwise uneconomic generation. Most U.S. regional transmission operators (RTOs) currently operate such markets or impose rules whose effects are similar. In practice, capacity charges are often substantial percentages of consumer bills, and their sizes have become political issues. In the northeastern United States, the governments of New Jersey and Maryland have mandated ratepayer-subsidized generation investments, ostensibly in order to reduce capacity charges.

This paper reviews the rationales for capacity markets and applies them to the Electricity Reliability Council of Texas (ERCOT), which requires only that distributors obtain sufficient energy to meet their load-serving obligations. Recent perceived slowdowns in generation investment in Texas are said to threaten ERCOT with dangerously low reserve margins, a situation compounded by high load growth and an increasing presence of intermittent wind generation that requires substantial support from conventional generators. Further, unlike other RTOs that can reach beyond their territories for additional supplies, ERCOT's operations are largely confined to Texas. This crisis, if indeed it is one, has divided policy analysts. Some see the current "energy-only" regime as sufficient to incentivize adequate and timely investments in generation, while others claim that capacity markets or similar interventions are necessary if competition and reliability are to be maintained.

Our examination of ERCOT's history and operation brings a conclusion that the costs of instituting capacity markets in its territory will almost surely exceed any benefits they might bring.

The most important rationales for capacity markets assume "market failures" that leave the "private" returns to generation investment lower than the "social" returns such as systemwide reliability that markets do not properly price. We find this analysis overly narrow because it does not examine the ability of peaking plants to supply ancillary services such as reserve capacity. Its logic also depends critically on an assumption that demand-side response in power markets is insufficient to prevent blackouts when electricity demand exceeds supply. In reality, demand response has become an important and growing force in power markets.

Our examination of ERCOT's history and operation brings a conclusion that the costs of instituting capacity markets in its territory will almost surely exceed any benefits they might bring. Those costs have proven substantial in other regions. In the Pennsylvania-New Jersey-Maryland Interconnection (PJM), capacity charges in 2010 added $140 per year to an average residential electric bill and $1,000 to that of a retail store. From the capacity market's 2007 inception through 2011, PJM retail customers paid over $50 billion in capacity charges, 93 percent of which went to owners of existing generation and only 1.8 percent to new and reactivated units. Had they been spent directly on new capacity, the funds could have purchased 129 combined-cycle gas-fired generators, each with 400 megawatts (MW) of capacity (American Public Power Association 2012, 1). Important differences between PJM and ERCOT, however, make it hard to forecast the latter's performance with a capacity market in place. ERCOT, for example, has yet to implement demand management on the scale of PJM. In addition, certain rules unique to ERCOT have reduced incentives for generation investment, as they have resulted in the use of reserve deployments to lower energy market prices at times when the economic scarcity of generation warrants very high ones.

The next section summarizes some characteristics of electricity markets that allegedly differentiate them from those for other goods and services and the economic theory that has been used to rationalize capacity markets as remedies for inefficiencies in energy markets and inadequate generation investments. Section III describes the difficulties that have been encountered in designing capacity markets and how their design and operation reflect the underlying regulatory politics. Section IV summarizes the relevant market institutions in ERCOT and how they operate in practice. Section V looks in more detail at allegations that

ERCOT faces major shortfalls in capacity investment and evaluates calculations purporting to show that generation investments are unprofitable. In fact, a more appropriate model of generator decisions shows that they are plausibly profitable. Section VI provides our conclusions.

II. Why Capacity Markets?

Is Electricity Different?

Capacity markets do not exist for goods other than electricity. There is, for example, no system of payments to pizza restaurants for merely having pizza ovens available. Milk drinkers do not pay surcharges to farmers to ensure that "cow capacity" is available. Someone who wants a hotel room for a certain night in the near future makes a reservation, but there is no payment for the hotel's capital costs. Any proposed justification of capacity markets for electricity should also explain why they are unnecessary or inefficient elsewhere.

The possible rationales for capacity markets rest on three physical properties of electricity. First, it cannot be stored at reasonable cost (except in hydroelectric facilities). Second, a grid operator must match production and demand instantaneously, either by altering generator outputs or by curtailing customers. A surplus of production over demand for a fraction of a second will overload lines, a deficit will produce instability, and either can blackout an entire region. Further, because electrical load varies greatly over the day in most locales, the operator must be able to control numerous power plants in anticipation of changes. The system operator must also make adjustments as system conditions unfold and have reserve generation either operating or ready to operate in the event a transmission line fails or a generator trips off the system. Third, unlike water or gas, electricity cannot easily be confined to a subset of the grid or forced to move on a particular line. Instead it flows over all interconnected lines in accordance with their relative resistances (impedances) at the speed of light, a phenomenon known as loop flow or parallel flow. Without centralized control, loop flows may overload some lines and underload others, again with a risk of blackouts.

Under electricity's traditional regulatory regime a monopoly utility was responsible for both day-to-day reliability and for investing prudently in generation for the future. Regulation promised cost recovery and reasonable returns to investors thanks to the utility's legal monopoly.

Under electricity's traditional regulatory regime, a monopoly utility was responsible for both day-to-day reliability and for investing prudently in generation for the future. Regulation promised cost recovery and reasonable returns to investors thanks to the utility's legal monopoly. Adequate investments, however, had to include high cost marginal generation capacity that would operate only during a handful of annual peak hours. A utility that owned a full fleet of generators could recover peaking costs as part of its overall regulatory "revenue requirement." In an unregulated system, however, private investors in peaker units must depend on high prices in a small number of hours to recover and earn a return on their capital. Random factors, such as weather and market conditions, may only add to the risks of these investments. Generation investors bear these risks, but inadequate capacity also leaves end-users at risk of blackouts that raise the costs of their own consumption and production.

On the other hand, most power markets have RTO-imposed or regulatory caps on prices and allowable supply bids by generators. At system peaks, price ceilings may mitigate the harm that results if marginal generators choose to exert their market power and raise the price of power from all operating generators. If peaks occur infrequently, however, price controls reduce generator revenues in periods of real resource scarcity and thereby reduce their investment incentives. Some have claimed that such revenue shortfalls (in industry parlance, "missing money") constitute another rationale for corrective intervention in the form of a capacity market.

Most U.S. RTOs have chosen to impose some type of capacity requirement on retail energy providers (REPs), whether they are competitive distributors or divisions of vertically integrated utilities. One variant is a "Resource Adequacy" requirement like that of the California Public Utilities Commission, which compels every REP to own or have contractual access to capacity that will cover its annual peak plus an additional margin of worst-case reserves. Another variant is found in RTOs that include PJM and the New York Independent System Operator (NYISO), which operate forward "installed capacity" (ICAP) markets where resource-deficient REPs can obtain capacity for compliance at defined future dates. Unlike either of these, ERCOT has neither a resource adequacy requirement nor a capacity market. REPs and generators are free to contract bilaterally as they wish, but if they do not they can transact energy at spot market prices. The non-existence of capacity requirements in ERCOT offers a unique opportunity to analyze their possible value by examining what happens in their absence.

Any economic rationale for capacity markets or resource adequacy programs must rest on a finding that absent such institutions markets will "fail" and yield economically inefficient outcomes.

Risk, Externality, and Strategy as Rationales

Any economic rationale for capacity markets or resource adequacy programs must rest on a finding that absent such institutions, markets will "fail" and yield economically inefficient outcomes. Further, there should be a showing that the choice of a capacity market or adequacy requirement is superior to economically feasible alternatives because it maximizes net economic benefits. Any chosen intervention must take account of electricity's physical properties that we previously discussed. The physical difficulties are compounded by economic inefficiencies in markets that have not been allowed to innovate and mature because of decades of heavy regulation.

For instance, although "smart metering" technologies are emerging, in most jurisdictions only the very largest users see and pay the instantaneous cost of their power. A lack of communications technology also makes it impossible for most customers to inform providers about their willingness to pay for reliable service and the compensation that would make lower consumption tolerable. Even if communication is possible, only the newest technologies allow providers to selectively disconnect and reconnect customers. The utility whose only alternative is a widespread blackout creates an economic misallocation because customers with a high willingness to pay for reliable service will lose it, while others whose lights stay on may have required little compensation to tolerate some darkness (Cramton and Ockenfels, 2011). Finally, in competitive energy markets the "missing money" problem may suppress generation investment. If market price of power cannot rise above the marginal cost of the least efficient generator, that unit will be unable to recover its fixed costs. In this extreme case, the supply curve will unravel as inadequate revenues drive peaking generators out of the market.

In the following sections we argue that *all* of these rationales are losing their relevance, and that even if relevant, there are generally less drastic remedies than capacity markets. Before examining them we note that some other possible justifications have also lost some of their force. The first of these is risk aversion. Joskow (2008), among others, has said that investors in an energy-only market may under-build peaking capacity because they must rely on rare and unpredictable price spikes for recovery of their capital. In most other markets, however, investment and the returns to investors will account for risks. As capacity runs short the increasing severity and frequency of price spikes will ultimately induce investment. Given the greater risks, investors will have to be compensated with higher average returns. There are few obvious differences between the types of risk faced by generation builders and investors in industry-specific capital elsewhere. Further, numerous institutions facilitate the reallocation of risks among investors, most obviously portfolio diversification. A capacity requirement may smooth income streams but allocates investments to capacity that would otherwise go to more valuable sectors.

Another argument for capacity markets rests upon differences between the private and social returns to generators. Assume that an RTO allows competitive retailers to choose their individual mixes of generation and reserves by vertical integration or contract. All retailers but one have made resource arrangements that suffice for their loads and reserve requirements. The private and social returns differ because the remaining REP can under supply itself in the knowledge that the RTO will allow it to free ride on the excess resources of others, a situation that will tempt every retailer to be a free rider and raise the risk of blackouts.[1] In reality, there are numerous ways to resolve this problem. Retailers may be bound by quantitative rules if they are to participate in the market at all, a liability rule can allow the other retailers to recover damages from the deviant, or the RTO can operate a balancing market like ERCOT's in which REPs can make up their shortages and unload their surpluses. A final possible justification for capacity markets comes from the economic theory of oligopoly. Assume that with oligopolistic interdependence a single large generation investment can affect market price and the profits of all sellers. If so, the timing of investments and evolution of the industry's capital stock may be inefficient relative to a competitive market.[2] The details depend on the underlying assumptions. One possibility is that investments will be delayed because each seller understands how its actions affect price. Another is that a dominant seller may invest at an inefficiently early date to deter the entry of competitors. In reality, most generation markets pass antitrust screens for competitiveness, but for those that fail, a capacity market is only one of many policy options.[3]

In reality, most generation markets pass antitrust screens for competitiveness, but for those that fail, a capacity market is only one of many policy options.

Supply Shortages

At any moment production of some good may be less than demand, but if price can adjust upward the shortfall will be transitory. Quantity demanded will fall and quantity supplied will rise to restore equilibrium. In electricity, however, most producers can see and respond to changing prices but consumers cannot. Regulated rates to consumers will be set to recover average costs and not vary with marginal costs, and consumers may not be able to see prices and adjust consumption in real time as they can in many other markets. If peak prices cannot ration demand the RTO may have to cut off geographic blocks of consumers without regard for their valuations of reliable power. Those with a high willingness to pay may be cut off while those less willing enjoy power that they would not have consumed had prices been visible and flexible. The economic loss ("deadweight loss") from such an allocation of power that does not accord with consumers' valuations might then be smaller were each retailer required to hold a quota of capacity, possibly purchased in an RTO-organized market. An always-adequate supply of power eliminates inefficient service cutoffs, but at the cost of acquiring and holding a substantial amount of seldom-used capacity.

The appeal of this reasoning about misallocations and capacity requirements is declining as technologies and markets change. It assumes that few customers can change their consumption on short notice, when in reality increasing numbers of them can observe and adapt to time-varying prices. New rate designs can induce lower consumption when capacity is constrained. In some jurisdictions consumers can choose among types of interruptible service with differing notice provisions and ceilings on annual interruptions. Some interruptible rate schedules include "ride-through" provisions that allow continuous service for a large premium. As the Smart Grid[4] becomes operational, some regulators are allowing utilities and competitive retailers to offer optional rates that allow them to remotely control customers' appliances. Efficiency does not require universal quick response capabilities that may not be cost-effective for small users. Supply/demand gaps are usually small percentages of load and happen infrequently. As demand response becomes more substantial and timely, inefficient rationing of shortages becomes a less persuasive rationale for capacity markets.

Market Power, Price Caps, and Missing Money

Around system peaks, most available capacity in an RTO market will either operate or be committed to ancillary services. In such circumstances generators may be able to earn supernormal profits by bidding more than marginal cost, particularly if demand response is not an important force. The RTO thus faces a dilemma. It wants to see prices that do not evidence short-term monopolistic profits, while at the same time ensuring that returns are high enough to induce investment, all in an environment where the number of peak hours cannot be reliably predicted. Thus, virtually all RTOs have chosen to impose price caps as attempted compromises between efficient and monopolistic incentives. PJM, for example, has a price cap of $1,000/MWh.

At system peaks, price ceilings may mitigate the harm caused if marginal generators choose to exert their market power (and raise the energy price paid to all operating generators). Price caps, however, will produce less revenue for generators in situations of real resource scarcity and thereby reduce their investment incentives. This "missing money" problem caused by the intervention of ceilings may itself require correction by some other intervention, such as a capacity market that will compensate generators for energy market revenue shortfalls. This argument that capacity markets are necessitated by price caps rests upon the assumption that price caps are binding and have a significant effect on generation revenues. ERCOT, however, is in the process of raising its cap to $9,000/MWh. Thus, this aspect of "missing money" appears to apply far less to ERCOT than to other RTOs.

Another form of the missing money argument rests on assumptions about cost characteristics of generators. A representative theoretical model assumes a competitive energy market supplied by baseload generators with lower energy costs and higher capital costs than the peaking units that supply the remainder.[5] All plants of a given type have equal marginal costs which are constant up to capacity. In high demand situations both types are dispatched. At the market price a baseload unit will receive more than its marginal cost for each kwh produced, but a peaking unit will earn no more than its marginal cost (absent spikes due to overall capacity limits). In low-demand situations, peaking units do not operate and baseload plants recover only their marginal costs. In this model investment incentives are weak or nonexistent for peaking plants. One remedy is familiar: put all generation under a regulated monopoly whose rate structure brings in sufficient revenue to collect total cost, including a fair rate of return. Another possibility is to separate energy and capacity markets, with prices in the latter set by regulators to recover capital costs of the various plants.[6]

The relevance of this aspect of missing money depends on both technology and regulation. It is most likely if there are only a few generator types and the marginal cost of a given type does not increase significantly with its output. In practice, most RTOs operate with convex bid stacks that slope shallowly upward when loads are low and become significantly steeper when they are high.[7] In addition, typical missing money models unrealistically restrict the revenues available to generators. Considering only short-term markets, all RTOs allow bidding to supply ancillary services such as reserves and load following (regulation). A generator whose bid is accepted receives the market-clearing hourly price (which gives the RTO the right to dispatch it) and the spot price of energy if called upon to run. Payments for ancillary services may allow a generator whose marginal costs exceed the market price of energy to earn a profit.

III. Political Realities and Economic Models

Institutions and Valuation

Most markets evolve without central guidance as producers and consumers devise exchanges that are in their mutual interest. A capacity market is quite different because it must be imposed on parties who would otherwise not transact in it. Because capacity markets do not "naturally" emerge from an energy-only regime, they must be imposed. One cannot, however, assume that if expert opinion favors a capacity market the institutions of RTO governance will produce a successful design. All interested parties may sincerely intend to put the most efficient market institutions in place, but information and foresight are inaccurate. The various persons charged with designing the market may choose to use the process to advance their individual interests. Capacity markets cannot be built on a foundation of hope that an RTO's actual decision makers will uncritically accept the recommendations of experts rather than pursuing their individual interests. Nevertheless this appears to be the hope of some proponents.

> …the administrative targets must, of course, be set by sophisticated and politically and economically independent planning committees consisting of experienced engineers and economists.[8]

RTOs are nonprofit organizations whose books will register zero profits, but they are also institutions whose governing interests will seek to advance their individual interests. Policymaking must somehow aggregate the preferences of individuals whose interests are strongly opposed. The economic theory of collective decision making reaches one broad conclusion: in all but some quite special situations, there is little reason to expect that the policies chosen will be economically efficient (Michaels, 1999). If so, there is no clear reason for optimism that a system that includes a capacity market will outperform one of energy transactions and bilateral contracts that emerged by voluntary agreements among individual agents. If political governance is highly imperfect it is quite possible that leaving capacity decisions to unplanned markets is the better policy, even if the asserted market failures actually apply, and there is no reason to expect that either of the imperfect regimes provides more benefits than the other.

RTOs are nonprofit organizations whose books will register zero profits, but they are also institutions whose governing interests will seek to advance their individual interests.

Implementing Capacity Markets

Markets generate prices whose movements convey information about shifts in consumers' valuations and producers' opportunity costs. Because prices impact the returns to alternative choices, they induce resource owners to shift toward more profitable activities and consumers to economize on goods whose relative prices have risen. Capacity prices, however, do not emerge from a market but are instead deduced from a model of a "demand curve" that planners have devised in order to yield predictable returns to investors in generation. The construct bears no relation to the demand curve of elementary economics, which summarizes the valuations of voluntary purchasers in a market. This artificiality means that capacity markets will be indicators of economic scarcity only by accident. In principle, the price of a unit of capacity should measure the value of improved reliability that will result from investment in it. Value of lost load, however, is difficult to even conceptualize (an outage allows some activities to be postponed while others are lost forever) and it differs among consumers and over time.[9] Whether to improve reliability depends on marginal costs and benefits of doing so, and whether to use a capacity market depends on the costs and benefits of the alternatives. As technologies, costs, and demand change, so must the calculations.

Because a wide variety of resources might serve the capacity function, the measurement and enforcement of compliance are inherently complex. The more the generation and load management options, the greater the importance of accurately determining the relative values of different technologies and of load management relative to generation. Valuation is particularly difficult because capacity markets only exist by regulatory order, and if they vanish, the values of assets traded in them will fall. Capacity is most valuable if it is instantly available and operable with high probability. Operational delays and delivery risks, however, are matters of degree. The capacity equivalence of wind power and other intermittent renewables is a contentious issue because even aggregation over large areas may not greatly reduce intermittency.[10] Capacity can also take the form of callable demand reductions, which have recently come to dominate PJM's auction (Pfeiffenberger and Newell, 2012).

In practice, capacity market prices make only rudimentary distinctions based on objective criteria of availability and operability. A capacity price should be uniform and independent of market conditions if it is to smooth the income streams of generators and ensure recovery of fixed costs.[11] Uniformity, however, does not provide generators with marginal incentives to maintain readiness when it is most valuable. Because rewards are not price-based, RTOs have had to resort to complex and contentious administrative rules that define and govern availability. The existence and value of availability, however, can depend on the particulars of units and market conditions. A generator's willingness to make itself available can depend on fuel prices, and on both the value of startup costs and the time necessary to reach full operation. Operations of some generators are limited by environmental rules, or by maximum allowable water flows for hydroelectricity. As intermittent wind and solar capacity increase, more gas-fired plants must be on standby; but if gas pipeline capacity is constrained on winter peak days, regulators may give priority to residential heating rather than to the gas generators needed for reliability. Such considerations become more important if fluctuations in the output of wind generators in an area are positively correlated.

A capacity market is an attempt to impose a complex global solution on a relatively simple local problem. However relevant the missing money model may be in reality, it points directly back to localization.

Geography and transmission availability further complicate the measurement and incentivization of compliance. Any generator that is callable as capacity can randomly fail or become unavailable. In addition, availability also depends on transmission capacity and reliability. This implies that a generator in or adjacent to a load pocket will likely contribute more to reliability than a similar but more distant generator. This yields another pricing problem: the capacity market's administrators must somehow assign a capacity value to merchant transmission. If the structure of the grid changes, however, that value should properly change. Unfortunately, meeting the cost recovery objectives of a capacity market will leave transmission like it leaves generation—carrying a fixed price that bears little relation to its actual economic value.

Whatever their basis in economic theory, prices in the capacity market impact the relative profitability of investment in different types of generation and load management. Those prices, however, can also provide perverse incentives; for example, maintaining obsolete generation in order to capture capacity market revenues rather than to retire and replace it with more efficient units. Further, market participants may have risk management tools other than capacity ownership at their disposal. Fuller development of these options can mean that capacity markets will not fade away after they have outlived their usefulness. Owners of otherwise uneconomic generators that remain operable will attempt to protect their income streams, and the politics of RTO governance may allow them to survive.

A capacity market is an attempt to impose a complex global solution on a relatively simple local problem. However relevant the missing money model may be in reality, it points directly back to localization. There is no reason—either in theory or in history—to assume that the missing money will be important for every type of generator at some points in its life. There appears to be general agreement that if generation investment in Texas is in fact inadequate, that problem is with a small set of peaking generators and it exists for no more than 100 to 200 hours per year. If these are the problem units, any capacity policy should be directed toward *them* rather than instituting a vastly more complex policy that affects *all* generators along with demand response. In our next section we show that there are strong reasons to doubt the common assertion that investment in peaking generators in ERCOT is intrinsically unprofitable.

A capacity market can also affect outcomes in less regulated markets by affecting the rewards to investments in them. Some advocates of capacity markets view reduced energy price volatility as a virtue. If, however, a capacity market reduces price fluctuations relative to an energy-only regime, investments by users to reduce peak consumption become less valuable. If there are reasons (e.g., environmental) for investments that reduce peak demand, the operators of a capacity market will again need to formulate administrative rules to restore desired levels of demand response, such as allowing it to be bid in as a capacity resource. Our problem is that prices in a capacity market are of necessity determined by those same administrative rules and may have little relation to actual scarcities in either the short run or the long.

Advocates of capacity markets believe that RTOs that operate energy-only markets display performance inferior to those with integrated energy and capacity markets.

IV. Investment and Profitability in ERCOT

Markets in ERCOT

1. Hypotheses to Test

Advocates of capacity markets believe that RTOs that operate energy-only markets display inferior performance to those with integrated energy and capacity markets. The theoretical discussion above suggests several dimensions along which energy-only systems may fall short. The most obvious analysis would compare retail bills over the long term in energy-only systems and those with both energy and capacity markets. Such a comparison is currently impossible in Texas because retail service is highly competitive, with numerous suppliers offering many different programs. They differ in ways that include contract duration, wholesale price passthrough, resource mixes, termination and switching provisions, and options to buy conservation services, among others. Unfortunately for researchers, contracts between retailers and generators are confidential, as are data on the numbers of customers taking service under different rate schedules and their load profiles.

The available data, however, allow us to examine two oft-cited virtues of RTOs with capacity requirements that may be less evident in ERCOT. First, if the missing money is indeed a problem, we should see boom-bust cycles in investment in ERCOT. There are numerous potential scenarios, many centering on imperfect foresight by investors that could in principle be corrected by introducing a capacity market. At the start of a typical scenario a dearth of new generation pushes energy prices upward, but risk-averse investors do not respond until prices are well above their steady-state values. These uncoordinated investors then binge on a mass of new plants. Prices fall, investment falls, scarcity increases with the passage of time, and the story begins again.[12] Under a boom-bust model we should see alternation between relatively long intervals of excess reserves (relative to engineering estimates of adequacy) followed by intervals of shortage whose duration reflects both investor inertia and delays in planning and construction. In some RTOs with capacity markets, administrative forecasting and planning errors may in themselves lead to poorer performance.[13] Because an energy-only market aggregates the expectations of individual investors without forcing them to invest, it is possible that it better serves the interests of long-term resource adequacy. Competition among investors rewards those who can more accurately anticipate near-term resource shortfalls and may suffice to maintain more stable reserve margins over the long term.

The second potential difference between ERCOT and RTOs with capacity rules is also a consequence of the missing money. In years when excess generation capacity prevails, energy prices should cover little more than the variable cost of the marginal generator operating. If so, recovery of capital happens only when generation is in short supply. Such periods are not predictable and returns to a generator will be heavily influenced by randomness in weather and operating conditions. If this reasoning is correct we should see that in most years generators will not recover their annualized capital costs. Rather the recovery will occur, if at all, in unpredictable years with unpredictable price spikes. Thus, a showing that capacity market regimes are superior requires a preliminary showing that generators in energy-only markets fall chronically short of capital recovery. An accurate estimate, however, will require a fuller accounting for possible sources of generator income than has hitherto been made. Understanding the determinants of generator income requires knowledge of how ERCOT's markets function, a task we undertake next.

Between 85 and 95 percent of ERCOT's daily throughput moves under contracts between REPs and generators, sometimes intermediated by marketers. Generation owners and certain large loads are known as "Qualified Scheduling Entities" (QSEs).

2. ERCOT Markets and Operations

Between 85 and 95 percent of ERCOT's daily throughput moves under contracts between REPs and generators, sometimes intermediated by marketers. Generation owners and certain large loads are known as "Qualified Scheduling Entities" (QSEs). They can bid into energy and ancillary services markets, and may contractually commit themselves to produce or purchase power intended for resale to end users. Contract terms are confidential and unavailable to the PUCT or ERCOT. An REP assembling its power supply must choose base load, intermediate, and peaking resources, while factoring in the possibility of transactions at spot market prices to make up imbalances between contracted supplies and load.

Non-contract power trades in a "Day-Ahead Market," (DAM), a "Balancing Market," and in markets for several "Ancillary Services" required to maintain reliability. The DAM acts very much like a one-day forward market for power, in which participation is entirely voluntary. REPs submit their expected loads (which may vary with prices), while suppliers submit bids. A computer algorithm clears the DAM simultaneously with the ancillary services markets to minimize total cost. The Balancing Market sets prices for flows that result from unplanned supply-demand imbalances. Generators can also sell power into the balancing market and REPs can take the risks of under scheduling and seek bargains there. Balancing market prices are set every 15 minutes. ERCOT also operates markets for three ancillary services:[14]

1. ***Regulation Reserve*** is generating capacity that follows instantaneous load changes to maintain system frequency of 60 Hertz. Generators respond automatically to telemetered signals from ERCOT. There are separate markets for "Regulation up" (to increase output) and "Regulation down" (to reduce output).

2. ***Responsive Reserves*** are operating generators available to increase output when a generation or transmission failure occurs in a 10-minute period. ERCOT also allows certain loads to serve as reserves if they are willing and capable of being dispatched.

3. ***Non-Spinning Reserves*** ("Non-Spin") are generators not currently operating ("spinning") that can be ramped to a specified output within 30 minutes, or large loads eligible to act as reserves that are interruptible on 30 minutes' notice. Non-spin can replace lost generating capacity and also compensate for uncertainty when large amounts of other reserves are unavailable on-line (ERCOT Nodal Market Guide 2010, 15).

Generators who win the bidding are paid for allowing ERCOT to utilize their capacity as needed. If called upon to operate, the generator receives the balancing market price for its output, which may be above or below the generator's marginal cost. Adjustments to new information continue until actual operation.

In December 2010, ERCOT's system of four zonal prices was replaced by "Locational Marginal Pricing" (LMP), also known as Nodal Pricing, which recalculates prices every five minutes at 600 locations.[15] It uses generator bids and transmission congestion data to calculate the incremental cost of supplying an additional megawatt at each node. Nodal pricing can incentivize efficient patterns of investment. A price difference between two nodes often signals that an increase in transmission capacity between them can reduce total costs.[16] Such a difference could also indicate advantages to building low-cost generation or reducing loads at the higher-price node.

V. Generation Investment in ERCOT

Resource Adequacy and Investment Incentives

Supporters of a capacity market see recent events in ERCOT as illustrating the difficulties of ensuring generation adequacy in an energy-only market. ERCOT's December 2011 Capacity, Demand, and Reserves (CDR) report predicted manageable summer reserve margins of 12.1 percent in 2012 and 2013, but estimated that by 2014 they would fall to an unmanageably low 4 percent (*see Figure 1*). By May 2012, delays in bringing a new coal-fired plant online and the shutdown of 2,000 MW of coal capacity in expectation of new EPA rules had reduced the anticipated 2013 reserve figure to less than 10 percent. Seeing these figures, some regulators, politicians, and interested generation owners expressed support for the formation of forward capacity markets like those in PJM and NYISO in hopes of ensuring sufficient investment.

Figure 1
Projected Reserves from CDR Reports

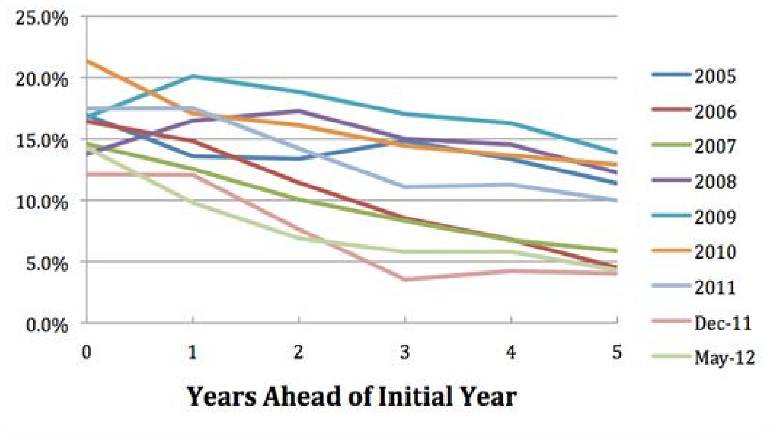

Figure 1 was compiled from CDR reports, which project reserve margins for future years. The zero point on the horizontal axis shows the CDR report's current reserve margin for the year's upcoming summer. Each year's line shows CDR projections for the next five years forward, which generally count only plants under construction or licensed. Perhaps unsurprisingly, they almost all slope downward and indicate significant worsening of reserve positions over the relatively near future. Figure 1 shows that the situation in early 2012 was hardly exceptional. Four of the CDR reports since the start of the series have projected re-

serve margins between 4 and 6 percent five years ahead. Their dates are 2006, 2007, December 2011, and May 2012.[17] Regarding the 2006 and 2007 reports, actual reserves five years later met ERCOT's adequacy criteria.

There is no clear way to distinguish in advance those resources that will be operational by a given date from those that will not, and risk-averse system planners may have reasons for conservatism. As one important example, the CDR reports identify "Other Potential Resources" not included in the projections. They include: 1) mothballed fossil fuel capacity, 2) 50 percent of available nonsynchronous DC ties with outside systems, and 3) planned generators in the full interconnection study phase.[18] As of May 2012, the projected 2016 total of these was 7,409 MW. Even if no other units are built between now and 2016, the 5,369 MW of additional generation required could suffice to restore a 13.75 percent reserve margin. The projected margin is also sensitive to assumptions about load growth and future demand management. The rates at which potential resources become actual ones also depend on assumptions about the prices of power and fuels, but the CDR reports do not consider alternative price scenarios.

The average U.S. electricity consumer can expect to lose 100 minutes of service per year due to outages from major storms. Sixty-seven percent of outage minutes were weather-related, generally affecting small areas.

The confluence of several unusual events rendered ERCOT's 2012 capacity situation somewhat extreme, but its dimensions differed little from those seen in some earlier years. Also like earlier times, by the end of summer 2012, the problem had greatly diminished. A court stayed EPA's rules and nearly 2,000 MW of mothballed gas-fired units were brought back into service. A more recent report by a PUCT Commissioner (Anderson, 2012) noted that since the start of 2012, 4,318 MW of new generation had been announced, most of which had either obtained financing or started construction.[19] Using these figures and assuming a low demand forecast (possibly reflecting growth in demand management) rather than ERCOT's high demand forecast would produce margins of 19.6 percent in 2013, 16.7 percent in 2014, and 13.2 percent in 2018, not counting any additional capacity that might materialize between now and those years.

We are left with several interim conclusions. First, assumptions about reserves in future years are likely to be quite conservative, disregarding both new investments and the convertibility of other resources such as mothballed plants. Second, there is no evidence that the behavior of investors is inherently destabilizing and that absent capacity requirements economically warranted power plants will go unbuilt. Similarly, the early years of ERCOT's existence were marked by excess capacity, in part a consequence of the regime it replaced. Third, there is no plausible way that a capacity market could have foreclosed the events that led to the shortfalls of early 2012. The same holds for its 2006 and 2011 episodes of rotating outages, both of which occurred with reserve margins above 15 percent (Anderson, 2012: 3). Fourth, policymakers must bear in mind that any reliability crisis is primarily the consequence of inadequate peaking, rather than baseload, resources (Anderson, 2012: 5). If there are valid reasons to subsidize peaking plants, they need not necessarily apply with equal force to baseload units.

There are great difficulties in forecasting capacity investments and evaluating their adequacy. Both advocates and opponents of capacity markets, however, acknowledge the importance of the logically preliminary question of determining a standard for adequacy. Unfortunately, there is no clear methodology for determining adequacy. The costs of ERCOT's 13.75 percent reliability standard may well exceed the economic value of lost load.[20] The criterion of “1 day in 10 years” is unclear and highly dissimilar load losses can satisfy it. If it means “one outage event in 10 years” the margin should be 14.5 percent, but if it means “24 outage hours in 10 years” that figure drops to 10 percent.[21] The actual value of lost load differs over geography and among customer types and little consensus exists regarding its value (London Economics, 2011). In Texas, as elsewhere, the bulk of outages occur on lower-voltage distribution lines rather than in the high-voltage system that a capacity market would impact. The average U.S. electricity consumer can expect to lose 100 minutes of service per year due to outages from major storms. Sixty-seven percent of outage minutes were weather-related, generally affecting small areas (Mansoor, 2013: 27). As noted above, there have been only two system-level emergencies since ERCOT’s founding, neither attributable to inadequate reserves (Anderson, 2012).

Generation Revenues in ERCOT

1. Peaker Net Margin

Any case for a capacity market or a resource adequacy requirement requires evidence that generator revenues without it are either insufficient on average or too risky to induce efficient levels and types of investment. In principle profitability is easy to determine. Assume for simplicity that the outlay for a generator is immediate while predictable operating costs and energy revenues accrue over the future. If the net present value of this stream (discounted at the risk-adjusted cost of capital) exceeds the initial outlay the investment is economically profitable. It is easy to find or approximate a plant’s capital, balancing energy and ancillary services prices, and market fuel prices. Unfortunately, 90 percent of the typical generator’s production is transacted under confidential contracts whose provisions are unknown to the public, regulators, and ERCOT’s market monitor.

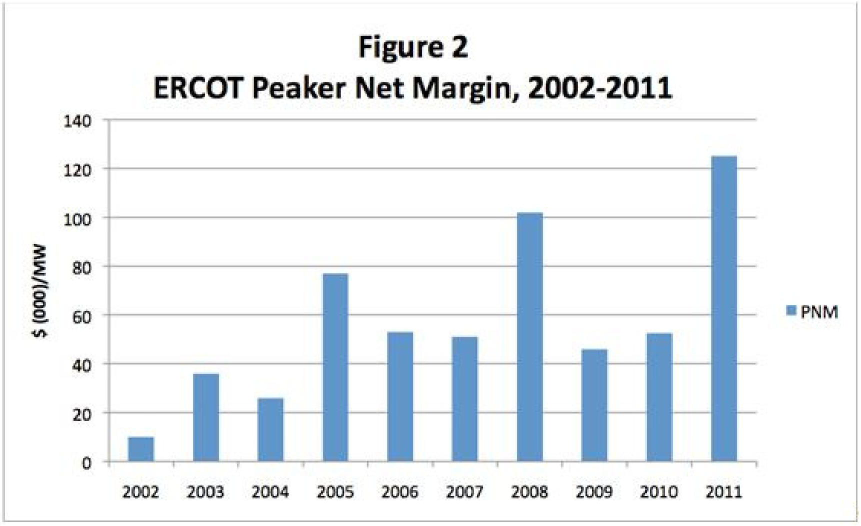
Figure 2
ERCOT Peaker Net Margin, 2002-2011

ERCOT’s monitor, however, is peculiarly constrained because the ground rules for its profitability calculation are embodied in regulations that do not allow the use of potentially relevant public data. The PUCT’s “Scarcity Pricing Mechanism” requires that the market monitor calculate Peaker Net Margin (PNM), an estimate of the net revenue a hypothetical peaking unit can earn from energy sales that can be compared with its operating cost.[22] PNM is the cumulated hourly difference between real-time energy price at an ERCOT hub and operating cost derived from a gas price index. The annual calculation begins on January 1 and is incremented by that difference for every hour that price exceeds cost (otherwise the increment is zero). ERCOT’s market monitor then compares PNM with its capital cost. **Figure 2** shows the end-of-year PNM for all years since the opening of the markets.

Because nuclear and coal-fired units with low marginal costs generally operate as baseload, the generation critical for reliability consists of gas-burning combined cycle plants and combustion turbines. The cost levels at which new units become “economic” have varied. ERCOT estimated that between 2002 and 2007 a combustion turbine with a heat rate[23] of

10 could cover all of its fixed costs (including carrying costs) by earning a margin of $60,000 to $85,000 per MW-year, a condition that held only in 2005.[24] In 2008 and 2009, the lower limit on viability ranged from $70,000 to $95,000 per MW-year.[25] The 2008 figure indicates profitability, but the market monitor attributes that return to inefficient transmission management and inefficient pricing of non-spinning reserves, both consequences of rules that no longer exist.[26] For 2010 and 2011, ERCOT raised the PNM threshold to between $80,000 and $105,000 per year.[27] The 2011 PNM surpassed this threshold, but ERCOT views it as an exceptional deviation due almost entirely to extreme weather during July and August.[28] For a combined cycle plant with a heat rate of 7, ERCOT estimated a 2010-2011 revenue requirement of $105,000 to $135,000 per MW-year.[29]

At this point we have a paradox. By ERCOT's calculations little if any new capacity should have been built over 2002-2011, but in reality fossil fuel capacity growth has kept pace with load. ERCOT's markets opened with a substantial surplus of generation, but if PNM calculations were in fact relevant, they indicate that between then and now its reserve position should have gone from adequate to disastrous. There is no discernible uptrend in bankrupt Texas generators or sales of generation assets at heavy discounts.

2. The Ancillary Services Option

To resolve the investment paradox we examine factors not accounted for by the PNM calculations. Most importantly, generators in ERCOT can supply *both* its balancing and ancillary services markets. The PNM formula calculates only net income from the balancing market when price there exceeds operating cost. Payment for ancillary services, however, comes in two parts. First, a successful bidder receives the market clearing per-hour capacity payment for the type of service being supplied.[30] Second, if that capacity is called upon to run its net income will be the difference between the real-time balancing market price and marginal cost. If the unit runs and price is below cost its owner is forced to take a loss. To simplify at the outset assume the generation owner has perfect foresight, i.e., it knows today's and tomorrow's balancing energy prices and the probability that its unit will be called. For present purposes we disregard complications that might result from the day-ahead clearing of ancillary services (and some energy) markets, as well as later decisions that deal with the Reliability Unit Commitment process. The generator must compare three scenarios:

1. The generator will supply energy if: the difference between the balancing market energy price and marginal cost is positive, and if it exceeds the expected net income from committing to ancillary services.

2. The generator will supply ancillary services (here assumed to be non-spin) if: the expected net income from committing to ancillary services is positive and exceeds the net income it gets at the market-clearing price of energy, our generator supplies ancillary services.

3. The unit will stay idle if: both net income from the balancing market and expected net income from ancillary services are negative.

Financial Effects of an Option to Provide Ancillary Services

1. Assumptions

For each year 2008-2010, **Table 1** shows the results of five calculations. Four are for a plant

Table 1: Generator Operating and Nonspin Reserve Hours

Year	Fuel Price	Heat Rate	Prob. Called	Energy Hours %	Nonspin Hours %	Idle Hours %
2008	Monthly	7	0.1	37.26%	19.72%	43.02%
2008	Monthly	7	0.2	36.66%	16.86%	46.48%
2008	Monthly	7	0.05	37.52%	22.55%	39.94%
2008	Ann. Avg.	7	0.1	35.35%	17.11%	47.54%
2008	Monthly	10	0.1	8.12%	29.16%	62.72%
2009	Monthly	7	0.1	28.68%	33.82%	37.50%
2009	Monthly	7	0.2	27.93%	25.55%	46.52%
2009	Monthly	7	0.05	29.02%	40.83%	30.15%
2009	Ann. Avg.	7	0.1	27.74%	34.27%	37.98%
2009	Monthly	10	0.1	7.75%	30.04%	62.21%
2010	Monthly	7	0.1	27.21%	30.63%	42.16%
2010	Monthly	7	0.2	26.84%	26.77%	46.39%
2010	Monthly	7	0.05	27.26%	33.39%	39.35%
2010	Ann. Avg.	7	0.1	19.38%	25.60%	55.02%
2010	Monthly	10	0.1	18.05%	30.69%	51.25%

Source: Authors' calculations.

with a heat rate of 7 whose marginal cost per MWh is 7 times (i.e., the heat rate) the price of gas per MMBtu, plus $4 in variable operation and maintenance costs, as assumed in the market monitor's calculation of PNM. The fifth calculation assumes a heat rate of 10, the same as the market monitor uses when calculating PNM.[31] We expect that net income will be sensitive to assumptions about fuel prices and about the probability that a generator that clears the ancillary services auction will be called upon to operate. We assume a base probability of 0.1 and analyze the effects of reducing it to 0.05 and raising it to 0.2.[32] Absent daily fuel price data from the PNM calculation, we consider two alternatives. The first assumes that the annual average gas price used by the market monitor prevails every hour of the year, and the second uses the corresponding monthly prices.[33] ERCOT's PNM calculation assumes that a generator will be unavailable due to scheduled or unavoidable maintenance 10 percent of the time. The division between these two types of maintenance, however, casts doubt upon a blanket 10 percent assumption.

In all of our estimates there are intervals of two weeks or longer in which the generator's best strategy is to remain idle, and during which normal maintenance may be feasible. We do not, however, have usable data on forced outages and their dependence on generator characteristics. Our estimates for 2008 and 2009 assume, perhaps unrealistically, that the generator is available to operate whenever doing so would be profitable. (Though calculations of PNM make the same assumption with respect to availability for energy markets.) The 2010 estimate uses only data from January through November. ERCOT has not yet produced comparable data for December and subsequent months, the period during which nodal pricing went into effect.

Table 2: Generator Net Income From Energy and Non-Spin Markets

Year	Fuel Price	Heat Rate	Prob. Called	Energy Income	Nonspin Income	Total Income	Nonspin Income %	Income Viability Range
2008	Monthly	7	0.1	$ 181,905	$17,352	$199,258	8.71%	$60,000 to $85,000
2008	Monthly	7	0.2	180,891	17,491	198,381	8.82%	
2008	Monthly	7	0.05	182,280	17,750	200,030	8.87%	
2008	Ann. Avg.	7	0.1	199,740	15,964	201,239	7.93%	
2008	Monthly	10	0.1	121,959	23,279	145,339	16.02%	
2009	Monthly	7	0.1	$65,751	$12,964	$78,715	16.47%	$70,000 to $95,000
2009	Monthly	7	0.2	65,249	12,302	77,551	15.86%	
2009	Monthly	7	0.05	65,992	13,734	79,727	17.23%	
2009	Ann.Avg.	7	0.1	66,765	13,281	80,047	16.59%	
2009	Monthly	10	0.1	47,104	12,990	60,095	21.62%	
2010	Monthly	7	0.1	$79,482	$20,129	$99,611	20.21%	$80,000 to $105,000
2010	Monthly	7	0.2	78,660	19,497	98,158	19.86%	
2010	Monthly	7	0.05	79,677	20,905	100,583	20.78%	
2010	Ann. Avg.	7	0.1	60,179	25,222	85,401	29.53%	
2010	Monthly	10	0.1	57,635	19,852	77,487	25.62%	

Source: Authors' calculations.

2. Operating Hours and Net Income

Table 1 shows the percentages of total annual hours under our various assumptions during which a generator will produce energy, make itself available to provide non-spin service, or remain idle because neither is profitable. In 2008, 2009, and under one set of assumptions for 2010, the generator with a heat rate of 10 will unsurprisingly be idle more hours than one with a heat rate of 7.[34] The reduction in hours that the unit with a heat rate of 10 produces energy is very pronounced in 2008 and 2009, but the percentage of hours during which it supplies nonspin exceeds that of a unit with a heat rate of 7 in 2008 and falls short of it in 2009. The less efficient unit will have lower or negative margins in producing energy but if the probability of being called is low it can still earn a positive expected margin in ancillary services.

Table 2 shows generator net income for each of the three years under the same assumptions about heat rates, probability of call, and fuel prices used in Table 1. The total consists of net income from energy sales and from the provision of non-spinning capacity and energy when the unit is called on. Taken as a group they indicate that estimates of profitability can be quite sensitive to what is assumed about the range of markets open to a generator. While expanding that range improves the generator's revenue picture, our calculations are unexpectedly insensitive to the probability that a unit that clears the ancillary services auction will operate. Varying the probability of a call between 0.05 and 0.2 has a negligible effect on both income from supplying non-spin and income from producing energy in all years, and only in 2009 does the percentage of hours supplying non-spin vary substantially with the probability of a call, from 41 percent of the total (if probability = 0.05) to 26 percent (if probability = 0.2).

In 2008, under all assumptions the combined cycle plant with a heat rate of 7 exceeds the Market Monitor's estimate of required peaker net market margin. As noted above, ERCOT's market monitor views 2008 as an outlier, its high figures the result of inefficient transmission management and inefficient pricing of non-spinning reserves which are peculiar to that year. In 2009, revenues from the energy-only market are insufficient to meet the "adequate" PNM level. Once revenues from non-spin services are added in, however, revenues exceed the lower bound of $70,000 of the estimated PNM. The same holds true for 2009, where revenues from energy alone are insufficient to meet the market monitor's threshold, but sufficient revenues are available once non-spin opportunities are included. The net income calculation is only sensitive to the choice of monthly versus annual average fuel cost in 2010, for reasons we cannot ascertain.

Computations that assume an ancillary services option and rational bidding behavior cast new light on the viability of ERCOT's energy-only market. Peaker Net Margin is an administrative creation that does not fully reflect the economic opportunities open to generators. Our calculations show that adding the option of producing non-spin to that of the balancing market can raise a peaking generator's net margin in ERCOT into the range of economic viability. Checking and extending these conclusions will require inclusion of such additional details as the effects of being able to supply several possible ancillary services and the consequences of LMP pricing.[35]

Operators of electricity markets have devoted great effort to ensuring the efficient dispatch of generation that will produce the quantity that users demand at minimum cost.

We note that the 2010 and 2011 PNM figures are lower than they otherwise would be due to ERCOT actions in calling non-spin generation capacity. The problem arose because ERCOT was repeatedly bringing non-spin capacity into the energy market at an effective bid price of zero to correct forecast errors. This behavior has resulted in "price reversals" that at times depress price significantly, at times when scarcity should be driving it upward. Available data indicate that price suppression due to deployment of non-spin has hardly been unusual. While the exact definition of a relevant interval for price suppression is to some extent arbitrary, independent power producer Calpine has assembled a list of 54 incidents of non-spin deployment between December 6, 2010 and May 28, 2011 (Calpine, 2011). These contain 15-minute balancing market prices before, during and after the non-spin deployment, as well as quantities of capacity called upon. Over this six month period, non-spin was deployed for a total of 184.7 hours. The fall in market price between the last period without non-spin and the first period without it averaged $134/MWh. After non-spin deployment begins, market price generally changes by little during the period it is in operation. The PUCT addressed this problem in late 2011 by requiring non-spin capacity to be brought into the market at a bid price of at least $120/MWh. Whether $120/MWh is the appropriate price for such resources is difficult to analyze. Moving forward, however, it is clear that the PUCT has at least partially improved the efficiency properties of its market.

Demand-side Participation

Operators of electricity markets have devoted great effort to ensuring the efficient dispatch of generation that will produce the quantity that users demand at minimum cost. Until recently, however, the price that consumers paid for that quantity at any instant might be only distantly related to the marginal cost of producing it. Today's metering and communication technologies allow consumers to see and respond to variation in marginal costs over the day and seasons. Programs and rate designs are encouraging demand response from those customers who can quickly and substantially affect system loads. As they proliferate, demand and supply will have the same symmetric roles in price determination that they have in markets for other goods. If both producers and consumers can respond to changing scarcities of power, some of the key arguments in favor of capacity markets will lose their relevance. Already most RTOs with capacity markets allow the use of verifiable demand response as a capacity resource.[36]

Demand response in ERCOT consists of contributions to various reserve services, some of which existed prior to restructuring.[37]

Regulation (up and down): These are loads that are automatically controllable by ERCOT, and require telemetry and four-second responses. Qualifying loads are also eligible to provide non-spin service.

Responsive Reserves: ERCOT allows up to 1,400 MW of loads as responsive reserves controlled by telemetry, but they cannot exceed 50 percent of the responsive reserve market. Suppliers must install underfrequency relays with instantaneous response and be able to manually interrupt their loads on 10 minutes' notice.

Non-spinning Reserves: Loads can participate as non-spinning reserves, and must also be callable by telemetry to supply the energy market with small increments of power known as "droop."

Emergency Response Services: ERCOT selects qualified loads, generators, and aggregations of loads and generators to supply incremental production and load reductions specifically for deployment in grid emergencies. Auctions take place every four months for supplies that will vary from about 200 MW for off-peak hours to about 1,800 MW on-peak. ERS may indeed bring the benefits of more load participation, but its operating procedures allow the potentially inefficient payment of different prices to loads and generators.

Pursuant to PUCT policy the transmission and distribution companies in ERCOT are currently installing "smart" meters for all retail consumers, which will further increase the potentials for efficient pricing and retail demand response. We cannot yet project the volume of consumer reaction to the new options, but note that ERCOT is currently preparing for substantial response.[38] If wholesale price volatility increases, this service should become more attractive. Unfortunately, the price distortions inherent in a capacity market or resource adequacy requirement would artificially reduce volatility and blunt incentives for more demand response.

The theoretical case for capacity markets is weak at best. Many of its arguments depend on oversimplified assumptions that are at variance with reality, particularly those that are necessary to produce the "missing money" phenomenon.

VI. Summary and Conclusions

The theoretical case for capacity markets is weak at best. Many of its arguments depend on oversimplified assumptions that are at variance with reality, particularly those that are necessary to produce the "missing money" phenomenon. Other possible market failures including the inefficiencies of nonprice rationing during shortages are becoming less relevant as markets develop, more users see prices based on marginal cost, and demand management becomes more widespread. A capacity market is an institution in which people have no choice but to trade a contrived good that has little or no economic value. This fact implies that the prices that prevail for capacity and amounts to be invested in it will be administratively set and have only a tenuous connection with economic efficiency. The "demand curves" seen in northeastern capacity markets are unrelated to those that measure consumer valuations in ordinary markets. There are great difficulties in ascertaining the contributions of various types of capacity to reliability, and for determining the value of deliverability. Even the most vocal of advocates for these markets have stated that their institutions and quantitative specifications must be determined by experts without actual market interests who will put theoretical ideals in place, a process quite at variance with the realities of RTO governance.

An examination of ERCOT's current state does not provide coherent support for radical change in its markets. Critics claim that ERCOT is falling behind on investments needed to maintain its reserve margin. In reality, the 2011-12 shortfalls are largely explicable as idiosyncratic, the results of political, regulatory, and weather events rather than economic ones. In most years of its existence, a three- or five-year projection would show ERCOT falling dangerously short of reserves, but market forces have invariably succeeded in restoring their generation adequacy. The most recent reports also indicate that market forces continue to operate, and that ERCOT is taking advantage of other options such as de-mothballing generation and augmenting demand response. Claims by critics that investment is persistently unprofitable in ERCOT's energy-only markets rest on a regulator-determined formula (Peaker Net Margin) whose definition deals solely with Balancing Market revenues and costs. Adding in potential revenues from the sale of ancillary services leads to a conclusion that peaking units are often economically viable investments. On the surface it appears odd that generation investment in ERCOT continues apace despite official calculations of its unprofitability. In reality a more detailed picture of the choices available to generators shows that building them for ERCOT's actual markets is often profitable.

The answer is not to abandon the market in favor of even more intervention, but to lessen intervention to allow the market to work more efficiently.

The problems that exist in ERCOT have multiple sources, two of which we examined in this paper. The first consists of certain rules that used reserve prices to reduce prices during periods of scarcity. The PUCT has partially addressed this question. The second is that demand management has yet to grow the institutions and attain the scale that would make it truly symmetric with supply in setting prices. The importance of both these problems should not be minimized.

However, these problems do not stem from any inherent flaws in electricity markets that render them incapable of functioning properly. Instead, they are a result of intervention that has inhibited—or prohibited—innovation and kept the market from developing solutions to these highly complex issues. The answer, then, is not to abandon the market in favor of even more intervention, but to lessen intervention to allow the market to work more efficiently. Some will dismiss this approach. But as we have shown, a realistic comparison of the performance of the energy-only market with that of an actual (rather than theoretical) capacity market performance proves that the case for retaining ERCOT's energy-only regime is a strong one. ✯

Endnotes

[1] Regionwide blackouts are actually rare. Most outages occur on low voltage distribution lines and cannot be prevented by adding generation. As noted below, ERCOT has never had a system-wide grid collapse in its history, and a capacity market would not have affected its only two episodes of rotating outages.

[2] For examples of such models, *see* Milstein and Tishler (2012), and Murphy and Smeers (2005).

[3] FERC only allows market-based wholesale prices by a single producer (as opposed to regulated cost-based ones) if it has found an absence of market power using measures of concentration and pivotal supplier positions. Order 697-B, Market-Based Rates for Wholesale Sales of Electric Energy, Capacity and Ancillary Services by Public Utilities, 125 FERC ¶61,326 (12 Dec. 2008).

[4] For a discussion of the Smart Grid, *see* Blumsack and Fernandez (2012).

[5] Cramton and Ockenfels (2011). Similar models with different assumptions about technical change and investment may sustain a competitive equilibrium. For an example *see* Olsina and Garces (2006).

[6] Note that demand response does not cope with the missing money problem in the model with two generator types and constant marginal costs. Whether price is high or low, some generator will have a problem breaking even.

[7] *See*, e.g. the illustration of a typical bid stack in PJM Interconnection (2012) 11.

[8] Cramton and Ockenfels 2011, 21. This is an instance of what has come to be known as the "nirvana fallacy," an assumption that if one can envision perfect policies they will necessarily come into being through existing institutions, and have the desired effects without unintended adverse consequences. *See* Demsetz 1969.

[9] For an introduction to the difficulties of estimating value of lost load, *see* London Economics (2011).

[10] For system planning ERCOT rates an effective MW of wind capacity at 8.7 percent of its nominal value. How to quantify the energy value of wind capacity is still a matter for debate. *See*, e.g. *Forbes*, et al. (2012).

[11] However desirable a steady price for capacity, the actual history of capacity clearing prices in PJM's base residual auctions has fluctuated unpredictably by several hundred percent between its 2007 inception and the latest 2015 allocation. *See* Wilson (2012) 6.

[12] Note that this scenario is so general that it might apply to almost any capital-intensive industry, raising the question of why electricity alone is worthy of interventions on the scale of capacity markets.

[13] Wilson (2012, 12-14) claims that overforecasting of loads has been a chronic problem in the implementation of PJM's capacity market, with a corresponding overpricing of capacity in the farther term.

[14] Other ancillary services are procured directly by contracts between suppliers and ERCOT. They include "black start" generating capability to reinitiate operation after a blackout, "Reliability Must-Run" generation whose costs are above-market but must operate at certain times to maintain system stability, voltage support services, and emergency interruptible load service in which contracted loads allow themselves to be terminated in a system emergency. *See* ERCOT Nodal Market Guide (2010) 15.

[15] Hogan, 1992. For a simplified introduction *see* ERCOT Nodal Market Guide (2010) 18.

[16] There is a potential problem here: If transmission is most economically built in large increments, the price difference between the nodes will fall to zero and investment will either not take place or be inefficiently small. For a discussion of policy incentives to invest in this case *see* Hayden and Michaels (2006) and Doucet et. al. (2013).

[17] CDR Reports are normally annual. The six months between December 2011 and May 2012 are the only such interval with two reports.

[18] The remaining 50 percent of those ties are already defined as resources and included in CDR reports.

[19] This generation was in addition to capacity from returning mothballed units.

[20] Another argument is that investment may be low because the current period is one of greater uncertainty than has prevailed in the recent past, due to factors that include environmental regulations, the future of shale gas, renewable quotas, and other regulatory issues. If so, the option value of investment deferral may be particularly high today. *See* Michaels (2012).

[21] *Brattle Report 2012*, 100-101. The figures are based on a study of a 40,000 MW system, but RTOs also use a range of other criteria.

[22] PUCT Substantive Rules 2010, § 25.505, 6-197.

[23] The heat rate times the price of natural gas equals the marginal cost of the generation facility.

[24] 2007 State of the Market Report, 47.

[25] 2009 State of the Market Report, 67.

[26] 2008 State of the Market Report, 66-67.

[27] 2010 State of the Market Report, 45.

[28] 2011 State of the Market Report, 82-83.

[29] 2010 State of the Market Report, 45. ERCOT assumed variable operation and maintenance costs of $4 per MWh for both types of plant, and that each would be unavailable for maintenance or outages 10 percent of the time.

[30] We have not explored the consequences of alternative ways to specify the generator's preferred constellation of ancillary services to offer and the effects of the recently introduced "co-optimization" process for pricing the various services in accordance with their quality. In particular, we do not consider whether a generator would wish to supply regulation or responsive reserves. Prices of capacity held for these services normally exceed those of non-spin capacity, but non-spin has a higher probability of being called. One expects that competition will equalize the net returns among the various services, but evidence is currently lacking. We also disregard the possible effects of ramp rate constraints and the costs of starting and restarting. Similar limitations, however, also apply to estimates of revenues from supplying energy.

[31] The $4 operation and maintenance costs are assumed in the State of the Market Report's all-in price per MWh for all years. (e.g. 2010 State of the Market Report, 44) The Market Monitor's calculation of Peaker Net Margin, however, assumes a combustion turbine with heat rate of 10 and no non-fuel variable operating or startup costs. *See* 2010 State of the Market Report, footnote 15.

[32] Between 2003 and 2007 the State of the Market Reports show that annual percentages of non-spin actually deployed ranged from 3.2 to 6.5 percent. No later figures are readily available.

[33] Average annual gas price in 2008 was $8.50/mmbtu (2009 State of the Market Report, iv), in 2009, $3.74 (2009 State of the Market Report, iv) and in 2010, $4.34 (2010 State of the Market Report, iii). Monthly amounts were estimated visually from all-in price graphics (2008 State of the Market Report, viii; 2009 State of the Market Report, ix; and 2010 State of the Market Report, 2). 2008 prices ranged from approximately $12.00 in June to $5.20 in December, and the corresponding 2010 figures were $3.40 and $5.80.

[34] The exception in 2010 comes when we assume that fuel for a unit with a heat rate of 7 is priced at its average annual value on all days. The seeming anomaly may reflect no more than idiosyncratic data.

[35] ERCOT has not yet released any data at the necessary level of detail since the inception of nodal pricing in December 2010.

[36] Pfeiffenberger and Newell (2011) document that responsive load has become the largest component of new capacity in PJM auctions.

[37] Load Participation in the ERCOT Nodal Market (11 July 2007) 14-16.

[38] ERCOT staff appear to be preparing for a significant amount of this type of demand response. *See* Wattles and Farley (2012).

References

American Public Power Association. 2012. "Money for Nothing in the Power Supply Business," Issue Brief (Mar. 2012).

Anderson, Kenneth W. Jr. 2012. "Resource Adequacy in ERCOT," presentation graphics submitted in PUCT Project No. 40000 (Nov. 12, 2012).

Blumsack, Seth and Alisha Fernandez, "Ready or Not, Here Comes the Smart Grid," *Energy* 37:1, 2012.

Brattle Group. 2012. "Ercot Investment Incentives and Resource Adequacy" ["Brattle Report"], prepared for ERCOT. (June 1, 2012).

Briggs, R. J. and Andrew N. Kleit, *Resource Adequacy and the Impacts of Capacity Subsidies in Competitive Electricity Markets* (Oct. 22, 2012).

Cramton, Peter and Axel Ockenfels. 2011. "Economics and Design of Capacity Markets for the Power Sector," Working Paper, University of Maryland (Oct. 30, 2011).

Demsetz, Harold. 1969. "Information and Efficiency: Another Viewpoint," *Journal of Law and Economics* 12 (1): 1-22.

Doucet, Joseph, Andrew Kleit and Serkan Fikirdanis, "Valuing Electricity Transmission: The Case of Alberta" (with Doucet and Fikirdanis). Forthcoming, *Energy Economics*.

Electricity Reliability Council of Texas (ERCOT) "Texas Nodal Market Guide, Version 3.0" (Dec. 2010).

ERCOT. Various years. "Report on the Capacity, Demand, and Reserves in the ERCOT Region." CDR Reports (Dec.).

Forbes, Kevin, et al. 2012. "Are Policies to Encourage Wind Energy Predicated on a Misleading Statistic?" *Electricity Journal* 25 (2): 42-54.

Harvey, Scott. ERCOT Market Design, Capacity Markets and Resource Adequacy, Presented at Gulf Coast Power Association, Workshop on Resource Adequacy in ERCOT, Austin (May 4, 2012).

Hayden, J. Jolly and Robert J. Michaels. 2006. "Merchant Transmission Redux," *Public Utilities Fortnightly* (Sept.) 58-61.

Hobbs, Benjamin, et al. 2001. "Installed Capacity Requirements and Price Caps: Oil on the Water, or Fuel on the Fire?" *Electricity Journal* 14 (July) 23-34.

Hogan, William W. 1992. "Contract Networks for Electric Power Transmission," *Journal of Regulatory Economics* 4 (3): 211-242.

Joskow, Paul L. 2008. "Capacity Payments in Imperfect Electricity Markets: Need and Design," *Energy Policy* 16 (1): 159-170.

Kleit, Andrew. "Market Monitoring, ERCOT Style," *Electricity Restructuring: The Texas Story* (Kiesling and Kleit, editors), American Enterprise Institute (2009).

Load Participation in the ERCOT Nodal Market (July 11, 2007).

London Economics, Estimating Value of Lost Load (VOLL), Final Report to OFGEM (July 5, 2011).

Maggio, David. 2011. "ERCOT Nodal—How's it Going?" Texas Renewables Conference (Nov. 7, 2011).

Mansoor, Arshad. 2013. "Emerging Technologies Enable 'No Regrets' Energy Strategy," *Power* 157 (1): 20-29.

Michaels, Robert J. 1999. "The Governance of Transmission Operators," *Energy Law Journal*, 20 (2): 233-262.

Michaels, Robert J. 2008. "Electricity Market Monitoring and the Economic Theory of Regulation," *Review of Industrial Organization* 32 (2): 197-216.

Michaels, Robert J. 2012. "ERCOT: The Past and Future of 'Energy-Only' Markets," presented at Gulf Coast Power Association forum on Resource Adequacy in ERCOT (May 4, 2012).

Milstein, Irena and Asher Tishler. 2012. "The Inevitability of Capacity Underinvestment in Competitive Electricity Markets," *Energy Economics* 34 (1): 64-77.

Murphy, Frederic and Yves Smeers. 2005. "Generation Capacity Expansion in Imperfectly Competitive Restructured Electricity Markets," *Operations Research* 53 (6): 646-661.

"NJ Ratepayers to Pay Subsidy Under Contract," *Electric Power Daily* (May 31, 2012).

Olsina, Fernando and Francisco Garces. 2006. "Modeling Long-Term Dynamics of Electricity Markets," *Energy Policy* 33 (4): 1411-1433.

Pfeiffenberger, Johannes, Sam Newell. 2011. "Trusting Capacity Markets," *Public Utilities Fortnightly* (Dec.) 34-40.

PJM Interconnection. 2012. January-June Quarterly State of the Markets Report.

Potomac Economics, Ltd. (annual 2001-2011). "2010 State of the Markets (SOM) Report for the ERCOT Wholesale Electricity Markets."

Public Utility Commission of Texas, Various Dates. Electric Substantive Rules. PUCT Substantive rules § 25.505; ERCOT protocols (Sept. 1, 2010) 6-197.

Wattles, Paul and Karen Farley, "Price Responsive Load: Next Steps—Data Collection" (Oct. 16, 2012).

Wilson, James F. "Fundamental Capacity Market Design Choices: How Far Forward? How Locational?" Presentation Graphics, EUCI Capacity Markets Conference, Indianapolis (Oct. 2-3, 2012).

About the Authors

Dr. Robert J. Michaels, senior fellow of IER, is Professor of Economics at California State University, Fullerton and an Adjunct Scholar of the Cato Institute.

Dr. Michaels holds an A.B. from the University of Chicago and a PhD from the University of California, Los Angeles. His expertise is in the economics of industrial organization, and his research is centered on deregulation and the emergence of competitive markets in electricity and natural gas. He has been named Outstanding Professor in the Mihaylo College of Business and Economics and has served as Co-Editor of *Contemporary Economic Policy*, a major peer-reviewed journal. His research regularly appears in academic, industry and legal journals, including *Public Utilities Fortnightly, The Electricity Journal* and *Energy Law Journal*.

He is also a consultant who has advised and provided expert testimony on behalf of independent power producers, natural gas producers, power marketers, industrial electricity users, domestic and foreign electric utilities, regulatory commissions and public interest organizations (including IER). He has testified before the Federal Energy Regulatory Commission, California Public Utilities Commission, and other regulatory bodies, as well the U.S. House of Representatives. He frequently speaks on emerging economic and political issues at corporate and industry events. His column, "Power Moves" appears biweekly in Scudder Publications' *Energy Metro Desk*.

His more recent research includes work on electricity market monitoring. He has presented invited testimony before the Federal Energy Regulatory Commission in its ongoing rulemaking on market monitors. Other research includes several publications on "renewable portfolio standards" that will require utilities to purchase certain quotas of power from unconventional generation sources.

Dr. Andrew N. Kleit is a Professor of Energy and Environmental Economics and MICASU Faculty Fellow in Energy, Environmental and Mineral Economics in the Department of Meteorology in Penn State's College of Earth and Mineral Sciences.

As a professor in the department, Dr. Kleit teaches classes in environmental economics, energy markets, corporate finance, and financial risk management, and engages in research on environmental, energy, health care, and antitrust issues, as well as weather economics. Dr. Kleit is the Program Officer of the EMS College's undergraduate major in Energy Business and Finance, and the minor in Global Business Strategy. He was the faculty member in charge of establishing the EBF major. Begun in May 2004, the EBF major is an interdisciplinary course of study that draws on classes in economics, business, finance, and the earth sciences.

The major currently has approximately 230 students, making it the second largest major in the College of Earth and Mineral Sciences. Dr. Kleit received his B.A. *cum laude* in Mathematics and Political Science from Middlebury College in 1982 and his Ph.D in Economics from Yale University in 1987. He also received from Penn State a Master of Arts in 1983 and a Master of Philosophy in 1987.

Texas Public Policy Foundation

The Texas Public Policy Foundation is a 501(c)3 non-profit, non-partisan research institute.

The Foundation's mission is to promote and defend liberty, personal responsibility, and free enterprise in Texas by educating and affecting policymakers and the Texas public policy debate with academically sound research and outreach.

Funded by hundreds of individuals, foundations, and corporations, the Foundation does not accept government funds or contributions to influence the outcomes of its research.

The public is demanding a different direction for their government, and the Texas Public Policy Foundation is providing the ideas that enable policymakers to chart that new course.

Texas Public Policy
FOUNDATION

900 Congress Ave., Suite 400 | Austin, Texas 78701 | (512) 472-2700 phone | (512) 472-2728 fax | www.TexasPolicy.com