American Economic Review 2022, 112(2): 409–441
https://doi.org/10.1257/aer.20172034

Imperfect Markets versus Imperfect Regulation in US Electricity Generation[†]

By Steve Cicala*

This paper evaluates changes in electricity generation costs caused by the introduction of market mechanisms to determine production in the United States. I use the staggered transition to markets from 1999 to 2012 to estimate the causal impact of liberalization using a differences-in-difference design on a comprehensive hourly panel of electricity demand, generators' costs, capacities, and output. I find that markets reduce production costs by 5 percent by reallocating production: gains from trade across service areas increase by 55 percent based on a 25 percent increase in traded electricity, and costs from using uneconomical units fall 16 percent. (*JEL* L51, L94, L98, Q41, Q48)

When regulation brings its own host of distortions and inefficiencies, the mere existence of a market failure is insufficient to ensure government intervention will improve welfare. Instead, by comparing the distortions under potential regulatory regimes, one can identify superior policies as those with relatively fewer imperfections (Demsetz 1969, Kahn 1979, Joskow 2010). This paper undertakes such an evaluation in the context of US wholesale electricity markets, which have replaced command-and-control-type operations for over 60 percent of US generation capacity.

To do so I construct a virtually complete hourly characterization of supply and demand of the US electrical grid from 1999 to 2012. I observe consumption (or "load") and which generating units were chosen to operate to meet this demand at any moment in time for each of the 98 authorities charged with making production decisions in the United States (referred to as "power control areas" or PCAs). I combine generation unit-level hourly production with data on fuel costs, capacities, and heat efficiency to calculate the hourly variable cost of generation throughout the continental United States over fourteen years.

*Department of Economics, Tufts University (email: scicala@gmail.com). Liran Einav was the coeditor for this article. I am grateful to Gary Becker, Jim Bushnell, Thom Covert, Tatyana Deryugina, Edward Glaeser, Michael Greenstone, David Hémous, Koichiro Ito, Lawrence Katz, Ryan Kellogg, Erin Mansur, Kevin Murphy, Erica Meyers, Morten Olsen, Mar Reguant, Jim Sallee, Andrei Shleifer, Chad Syverson, Roberton Williams III, Matthew White, and seminar participants at Harvard, MIT, Cornell, Chicago, Yale, the UC Energy Institute, the EEE Session of the 2015 NBER Summer Institute, UIUC, Wharton, Brown, the 2017 European Econometric Society Meetings, Tufts, the University of Alberta, the 2019 American Economic Association Meetings, the University of Maryland, and Washington University for helpful comments and suggestions. Iván Higuera-Mendieta, Sébastien Phan, Julien Sauvan, Songyuan Ding, Enrique Chazaro-Acosta, Xianying Fan, Mary Vansuch, and Dan Pechi provided excellent research assistance. This paper has been reviewed by the Energy Information Administration to ensure no confidential data have been disclosed. All errors remain my own.

[†]Go to https://doi.org/10.1257/aer.20172034 to visit the article page for additional materials and author disclosure statement.

The abrupt and staggered manner in which some PCAs adopted market mechanisms while others did not, suggests that the impact of markets on costs can be evaluated with a straightforward difference-in-difference (DD) analysis. The invariance of load and generation capacity (in the short run) reinforces the credibility of such an exercise.

Trading power between PCAs makes production costs in isolation a problematic measure to evaluate the impact of markets. It is not possible to determine which generators are operating for the consumption of the local PCA, or for transmitting elsewhere. It will therefore appear that total costs are rising in PCAs that increase their exports after switching to markets. In PCAs that increase their imports, on the other hand, total observed costs will fall by the full value of reduced generation if one fails to account for the generation costs of imports that are borne elsewhere. Imbalanced changes in trade in the treatment group will yield a biased estimate of the net impact of markets on costs.[1]

To address this issue I decompose generation costs in a way that allows me to account for trade across PCAs. The central measure that makes this possible is what is known in the electricity sector as the "merit order." This is an idealized power supply curve that sorts installed generation capacity in order of increasing marginal cost regardless of whether the unit is operating or not. I show that realized costs can be separated into parts that can be calculated PCA by PCA relative to the merit order without knowing from where power is coming nor where it is going. I then perform my analysis on the components of this cost decomposition.

The first component I evaluate in the empirical analysis is the difference between observed generation costs and those of the lowest marginal cost units among installed capacity up to the quantity generated within each PCA. These are costs in excess of the merit order cost of generation, and are referred to as "out-of-merit" costs. Out-of-merit costs occur for a variety of reasons: when economical units go down for maintenance, when transmission lines are congested, or when peaks in demand are too brief to justify paying the start-up costs of a low marginal cost unit, among other reasons discussed in Section II.

Having isolated the part of costs that exceed the merit order, it is then possible to account for the costs and surpluses of trading power across PCAs using each PCA's merit order instead of estimating empirical supply curves. This simplifies the exercise because the merit order is straightforward to calculate from observed data, monotonically increasing, and always defined over quantities required to compute costs in autarky (which are necessary to measure gains from trade).[2] I refer to the savings from trading power according to the merit order as "gains from trade," and this is the second component of cost whose response to market dispatch I evaluate. To the extent that trade flows come from out-of-merit generation, the total surplus from trading power will entail a combination of net changes in out-of-merit costs and merit order gains from trade.

[1] This will occur when the counterparty to a trade is outside of the system (i.e., Canada), or outside of the short-term window for market onset (i.e., incumbent market regions or nonmarket regions).

[2] It is always possible to calculate merit order cost under autarky thanks to requirements that each PCA's installed capacity exceed peak load to avoid blackouts in the event of transmission failures.

I connect out-of-merit costs and gains from trade defined this way to the prior literature that evaluates the key tension in liberalization: the potential for increased efficiency versus the risk of deadweight loss due to market power. In electricity markets, market power manifests itself as out-of-merit generation when an economical unit is taken offline during moments of peak demand (Wolfram 1999; Borenstein, Bushnell, and Wolak 2002; Mansur 2008). It might also be exerted by an intermediate PCA blocking transmission access, reducing trade surplus. If these forces overwhelm any cost savings, I would observe a net increase in costs as measured by these two components. The purpose of the decomposition is to generate outcomes to estimate the causal impact of markets on costs, not to separate the change in costs into mechanisms. Increased coordination across PCAs may reduce out-of-merit costs, while the exertion of market power within PCAs may prevent trade flows.

I employ a DD framework to estimate changes in gains from trade and out-of-merit losses caused by the adoption of market dispatch. This approach finds gains from trade increase by 55 percent along with a 25 percent increase in electricity traded. There is also a 5 percent decrease in out-of-merit operations, and out-of-merit costs fall by 16 percent. I further evaluate the impact on nonmarket PCAs of having a neighboring PCA adopt market dispatch. This accounts for the possibility that control PCAs might be contaminated with new trading opportunities. I find the primary estimates for the effects of market dispatch are largely unaffected when accounting for these spillovers, which are small and statistically indistiguishable from zero. Evaluating the components of costs indicates markets caused a roughly 5 percent reduction in expenditures overall, which is less than the 8 percent one would find when ignoring trade flows.

It should be noted at the outset that my estimates measure short-run changes in how output is allocated given the installed capacity, costs, and patterns of demand. It would not be unreasonable to suspect that market dispatch has affected investment incentives in the longer run, which are likely to be an important source of welfare changes. In addition, my estimates measure the average effect of market dispatch, which itself has been heterogeneous both with respect to preexisting institutions (i.e., power pools, bilateral markets, or smoke-filled rooms), and with respect to the rules of the markets implemented (uniform or locational marginal prices, virtual bidding, market monitors, etc.). However, given the even greater differences between market and traditional dispatch methods, these estimates should be informative regarding the net impact of liberalization on allocative efficiency thus far.

The paper is organized as follows. I describe the structure of electricity generation and transmission in the United States, and the institutional details that will facilitate estimation in Section I. Section II describes how observed patterns of production can be decomposed into out-of-merit costs and gains from trade, which can be calculated without knowing the origin or destination of traded power. Section III describes the data. Section IV presents an estimation strategy motivated by this setting, using the components of the decomposition in Section II as outcomes. Section V presents causal estimates of the impact of markets on gains from trade and out-of-merit costs. Section VI concludes. There are two online appendices: the first provides greater detail on data assembly; the second evaluates the potential role of confounders and contains further sensitivity analyses.

I. Background on PCAs and Dispatch in the United States

The US electricity grid developed over the twentieth century based on a mix of investor owned utilities (IOUs), government-owned utilities (municipal, state, and federal), and nonprofit cooperatives. All of these organizations tended to be vertically integrated, so they owned the power plants, the transmission system, and the delivery network within their respective, exclusively operated territories. As either government-run or regulated monopolies, I use the term "command-and-control" to refer to the methods of dispatch under this industrial structure. The entity that has historically determined which power plants operate to meet demand is called a "balancing authority." A single balancing authority controls the transmission system and dispatches power plants within a PCA. When vertically integrated, the balancing authority and utility have often been one and the same, as with the service territory and the PCA.[3] These areas operate with relative autonomy over their assets, and transmission lines that connect areas enable flows between them.

The national grid consists of three large interconnections: East, West, and Texas (with relatively little capacity to transmit power between them). Figure 1 shows the approximate configurations of the US electricity grid in 1999 and 2012.[4] The boundaries between interconnections are denoted in panel A by the thick black lines separating Texas and the West (unchanged over the period). Each color family identifies separate regions of the North American Electric Reliability Corporation (NERC) that have historically coordinated operations in order to preserve the stability of the transmission system (when large plants go down for maintenance, for example). The tangle of power control areas delineated by white boundaries and separate shades within each NERC region reflects the legacy of local monopolies that have been the principal architects of the US electricity grid.

Although the Public Utility Regulatory Policies Act of 1978 opened the door for independent power generation (by requiring IOUs to buy their output at "avoided cost"), the growth of such producers was impeded by discriminatory transmission practices (Joskow 2000). Because the IOUs owned the transmission system, they could effectively shut independent producers out of wider markets by denying transmission access.[5] This began to change with the Energy Policy Act of 1992, which required the functional separation of transmission system owners and power marketers–they were no longer allowed to use their wires to prevent or extract the surplus from trades across their territory. These changes were codified on April 24, 1996 with Federal Energy Regulatory Commission (FERC) orders 888 and 889, which required open-access, nondiscriminatory tariffs for wholesale electricity transmission.

[3] Exceptions include the New York and New England Power Pools, which formed in response to the Great Northeast Blackout of 1965, as well as smaller utilities that do not control dispatch directly. Regional reserve margin coordination was also formalized during this time with the establishment of the National Electric Reliability Council.

[4] The exact geographic boundaries of PCAs often defy straightforward demarcation. This map is based on US counties, with the predominant PCA receiving assignment of the entire county—and is therefore approximate for visualization purposes. In addition, a number of small or hydro-only PCAs are merged with the larger neighboring areas that provide the majority of their (fossil-based) energy.

[5] Examples of IOUs exercising such market dominance can be found in Appendix C of FERC Order 888.

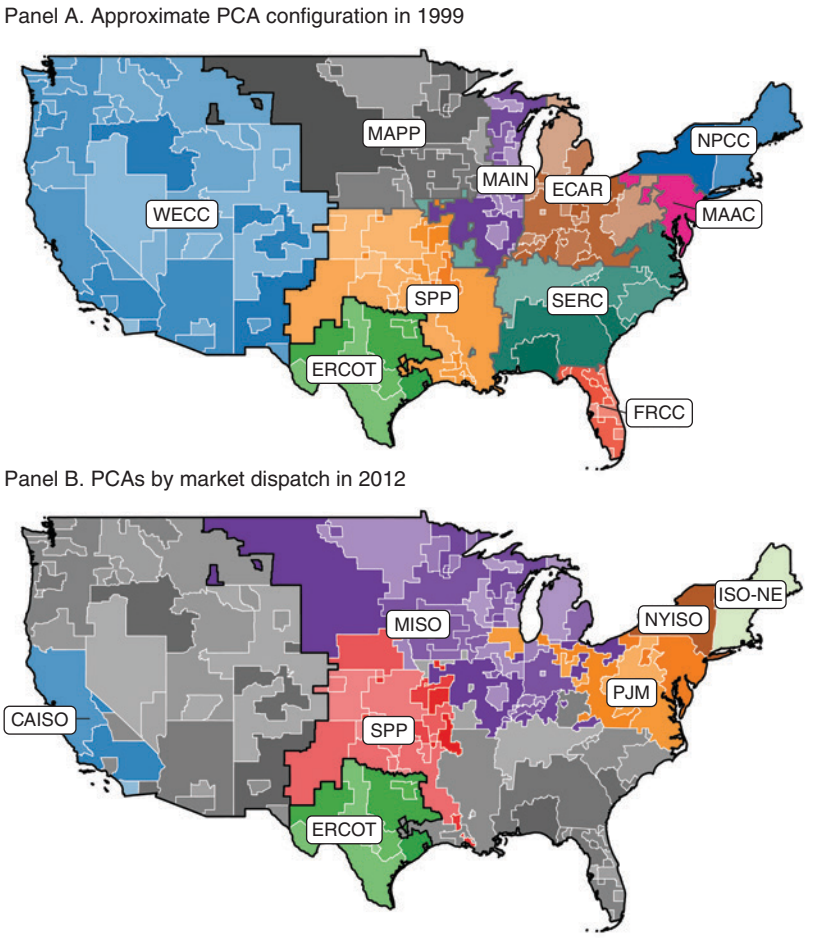

FIGURE 1. US ELECTRICAL GRID AS PCAS

Notes: Thick black lines identify interconnection boundaries. In panel A, white borders delineate PCAs and common color families denote NERC reliability regions. In panel B, common color families distinguish separate wholesale markets. Boundaries are approximate.

Open access created greater potential for wholesale electricity markets, which were initially conducted through bilateral contracts for power. In this decentralized setting, contracts would typically specify the amount of electricity to be generated by one utility under a set of conditions, transmitted across a particular area, and withdrawn from the system by the purchasing utility. Mansur and White (2012) give examples showing why the nature of congestion in electricity transmission networks renders decentralized markets particularly poorly suited for identifying all of the potential gains from trade. In particular, transmission lines are constrained by net flows of power. When this is the case, there are production externalities that may allow otherwise infeasible bilateral trades to occur by coordinating offsetting transactions to keep net flows below transmission capacity. Identifying such potential trades in this type of decentralized market is a challenge akin to coordinating simultaneous multilateral exchanges (Roth, Sönmez, and Ünver 2004).

Operationally, balancing authorities have relied on engineering estimates of costs to devise dispatch algorithms to determine which plants within the PCA operate, and separately schedule any other operations requested by utilities (for bilateral trades). Centralized wholesale electricity markets ("market dispatch") integrate dispatch operations into an auction for electricity. In day-ahead auctions, for example, generators submit bids to produce electricity, and only those below the price needed to meet projected demand are called on to operate. These auctions incorporate feasibility constraints, so calling on higher-priced units to operate due to transmission congestion allows for the direct revelation of the cost of shortcomings in the transmission system.[6] Day-ahead markets establish financial obligations to produce, which are subsequently either met with production in the real time market or unwound by buying back one's allocated output at the real time price (Wolak 2000, Hortaçsu and Puller 2008, Ito and Reguant 2016, Jha and Wolak 2013, Borenstein et al. 2008, Cramton 2003, among others).

As of 2012, 60 of the 98 PCAs operating in 1999 had adopted market dispatch, either during the initial creation of a new market or as part of the expansion of an existing market. Adopting market dispatch is a discrete change in the decision algorithm that allocates output to generating units: the local PCA cedes control of their transmission system to an independent system operator (ISO), who conducts the auctions.

All told, there were 15 distinct events in which PCAs have transitioned to market dispatch overnight between 1999–2012. Figure 2 denotes each of these events with a vertical red line, and shows that over the period of study markets have expanded from covering about 10 percent of US generation capacity to roughly 60 percent. The remaining areas have retained their traditional dispatch methods, though a number have continued to explore the possibility of joining existing markets.[7] This variation in market adoption forms the basis of the empirical strategy for causal estimates by allowing the comparison of changes in allocative efficiency following the transition to market dispatch relative to areas that have not undergone such changes over the same period.

The transition from command-and-control to market dispatch is related to, but distinct from the movement toward restructured electricity markets in the United States (Joskow and Schmalensee 1988). In particular, the changes to dispatch and transmission described thus far were undertaken by the federal government.[8] The end of cost-of-service regulation of vertically integrated IOUs was initiated by states. These state-led initiatives halted after the California electricity crisis, while the adoption of market dispatch has continued through the 2000s. It is important

[6] In particular, auctions using the "standard market design" yield "locational marginal prices" (LMPs) which denote the market-clearing price at each of the points of withdrawal from the system. When LMPs are identical everywhere, the system is said to be uncongested.

[7] For example, the East Kentucky Power Cooperative joined Pennsylvania-Jersey-Maryland (PJM) on June 1, 2013; there was a major southern expansion of the Midcontinent ISO (MISO) on December 18, 2013; and Pacificorp has formally begun to explore the possibility of joining California ISO (CAISO), while participating voluntarily in a real-time imbalance market.

[8] The Electric Reliability Council of Texas (ERCOT) system is the exception because this interconnection does not cross state lines, and is therefore not subject to FERC jurisdiction on many matters. However, Texas does participates in the NERC, which has been designated by the FERC as the electricity reliability organization for the United States.

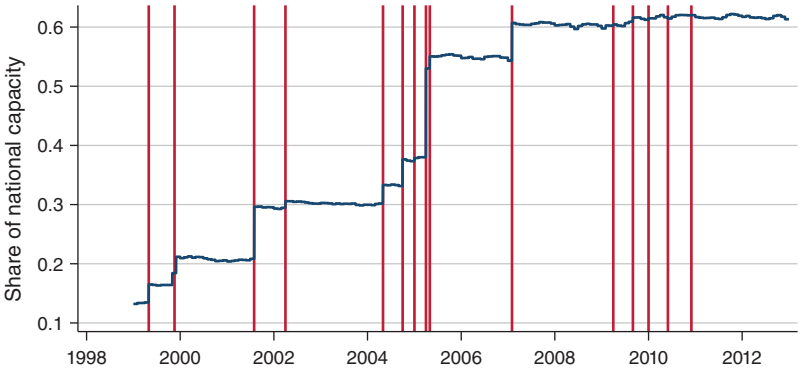

FIGURE 2. SHARE OF US GENERATING CAPACITY DISPATCHED BY MARKETS

Note: Vertical red lines indicate dates of transition to market-based dispatch.

to distinguish between these developments, for although all states that adopted restructuring legislation eventually adopted market dispatch, many areas began participating in these markets while preserving their traditional regulatory framework.[9] I therefore focus my attention on the cost of generating electricity, rather than the retail price of power delivered to consumers, whose relationship with their local utility may or may not have changed over this period.

Vulnerability to the exercise of market power has been a primary focus of the research on wholesale electricity markets to date. From the United Kingdom (Green and Newberry 1992, Wolfram 1999, Wolak and Patrick 1997, Sweeting 2007), Spain (Reguant 2014, Ito and Reguant 2016), New Zealand and Australia (Wolak 20142) abroad; to California (Borenstein, Bushnell, and Wolak 2002; Borenstein 2002; Joskow and Kahn 2002; Bushnell, Mansur, and Saravia 2008; Puller 2007), PJM (Mansur 2001, 2008), plus New England (Bushnell, Mansur, and Saravia 2008) and Texas (Hortaçsu and Puller 2008) in the United States, one could fairly characterize these vulnerabilities as robust. Against these losses, there is sparse evidence of allocative efficiency gains from market dispatch, with the notable exception of Mansur and White (2012) who study one of the 15 market expansion events described above. Instead, liberalization studies have focused on state-led deregulatory events to estimate within-plant changes, and have found substantial cost reductions: 5–10 percent reduced maintenance time (Davis and Wolfram 2012, Kabir et al. 2011), 5–15 percent fewer labor and fuel costs (Fabrizio, Rose, and Wolfram 2007; Cicala 2015), and less capital-intensive pollution abatement equipment (Fowlie 2010, Cicala 2015). On the other hand, the actual rate at which heat is converted to electricity (heat rate) has proven largely unaffected by the nature of

[9] Examples include Indiana, West Virginia, and parts of Kentucky in the PJM Interconnection, most of the MISO, and all of the Southwest Power Pool (SPP). Additionally, the timing of power plant divestitures and expansion of nonutility generators do not line up with the transition to market dispatch: most divestiture was completed by 2001 when state-level restructuring stalled, while the median transition to market dispatch occurred in 2005. In fact, over one-half of PCAs adopt market dispatch while retaining their traditional cost of service structure for determining the retail price of electricity. See online Appendix B for an event study-type figure.

regulatory oversight (Fabrizio, Rose, and Wolfram 2007; Wolfram 2005; Kabir et al. 2011). Borenstein and Bushnell (2015) and Kwoka (2008) provide reviews of the various forms restructuring of the electricity sector has taken, with an emphasis on consumer prices.

While market imperfections are certainly cause for concern, evidence of their existence is not proof of the market's inferiority (Joskow 2010). The relevant question for policymakers is, have markets (including all of their flaws) outperformed the alternative methods for deciding which plants should operate in order to satisfy demand for electricity?

II. Decomposing Production Costs

Evaluating how markets have affected the cost of keeping the lights on is complicated by power flows between PCAs. Trading power creates surplus from the difference in marginal costs between the importing and exporting regions. Attributing individual generators' costs to the PCAs where their power is being consumed is a complex problem beyond the scope of this paper.

In this section I show how aggregate production costs can be separated into objects that account for the fact that power is traded between PCAs, yet can be calculated for without knowing the origin or destination of electrons. I begin by defining "out-of-merit" costs, which measure how closely observed production utilizes the lowest marginal cost units from each PCA's installed generation capacity. I then define "gains from trade," which calculate the savings from reallocating power between PCAs relative to utilizing the lowest marginal cost units in autarky. These will be the outcomes I evaluate empirically in Section IV. I show how these objects relate to observed production costs, and how changing patterns of production across PCAs change the components of the decomposition to yield the total net surplus from trading power.

To fix ideas, suppose PCA p in hour t has N_{pt} MW (megawatts) of total capacity installed. Each generation unit has a known capacity and marginal cost of generation, so one can construct what is referred to as its "merit order" by lining up each of the generators in increasing order of marginal cost (ignoring ramping and start-up constraints). The merit order is indexed by i, and $q_{pt}(i) = \{0, 1\}$ indicates generation from the ith MW of the merit order.[10] The marginal cost of generation from the ith MW of the merit order is $c_{pt}(i)$. Total generation from the PCA in hour t is the tally of each active MW of installed capacity, $Q_{pt} = \sum_{i=0}^{N_{pt}} q_{pt}(i)$. We can then define

$$\text{Observed Costs:} \qquad C_{pt}(Q_{pt}) = \sum_{i=0}^{N_{pt}} c_{pt}(i)\, q_{pt}(i)$$

$$\text{Merit Order Costs:} \qquad C^{*}_{pt}(Q_{pt}) = \sum_{i=0}^{Q_{pt}} c_{pt}(i).$$

[10] The unit dispatch problem partitions the N_{pt} MW of capacity into distinct units and chooses how much to generate from each unit subject to nameplate rating constraints. While this is an identical problem, indexing MW according to i creates a stable metric of the merit order, while indexing units themselves may shuffle as fuel prices vary.

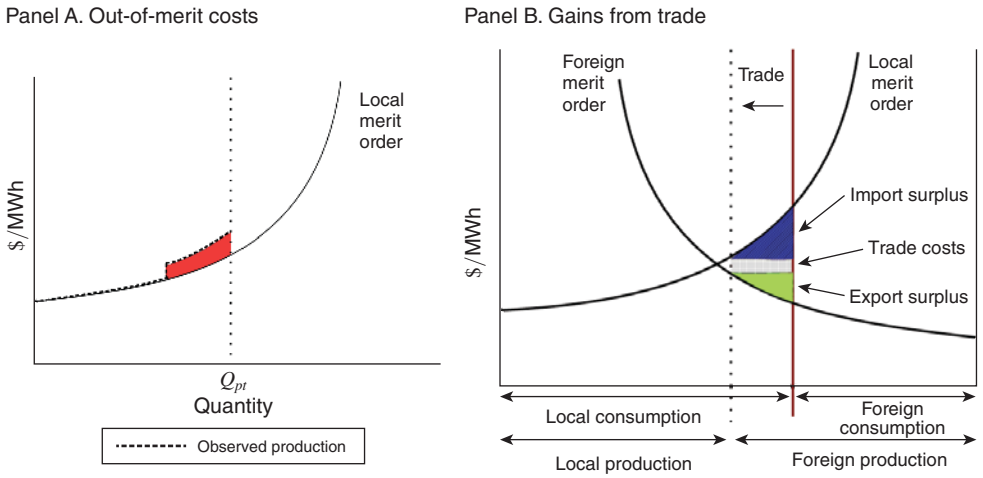

FIGURE 3. DECOMPOSING COSTS IN ELECTRICITY MARKETS

Observed costs are calculated by adding up the cost of generation from each MW that is active $(q_{pt}(i) = 1)$, while the merit order costs add up the cost of generation from the cheapest Q_{pt} MW of installed capacity.

A. *Out-of-Merit Costs*

A unit operates "out of the merit order" when it is called on to help meet Q_{pt} MW of demand although it is not one of the Q_{pt} cheapest MW of PCA p's installed capacity based on its marginal cost. Out-of-merit costs, denoted $O_{pt}(Q_{pt})$, are defined as the difference between a PCA's observed and merit order costs: $O_{pt}(Q_{pt}) = C_{pt}(Q_{pt}) - C^*_{pt}(Q_{pt})$. This is illustrated in panel A of Figure 3. The dashed line orders operating units according to their marginal costs, and Q_{pt} is being generated at this particular moment. Weakly below the dashed line is the merit order, which represents the marginal cost of installed generating capacity, whether those units are operating or not. The cost of dispatching units out-of-merit is simply the additional cost of output from those higher-cost units relative to dispatching the lowest marginal cost units installed in the area. This is depicted as the shaded red area. Changing generation quantities has an ambiguous impact on out-of-merit costs: increasing output from in-merit generators pushes the dashed line that is above the merit order to the right, reducing out-of-merit costs. Increasing output from additional out-of-merit generators will extend the dashed line, increasing out-of-merit costs with additional shaded area above the merit order.

There are a number of reasons that the true cost-minimizing allocation of output requires generation from units that are out-of-merit: Plants must occasionally go offline for maintenance, or are forced to shut down unannounced, causing more expensive units to fill the gap. Transmission constraints may make it infeasible for the least-cost units to meet local demand. Large units require time and fuel to substantially change their output (ramping and start-up costs) which may exceed the cost of firing up a more nimble out-of-merit unit (Reguant 2014, Cullen 2011,

Mansur 2008). Large units may also continue operating when out-of-merit to prevent having to pay larger start-up costs from a cold start (idling). These are all real physical constraints that make some positive amount of out-of-merit operation the true cost-minimizing allocation of output. That is, suppose the true cost-minimizing allocation that incorporates all of these constraints is $\hat{C}_{pt}(Q_{pt})$. The cost of these constraints can be measured by the out-of-merit cost of this output allocation: $\hat{C}_{pt}(Q_{pt}) - C^{*}_{pt}(Q_{pt})$.[11]

Out-of-merit costs are also the losses borne when a firm exerts market power (Wolfram 1999; Borenstein, Bushnell, and Wolak 2002; Mansur 2001). Taking an economical unit "down for maintenance," means the market prospectively clears on a higher-priced generator and allows the firm to collect rents on co-owned inframarginal units. Because demand is completely inelastic (in real-time operations), the welfare loss from this strategy is the incrementally higher operating costs caused by taking economical units offline.

It should be clear that legitimate maintenance, congestion, etc. are observationally equivalent from a cost perspective to the exertion of market power—they differ by intent only. Mansur (2008) and Reguant (2014) note that failing to account for start-up and ramping costs will lead one to overattribute the gap between the merit order and observed dispatch to market power when only accounting for normal maintenance and outages. The same is true when failing to account for transmission constraints (Ryan 2013; Borenstein, Bushnell, and Stoft 2000). This means that *levels* of out-of-merit costs are problematic for measuring the costs of market power. Whether *changes* in out-of-merit costs can be interpreted as cost reductions depends on coincident changes in ancillary costs: an allocation may be closer to the merit order, but requiring greater start-up or ramping expenditures, for example.[12] I test for this explicitly in online Appendix B to rule out the possibility that costs might be rising outside of the key metrics of allocative efficiency.

B. *Gains from Trade*

When transmitting electricity between areas, costs are reduced by supplanting generation in one PCA with output from a lower-cost unit in another, holding total production fixed. Panel B of Figure 3 considers the gains from trade between two areas according to their merit orders. The two merit orders line up each PCA's installed capacity in increasing order of marginal cost.[13] The red line represents demand in the "local" PCA of panel A, and is perfectly inelastic to avoid blackouts. Superimposed on this is the mirror image merit order and demand figure from a "foreign" PCA. The width of the x-axis is the sum of the demand of the two areas. If the two areas were to operate in autarky, the merit order cost of meeting this demand would be the area under the upper envelope of the supply curves, meeting

[11] For example, if a transmission constraint requires 100 megawatt hours (MWh) of output from a unit whose marginal cost is $100 per MWh instead of an available in-merit generator whose marginal cost is $50 per MWh, the cost of the transmission constraint is $5,000.

[12] Mansur (2008) notes if these costs are time invariant, then changes in out-of-merit costs give an accurate estimate of overall cost reductions.

[13] Accounting for trade with empirical supply curves that may deviate from the merit order will be addressed in the following subsection.

at the solid demand line. Total cost would be the lower envelope of the curves if they instead traded up to the point of intersection between their merit orders, and the gains from trade would be the difference between costs under autarky and this lower envelope. The marginal cost of barriers to trade (such as transmission constraints and line losses) is reflected by the height of diverging marginal costs between the two areas.

The volume of trade for PCA p is the absolute difference between its total generation, Q_{pt} and load, L_{pt}. The surplus accrued by this trade is depicted as a wedge for each PCA in panel B of Figure 3. Denoted as $G^*_{pt}(L_{pt}, Q_{pt})$, it is calculated as the area between the merit order curve and the marginal cost of generation according to the merit order ($c_{pt}(i = Q_{pt})$), over the domain of traded quantities:[14]

$$(1) \qquad G^*_{pt}\big(L_{pt}, Q_{pt}\big) = C^*_{pt}\big(L_{pt}\big) - C^*_{pt}\big(Q_{pt}\big) + c_{pt}\big(i = Q_{pt}\big) * \big[Q_{pt} - L_{pt}\big].$$

In an importing area, $C^*_{pt}(L_{pt}) - C^*_{pt}(Q_{pt})$ is positive as the area under the merit order between load and production. From this, one subtracts the rectangle between L_{pt} and Q_{pt}, at the height of the merit order at the amount being generated, $c_{pt}(i = Q_{pt})$. The remaining wedge is the area below the merit order, above c_{pt} $(i = Q_{pt})$, between load and generation.

In an exporting area, the rectangle $c_{pt}(i = Q_{pt}) * [Q_{pt} - L_{pt}]$ is positive, and from this one subtracts the merit order cost of the generation that is being exported, from L_{pt} to Q_{pt}. The remaining wedge above the merit order curve and below the merit order marginal cost of generation represents the gains from trade accrued in the exporting PCA.

Gains from trade are calculated as wedges accrued PCA by PCA rather than complete trapezoids (which would include the crosshatched area of Figure 3, panel B). This distinction attributes the remaining barriers to a single marginal cost from prevailing as due to transmission costs. The inclusion of trade costs (broadly defined) provides a natural explanation for marginal cost variance: importers buy power up to the point that the delivery-inclusive price equates their own marginal cost of production, and likewise for exporters. The crosshatched areas would represent additional savings if not fully dissipated as iceberg costs.

C. *Decomposing Production Costs*

The definition of gains from trade above may differ from what comes to mind when thinking in empirical supply curves: when an area exports an additional MW, its marginal cost is the most expensive unit it is operating, not its most expensive unit in the merit order. If an exporting area increases its production from a \$50 per MWh unit by 1 MWh to offset the generation of a \$100 per MWh unit elsewhere, the gains from this trade as typically conceived should be \$50. However, measuring gains from trade based on empirical curves in this setting often leads to

[14] The asterisk is included to highlight that this calculation is relative to the merit order, in contrast to the total surplus accrued from trading power, which may entail a net change in out-of-merit generation, discussed in greater detail in the following subsection.

puzzles: how can one rationalize the operations of an area that is simultaneously exporting and also running exceptionally expensive units out-of-merit? If the curves in panel B of Figure 3 were empirical supply curves instead of merit order curves, one would often observe outcomes to the left of the point of intersection, with higher marginal costs in the exporting region than the importing region. It would appear that integrating between the curves is measuring the losses from excessive trade that would be reduced if the areas were closer to autarky. The problem is in separating generation used for internal load balancing from that which is transmitted between PCAs.

The approach I employ decomposes observed costs into mutually exclusive components: the cost of generating observed quantities that exceed the merit order cost, and the merit order savings of trading power relative to autarky. Continuing the example, suppose the marginal costs *according to the merit orders* in the exporting and importing regions were \$30 and \$75, respectively—so that the trade described above involved out-of-merit generation on both ends. This means the trade entails an increase in out-of-merit costs of \$20 in the exporting region (it is using a \$50 per MWh generator when the merit order marginal cost is \$30 per MWh), and a decrease in out-of-merit costs in the importing region of \$25 (the unit ramping down cost \$100 per MWh when the merit order marginal cost was \$75). The total surplus of this trade following the definitions above (and ignoring transmission costs), is a net \$5 reduction in out-of-merit generation, and a \$45 gain from trade between the two merit orders (\$75 – \$30), arriving back at the total surplus of \$50. In other words, the total surplus is a mix of net out-of-merit costs and gains from trade according to the merit order.

To operationalize this approach to decompose aggregate costs, measured PCA by PCA, let $C_t(Q_t) = \sum_p C_{pt}(Q_{pt})$ denote the observed cost of generating aggregate national quantity Q_t by adding up all active generation costs in hour t. By adding and subtracting the sum of merit order costs for each PCA's production ($C^*_{pt}(Q_{pt})$) and those under autarky ($C^*_{pt}(L_{pt})$), total costs over all PCAs become

$$\sum_p C_{pt}(Q_{pt}) = \underbrace{\sum_p \left[C_{pt}(Q_{pt}) - C^*_{pt}(Q_{pt})\right]}_{\text{Out-of-Merit Costs}} + \underbrace{\sum_p \left\{ C^*_{pt}(L_{pt}) - \left[C^*_{pt}(L_{pt}) - C^*_{pt}(Q_{pt})\right] \right\}}_{\text{In-Merit Costs Relative to Autarky}}. \tag{2}$$

The first term holds each PCA's total production fixed, and aggregates out-of-merit costs as defined above. The second term measures the aggregated extent to which the merit order costs of the observed production quantities are below those of autarky. Let $c_t(i = Q_t)$ denote the marginal cost in the national merit order for total production in hour t.[15] Noting that total supply equals total demand at any moment in time (i.e., $\sum_p Q_{pt} - L_{pt} = 0$), I add and subtract the rectangle of net exports

[15] Analogous to how the merit order for a PCA is constructed, the national merit order lines up all installed capacity in the country in increasing order of marginal cost regardless of PCA.

times the merit order marginal cost of generation ($c_{pt}(i = Q_{pt}) * [Q_{pt} - L_{pt}]$). Cost relative to autarky according to the merit order from trading becomes

$$(3) \quad \sum_p C^*_{pt}(L_{pt}) - \sum_p C^*_{pt}(Q_{pt}) = \sum_p \left\{ C^*_{pt}(L_{pt}) - C^*_{pt}(Q_{pt}) + c_{pt}(i = Q_{pt}) * \left[Q_{pt} - L_{pt} \right] \right\} - \sum_p \left[c_{pt}(i = Q_{pt}) - c_t(i = Q_t) \right] * \left[Q_{pt} - L_{pt} \right].$$

The right side of the first line of equation (3) aggregates the gains from trade that accrue to each PCA according to the merit order, and are denoted as $G^*_{pt}(L_{pt}, Q_{pt})$ in equation (1). These correspond to the wedges of Figure 3, panel B. The second line adds up implied transmission costs when barriers to trade are borne as iceberg costs: by how much the marginal merit order unit differs from the national merit order (i.e., the intersection of marginal cost curves), times trade volumes. Note that any constant times the sum of trade flows will add up to zero, so the national marginal merit order cost simply facilitates interpretation. This term is ultimately a measure of by how much marginal costs in importing areas exceed those of exporting areas and when added up becomes the multilateral analog of the crosshatched areas in Figure 3, panel B.

Putting everything together, total costs can be written as

$$(4) \quad \sum_p C_{pt}(Q_{pt}) = \sum_p O_{pt}(Q_{pt}) - \sum_p G^*_{pt}(L_{pt}, Q_{pt}) + \sum_p C^*_{pt}(L_{pt}) + \sum_p \left[c_{pt}(i = Q_{pt}) - c_t(i = Q_t) \right] * \left[Q_{pt} - L_{pt} \right].$$

Out-of-merit costs and gains from trade as defined above therefore follow directly from a decomposition of observed costs. Importantly, these measures can be calculated separately for each PCA in a multilateral setting in which it is impossible to determine the origin or destination of electrons.[16] Another advantage of using these outcomes is that they provide a framework that can seamlessly accommodate a wide range of institutions. Mansur and White (2012), for example, use the convergence of market prices to infer gains from trade. Such an analysis is only possible in settings where some form of market mechanisms (with price formation) is already in use.

Decomposing cost levels relative to the merit order forgoes identifying which units are actually generating power for trade. This means that the total surplus from increased trade will typically entail changing a combination of both objects of the decomposition. To make this more explicit and complete the analogy to the example that opens this subsection, consider how out-of-merit costs change between two

[16] Note that gains from trade as $G^*_{pt}(L_{pt}, Q_{pt})$ can be calculated for PCA p at time t using only data observed for that particular PCA without reference to traded power's origin or destination PCA—it requires information on the merit order, load, and the PCA's total production only.

periods, $t = 1,2$ in which load and merit orders are fixed, but PCAs change the quantities they produce from Q_{p1} to Q_{p2}:

$$(5)\quad \sum_p O_{p2}(Q_{p2}) - \sum_p O_{p1}(Q_{p1}) = \underbrace{\left[\sum_p O_{p2}(Q_{p2}) - \sum_p O_{p1}(Q_{p2})\right]}_{\text{Within PCA Change Holding Quantities Fixed}} + \underbrace{\left[\sum_p O_{p1}(Q_{p2}) - \sum_p O_{p1}(Q_{p1})\right]}_{\text{Net Out-of-Merit Costs from Changing Quantities}}.$$

There are two ways that out-of-merit costs may change between periods. First, PCAs may generate closer to (or farther from) their respective merit orders holding the quantities they produce fixed: units may go down for maintenance, there may be more internal transmission congestion, or stronger incentives to exert market power without changing how much power is traded across PCAs. Second, changing production across PCAs may also cause net changes in out-of-merit costs. Greater trade may be used to reduce out-of-merit generation, or the exertion of market power may encourage more trade to reduce its impact on prices. As in the example above, marginal costs may be sufficiently different across PCAs that it is worthwhile to increase out-of-merit generation to reduce even higher cost generation elsewhere. These net cost changes are part of the total surplus of trading electricity across PCAs in any case. When interpreting changes in out-of-merit costs it is important to keep in mind these two forces: shifts in the supply curves themselves, as well as movements along them as PCAs engage in more or less trade.

III. Data

This study draws from a disparate and incongruous set of data sources to synthesize an essentially complete characterization of US electricity production at the hourly, generating unit level from 1999 to 2012 (over 530 million unit-hour observations). This section presents an overview of the data, while the details of data assembly can be found in online Appendix A.

A. *Hourly Load Data*

The demand side consists of a balanced panel of hourly load (consumption, including line losses) from all US PCAs that dispatched power plants in 1999 to meet demand. These data have been reported annually to the FERC on form 714, "Annual Electric Balancing Authority and Planning Area Report." Record-keeping challenges at the FERC requires these data to be supplemented with equivalent data from regional authorities and markets. When there are gaps or reporting changes, I employ LASSO to estimate demand based on weather, population, and employment. Combined with cross validation to maximize out-of-sample accuracy, this procedure delivers predictions within 4 percent of the realized values on average (see the online data Appendix). Small municipal authorities that do not actually conduct dispatch of fossil- or nuclear-powered plants are added to the load of their principal suppliers or customers, yielding 98 total PCAs.

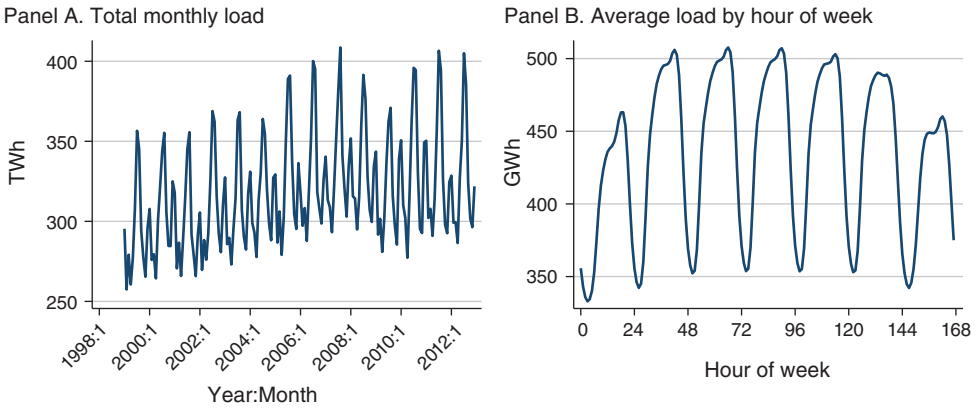

FIGURE 4. ELECTRICITY LOAD OVER TIME

Figure 4 summarizes the electricity load data. The United States consumes a bit less than 4,000 TWh (billions of kilowatt hours) annually. Panel A shows that electricity consumption increased from 1999 until the Great Recession, and was relatively flat through 2012. Panel A also highlights the seasonal nature of electricity usage: summer cooling and winter heating can increase usage by over one-third of temperate seasonal usage on a month-to-month basis, with much larger swings during peak usage. Panel B plots hourly usage over the course of the week, averaged over the 14-year study period. Here too, there are large swings in usage both over the course of the day and the week. The key fact to remember when interpreting these figures is that production must be exactly synchronized with these demand swings, and that utilities must have enough generation capacity to meet demand at the moment of peak usage. Thus every downward swing also represents vast quantities of generating capacity becoming idle.

As a demonstration of real-time patterns of demand, I have animated one year's worth of hourly load in a short video.[17] This animation shows the East-to-West flow of electricity demand as usage follows local clocks. It also reflects the daily and seasonal patterns shown in Figure 4, while highlighting the substantial variation around these averages: peak demand can be as much as 2.5 times average annual usage, can be quite persistent during summer months in the South and Southern Plains, and generally varies less in temperate areas of the Pacific Northwest.

B. *Hourly Generation Data*

The supply side is based on data from the Energy Information Administration (EIA), merged with hourly gross generation reported to the Environmental Protection Agency (EPA) with Continuous Emissions Monitoring Systems (CEMS), as well as daily production at nuclear-powered units from the Nuclear Regulatory Commission. Boilers from the EPA are matched to generators' monthly net generation and heat rates via forms EIA-767 and EIA-923, "Annual

[17] *Hourly Load 1999,* https://youtu.be/QRMPUqMeNIw (posted March 29, 2017)

Steam-Electric Plant Operation and Design Data" and "Power Plant Operations Report." Hourly production in the data is the gross generation from CEMS scaled by the ratio of monthly gross-to-net generation from EIA at the unit level. I then merge this data on heat rates and hourly production with coal and oil fuel costs under a nondisclosure agreement with the EIA (from forms EIA-423, "Monthly Report of Cost and Quality of Fuels for Electric Plants," EIA-923, and form FERC-423, "Monthly Report of Cost and Quality of Fuels for Electric Plants"). These are shipment-level data, reported monthly by generating facilities with a combined capacity greater than 50 MW. I use estimated spot-market coal prices to measure the opportunity cost of coal burned rather than contract prices. Natural gas prices are from 65 trading hubs around the country reported by Platts, Bloomberg, and Natural Gas Intelligence (not EIA), and are quoted daily. Plants are linked to their nearest trading hub along the pipeline network. Areas with emissions markets for sulfur and nitrogen oxides include the cost of pollution based on measured emissions and monthly market prices from BGC Environmental Brokerage Services.

Generation from hydropowered units either comes directly from the source (i.e., Tennessee Valley Authority, US Bureau of Reclamation, etc.), or is based on the streamflow of the nearest downstream gage from the US Geological Survey's Streamgage Network (linked through analysis of the National Hydrography database). Because the cost of reservoir-based hydropower is the opportunity cost of the water, I price hydropower-based power on the marginal cost of fossil generation in the merit order that is being supplanted.[18] Run-of-river hydro is priced at zero. Hydropower units over ten MW were classified as reservoir or run-of-river based on internet searches and/or satellite images.[19]

Hourly generation is unavailable from a number of smaller fossil-fired units (whose net generation rarely exceeds 3 percent by NERC region-year). Power from these units is distributed across the hours of the month in an intuitive manner: having produced nW MWh in a month, where W is the unit's nameplate capacity, I assume that the unit produced at maximum capacity during the n hours of highest demand observed over the course of the month. This replicates the behavior of a dispatcher who employs a threshold rule of when to generate from a unit (assuming no start-up costs or ramping constraints), while allowing observed behavior to dictate what threshold was employed each month.

Figure 5 presents the aggregate annual statistics for electricity generation in the United States. Roughly one-half of the electricity generated from 1999 to 2012 was powered by coal, with a declining share since 2007. From that time, natural gas has grown from rough parity with nuclear (20 percent) to 30 percent, almost entirely at the expense of coal-fired generation. Following a nearly threefold increase from 1999 to 2008, Panel B shows that fossil fuel expenditures fell by nearly 50 percent from the peak in 2008 from the combined effects of reduced demand overall and the massive reallocation of output to units burning cheaper gas thanks to the advent of

[18] Robustness to this assumption is shown in online Appendix B.

[19] Pricing hydropower at the merit order marginal cost assumes that it is never operated out-of-merit, so all dynamics of hydro production are orthogonal to out-of-merit results: the difference between observed and merit order production costs will net out hydro at any price. Pricing reservoir hydropower in this manner infers that the value of storing one MWh worth of power (either for production later, or flood management, irrigation, etc.) is equal to the marginal cost of fossil power required to offset hydro production. See Archsmith (2017) for a recent application.

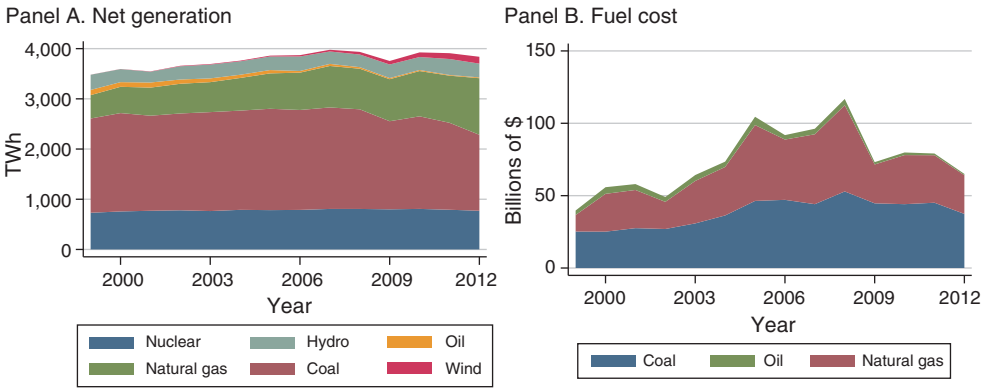

FIGURE 5. ANNUAL NET GENERATION AND FUEL COST BY SOURCE

hydraulic fracturing (Hausman and Kellogg 2015; Linn, Muehlenbachs, and Wang 2014; Knittel, Metaxoglou, and Trindade 2014). Fossil fuel expenditures averaged about $75B per year over these 14 years; thus the complete dataset tracks the burning of $1T of fuel at the plant-generation unit (or prime mover)-hour level.

C. *Matching Supply and Demand*

Because the supply data are built up from microdata independently from the demand side, it is important to ensure congruence between the data sources—there is nothing institutional about their reporting to ensure they agree. Beginning with the 1999 configuration of the electrical grid, I match plants to their initial PCA from the EPA Emissions and Generation Resource Integrated Database (eGRID). New capacity since that time is matched to PCAs either directly or based on historical utility service territory in the case that the PCA territory has changed. These associations are then checked against power plant names reported by PCAs in FERC 714. I then compare the implied monthly totals from the supply side of the data against those reported by the PCAs to FERC. In total, about 99 percent of reported generation from FERC 714 can be accounted for in the supply-side data. About 3 percent of net generation does not fit neatly into a single power control area because multiple PCAs report a share of output from large plants as their own. In these cases, the plant is assigned to the PCA with greatest dispatch authority.

Figure 6 breaks down generation by data source, and shows the quality of the match between supply and demand. The top black line in panel A is identical to the total monthly load shown in panel A of Figure 4. After totaling the generation observed (or calculated) based on high-frequency data, the remaining numbers reported at the monthly level result in totals that almost exactly match the demand side of the data. Panel B gives a closer view of what is missing by calculating the gap (as imports or exports) for every hour across PCAs, then adding them separately up to the monthly level, measuring the volume of trade across areas. The first striking statistic is that roughly 90 percent of generation is effectively consumed in its local PCA—while PCAs are interconnected, they continue to largely produce energy for their own consumption. To my knowledge, these statistics are new: regulatory bodies typically

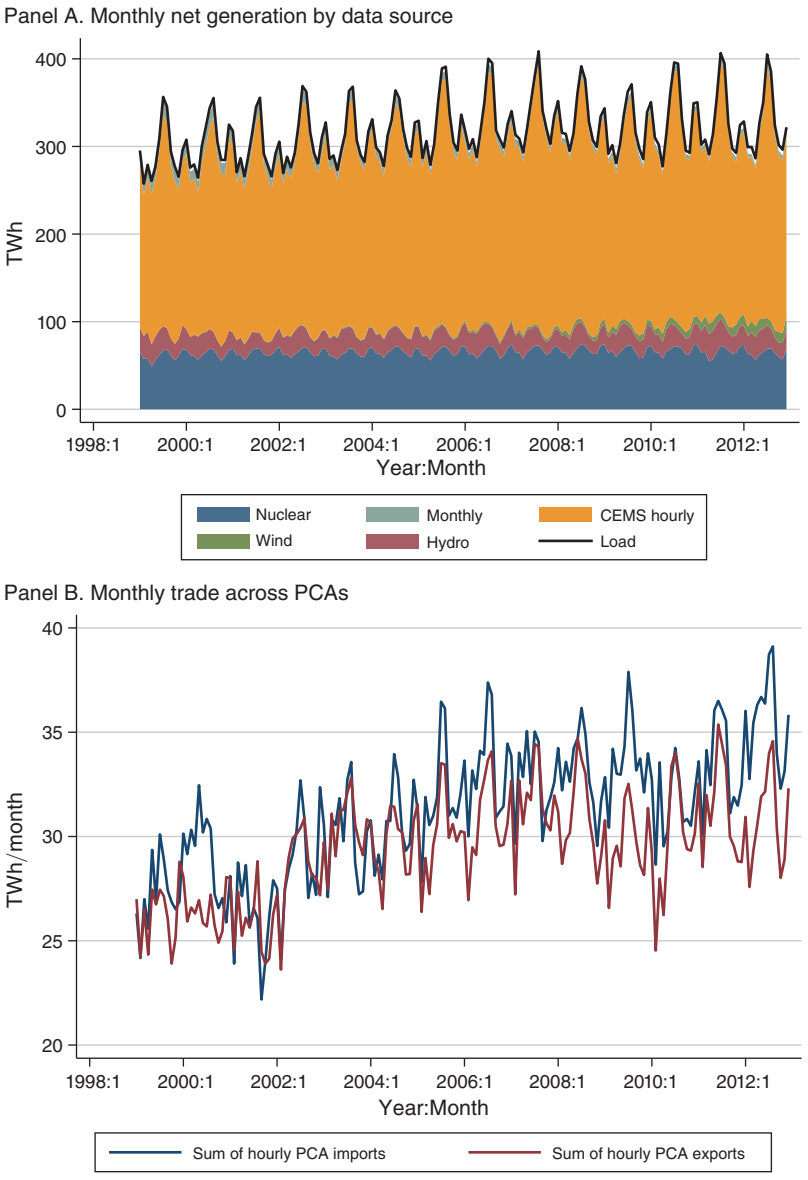

Figure 6. Monthly Net Generation by Data Source and Trade across PCAs

report the monthly *net* flow of electricity between areas, which fails to reflect the real-time interdependence among PCAs (or lack thereof).[20]

The remaining gap between imports and exports as I observe them is due to imports from outside of the United States (which have grown since 2004 to about 1 percent of supply (EIA 2012, Table 2.13). Based on the framework presented in Section II, not observing this generation effectively treats it as an import from outside of each PCA, which is valued as displaced local generation. The production

[20] See, e.g., https://www.eia.gov/todayinenergy/detail.php?id=4270 (accessed December 14, 2022).

costs and exporter surpluses (mostly from Canada) are outside of the data. Online Appendix A.4 validates the data further by merging the gap between load and generation with monthly net export statistics from the Canadian National Energy Board and shows a high degree of congruence between the time series when calculated as monthly net flows.

While Figure 1, panel B makes clear that markets were not randomly assigned across space, the summary statistics of Table 1 show broad similarities in mean composition of areas that adopted markets and those that did not. They have similar mean load, generation, and capacity. Areas that became markets were initially more gas- and less hydro-intensive. Load growth over the 14-year period was also similar, though the signs have flipped in differences in trade volumes and out-of-merit generation, with market areas now trading more and generating less out-of-merit. Mean trade volumes in nonmarket areas were unchanged from 1999 to 2012 in spite of a 10 percent growth in load. One major difference between the areas is that market PCAs have had more nonutility generation, starting at baseline, and growing over the study period. Previewing results, there has been a striking change in the components of the cost decomposition. Market areas initially had roughly one-third fewer gains from trade in 1999, but had over one-third more by 2012. Their out-of-merit costs were approximately 50 percent higher in 1999, and were 10 percent lower by 2012.

IV. Estimation Strategy

I use the staggered timing of market creation and expansion to arrive at an estimate of the causal impact of the transition to market-based electricity dispatch. These events are defined as the PCAs' formal cession of control of their transmission system to an ISO, who conducts auctions to allocate output to generating units. As demonstrated in Figure 2, these are discrete events—typically demarcated prominently in the history of each market. These events suggest a DD approach, using areas without regulatory change to estimate counterfactual outcomes after one has adjusted for common shocks and time-invariant differences.

While the impact of markets on the cost of meeting load is the central focus of this paper, evaluating production costs as an outcome directly is problematic. If a PCA increases its imports when markets begin, its costs fall by the full value of reduced production. This imported power must come from somewhere, and is costly to produce. If it comes from a PCA that joined the market at the same time, then regressing production costs on market onset will estimate the net savings from the increased power flows, which is the object of interest. Departures from this specific reallocation will yield a biased estimate of cost changes. If the power comes from a long-term incumbent market (i.e., a PCA that has used markets for at least two years), then the cost of the incoming power will not be incorporated in estimating the short-term effect. The bias reverses for PCAs that increase exports during market expansions. If imported power comes from Canada, the regression coefficient would treat the incoming power as free.

Evaluating out-of-merit costs $(O_{pt}(Q_{pt}))$ and gains from trade $(G^{*}_{pt}(L_{pt}, Q_{pt}))$ as defined in Section II instead of observed costs $(C_{pt}(Q_{pt}))$ reduces the extent of this problem by valuing traded power at the merit order marginal cost of production

for each PCA. The strategy described below estimates the extent to which these costs and surpluses change in a PCA that joins a market. While it does not estimate changes that accrue to market PCAs outside of their initial two-year window, it also does not treat power that comes from them as free, nor power sent to them as without surplus accruing to the exporter.

This section describes the sources of potential confounding in the context of electricity markets, the assumptions required to interpret results as unbiased estimates of causal effects, and tests of the validity of these assumptions, when possible.

Both gains from trade and out-of-merit costs depend critically on fuel prices. For units that burn the same fuel, for example, the merit order is determined by each unit's respective efficiency in converting fuel into electricity. The value of generating from a more efficient generator then scales directly with the cost of the fuel saved: it is high when fuel is expensive, and low when fuel is cheap.[21] The fact that fuel prices change over time means these measures of allocative efficiency will change even when patterns of production have not. This suggests it is important to control for contemporaneous differences across areas. However, the impact of fuel prices on savings (and operations) depends on the installed capacity that is specific to each area: a gas-intensive PCA will find its merit order is unaffected by shocks to coal prices, but may find demand for exports increase when coal prices are relatively high. This can make contemporaneous shocks in other PCAs deliver poor counterfactuals for what shocks would have been if not for treatment (i.e., parallel trends are violated).

To account for PCA-specific time-varying shocks, I control for the cost of meeting load according to the merit order, $C^*_{pt}(L_{pt})$.[22] Although it does not directly account for what outcomes would have been in the absence of treatment, it traces through the merit order given observed generation capacities and fuel prices, reaching more expensive units when load is high, and isolating shocks to low-cost units when load is low. Importantly, it depends solely upon variables assumed to be exogenous: $C^*_{pt}(L_{pt})$ is calculated from generation capacity, heat rates, fuel and emissions prices, and load. By contrast, quantities actually generated may respond directly to treatment, so any function of Q_{pt} is endogenous.

Using the out-of-merit cost and gains from trade metrics from Section II as outcomes, I estimate equations of the form

$$(6)\quad y_{pt} = \tau D_{pt} + \gamma_{pm} + \delta_{tr} + \lambda_{pm} Log\big(L_{pt}\big) + \kappa_{pm} Log\Big[C^*_{pt}\big(L_{pt}\big)\Big] + \eta \chi_{pt} + \varepsilon_{pt},$$

where y_{pt} is the logged value of the outcome variable for PCA p in date-hour t, and D_{pt} is an indicator of market dispatch. Separate fixed effects and slopes for load and merit order costs are estimated by PCA-month of year (i.e., New York in May). These account for the fact that PCAs vary in how they respond to load in a time-invariant manner, and that there are also persistent seasonal differences across areas, particularly with respect to how maintenance and refueling downtime is scheduled. The time fixed effects δ_{tr} are included at the date-hour-region level

[21] Generating from a unit that requires 10 metric million British thermal units (MMBTU) of gas per MWh instead of a unit that requires 11 MMBTU per MWh saves 1 MMBTU. The value created by using the more efficient generator is the price per MMBTU of gas.

[22] As defined in Section II, this means lining the generating units up in order of marginal cost, and tallying the cost of the L_{pt} lowest-marginal cost generators: $C^*_{pt}(L_{pt}) = \sum_{i=0}^{L_{pt}} c_{pt}(i)$.

to account for high-frequency spatial and time-varying unobservables. Because long-term responses to treatment may be confounded with other trends not isolated with this estimation strategy, I also include annual event-time dummies for periods greater than two years prior and two years after the adoption of market dispatch. These are represented by χ_{pt}. All regressions weight by each PCA's mean load in 1999. Changing the use of a single generator can yield large proportional impacts in small areas, so giving more weight to larger areas ensures estimates are representative of the grid at large and helps smooth transitory shocks.

The variable τ measures the short-run (two years) average effect of market dispatch, and should be interpreted as an average treatment on the treated (ATT)—it measures the effect in the areas that have adopted market dispatch. Interpreting this as an average treatment effect requires the stronger assumption that PCAs in the South and West have the same potential benefits from market integration—rather than the continued business-as-usual assumption required for the validity of the ATT. One should keep in mind that markets themselves are heterogeneous, and their rules change over time. Thus a single "treatment effect" of markets as conceived here takes the average of these various institutional changes, compared to the various institutions that preceded the transition to market dispatch.

I evaluate four main PCA-level outcomes of two quantity measures and two cost measures using an hourly panel. The quantity measures are trade volumes and MW of out-of-merit generation. Trade volumes are calculated as the absolute value of the difference between PCA generation and load, $Q_{pt} - L_{pt}$. The quantity of out-of-merit generation is the number of MW generated from installed capacity that is higher than Q_{pt} in the merit order. This is calculated as $\sum_{i=Q_{pt}+1}^{N_{pt}} q_{pt}(i)$ using the notation introduced in Section II. The cost outcomes are the gains from trade defined as $G^*_{pt}(L_{pt}, Q_{pt})$ in equation (1) and out-of-merit costs, defined as $O_{pt}(Q_{pt})$ in Section II.

There are a number of potential threats to the validity of this research design. First and foremost, the stable unit treatment value assumption (SUTVA) requires that the treatment status of PCAs that become markets does not affect the outcomes of other areas. This will be violated, for example, if the expansion of markets in Ohio facilitates the delivery of electricity from the Tennessee Valley Authority (TVA), which is not dispatched by markets. Using TVA as a control PCA will understate the true effect of market dispatch when their exports change due to the policy change. Assuming that PCAs neighboring areas with market dispatch are those most likely to be affected, SUTVA violations are testable by considering such proximity a separate treatment. I measure the extent to which the initiation of market dispatch along one's border affects outcomes by using farther away PCAs as controls.

This estimation framework also assumes that outcomes change immediately with the change in treatment status. However, sudden massive changes tend not to be conducive to keeping the lights on. The preperiod may be contaminated if PCAs began to change their dispatch policies in preparation for the transition to markets. On the other hand, the treatment effect may take time to fully manifest itself as PCAs learn how to use the market to improve their operations (or exert market power). Such dynamics should become apparent in event study-type figures.

Coincident treatments are another potential threat to the research design. If nonutility generation was expanding in areas that adopted markets (through power plant divestiture, for example), one may misattribute estimated changes to markets

Table 1—Summary Statistics for Power Control Areas by Eventual Market Adoption

	1999			2012		
	Adopt markets	No markets	Difference of means	Adopt markets	No markets	Difference of means
Quantities (*GWh*)						
Load	10.98	9.94	1.03	11.83	10.94	0.90
	[8.72]	[7.45]	(0.90)	[9.24]	[8.20]	(0.98)
Generation	10.50	10.49	0.01	11.08	11.63	−0.54
	[8.59]	[7.93]	(0.94)	[8.98]	[8.90]	(1.02)
Net trade volume	1.27	1.49	−0.23	1.76	1.49	0.28
	[1.31]	[2.03]	(0.13)	[1.79]	[2.17]	(0.16)
Out-of-merit	2.43	2.07	0.37	2.69	3.04	−0.36
Generation	[1.98]	[1.87]	(0.22)	[2.10]	[2.60]	(0.28)
Observations	525,600	332,880	1,719,312	527,040	333,792	1,719,312
Costs (*thousands of US$*)						
Observed	136.55	116.72	19.83	192.38	202.68	−10.30
	[122.39]	[102.32]	(12.09)	[170.39]	[178.21]	(19.57)
Out-of-merit	29.63	19.24	10.38	37.30	41.92	−4.62
Costs	[38.55]	[19.54]	(3.40)	[44.47]	[40.00]	(4.65)
Gains from trade	2.56	3.84	−1.27	8.74	6.45	2.29
	[10.36]	[10.46]	(0.57)	[54.80]	[30.43]	(2.44)
Capacity (*GW*)						
Total	19.69	16.50	3.19	23.70	23.17	0.53
	[16.20]	[11.76]	(1.65)	[18.22]	[18.03]	(2.17)
Coal	7.45	7.78	−0.33	6.86	7.20	−0.33
	[7.29]	[8.13]	(0.97)	[6.89]	[7.34]	(0.89)
Gas	4.96	3.24	1.72	10.40	10.72	−0.32
	[4.71]	[3.94]	(0.44)	[7.64]	[9.45]	(1.00)
Nuclear	3.60	3.27	0.33	3.66	3.40	0.26
	[4.62]	[2.79]	(0.44)	[4.61]	[2.98]	(0.44)
Hydro	0.47	1.15	−0.68	0.47	1.07	−0.60
	[0.98]	[3.00]	(0.20)	[0.90]	[2.82]	(0.19)
Nonutility	3.26	0.26	3.00	16.97	4.45	12.52
	[4.39]	[0.33]	(0.29)	[19.52]	[5.86]	(1.75)
Power control areas	60	38	98	60	38	98

Notes: Values are weighted by PCA mean load in 1999. All statistics are calculated from hourly data. Net trade volume is the absolute difference between hourly load and generation. Out-of-merit costs and gains from trade are calculated as defined in Section III. Standard errors clustered by PCA-month in parentheses and standard deviations in brackets.

when it was really ownership structure changes that had the salient impact. While Table 1 shows that nonutility generation grew disproportionately in market areas, I show in online Appendix B.1 that the timing of nonutility capacity growth does not line up with the introduction of market dispatch.

On interpretation, out-of-merit costs fall when operations more closely follow the merit order I construct. My estimates will be biased if market dispatch causes generators to move farther from the true cost-minimizing allocation in order to more closely follow the merit order. Intertemporal considerations are a prime example of how this might happen: generators may chase transitory peaks if they ignore the cost of starting up and ramping their units. I examine this possibility directly in online Appendix B.1, and do not find increases in such behavior that might meaningfully affect the main results. Insofar as unmeasured labor, operations and maintenance costs might be differentially affected by market dispatch, the small share of nonfuel costs multiplied

Table 2—Impact of Market Dispatch on Cost Components

	(1)	(2)	(3)	(4)
Panel A. log(observed costs)				
Market dispatch	−0.085	−0.077	−0.081	−0.083
	(0.012)	(0.012)	(0.009)	(0.009)
First neighbor Market dispatch				0.027 (0.009)
Second neighbor Market dispatch				−0.007 (0.008)
$log(L_{pt})$		Yes	Yes	Yes
$log\left(C^{*}_{pt}(L_{pt})\right)$			Yes	Yes
Clusters	16,464	16,464	16,464	16,464
PCAs	98	98	98	98
R^2	0.946	0.955	0.963	0.963
Observations	11,996,766	11,996,766	11,996,766	11,996,766
Panel B. log(gains from trade)				
Market dispatch	0.448	0.461	0.470	0.437
	(0.071)	(0.072)	(0.066)	(0.065)
First neighbor Market dispatch				0.032 (0.079)
Second neighbor Market dispatch				0.011 (0.072)
$log(L_{pt})$		Yes	Yes	Yes
$log\left(C^{*}_{pt}(L_{pt})\right)$			Yes	Yes
Clusters	16,412	16,412	16,412	16,412
PCAs	98	98	98	98
R^2	0.501	0.559	0.582	0.583
Observations	8,475,828	8,475,828	8,475,828	8,475,828
Panel C. log(out-of-merit costs)				
Market dispatch	−0.130	−0.114	−0.155	−0.180
	(0.029)	(0.028)	(0.025)	(0.026)
First neighbor Market dispatch				−0.008 (0.032)
Second neighbor Market dispatch				−0.009 (0.025)
log(load)		Yes	Yes	Yes
log(load merit cost)			Yes	Yes
Clusters	16,437	16,437	16,437	16,437
PCAs	98	98	98	98
R^2	0.862	0.870	0.879	0.880
Observations	11,618,837	11618,837	11,618,837	11,618,837

Notes: All specifications include PCA-month of year and region-date-hour fixed effects. Controls for the logarithm of load L_{pt} and its merit order cost $C^{*}_{pt}(L_{pt})$ are estimated with separate slopes by PCA-month of year. Standard errors clustered by PCA-month in parentheses.

by the modest impact of restructuring found in Fabrizio, Rose, and Wolfram (2007) renders the potential magnitude of such bias quite small.

Regarding inference, estimates using this approach are presented with standard errors clustered at the PCA-month. This reflects the thought experiment that the observed data (a complete census of operations) are drawn from a super-population of operations—and that each months' fluctuations in demand and costs allow for an independent observation for each PCA. If one believes that there are only 98 (PCA) independent observations, the reported standard errors roughly double.

Table 3—Impact of Market Dispatch on Quantities

	(1)	(2)	(3)	(4)
Panel A. log(trade volume)				
Market dispatch	0.168	0.149	0.211	0.226
	(0.033)	(0.033)	(0.031)	(0.031)
First neighbor Market dispatch				0.044
				(0.036)
Second neighbor Market dispatch				0.009
				(0.032)
$log(L_{pt})$		Yes	Yes	Yes
$log\left(C^{*}_{pt}(L_{pt})\right)$			Yes	Yes
Clusters	16,464	16,464	16,464	16,464
PCAs	98	98	98	98
R^2	0.537	0.568	0.584	0.585
Observations	12,004,719	12,004,719	12,004,719	12,004,719
Panel B. log(MWh out-of-merit)				
Market dispatch	−0.072	−0.073	−0.054	−0.055
	(0.013)	(0.013)	(0.013)	(0.014)
First neighbor Market dispatch				−0.023
				(0.016)
Second neighbor Market dispatch				0.026
				(0.013)
$log(L_{pt})$		Yes	Yes	Yes
$log\left(C^{*}_{pt}(L_{pt})\right)$			Yes	Yes
Clusters	16,440	16,440	16,440	16,440
PCAs	98	98	98	98
R^2	0.890	0.896	0.901	0.901
Observations	11,625,543	11,625,543	11,625,543	11,625,543

Notes: All specifications include PCA-month of year and region-date-hour fixed effects. Controls for the logarithm of load L_{pt} and its merit order cost $C^{*}_{pt}(L_{pt})$ are estimated with separate slopes by PCA-month of year. Standard errors clustered by PCA-month in parentheses.

V. Results

Tables 2 and 3 present the main results as ATT estimates to measure the short-run impact of market dispatch on costs. As described in Section II, net trade volume is measured hourly, PCA by PCA, as the absolute difference between generation and load. Out-of-merit generation is the quantity of MW produced from generators who are out-of-merit relative to installed capacity given that PCA-hour's generation. Out-of-merit costs correspond to the shaded red areas in panel A of Figure 3: the difference between observed generation costs and merit order costs. Gains from trade are calculated according to equation (1) and correspond to either the blue or green wedges in panel B of Figure 3 depending upon whether the area is a net importer or exporter in that hour, respectively.

The first columns of the tables are based on DD estimates that include date-hour-region and PCA-month of year fixed effects. The second column adds PCA-month of year-specific slopes for load, and the third column further adds analogous controls for the merit order cost of meeting load, $C^{*}_{pt}(L_{pt})$. These permit each area to have persistent idiosyncratic relationships between demand, fuel prices, and how it goes about meeting that demand with out-of-merit generation and trade. The fourth

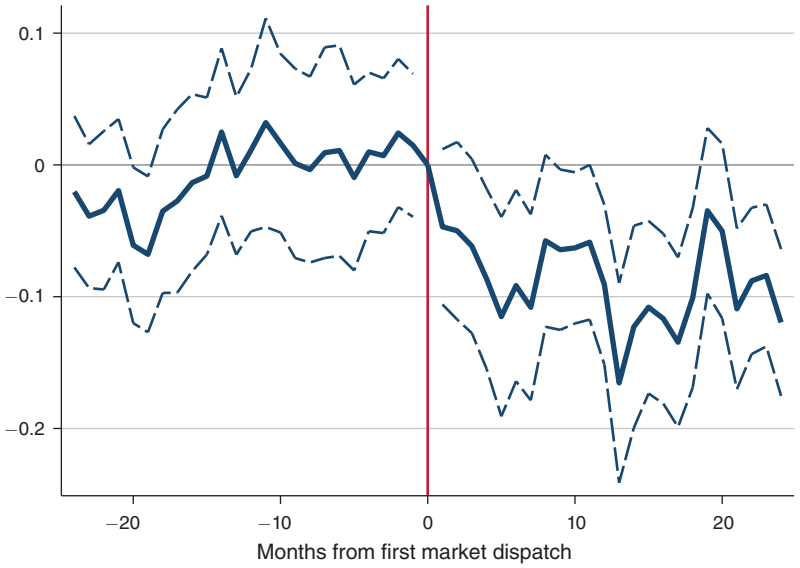

FIGURE 7. TREATMENT EFFECTS BY MONTHS TO MARKET: *Log*(Observed Generation Costs)

Notes: This figure is based on regressing the logged outcome on a set of indicator variables for each month until (after) the transition to market dispatch. The specification corresponds with column 4 of Table 2, where observations are weighted by mean PCA load in 1999. The month prior to treatment is normalized to zero. The 95 percent confidence intervals in dashed lines are based on clustering at the PCA-month level.

column adds estimates for the impact of market dispatch on first- and second-order neighbors who have not adopted markets to measure potential contamination of the control group. Changes in observation counts between outcomes indicate the extent to which PCAs operate exactly according to my measure of the merit order, or do not benefit from trade: zeros are dropped in the logarithmic specifications when the merit order is followed so that no generation is out-of-merit.

Table 2 estimates the cost savings in the first two years following the adoption of market dispatch. Panel A presents the results using the logarithm of observed costs as the outcome variable. These estimates show an 8 percent decline in costs, but fail to account for the cost of production for power that comes from incumbent market PCAs or outside of the United States. Figure 7 shows an abrupt drop in costs that corresponds with the initiation of market dispatch. This and the subsequent event-time figures are based on the model of column 4 to control for load, merit order costs, and potential spillovers. Instead of a single treatment effect for the two years following treatment, it includes separate dummies for each month measuring the time until (or since) market dispatch adoption. Note that this specification only measures the effect for the initial transition to market dispatch: performance changes among incumbents (with whom the area is trading) following market expansion are not included. The potential bias in these estimates motivate the analysis of the key components of the cost decomposition.

Panels B and C of Table 2 find considerable cost savings from market dispatch: 43 log points for gains from trade (55 percent), and an 18 log point (16 percent) reduction in out-of-merit costs. Nonmarket neighbors do not appear to be significantly impacted, and accounting for potential spillovers has little impact on

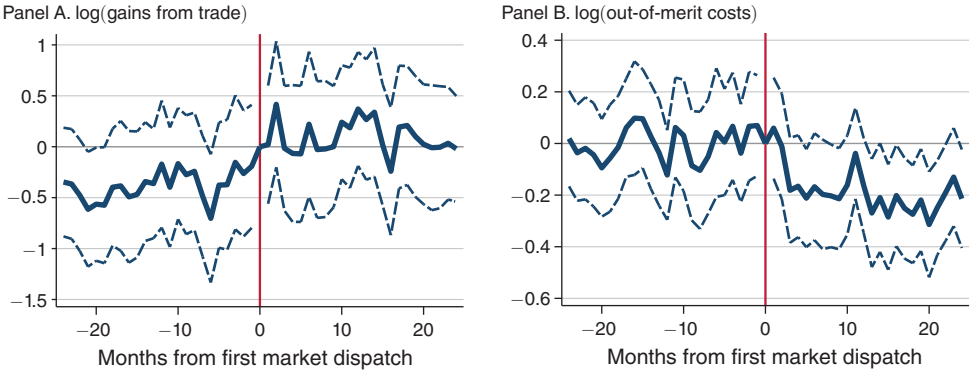

FIGURE 8. TREATMENT EFFECTS BY MONTHS TO MARKET: GAINS FROM TRADE AND OUT-OF-MERIT COSTS

Notes: These figures are based on regressing logged outcomes on a set of indicator variables for each month until (after) the transition to market dispatch. The specification corresponds with column 4 of Table 2, where observations are weighted by mean PCA load in 1999. The month prior to treatment is normalized to zero. The 95 percent confidence intervals in dashed lines are based on clustering at the PCA-month level.

the main coefficients. One should note the substantial difference in observations between the two panels. This is because the specifications in panel B condition upon positive gains from trade: in roughly 25 percent of PCA-hours, there is sufficiently little trade that both supply and demand land on the same generator, which yields zero surplus of this form. (As described in Section II, traded power may still be creating surplus by reducing out-of-merit costs in other PCAs.) In online Appendix Tables B.3, B.4, and online Appendix Figure B.6, these results are presented using the inverse hyperbolic since (IHS) transformation, which allows the inclusion of these zeros. Under this alternative transformation other outcomes are broadly similar while gains from trade increase to an approximate doubling due to markets.

Figure 8 presents the main results on cost reductions relative to the onset of treatment. While the drop in out-of-merit costs is quite stark, the change in gains from trade is a bit smoother in the few months leading up to the opening of the markets. It appears this is due to a mix of hours with zero gains from trade and differentially changing fuel prices: the break is sharper in the figure for trade volumes below (the base for measuring the gains from trade wedges), and alternative fuel price assumptions and the IHS transformation in online Appendix B also yield somewhat cleaner breaks.

In terms of magnitudes, gains from trade respond more strongly in proportions, but began from a much lower baseline, as presented in Table 1. Using 1999 as the base, the estimates imply an average increase in gains from trade of about $700 million per year and a reduction in out-of-merit costs of about $2.2 billion per year. The implied impacts in 2012 grow to $1.7 billion for gains from trade and $3.2 billion in reduced out-of-merit costs.[23] Annualizing the hourly mean observed costs for market areas in Table 1, these savings represent about 5 percent of total variable

[23] For the 1999 base, these totals are calculated by multiplying the relevant mean in 1999 for PCAs that adopted markets by the estimated treatment effect in percent changes $(e^{\tau} - 1)$, and multiplying by 60 PCAs and 8,760 hours per year. For the 2012 base, the relevant mean is multiplied by $(1 - (1/e^{\tau}))$ and multiplying similarly to calculate the savings against the values that would have prevailed but for treatment.

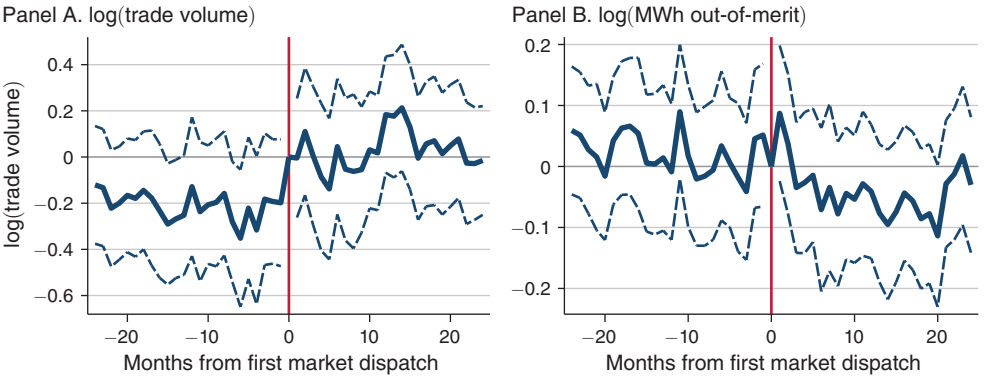

FIGURE 9. TREATMENT EFFECTS BY MONTHS TO MARKET: QUANTITIES

Notes: These figures are based on regressing logged outcomes on a set of indicator variables for each month until (after) the transition to market dispatch. The specification corresponds with column 4 of Table 3, where observations are weighted by mean PCA load in 1999. The month prior to treatment is normalized to zero. The 95 percent confidence intervals in dashed lines are based on clustering at the PCA-month level.

costs. As a point of comparison with the changes in raw means, the relative change between market and nonmarket areas from 1999 to 2012 has been a $2.0 billion increase in gains from trade, and a $7.6B decline in out-of-merit costs. These results suggest that accounting for the costs of traded power yields an estimated impact of markets that is smaller than one would find by simply analyzing observed costs.

The cost savings found here come from moving power: between PCAs, and between generators within PCAs. Examining the changing pattern of quantities, Panel A of Table 3 indicates a 23 log-point (25 percent) increase in traded volumes following the adoption of market dispatch. Because the data are structured PCA by PCA, an increase in exports in one area will be complemented with increases in imports in other areas. In a DD framework, this yields an underestimate of the true treatment effect if that power is being sent to a control PCA in a way that increases its trade volumes (netting out the increased trade between treatment and control). Increases in trade between market PCAs are not double counted in coefficient estimates, as their changes are being compared to changes in control PCAs. Column 4 of panel A suggests that neighboring nonmarket PCAs do not significantly increase trade when markets open nearby. Accounting for these potential spillovers has a negligible effect on the main estimate. This null finding for spillovers also holds for out-of-merit costs, where one might be concerned that reductions in treatment PCAs are offset with increases in control PCAs, which would yield an overestimate of the true cost reductions.

Panel B of Table 3 uses the same framework to evaluate the volume of out-of-merit generation. This measure does not have a direct correspondence with allocative efficiency—it is possible to reduce out-of-merit costs while increasing the amount of out-of-merit generation depending on relative costs.[24] The quantity of

[24] For example, a PCA could generate more power at a lower cost by generating two MW from a unit marginally out-of-merit than one MW from a much more expensive unit. The difference between the former and the latter allocations would be a decrease in out-of-merit costs, but a one MW increase in out-of-merit generation.

out-of-merit generation can, however, suggest mechanisms at work. I find here that market dispatch reduces out-of-merit generation by about 5 percent.

Figure 9 presents the analogous event-time plots for changes in quantities. Any potential pretrends are economically small compared to the relative sharp changes in outcomes around the adoption of market dispatch. There are a number of reasons the timing of these changes might not line up exactly with the official opening of the markets. ISOs often have training periods in advance to adjust to the new environment, markets tend to start during mild seasons so that peak demand arrives after a few months of experience, and firms may require time to learn how to best operate (Doraszelski, Lewis, and Pakes 2018). In any case, the rise in trade volumes and reduction in out-of-merit generation do not appear to be continuing trends that predated the transition to market dispatch.

A. *Heterogeneity*

The richness of the data allows for the examination of heterogeneous treatment effects to provide suggestive evidence of the forces driving the overall point estimate. Table 4 reproduces the specifications of the fourth columns of Tables 3 and 2, interacting the indicator for within two years posttreatment with prospective sources of differences in the impact of market dispatch: across the various markets, by tercile of natural gas prices, and by month of year. This is exploratory analysis that should be interpreted with a degree of caution: each of these dimensions may be correlated with true drivers of heterogeneity without actually directly affecting the outcome.

Panel A breaks estimates out by market,[25] which vary along a number of dimensions. New York and New England, for example, consist of single PCAs that have not been integrated with other areas. The MISO integrated 33 former PCAs, many of whom have retained their traditional vertically integrated utility structure. The SPP introduced a real-time "energy imbalance" market that did not offer day-ahead scheduling, which helps larger units ensure it is worthwhile to pay substantial start-up costs. Its constituent utilities have also retained their traditional structure. It appears the largest response to markets was in New England, though such a conclusion would require assuming away the potential importance of contemporaneous events that would get averaged out in larger areas. Instead, panel A is useful for demonstrating that the main results are not driven by a single PCA: the largest markets (PJM and MISO) each have coefficients that are close to the overall estimate. SPP's estimates are generally smaller in magnitude, while the impact in Texas is often of the opposite sign and mostly not statistically different from zero.

For panel B I use the variance in daily natural gas prices to estimate separate coefficients by fuel price tercile. Each PCA-year has an equal number of observations in each bin, so these coefficients do not depend on which areas became markets during periods of high versus low gas prices over the entire sample period.[26] Instead, these

[25] CAISO began before 1999, so the absence of a preperiod prevents the estimation of separate coefficients.

[26] Coal prices vary at too low a frequency to avoid making comparisons over the entire sample period with such an exercise.

Table 4—Heterogeneous Effects of Market Dispatch

	Trade volume	Out-of-merit generation	Gains from trade	Out-of-merit costs
Panel A. Markets				
Texas	−0.060	0.088	0.174	0.062
	(0.061)	(0.026)	(0.109)	(0.052)
ISO New England	0.895	−0.298	2.540	−0.217
	(0.078)	(0.044)	(0.158)	(0.079)
New York ISO	−0.048	−0.134	−0.162	−0.274
	(0.067)	(0.031)	(0.162)	(0.060)
Midcontinent ISO	0.264	−0.077	0.476	−0.125
	(0.041)	(0.022)	(0.085)	(0.039)
Pennsylvania-Jersey-Maryland	0.382	−0.044	0.420	−0.321
	(0.050)	(0.021)	(0.097)	(0.039)
Southwest Power Pool	0.089	−0.103	0.133	−0.167
	(0.040)	(0.025)	(0.084)	(0.042)
Panel B. Gas prices				
Low	0.181	−0.046	0.329	−0.218
	(0.038)	(0.018)	(0.077)	(0.031)
Medium	0.232	−0.062	0.461	−0.149
	(0.037)	(0.016)	(0.078)	(0.029)
High	0.252	−0.063	0.505	−0.154
	(0.043)	(0.018)	(0.087)	(0.036)
Panel C. Month of year				
January	0.207	−0.026	0.464	−0.156
	(0.064)	(0.036)	(0.131)	(0.060)
February	0.140	−0.014	0.309	−0.129
	(0.074)	(0.039)	(0.156)	(0.066)
March	0.200	−0.032	0.372	−0.100
	(0.087)	(0.034)	(0.184)	(0.061)
April	0.315	−0.132	0.519	−0.150
	(0.078)	(0.045)	(0.158)	(0.071)
May	0.354	−0.085	0.641	−0.151
	(0.081)	(0.039)	(0.151)	(0.072)
June	0.269	−0.057	0.509	−0.151
	(0.080)	(0.033)	(0.181)	(0.067)
July	0.227	−0.063	0.428	−0.138
	(0.075)	(0.029)	(0.155)	(0.053)
August	0.244	−0.026	0.517	−0.168
	(0.084)	(0.031)	(0.187)	(0.051)
September	0.353	−0.063	0.669	−0.250
	(0.084)	(0.038)	(0.189)	(0.062)
October	0.142	−0.076	0.257	−0.280
	(0.082)	(0.041)	(0.174)	(0.087)
November	0.065	−0.074	0.275	−0.212
	(0.099)	(0.045)	(0.197)	(0.078)
December	0.154	−0.051	0.195	−0.229
	(0.091)	(0.035)	(0.161)	(0.061)

Notes: Each panel-column reports a separate weighted regression, with postmarket dispatch interacted with the indicated variables. Gas price categories are terciles by PCA-year. All specifications are based on column 4 of Tables 2 and 3 with outcomes in logarithms. Standard errors clustered by PCA-month in parentheses.

coefficients represent the differential effect of markets during the days of the year when gas was relatively cheap or expensive. Confirming the intuition described in Section III, gains from trade are higher when fuel is more expensive. It appears the

propensity to trade is also higher, while reductions in out-of-merit costs are somewhat stronger when gas prices are low.

Panel C breaks out estimates by month of year, and presents evidence of seasonality as well as complementarity between the measures of allocative efficiency. The largest impacts occur during the mild, low-demand periods that generators typically use for seasonal maintenance. This pattern suggests that markets keep utilities from favoring their own higher-cost units during maintenance, and instead coordinate supply of lower-cost power across PCAs. These results complement the prior findings of Davis and Wolfram (2012), who show that merchant nuclear units reduce their down time. The timing of those divestitures largely precede the transition to market dispatch, so the results presented here should be interpreted as mostly in addition to those found previously. Furthermore, the decompositions of equations (2) and (3) make clear that there is no a priori reason that these effects would go hand in hand—a PCA that reduces its out-of-merit generation may find itself less reliant on outside generation sources. The impact of greater trade on out-of-merit generation (and vice-versa) is ambiguous in general. For whatever temptations may exist to exert market power by taking economical units offline during moments of peak demand, these estimates find that markets (combined with diligent monitoring) have improved allocative efficiency of the generation sector throughout the year.

VI. Conclusion

In this paper I use the recent introduction of wholesale electricity markets in some areas as a natural experiment to evaluate the performance of markets relative to the policy-relevant counterfactual: centralized dispatch by a regulated private or government-owned local monopolist. These are starkly different mechanisms for balancing supply and demand within the boundaries of the 98 PCAs that have historically managed the US electricity grid. To evaluate how this change has affected the cost of meeting demand I construct a detailed 14-year panel of hourly load (i.e., consumption) for each PCA, and the costs, capacities, and operations of power plants.

It is insufficient to simply evaluate the cost of generation for each PCA in isolation because trading power between PCAs is a key mode of cost reduction. The bilateral flows of power are unobserved, so it is necessary to account for the costs and surpluses of traded power without knowing where imported power is coming from, nor where exported power is going.

I account for trade by decomposing observed production costs relative to those of the most economical units in each PCA. I use two components of this decomposition as outcomes to evaluate the impact of markets. The first, "out-of-merit costs," measure how closely production adheres to utilizing only the lowest-cost units. The second, "gains from trade" measures the surplus that accrues to each PCA from trading power according to this low-cost frontier.

I find that market-based dispatch has caused a 16 percent reduction in out-of-merit costs, while increasing gains from trade by 55 percent—a reduction in production costs of between $3 and $5 billion per year. These savings are worth roughly 5 percent of the total variable cost of generating electricity in market areas. This is less than the 8 percent one would measure from running the analysis on observed

costs alone, as the increase in trade in my sample has been relatively import intensive. Without accounting for trade, greater imports would appear to be an increase in free power, and therefore larger cost reductions.

While the estimated allocative efficiency improvements caused by market dispatch are substantial, they are likely part of a much bigger story. These short-run estimates are based on responses to institutional changes imposed on a grid that was built for reliability rather than massive transregional exchange. This inherently imposes an upper bound on the potential gains that might be observed with this estimation strategy, but is a constraint that may be relaxed over time as locational marginal prices reveal profitable transmission investments. Further, the focus on production costs leaves open the possibility that overall welfare may have been differentially impacted by the introduction of wholesale markets by shifting output to generators with more or fewer external costs (Palmer and Burtraw 2005).

About 40 percent of electricity in the United States continues to be generated by plants called upon to operate based on the decision-making of a local balancing authority. It is difficult to say whether market dispatch would similarly benefit the remaining areas that have chosen to retain their traditional dispatch methods. It is nonetheless important to understand the balance between market failures and regulatory shortcomings thus far. While market power is certainly a concern for market monitors (Wolak 2014 shows their work is critical), my results suggest the benefits realized by more efficient allocation of output though market-based dispatch have far outweighed such losses in areas that have adopted wholesale markets to determine production.

REFERENCES

Archsmith, James. 2017. "Dam Spillovers: Direct Costs and Spillovers from Environmental Constraints on Hydroelectric Generation." Unpublished.

Borenstein, Severin. 2002. "The Trouble with Electricity Markets: Understanding California's Restructuring Disaster." *Journal of Economic Perspectives* 16 (1): 191–211.

Borenstein, Severin, and James B. Bushnell. 2015. "The U.S. Electricity Industry after 20 Years of Restructuring." *Annual Review of Economics* 7 (1): 437–63.

Borenstein, Severin, James B. Bushnell, Christopher R. Knittel, and Catherine D. Wolfram. 2008. "Inefficiencies and Market Power in Financial Arbitrage: A Study of California's Electricity Markets." *Journal of Industrial Economics* 56 (2): 347–78.

Borenstein, Severin, James B. Bushnell, and Steven Stoft. 2000. "The Competitive Effects of Transmission Capacity in a Deregulated Electricity Industry." *RAND Journal of Economics* 31(2): 294–325.

Borenstein, Severin, James B. Bushnell, and Frank A. Wolak. 2002. "Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market." *American Economic Review* 92 (5): 1376–1405.

Bushnell, James B., Erin T. Mansur, and Celeste Saravia. 2008. "Vertical Arrangements, Market Structure, and Competition: An Analysis of Restructured US Electricity Markets." *American Economic Review* 98 (1): 237–66.

Cicala, Steve. 2015. "When Does Regulation Distort Costs? Lessons from Fuel Procurement in US Electricity Generation." *American Economic Review* 105 (1): 411–44.

Cicala, Steve. 2022. "Replication Data for: Imperfect Markets versus Imperfect Regulation in US Electricity Generation." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E146802V1.

Cramton, Peter. 2003. "Electricity Market Design: The Good, the Bad, and the Ugly." In *Proceedings of the 36th Hawaii International Conference on System Sciences*, edited by Ralph H. Sprague, 8. Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Cullen, Joseph A. 2011. "Dynamic Response to Environmental Regulation in the Electricity Industry." Unpublished.

Davis, Lucas W., and Catherine D. Wolfram. 2012. "Deregulation, Consolidation, and Efficiency: Evidence from US Nuclear Power." *American Economic Journal: Applied Economics* 4 (4): 194–225.

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

Doraszelski, Ulrich, Gregory Lewis, and Ariel Pakes. 2018. "Just Starting Out: Learning and Equilibrium in a New Market." *American Economic Review* 108 (3): 565–615.

Energy Information Administration (EIA). 2012. *Annual Energy Review*. Washington DC: US Energy Information Administration.

Fabrizio, Kira R., Nancy L. Rose, and Catherine D. Wolfram. 2007. "Do Markets Reduce Costs? Assessing the Impact of Regulatory Restructuring on US Electric Generation Efficiency." *American Economic Review* 97 (4): 1250–77.

Fowlie, Meredith L. 2010. "Emissions Trading, Electricity Restructuring, and Investment in Pollution Abatement." *American Economic Review* 100 (3): 837–69.

Green, Richard J., and David M. Newberry. 1992. "Competition in the British Electricity Spot Market." *Journal of Political Economy* 100 (5): 929–53.

Hausman, Catherine, and Ryan Kellogg. 2015. "Welfare and Distributional Implications of Shale Gas." *Brookings Papers on Economic Activity* 45 (1): 71–125.

Hortaçsu, Ali, and Steven L. Puller. 2008. "Understanding Strategic Bidding in Multi-unit Auctions: A Case Study of the Texas Electricity Spot Market." *RAND Journal of Economics* 39 (1): 86–114.

Ito, Koichiro, and Mar Reguant. 2016. "Sequential Markets, Market Power and Arbitrage." *American Economic Review* 106 (7): 1921–57.

Jha, Akshaya, and Frank A. Wolak. 2013. "Testing for Market Efficiency with Transactions Costs: An Application to Convergence Bidding in Wholesale Electricity Markets." Unpublished.

Joskow, Paul L. 2000. "Deregulation and Regulatory Reform in the U.S. Electric Power Sector." In *Deregulation of Network Industries: What's Next?* edited by Sam Peltzman and Clifford Winston, 113–88. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

Joskow, Paul L. 2010. "Market Imperfections versus Regulatory Imperfections." *ifo DICE Report* 8 (3): 3–7.

Joskow, Paul L., and Edward Kahn. 2002. "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market During Summer 2000." *Energy Journal* 23 (4): 1–35.

Joskow, Paul L., and Richard Schmalensee. 1988. *Markets for Power: An Analysis of Electrical Utility Deregulation.* Cambridge, MA: MIT Press.

Kabir, Malik, Maureen Cropper, Alexander Limonov, and Anoop Singh. 2011. "Estimating the Impact of Restructuring on Electricity Generation Efficiency: The Case of the Indian Thermal Power Sector." NBER Working Paper 17383.

Kahn, Alfred Edward. 1979. "Applications of Economics to an Imperfect World." *American Economic Review* 69 (2): 1–13.

Knittel, Christopher R., Konstantinos Metaxoglou, and Andre Trindade. 2014. "Dash for Gas: The Sequel." Unpublished.

Kwoka, John. 2008. "Restructuring the U.S. Electric Power Sector: A Review of Recent Studies." *Review of Industrial Organization* 32 (3): 165–96.

Linn, Joshua, Lucija Muehlenbachs, and Yushuang Wang. 2014. "How Do Natural Gas Prices Affect Electricity Consumers and the Environment?" Unpublished.

Mansur, Erin T. 2001. "Pricing Behavior in the Initial Summer of the Restructured PJM Wholesale Electricity Market." Unpublished.

Mansur, Erin T. 2008. "Measuring Welfare in Restructured Electricity Markets." *Review of Economics and Statistics* 90 (2): 369–386.

Mansur, Erin T., and Matthew W. White. 2012. "Market Organization and Efficiency in Electricity Markets." Unpublished.

Palmer, Karen, and Dallas Burtraw. 2005. "The Environmental Impacts of Electricity Restructuring: Looking Back and Looking Forward." *Environmental and Energy Law and Policy Journal* 1 (1): 171–219.

Puller, Steven L. 2007. "Pricing and Firm Conduct in California's Deregulated Electricity Market." *Review of Economics and Statistics* 89 (1): 75–87.

Reguant, Mar. 2014. "Complementary Bidding Mechanisms: An Application to Electricity Markets." *Review of Economic Studies* 81 (4): 1708–42.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2004. "Kidney Exchange." *Quarterly Journal of Economics* 119 (2): 457–88.

Ryan, Nicholas. 2013. "The Competitive Effects of Transmission Infrastructure in the Indian Electricity Market." Unpublished.

Sweeting, Andrew. 2007. "Market Power in the England and Wales Wholesale Electricity Market 1995-2000." *Economic Journal* 117 (520): 654–85.

Wolak, Frank A. 2000. "An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market." *International Economic Journal* 14 (2): 1–39.

Wolak, Frank A. 2014. "Regulating Competition in Wholesale Electricity Supply." In *Economic Regulation and Its Reform: What Have We Learned?* edited by Nancy L. Rose, 195–290. Chicago: University of Chicago Press.

Wolak, Frank A., and Robert H. Patrick. 1997. "The Impact of Market Rules and Market Structure on the Price Determination Process in the England and Wales Electricity Market." Unpublished.

Wolfram, Catherine D. 1999. "Measuring Duopoly Power in the British Electricity Spot Market." *American Economic Review* 89 (4): 805–26.

Wolfram, Catherine D. 2005. "The Efficiency of Electricity Generation in the U.S. After Restructuring." In *Electricity Deregulation: Choices and Challenges,* edited by James M. Griffin and Steven L. Puller, 227–55. Chicago: University of Chicago Press.