MARKETSHEARD ON THE STREET

Utilities Have a High-Wire Act Ahead

Rising fuel prices and interest rates could test utilities' ability to increase their earning potential without overly burdening customers

By Jinjoo Lee Follow

Updated Oct. 9, 2022 10:03 am ET

Utilities are meant to serve both the customers who pay the bills and the investors who fund them. For years, low interest rates and cheap natural gas made it easy to please both stakeholders. Today's environment could break down that win-win formula.

Thus far, high natural-gas prices have been a problem for consumers, not utilities, many of which automatically pass on the cost of fuel to customers. But trouble for utilities could start the next time they ask regulators for a bump in the revenue they can collect. In what is known as the rate-case process, a utility has to make the case for a rate increase that depends partly on what it costs to improve and maintain its service (say, a new transmission line) and partly on what it costs to fairly compensate investors.

The higher the burden on consumers, the bigger the risk that regulatory commissions will take a long, hard look at whether a rate increase is warranted. Utility regulators are typically elected or else appointed by elected officials, so they can be sensitive to ratepayer concerns.

In the last decade or so, rate cases have been a breeze for utilities. Low natural-gas prices meant they could get aggressive capital-spending plans approved without causing big utility bill shocks to customers, according to Lillian Federico, energy research director at S&P Global Commodity Insights. Those lower costs might also have helped utilities persuade regulators to keep approving attractive returns on equity rather than passing on the declining cost of capital to consumers.

A recent working paper by Karl Dunkle Werner and Stephen Jarvis published by the Energy Institute at Haas showed that the inflation-adjusted return regulators allow equity investors to earn has been steady over the past 40 years, even while various measures of capital cost— such as the U.S. Treasury yield—have been declining. The study found that utilities were quick to ask for increases on their return on equity when market measures of capital cost rose and regulators were quick to respond.

Conversely, when cost of capital measures declined, utilities were slower to adjust those rates. The researchers estimate that consumers might be paying anywhere from $2 billion to $20 billion a year more than they otherwise would if rates of return fell in line with capital-market trends.

In any other year, that could just be an interesting academic finding. But in an environment in which both fuel prices and interest rates are rising so quickly, it might give regulators pause. Given their record, utilities are likely to ask for higher returns on equity given rising interest rates, but getting approval might not be a breeze. Last year, electric and gas utilities tracked by S&P Global Commodity Insights collectively asked for $15 billion in rate increases, the biggest bump since 2000. So far this year, they have requested $12.4 billion in rate increases.

Of course, utilities might make the case that high natural-gas and coal prices are just the reason regulators should allow larger capital-spending plans for solar, wind and other grid improvements. Clean-energy incentives in the Inflation Reduction Act should also support such investments. But short-term shocks to customer bills could nevertheless make it hard for utilities to convince regulators. "The growth opportunity [for utilities] is even better today, but rising bills could be the thing that derails some of that," said Jay Rhame, chief executive of Reaves Asset Management, which manages utilities-focused funds.

For now, the fears of a recession seem to have overridden those concerns among investors. Utilities in the S&P 500 are down 11% year to date, outperforming the rest of the index by 13 percentage points.

Utilities are indeed more defensive than most other sectors, but no industry is a perfect shelter from disgruntled customers.

Write to Jinjoo Lee at jinjoo.lee@wsj.com

https://www.wsj.com/articles/utilities-have-a-high-wire-act-ahead-11665274525

MARKETSHEARD ON THE STREET

Utilities Have a High-Wire Act Ahead

Rising fuel prices and interest rates could test utilities' ability to increase their earning potential without overly burdening customers

By Jinjoo Lee Follow

Updated Oct. 9, 2022 10:03 am ET

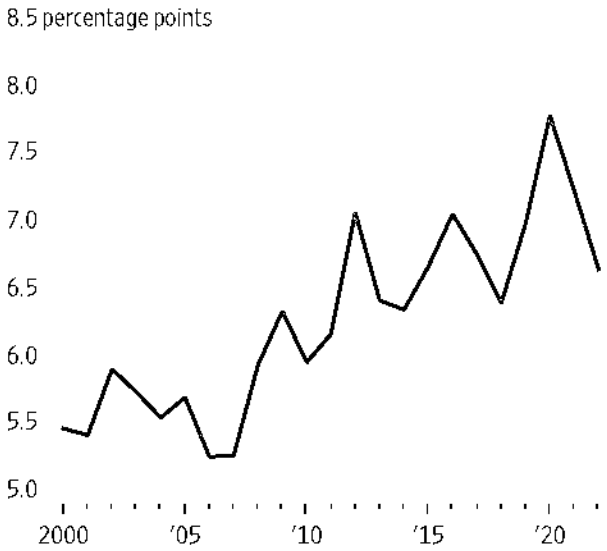
Utilities are meant to serve both the customers who pay the bills and the investors who fund them. For years, low interest rates and cheap natural gas made it easy to please both stakeholders. Today's environment could break down that win-win formula.

Thus far, high natural-gas prices have been a problem for consumers, not utilities, many of which automatically pass on the cost of fuel to customers. But trouble for utilities could start the next time they ask regulators for a bump in the revenue they can collect. In what is known as the rate-case process, a utility has to make the case for a rate increase that depends partly on what it costs to improve and maintain its service (say, a new transmission line) and partly on what it costs to fairly compensate investors.

The higher the burden on consumers, the bigger the risk that regulatory commissions will take a long, hard look at whether a rate increase is warranted. Utility regulators are typically elected or else appointed by elected officials, so they can be sensitive to ratepayer concerns.

Electric Returns

Average authorized return on equity for electric utilities, spread against the 30-year Treasury yield



Source: S&P Global Market Intelligence

In the last decade or so, rate cases have been a breeze for utilities. Low natural-gas prices meant they could get aggressive capital-spending plans approved without causing big utility bill shocks to customers, according to Lillian Federico, energy research director at S&P Global Commodity Insights. Those lower costs might also have helped utilities persuade regulators to keep approving attractive returns on equity rather than passing on the declining cost of capital to consumers.

A recent working paper by Karl Dunkle Werner and Stephen Jarvis published by the Energy Institute at Haas showed that the inflation-adjusted return regulators allow equity investors to earn has been steady over the past 40 years, even while various measures of capital cost—such as the U.S. Treasury yield—have been declining. The study found that utilities were quick to ask for increases on their return on equity when market measures of capital cost rose and regulators were quick to respond.

Conversely, when cost of capital measures declined, utilities were slower to adjust those rates. The researchers estimate that consumers might be paying anywhere from $2 billion to $20 billion a year more than they otherwise would if rates of return fell in line with capital-market trends.

In any other year, that could just be an interesting academic finding. But in an environment in which both fuel prices and interest rates are rising so quickly, it might give regulators pause. Given their record, utilities are likely to ask for higher returns on equity given rising interest rates, but getting approval might not be a breeze. Last year, electric and gas utilities tracked by S&P Global Commodity Insights collectively asked for $15 billion in rate increases, the biggest bump since 2000. So far this year, they have requested $12.4 billion in rate increases.

Of course, utilities might make the case that high natural-gas and coal prices are just the reason regulators should allow larger capital-spending plans for solar, wind and other grid improvements. Clean-energy incentives in the Inflation Reduction Act should also support such investments. But short-term shocks to customer bills could nevertheless make it hard for utilities to convince regulators. "The growth opportunity [for utilities] is even better today, but rising bills could be the thing that derails some of that," said Jay Rhame, chief executive of Reaves Asset Management, which manages utilities-focused funds.

For now, the fears of a recession seem to have overridden those concerns among investors. Utilities in the S&P 500 are down 11% year to date, outperforming the rest of the index by 13 percentage points.

Utilities are indeed more defensive than most other sectors, but no industry is a perfect shelter from disgruntled customers.

Write to Jinjoo Lee at jinjoo.lee@wsj.com

FINANCIAL FOCUS

Amid historic broad-market sell-off in H1, utilities maintain premium valuation

Thursday, July 7, 2022 8:52 AM ET

By Jason Lehmann
Market Intelligence

Despite an approximately 5% decline in June and a move into negative territory for the year, the S&P 500 Utilities group again outperformed broad indexes, which continued their decidedly downward trend en route to the worst first half-year for U.S. stock markets in more than a half century.

The Take

The S&P 500 Utilities index extended its valuation premium to approximately 24% over the S&P 500 in June on relative outperformance to other S&P 500 subsectors amid soaring inflation and growing concerns of the likelihood of another recession.

The S&P 500 Utilities index is down 2% year-to-date through June, second only to the S&P 500 Energy subsector, which remains up 29% year-to-date despite heavy selling during the month in tandem with the fall in oil prices. Selling pressure remains concentrated within the consumer discretionary, communications services and information technology subsectors, which are down 33%, 30% and 27% year-to-date, respectively.

Amid the dimming U.S. economic outlook, the pace of state-level utility rate case activity remains fairly robust. As of June 9, there were 109 electric and gas rate proceedings pending in 38 states.

To recover higher costs associated with inflation and potentially rising interest rates, RRA expects elevated rate case activity in 2022. Rising energy prices, which are placing additional pressure on customer bills, coupled with regulators' focus on rate affordability have the potential to challenge rate case outcomes in the near term. For additional detail, see the June 16 RRA Regulatory Focus report Electric, gas rate case activity remains robust amid dimming US economic outlook.

The S&P 500 Utilities was down 2% year-to-date through June. The S&P 500 was down 20.6% through the first six months of the year following June's 8.4% decline — its worst start since 1970, when the large-cap index dropped 21% in the first half of the year.

The Dow Jones Industrial Average and the tech-heavy Nasdaq composite index declined 15.3% and 29.5%, respectively, in the first half of 2022.

S&P 500 Utilities, S&P 500 YTD performance (%)

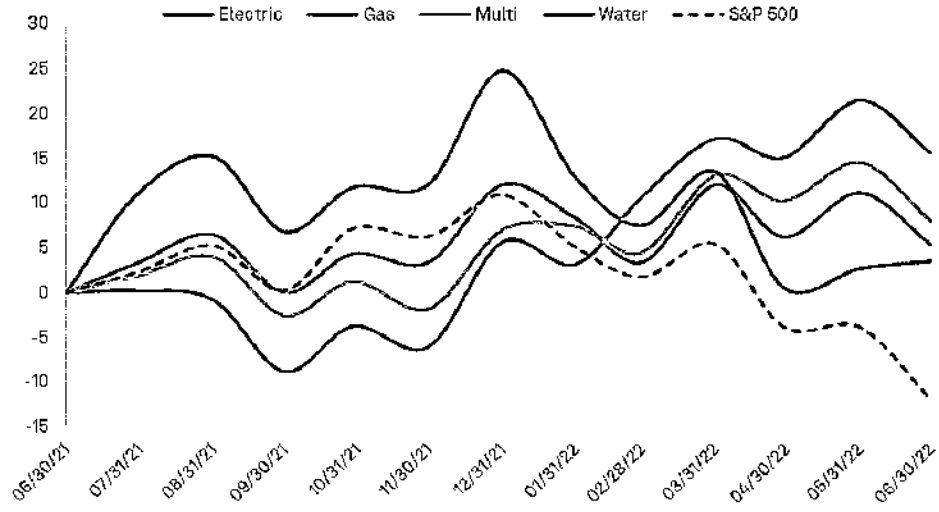
As of June 30, 2022, close.
Source: S&P Global Market Intelligence

Water utilities outperformed energy utility stocks in June, with investors likely taking advantage of recent share price weakness that has dampened sector valuations thus far in 2022. California Water Service Group, Middlesex Water Co. and American States Water Co. rose 3.5%, 3.1% and 2.9%, respectively, in June, yet remain the sector's worst performers year-to-date. Artesian Resources Corp. is the lone water utility stock in positive territory in 2022, up approximately 6%.

Simmering M&A activity may also be at play in investors' move into water utility equities in June; numerous transactions are pending in states spanning nationwide. There has also been increased diversity in the acquirers of water and wastewater systems to include not only smaller investor-owned and private-equity-funded water utilities but also larger investor-owned entities, including some outside the traditional water sector. For additional detail, see the June 23 Financial Focus report Pa. wastewater transactions dominate sector's acquisition market.

American Water Works Co. Inc. currently carries the highest forward share price-to-estimated EPS, or P/E, multiple, at 30.8x, above the group's 27.3x average. By comparison, average multi-utility, electric and gas utility forward P/Es stand at 19.1x, 18x and 17.8x, respectively.

Utility monthly average share price change (%)



Prices are through June 30, 2022.
Source: S&P Global Market Intelligence

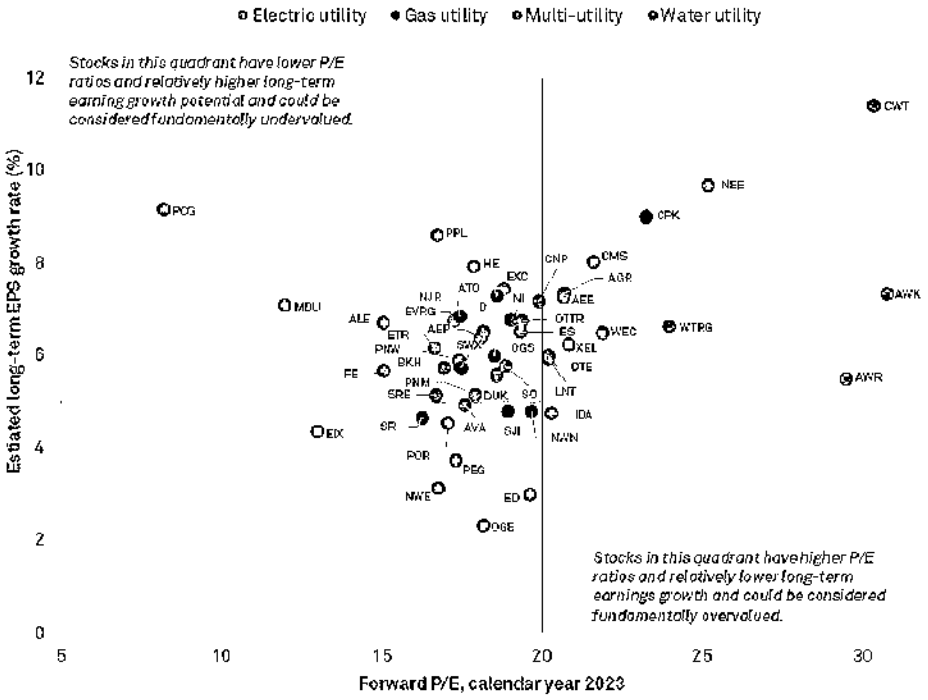
Only four energy utilities registered month-over-month share price gains in June, led by Otter Tail Corp., up 2.7% to pare its year-to-date loss to 6%. Shares had slumped through mid-June before recovering on heightened trading volume. Otter Tail shares currently trade at 19.4x the 2023 S&P Capital IQ consensus EPS estimate, above the electric utility group's 17.8x average. Recent share price appreciation may be attributed to the company's strong EPS outlook, with management recently increasing 2022 guidance to a range of $5.15 to $5.45 on record first-quarter results and expected performance from its plastics segment.

NextEra Energy Inc.'s forward P/E multiple increased 1.1% in June to 25.2x — the highest among electric utilities — after shares increased 2.3% following the company's investor and analyst day. On June 14, NextEra unveiled an ambitious decarbonization program and reassured investors that it can continue to hit growth and climate targets despite inflation and a potential recession. For additional detail, see the June 14 S&P Global Market Intelligence news article.

Within the electric utility sector, forward P/Es declined approximately 5% in June to 17.8x. Multi-utilities covered by RRA, a group within S&P Global Commodity Insights, saw forward P/Es decline 5.7% on average to just above 18x, and the average gas utility P/E stood at 18.8x, down 3.6% from May.

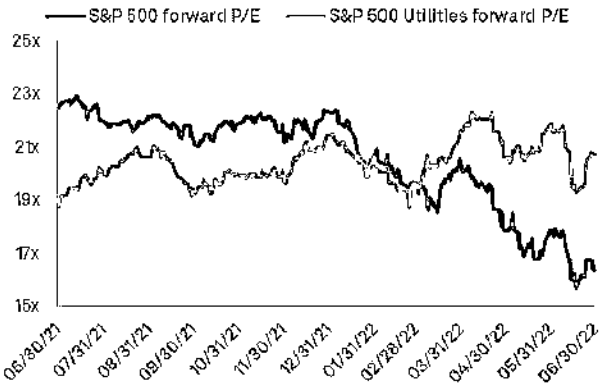
The quadrant chart below shows how Regulatory Research Associates' utility universe appears when comparing the P/E ratio and the estimated long-term earnings growth rate. A sizeable portion of electric utility P/E multiples has remained in the upper-left quadrant in 2022, suggesting the companies could be relatively undervalued considering their lower P/E values and long-term earnings growth potential.

Valuation quadrant: EPS growth forecast vs. forward P/E



As of June 30, 2022, close.
For the 12 months ending Dec. 31, 2023.
P/E = stock price-to-earnings ratio
Source: S&P Global Market Intelligence

S&P Utilities, S&P 500 next-12-months P/E estimates



As of June 30, 2022, close.
P/E = stock price-to-estimated EPS multiple
Source: S&P Global Market Intelligence

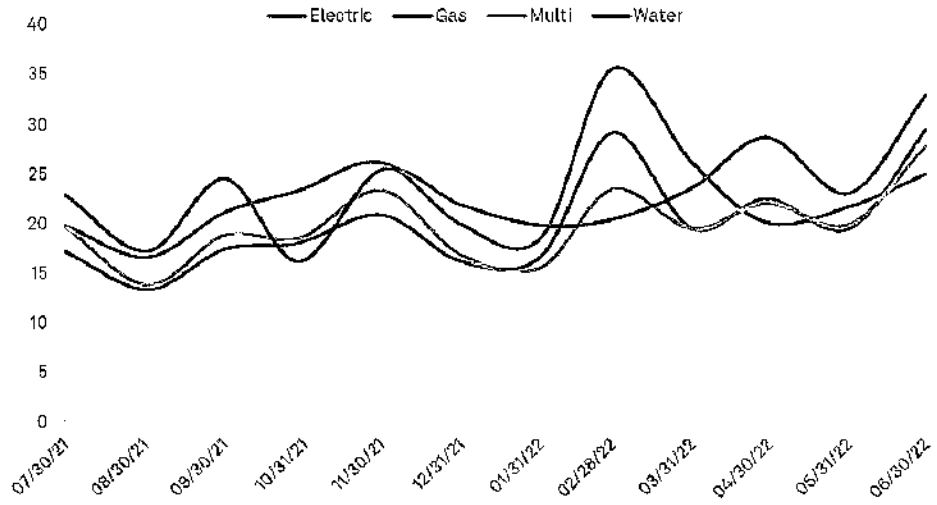
Share price volatility

Smaller-cap companies generally have lower trading liquidity, and, therefore, all other things being equal, tend to have more significant share-price swings than larger-cap equities. An analysis of the standard deviation of log-normalized daily price returns for utility stocks over the past year supports this thesis, with the smaller-cap water utility sector displaying the highest average price volatility.

Electric utility stocks saw the largest increase in average share price volatility — to 29.5% from 19.5% in May — led by OGE Energy Corp. The OGE shares declined 6.6% in June, essentially reversing May's gains.

The company's Oklahoma Gas and Electric Co. subsidiary is presently seeking a $163.5 million rate increase in Oklahoma, driven by the need for a higher return on equity, revised depreciation rates and an expansion of the grid enhancement mechanism to include certain weather hardening upgrades. Oklahoma Corporation Commission staff recently recommended a base rate hike that is about half of what the company requested earlier in 2022, premised upon an 8.75% ROE, which is well below the prevailing industrywide averages for electric utilities. Staff further recommended that OG&E's request to operate under a performance-based ratemaking framework be rejected. A final OCC decision is expected to be issued by September (Cause No. PUD202100164).

Utility monthly share price volatility (%)



Prices are through June 30, 2022.
Volatility is calculated as the annualized standard deviation of daily log-normal price returns over each month.
Source: S&P Global Market Intelligence

Regulatory Research Associates is a group within S&P Global Commodity Insights.

S&P Global Commodity Insights produces content for distribution on S&P Capital IQ Pro.

Mr. Buffett on the Stock Market The most celebrated of investors says stocks can't possibly meet the public's expectations. As for the Internet? He notes how few people got rich from two other transforming industries, auto and aviation.

By Warren Buffett; Carol Loomis
November 22, 1999
(FORTUNE Magazine) – Warren Buffett, chairman of Berkshire Hathaway, almost never talks publicly about the general level of stock prices--neither in his famed annual report nor at Berkshire's thronged annual meetings nor in the rare speeches he gives. But in the past few months, on four occasions, Buffett did step up to that subject, laying out his opinions, in ways both analytical and creative, about the long-term future for stocks. FORTUNE's Carol Loomis heard the last of those talks, given in September to a group of Buffett's friends (of whom she is one), and also watched a videotape of the first speech, given in July at Allen & Co.'s Sun Valley, Idaho, bash for business leaders. From those extemporaneous talks (the first made with the Dow Jones industrial average at 11,194), Loomis distilled the following account of what Buffett said. Buffett reviewed it and weighed in with some clarifications.
Investors in stocks these days are expecting far too much, and I'm going to explain why. That will inevitably set me to talking about the general stock market, a subject I'm usually unwilling to discuss. But I want to make one thing clear going in: Though I will be talking about the level of the market, I will not be predicting its next moves. At Berkshire we focus almost exclusively on the valuations of individual companies, looking only to a very limited extent at the valuation of the overall market. Even then, valuing the market has nothing to do with where it's going to go next week or next month or next year, a line of thought we never get into. The fact is that markets behave in ways, sometimes for a very long stretch, that are not linked to value. Sooner or later, though, value counts. So what I am going to be saying--assuming it's correct--will have implications for the long-term results to be realized by American stockholders.

Let's start by defining "investing." The definition is simple but often forgotten: Investing is laying out money now to get more money back in the future--more money in real terms, after taking inflation into account.
Now, to get some historical perspective, let's look back at the 34 years before this one--and here we are going to see an almost Biblical kind of symmetry, in the sense of lean years and fat years--to observe what happened in the stock market. Take, to begin with, the first 17 years of the period, from the end of 1964 through 1981. Here's what took place in that interval:
DOW JONES INDUSTRIAL AVERAGE Dec. 31, 1964: 874.12 Dec. 31, 1981: 875.00
Now I'm known as a long-term investor and a patient guy, but that is not my idea of a big move.

And here's a major and very opposite fact: During that same 17 years, the GDP of the U.S.--that is, the business being done in this country--almost quintupled, rising by 370%. Or, if we look at another measure, the sales of the FORTUNE 500 (a changing mix of companies, of course) more than sextupled. And yet the Dow went exactly nowhere.

To understand why that happened, we need first to look at one of the two important variables that affect investment results: interest rates. These act on financial valuations the way gravity acts on matter: The higher the rate, the greater the downward pull. That's because the rates of return that investors need from any kind of investment are directly tied to the risk-free rate that they can earn from government securities. So if the government rate rises, the prices of all other investments must adjust downward, to a level that brings their expected rates of return into line. Conversely, if government interest rates fall, the move pushes the prices of all other investments upward. The basic proposition is this: What an investor should pay today for a dollar to be received tomorrow can only be determined by first looking at the risk-free interest rate.

Consequently, every time the risk-free rate moves by one basis point--by 0.01%--the value of every investment in the country changes. People can see this easily in the case of bonds, whose value is normally affected only by interest rates. In the case of equities or real estate or farms or whatever, other very important variables are almost always at work, and that means the effect of interest rate changes is usually obscured. Nonetheless, the effect--like the invisible pull of gravity--is constantly there.

In the 1964-81 period, there was a tremendous increase in the rates on long-term government bonds, which moved from just over 4% at year-end 1964 to more than 15% by late 1981. That rise in rates had a huge depressing effect on the value of all investments, but the one we noticed, of course, was the price of equities. So there--in that tripling of the gravitational pull of interest rates--lies the major explanation of why tremendous growth in the economy was accompanied by a stock market going nowhere.

Then, in the early 1980s, the situation reversed itself. You will remember Paul Volcker coming in as chairman of the Fed and remember also how unpopular he was. But the heroic things he did--his taking a two-by-four to the economy and breaking the back of inflation--caused the interest rate trend to reverse, with some rather spectacular results. Let's say you put $1 million into the 14% 30-year U.S. bond issued Nov. 16, 1981, and reinvested the coupons. That is, every time you got an interest payment, you used it to buy more of that same bond. At the end of 1998, with long-term governments by then selling at 5%, you would have had $8,181,219 and would have earned an annual return of more than 13%.

That 13% annual return is better than stocks have done in a great many 17-year periods in history--in most 17-year periods, in fact. It was a helluva result, and from none other than a stodgy bond.

The power of interest rates had the effect of pushing up equities as well, though other things that we will get to pushed additionally. And so here's what equities did in that same 17 years: If you'd invested $1 million in the Dow on Nov. 16, 1981, and reinvested all dividends, you'd have had $19,720,112 on Dec. 31, 1998. And your annual return would have been 19%.

The increase in equity values since 1981 beats anything you can find in history. This increase even surpasses what you would have realized if you'd bought stocks in 1932, at their Depression bottom--on its lowest day, July 8, 1932, the Dow closed at 41.22--and held them for 17 years.

The second thing bearing on stock prices during this 17 years was after-tax corporate profits, which this chart [above] displays as a percentage of GDP. In effect, what this chart tells you is what portion of the GDP ended up every year with the shareholders of American business.

The chart, as you will see, starts in 1929. I'm quite fond of 1929, since that's when it all began for me. My dad was a stock salesman at the time, and after the Crash came, in the fall, he was afraid to call anyone--all those people who'd been burned. So he just stayed home in the afternoons. And there wasn't television then. Soooo... I was conceived on or about Nov. 30, 1929 (and born nine months later, on Aug. 30, 1930), and I've forever had a kind of warm feeling about the Crash.

As you can see, corporate profits as a percentage of GDP peaked in 1929, and then they tanked. The left-hand side of the chart, in fact, is filled with aberrations: not only the Depression but also a wartime

profits boom--sedated by the excess-profits tax--and another boom after the war. But from 1951 on, the percentage settled down pretty much to a 4% to 6.5% range.

By 1981, though, the trend was headed toward the bottom of that band, and in 1982 profits tumbled to 3.5%. So at that point investors were looking at two strong negatives: Profits were sub-par and interest rates were sky-high.

And as is so typical, investors projected out into the future what they were seeing. That's their unshakable habit: looking into the rear-view mirror instead of through the windshield. What they were observing, looking backward, made them very discouraged about the country. They were projecting high interest rates, they were projecting low profits, and they were therefore valuing the Dow at a level that was the same as 17 years earlier, even though GDP had nearly quintupled.

Now, what happened in the 17 years beginning with 1982? One thing that didn't happen was comparable growth in GDP: In this second 17-year period, GDP less than tripled. But interest rates began their descent, and after the Volcker effect wore off, profits began to climb--not steadily, but nonetheless with real power. You can see the profit trend in the chart, which shows that by the late 1990s, after-tax profits as a percent of GDP were running close to 6%, which is on the upper part of the "normalcy" band. And at the end of 1998, long-term government interest rates had made their way down to that 5%.

These dramatic changes in the two fundamentals that matter most to investors explain much, though not all, of the more than tenfold rise in equity prices--the Dow went from 875 to 9,181-- during this 17-year period. What was at work also, of course, was market psychology. Once a bull market gets under way, and once you reach the point where everybody has made money no matter what system he or she followed, a crowd is attracted into the game that is responding not to interest rates and profits but simply to the fact that it seems a mistake to be out of stocks. In effect, these people superimpose an I-can't-miss-the-party factor on top of the fundamental factors that drive the market. Like Pavlov's dog, these "investors" learn that when the bell rings--in this case, the one that opens the New York Stock Exchange at 9:30 a.m.--they get fed. Through this daily reinforcement, they become convinced that there is a God and that He wants them to get rich.

Today, staring fixedly back at the road they just traveled, most investors have rosy expectations. A Paine Webber and Gallup Organization survey released in July shows that the least experienced investors--those who have invested for less than five years--expect annual returns over the next ten years of 22.6%. Even those who have invested for more than 20 years are expecting 12.9%.

Now, I'd like to argue that we can't come even remotely close to that 12.9%, and make my case by examining the key value-determining factors. Today, if an investor is to achieve juicy profits in the market over ten years or 17 or 20, one or more of three things must happen. I'll delay talking about the last of them for a bit, but here are the first two:

(1) Interest rates must fall further. If government interest rates, now at a level of about 6%, were to fall to 3%, that factor alone would come close to doubling the value of common stocks. Incidentally, if you think interest rates are going to do that--or fall to the 1% that Japan has experienced--you should head for where you can really make a bundle: bond options.

(2) Corporate profitability in relation to GDP must rise. You know, someone once told me that New York has more lawyers than people. I think that's the same fellow who thinks profits will become larger than GDP. When you begin to expect the growth of a component factor to forever outpace that of the aggregate, you get into certain mathematical problems. In my opinion, you have to be wildly optimistic to believe that corporate profits as a percent of GDP can, for any sustained period, hold much above 6%. One thing keeping the percentage down will be competition, which is alive and well. In addition, there's a public-policy point: If corporate investors, in aggregate, are going to eat an ever-growing portion of the American economic pie, some other group will have to settle for a smaller portion. That would justifiably raise political problems--and in my view a major reslicing of the pie just isn't going to happen.

So where do some reasonable assumptions lead us? Let's say that GDP grows at an average 5% a year--3% real growth, which is pretty darn good, plus 2% inflation. If GDP grows at 5%, and you don't have some help from interest rates, the aggregate value of equities is not going to grow a whole lot more. Yes,

you can add on a bit of return from dividends. But with stocks selling where they are today, the importance of dividends to total return is way down from what it used to be. Nor can investors expect to score because companies are busy boosting their per-share earnings by buying in their stock. The offset here is that the companies are just about as busy issuing new stock, both through primary offerings and those ever present stock options.

So I come back to my postulation of 5% growth in GDP and remind you that it is a limiting factor in the returns you're going to get: You cannot expect to forever realize a 12% annual increase--much less 22%--in the valuation of American business if its profitability is growing only at 5%. The inescapable fact is that the value of an asset, whatever its character, cannot over the long term grow faster than its earnings do.

Now, maybe you'd like to argue a different case. Fair enough. But give me your assumptions. If you think the American public is going to make 12% a year in stocks, I think you have to say, for example, "Well, that's because I expect GDP to grow at 10% a year, dividends to add two percentage points to returns, and interest rates to stay at a constant level." Or you've got to rearrange these key variables in some other manner. The Tinker Bell approach--clap if you believe--just won't cut it.

Beyond that, you need to remember that future returns are always affected by current valuations and give some thought to what you're getting for your money in the stock market right now. Here are two 1998 figures for the FORTUNE 500. The companies in this universe account for about 75% of the value of all publicly owned American businesses, so when you look at the 500, you're really talking about America Inc.

FORTUNE 500 1998 profits: $334,335,000,000 Market value on March 15, 1999: $9,907,233,000,000

As we focus on those two numbers, we need to be aware that the profits figure has its quirks. Profits in 1998 included one very unusual item--a $16 billion bookkeeping gain that Ford reported from its spinoff of Associates--and profits also included, as they always do in the 500, the earnings of a few mutual companies, such as State Farm, that do not have a market value. Additionally, one major corporate expense, stock-option compensation costs, is not deducted from profits. On the other hand, the profits figure has been reduced in some cases by write-offs that probably didn't reflect economic reality and could just as well be added back in. But leaving aside these qualifications, investors were saying on March 15 this year that they would pay a hefty $10 trillion for the $334 billion in profits.

Bear in mind--this is a critical fact often ignored--that investors as a whole cannot get anything out of their businesses except what the businesses earn. Sure, you and I can sell each other stocks at higher and higher prices. Let's say the FORTUNE 500 was just one business and that the people in this room each owned a piece of it. In that case, we could sit here and sell each other pieces at ever-ascending prices. You personally might outsmart the next fellow by buying low and selling high. But no money would leave the game when that happened: You'd simply take out what he put in. Meanwhile, the experience of the group wouldn't have been affected a whit, because its fate would still be tied to profits. The absolute most that the owners of a business, in aggregate, can get out of it in the end--between now and Judgment Day--is what that business earns over time.

And there's still another major qualification to be considered. If you and I were trading pieces of our business in this room, we could escape transactional costs because there would be no brokers around to take a bite out of every trade we made. But in the real world investors have a habit of wanting to change chairs, or of at least getting advice as to whether they should, and that costs money--big money. The expenses they bear--I call them frictional costs--are for a wide range of items. There's the market maker's spread, and commissions, and sales loads, and 12b-1 fees, and management fees, and custodial fees, and wrap fees, and even subscriptions to financial publications. And don't brush these expenses off as irrelevancies. If you were evaluating a piece of investment real estate, would you not deduct management costs in figuring your return? Yes, of course--and in exactly the same way, stock market investors who are figuring their returns must face up to the frictional costs they bear.

And what do they come to? My estimate is that investors in American stocks pay out well over $100 billion a year--say, $130 billion--to move around on those chairs or to buy advice as to whether they should! Perhaps $100 billion of that relates to the FORTUNE 500. In other words, investors are dissipating almost

a third of everything that the FORTUNE 500 is earning for them--that $334 billion in 1998--by handing it over to various types of chair-changing and chair-advisory "helpers." And when that handoff is completed, the investors who own the 500 are reaping less than a $250 billion return on their $10 trillion investment. In my view, that's slim pickings.

Perhaps by now you're mentally quarreling with my estimate that $100 billion flows to those "helpers." How do they charge thee? Let me count the ways. Start with transaction costs, including commissions, the market maker's take, and the spread on underwritten offerings: With double counting stripped out, there will this year be at least 350 billion shares of stock traded in the U.S., and I would estimate that the transaction cost per share for each side--that is, for both the buyer and the seller--will average 6 cents. That adds up to $42 billion.

Move on to the additional costs: hefty charges for little guys who have wrap accounts; management fees for big guys; and, looming very large, a raft of expenses for the holders of domestic equity mutual funds. These funds now have assets of about $3.5 trillion, and you have to conclude that the annual cost of these to their investors--counting management fees, sales loads, 12b-1 fees, general operating costs--runs to at least 1%, or $35 billion.

And none of the damage I've so far described counts the commissions and spreads on options and futures, or the costs borne by holders of variable annuities, or the myriad other charges that the "helpers" manage to think up. In short, $100 billion of frictional costs for the owners of the FORTUNE 500--which is 1% of the 500's market value--looks to me not only highly defensible as an estimate, but quite possibly on the low side.

It also looks like a horrendous cost. I heard once about a cartoon in which a news commentator says, "There was no trading on the New York Stock Exchange today. Everyone was happy with what they owned." Well, if that were really the case, investors would every year keep around $130 billion in their pockets.

Let me summarize what I've been saying about the stock market: I think it's very hard to come up with a persuasive case that equities will over the next 17 years perform anything like--anything like--they've performed in the past 17. If I had to pick the most probable return, from appreciation and dividends combined, that investors in aggregate--repeat, aggregate--would earn in a world of constant interest rates, 2% inflation, and those ever hurtful frictional costs, it would be 6%. If you strip out the inflation component from this nominal return (which you would need to do however inflation fluctuates), that's 4% in real terms. And if 4% is wrong, I believe that the percentage is just as likely to be less as more.

Let me come back to what I said earlier: that there are three things that might allow investors to realize significant profits in the market going forward. The first was that interest rates might fall, and the second was that corporate profits as a percent of GDP might rise dramatically. I get to the third point now: Perhaps you are an optimist who believes that though investors as a whole may slog along, you yourself will be a winner. That thought might be particularly seductive in these early days of the information revolution (which I wholeheartedly believe in). Just pick the obvious winners, your broker will tell you, and ride the wave.

Well, I thought it would be instructive to go back and look at a couple of industries that transformed this country much earlier in this century: automobiles and aviation. Take automobiles first: I have here one page, out of 70 in total, of car and truck manufacturers that have operated in this country. At one time, there was a Berkshire car and an Omaha car. Naturally I noticed those. But there was also a telephone book of others.

All told, there appear to have been at least 2,000 car makes, in an industry that had an incredible impact on people's lives. If you had foreseen in the early days of cars how this industry would develop, you would have said, "Here is the road to riches." So what did we progress to by the 1990s? After corporate carnage that never let up, we came down to three U.S. car companies--themselves no lollapaloozas for investors. So here is an industry that had an enormous impact on America--and also an enormous impact, though not the anticipated one, on investors.

Sometimes, incidentally, it's much easier in these transforming events to figure out the losers. You could have grasped the importance of the auto when it came along but still found it hard to pick companies that would make you money. But there was one obvious decision you could have made back then--it's better sometimes to turn these things upside down--and that was to short horses. Frankly, I'm disappointed that the Buffett family was not short horses through this entire period. And we really had no excuse: Living in Nebraska, we would have found it super-easy to borrow horses and avoid a "short squeeze."

U.S. Horse Population 1900: 21 million 1998: 5 million

The other truly transforming business invention of the first quarter of the century, besides the car, was the airplane--another industry whose plainly brilliant future would have caused investors to salivate. So I went back to check out aircraft manufacturers and found that in the 1919-39 period, there were about 300 companies, only a handful still breathing today. Among the planes made then--we must have been the Silicon Valley of that age--were both the Nebraska and the Omaha, two aircraft that even the most loyal Nebraskan no longer relies upon.

Move on to failures of airlines. Here's a list of 129 airlines that in the past 20 years filed for bankruptcy. Continental was smart enough to make that list twice. As of 1992, in fact--though the picture would have improved since then--the money that had been made since the dawn of aviation by all of this country's airline companies was zero. Absolutely zero.

Sizing all this up, I like to think that if I'd been at Kitty Hawk in 1903 when Orville Wright took off, I would have been farsighted enough, and public-spirited enough--I owed this to future capitalists--to shoot him down. I mean, Karl Marx couldn't have done as much damage to capitalists as Orville did.

I won't dwell on other glamorous businesses that dramatically changed our lives but concurrently failed to deliver rewards to U.S. investors: the manufacture of radios and televisions, for example. But I will draw a lesson from these businesses: The key to investing is not assessing how much an industry is going to affect society, or how much it will grow, but rather determining the competitive advantage of any given company and, above all, the durability of that advantage. The products or services that have wide, sustainable moats around them are the ones that deliver rewards to investors.

This talk of 17-year periods makes me think--incongruously, I admit--of 17-year locusts [pictured below]. What could a current brood of these critters, scheduled to take flight in 2016, expect to encounter? I see them entering a world in which the public is less euphoric about stocks than it is now. Naturally, investors will be feeling disappointment--but only because they started out expecting too much.

Grumpy or not, they will have by then grown considerably wealthier, simply because the American business establishment that they own will have been chugging along, increasing its profits by 3% annually in real terms. Best of all, the rewards from this creation of wealth will have flowed through to Americans in general, who will be enjoying a far higher standard of living than they do today. That wouldn't be a bad world at all--even if it doesn't measure up to what investors got used to in the 17 years just passed.

The Size Premium in the Long Run

Ching-Chih Lu*

This Draft: December 25, 2009

Abstract

Contrary to the usual practice of including a size premium in a small firm's cost-of-equity estimation, this paper shows that there should not be such a premium in the long run because firm size is a changing characteristic. By tracking the return performance of firms in the same size group for a longer horizon, I find that the size premium wears off just after two years. This is much shorter than the general assumption used in the cost-of-equity estimation, so the role of the size premium in it should be reconsidered.

Keywords: Cost of Equity Capital, Size Premium, Size Effect, Regime Switching
JEL Classification: G12, G14

*Department of Finance, National Chengchi University, No. 64, Sec.2 Zhinan Rd., Mucha, Taipei 116, Taiwan. E-mail: cclu@nccu.edu.tw.

1 Introduction

In the field of business valuation, practitioners usually include a size premium in a small firm's cost-of-equity estimation to account for a risk source or risk sources that cannot be captured by usual risk factors.[1] That is, on top of the cost of equity a small firm gets from the estimation by the CAPM or other models, it is usually offered an extra premium to compensate for the higher risk it is taking.[2] This paper aims to examine its validity, and the finding suggests that this commonly accepted size premium is not appropriate.

Since Banz (1981) and Reinganum (1981) both demonstrated that small size firms on the New York Stock Exchange usually outperform big firms than what the asset-pricing model of Sharpe (1964), Lintner (1965) and Black (1972) would suggest, the existence of the size effect has come into consideration by standard practice in the finance industry and soon became one of the most exploited concepts in modern finance. This size anomaly leads to an assumption that it might stem from a risk source or risk sources which cannot be explained by the market factor. Berk (1995) explains in theory that market value is inversely correlated with unmeasured risk because investors pay a lower price for a company's stock if it bears a higher risk than its CAPM beta could measure. The seminal works of Fama and French (1993), and Fama and French (1995) also acknowledge another kind of size effect in which

[1] Although there are many ways to define the size of a company, I stick to the most popular criteria, the market value of its equity, to proceed the discussion.

[2] Other than the CAPM, the build-up method and the Fama-French 3-factor model are also popular approaches in business valuation. The build-up method is advocated by the Ibbotson Associates, now a part of Morningstar, Inc., which aims to break down the expected return of a firm into a risk-free rate, a premium for equity risk, a risk premium attributable to this company by the industry it is in, and another risk premium for smaller size if applicable. This size premium is added in practice no matter whether the CAPM model or the build-up method is used. Please see Pratt and Grabowski (2008) Chapter 12 for a thorough discussion. Such a size premium is not required in the Fama-French 3-factor model because size is a risk factor embedded in it already.

small firms usually outperform big firms in realize returns and they use the return differential between small and big stork portfolios (I call it "small stock premium" hereafter for convenience) as a risk factor (also known as *SMB*). If the CAPM holds well, the small stock premium should be proportional to the difference between the CAPM betas of small and big stock portfolios in cross section, and the size premium should not exist. However, empirical evidence shows that the small stock premium is usually much bigger than the CAPM could explain because small firms usually have a significant size premium, which links these two different perspectives of size anomalies together.

Besides serving as a measure of an alternative risk source, the idea of the existence of a small stock premium is often used in forming a trading strategy. Since the commence of the Dimensional Fund Advisors (DFA hereafter) in 1981, the strategy of overweighing small-cap stocks to exploit this small stock premium has been utilized extensively. This same concept is also used to construct ETFs featuring size as an important characteristic. There are currently at least 6 micro-cap and 40 small-cap ETFs trading on the U.S. stock exchanges.[3] The main attraction of these ETFs is to exploit their potentially higher returns over big firms or the market.

With all the acknowledgement from both academics and practitioners, however, there lies an inconsistency between these applications of the size effect. The usage of the *SMB* factor requires yearly rebalancing of the size portfolios, and a trading strategy related to firm size demands probably even more frequent position adjustments. However, the size premium added to a small firm's cost-of-equity estimation is based

[3]Size is an important characteristic of these ETFs. However, it may not be the "only" characteristic. For example, the Vanguard Group, a U.S. investment management company, has three ETFs related to small-cap firms. Their exchange ticker symbols are VB, VBR, and VBK, which account for a total of $2.79 billion capital at the end of 2007. VBK is the combination of small-cap and growth stocks, while VBR is a small-cap and value stock ETF.

on the assumption that a firm will carry this extra premium in its discount factor moving forward for an extended period of time. Fama and French (2007) explain that the small stock premium comes from small firms gaining market capitalization and subsequently becoming bigger firms, but a firm's size behaves more like a long-lasting characteristic in the size premium application, which contradicts the empirical evidence. Although we do not know for certain which small firm will move to a bigger size group because of its own success, we do know that firms shift between different size groups in subsequent years after they were first assigned to a certain size rank. The size premium of a firm should be time-varying even if the CAPM beta of the size portfolio is time-invariant, so the cost of equity capital estimation could or should be adjusted accordingly if size has to be taken into consideration.

The existence of the size effect is not always perceived with full faith. This issue has to be addressed first, otherwise the debate of the application of the size premium will become a vain attempt. In the early 1980s when a fierce debate was conducted about the existence and the explanation of the size effect, Roll (1983) and Blume and Stambaugh (1983) both question the empirical importance of this phenomenon because the magnitude of the size effect is too sensitive to the technique used to evaluate the risk-adjusted return. Keim (1983) and Reinganum (1983) show that most of the risk-adjusted abnormal return to small firms occurs in the first two weeks in January, thus makes this effect easily exploited. It was the evaluation and the existence of the size premium being challenged, but the small stock premium was mostly untouched. Fiercer challenges came in the late 1990s, when Booth, Keim, and Ziemba (2000) argue that the January effect is not significantly different from zero in the returns to the DFA 9-10 portfolio over the period 1982-1995,[4] and Horowitz,

[4]The DFA 9-10 portfolio includes stocks with the lowest 20% market capitalization according to NYSE breakpoints.

Loughran, and Savin (2000b) also claim that the size effect ceases to exist after it was made well known because its benefit has already been exploited. Small firms do not have higher returns over big firms from the early 1980s to the mid-to-late 1990s, so the existence of the size effect is in doubt and deserves a thorough examination.

In this paper I will show that the size effect in the traditional definition is still intact given a longer sample period. The disappearance of the size effect in the 1980s and 1990s probably stems from a sample selection bias because the effect re-emerged in the late 1990s. I also examine whether this sample selection anomaly is a recurring scenario with a longer history of stock prices and find that the similar event occurred from the 1940s to 1960s.

However, an analysis of the evolution of the size premium will show that it is inappropriate to attach a fixed amount of premium to the cost of equity of a firm simply because of its current market capitalization. For a small stock portfolio which does not rebalance since the day it was constructed, its annual return and the size premium are all declining over years instead of staying at a relatively stable level. This confirms that a small firm should not be expected to have a higher size premium going forward sheerly because it is small now.

The paper proceeds as follows. Section 2 introduces the data used in this study. All NYSE, AMEX and NASDAQ operating firms are included and they are sorted by their respective market capitalization to form size portfolios. I also examine whether the size effect disappeared during the 1980s and 1990s and discuss its possible impact in this section. Section 3 offers a forward looking perspective of the size effect in response to the assumption of Fama and French (2007) that the small stock premium mainly resulted from firms moving between different size groups. We can also see the evolution of the size premium of the small stock portfolio and find evidence to con-

clude that a small firm does not always have a larger size premium simply because of its current size. Section 4 provides a method to separate the size premium into different regimes with macroeconomic variables, which shows that it is also very difficult to estimate the size premium with a time-varying estimation. Section 5 offers concluding remarks.

2 Data Description and the Evidence of the Existence of the Size Effect

2.1 Data Description

Monthly stock return data used in this research are collected from the University of Chicago Center for Research in Security Prices (CRSP) database. All NYSE, AMEX and NASDAQ operating firms are included when they are available on the CRSP tape.[5] Unlike Fama and French (1992), this study does not exclude financial firms from the sample because financial leverage is not in discussion. Since the market capitalization of a firm is the only firm characteristic covered in this paper and I also do not incorporate the Compustat database for the book equity data of companies, the number of firms each year is also greater than research considering both size and book-to-market equity characteristics. This choice of sample also prevents the potential survival bias generated by the Compustat database, please see the discussion in Kothari, Shanken, and Sloan (1995). The sample period is from December 1925 to December 2008.

The market portfolio return used in this paper is the CRSP value-weighted return on all NYSE, AMEX, and NASDAQ stocks, and the risk free rate is the total return on 30-day Treasury bill calculated by Ibbotson Associates.

To sort firms into different deciles according to their relative size, I follow the Fama and French (1992, 1993) tradition to use a firm's market equity at the end of June each year as the measure of its size. A firm has to be on the CRSP tape in

[5] American Depository Receipts, closed-end funds, Real Estate Investment Trusts, and companies incorporated outside the U.S. are excluded, which means only firms with CRSP share code 12 or less are included in this research.

June of year t to be included in a size portfolio from July of year t to June of year $t+1$ and years after that.[6] All NYSE listed firms are ranked each year according to their June market value, then these firms are allocated equally into 10 size portfolios on the basis of their relative size, so each portfolio has the same number of NYSE firms. The breakpoints between size portfolios are extracted from these NYSE firms, and AMEX and NASDAQ firms are inserted into these portfolios according to their market capitalization relative to the portfolio breakpoints. The first decile (portfolio 1) contains the smallest firms and the 10th decile (portfolio 10) includes the largest firms. In December 2008, Portfolio 1 has 1,895 firms and portfolio 10 has 158.

2.2 Does the Size Effect Still Exist?

In response to the question raised by Horowitz, Loughran, and Savin (2000b) about whether the size effect still exists, some basic statistics are presented in Table [illegible] to show that the effect did disappear during the 1980s and the early 1990s, but it was intact in most of the other sample periods. The statistics from the full sample are shown in Panel A. They are consistent with early findings on the size effect: big firms report lower returns than small firms, and the CAPM beta is also negatively related to size. The size premiums in the last row of each panel are calculated as follows:

$$\begin{aligned} SP_{i,t} &= R_{i,t} - \left(R_{f,t} + \beta_i (R_{m,t} - R_{f,t})\right), \text{ and} \\ SP_i &= \frac{1}{T}\sum_{t=1}^{T} SP_{i,t} \qquad i = 1, \ldots 10. \end{aligned} \tag{1}$$

[6] Instead of the usual one-year holding period immediately following the size sorting date, I also extend the holding period to longer time spans to see how persistent the size premium is for the same group of firms.

where SP_i represents the average size premium of portfolio i which is shown in the table, $R_{i,t}$ and $R_{m,t}$ are monthly returns on size portfolio i and the market portfolio, respectively. R_f is the risk-free rate. β_i is the CAPM beta estimated by regressing $(R_i - R_f)$ on $(R_m - R_f)$ with the matching sample period. This size premium captures the part of the size portfolio return which cannot be explained by the CAPM. Practitioners usually add it to the cost-of-equity estimation of small-cap firms to compensate for their higher risks. Another way to estimate the size premium is through the estimation of the CAPM alpha. However, I will not adopt this approach because the sample period used by the regression to estimate CAPM coefficients and the one used by the realized return in equation (11) do not always match in this article.

[Insert Table 1 here.]

Panel B displays the statistics of the same variables with the sample period before June 1980, roughly when the size effect was made well known by academia. Although the statistics in the first two panels are not exactly the same, they look very much alike.

Panel C of Table 1 is consistent with the assertion of Horowitz, Loughran, and Savin (2000a) that there is no significant difference between the performance of different size portfolios during the period from 1980 to 1996.[7] The average returns on different size portfolios are no longer negatively related to their market capitalizations. From portfolio 1 to 4, the four smallest size portfolios, the average returns are increasing instead of moving in the opposite direction shown in the early years. The pattern of size premiums is also different from the ones shown in the previous two

[7]This period can be extended to 1998 and the results are still in the similar pattern to what one would get with sample period from 1980 to 1996, so this longer sub-sample period is chosen instead of the one used by Horowitz, Loughran, and Savin (2000a).

panels. For instance, portfolio 1 and 2 did not have the largest size premiums, they had biggest size "discounts" instead.

It is often suggested that pricing anomalies may disappear after they were made known to the public by researchers or financial practitioners if these anomalies were easily exploited. Horowitz, Loughran, and Savin (2000a) show that simply adding \$0.125 to the December 31 price of small stocks can easily lower their average January returns from over 8% to -0.37% during the 1982-1997 span. Since Keim (1983) and Reinganum (1983) showed that most of the size premiums to small firms occurred during the first two weeks in January, it is no surprise that the January effect could be totally wiped out just by informed investors flocking into the market to buy small firm stocks in December, and so goes the size premium.

Sixteen years of time is not short, but the recent development shows that the result in Panel C is more likely to be an aberration from the formerly established rule than a new norm. Panel D presents the statistics from the past 10 years and shows that the negative relation between firm size and equity return has been restored, with only a few exceptions from some mid-cap size portfolios. The inconsistency of the mid-cap portfolios probably arises because the sample period is too short to offer a robust pattern between a firm's size and its return. It has to be noted that the realized equity premium of the U.S. market during these 10 years is slightly below zero, which is significantly lower than the historical standard. This might contribute to the flat security market line, where the beta of size portfolios seems independent of their respective average return.

Another serious threat generated by the data from the 1980s and 1990s is that the return differential between small and big firm size portfolios, also known as SMB in the Fama-French 3-factor model, may have an insignificant or even a negative price

of risk. This implies that the SMB factor is either meaningless or has a negative effect on the stock return. We can use a simple cross-sectional regression to show how and why this matters.

[Insert Table 2 here.]

Table 2 displays price-of-risk estimations of the popular Fama-French factors with different sample periods. Following the Fama and MacBeth (1973) procedures, in each sub-sample period I run time-series regressions of each test portfolio return in excess of the risk-free rate ($R^e_{it} = R_{it} - R_{ft}$) on the excess market return ($R^e_{mt} = R_{mt} - R_{ft}$), the returns on the small size portfolios minus the returns on the big size portfolio (SMB), and the differential between the returns on high and low book-to-market equity firms (HML).[8]

$$R^e_{it} = \alpha_i + \beta_i R^e_{mt} + s_i SMB_t + h_i HML_t + \varepsilon_{it} \qquad t = 1, 2, \ldots, T, \forall i. \tag{2}$$

The test portfolios include 5-by-5 portfolios formed on book-to-market equity and size, and 17 industry portfolios.[9] Since there are missing observations in the return series of the portfolio with the highest book-to-market equity and the largest size, it is taken out of the test portfolios. These portfolios are chosen because they cover different aspects of security characteristics.

The next step is to regress the expected returns of test portfolios from each sample period on their respective risk loading estimates from the time-series regression. I

[8]Please refer to Fama and French (1993) for the detailed definition of SMB and HML. Data on these two variables are obtained from Professor Kenneth French's website at Dartmouth University.

[9]All the portfolio data are also acquired from French's website.

take the average return of each portfolio from the corresponding sample period as their return expectation. The cross-sectional regression is:

$$E_T(R_i^e) = \beta_i \lambda_1 + s_i \lambda_2 + h_i \lambda_3 + a_i, \qquad i = 1, 2, \ldots, N. \tag{3}$$

where λ_2 is the price of the risk represented by the size factor SMB. During the period from 1980 to 1998, the price of SMB is insignificantly different from zero and its magnitude is also comparably smaller than it is in the other sub-periods. The number is 0.29 before 1980 and 0.20 after 1998, but it is only 0.07 from July 1980 to June 1998. The other parameters do not change as dramatically over different sub-periods. The price of a risk factor being equal to zero discredits its explanatory power to the cross-sectional variability of returns, and this is exactly the case for the SMB factor from 1980 to 1998.

It may be too early to say that the explanatory power of the SMB factor fully recovers in the post-1996 or the post-1998 period, but it is clear that the zero or slightly negative SMB price during the 1980s and 1990s is not necessary a lasting problem.

2.3 Regime Shifts of the small stock premium

As mentioned earlier, the size premium and the small stock premium are related because the risk-adjusted abnormal return of small firms is an important part of the return differential between small and big stock portfolios. According to Table II Panel A, the small stock premium of portfolio 1 is 3.39%, which accounts for half of the return difference between portfolio 1 and 10. Since the size premium is highly dependent on the asset pricing model and the sample period it is using, I will focus

on the possible structural change or regime shift of the small stock premium in this section first.

Although the differential between the returns on size portfolio 1 and portfolio 10 is different from the definition of the SMB factor in the Fama and French 3-factor model, I will borrow this acronym to represent the small stock premium for the following discussion. Motivated by the earlier discussion of the disappearance of the small stock premium in the 1980s and 1990s and the reappearance in the following years, I believe that there may exist structural changes or regime shifts of the expected mean of SMB. Panel A of Figure □ exhibits the annual return differential between portfolio 1 and portfolio 10, in which we see annual SMB alternates between high and low values but certain persistency exists. From 1984 to 1998, the supposedly positive SMB is negative in most years except in 1988 and 1991 to 1993. The sample average of the equity risk premium during these 15 years is 10.53%, which is well above the historical average. Big firms performed exceptionally well while small firms did not during this period, so the disappearance of SMB should certainly came from the size premium, or lack thereof.

[Insert Figure □ here.]

Assuming that the expected mean and variance of SMB can be expressed by a two state Markov-switching model, so the state variable S_t, which governs the regime shift, takes a value of 1 or 2. When $S_t = 1$, the expected mean of SMB_t is in the state of a low value, while $S_t = 2$ represents the state when the expected mean of SMB_t is high.

$$y_t = \mu_k + \sigma_k \varepsilon_t \qquad \varepsilon_t \sim N(0,1). \tag{4}$$

where y_t represents SMB_t, μ_k and σ_k are state-dependent mean and standard deviation of SMB_t. k=1 or 2, which identifies the state SMB_t is in at time t.

The state variable S_t is assumed to follow a 2-state first-order Markov process with fixed transition probabilities as follows:

$$\begin{aligned} p &= \Pr(S_t = 1 | S_{t-1} = 1) \\ 1-p &= \Pr(S_t = 2 | S_{t-1} = 1) \\ q &= \Pr(S_t = 2 | S_{t-1} = 2) \\ 1-q &= \Pr(S_t = 1 | S_{t-1} = 2) \end{aligned} \tag{5}$$

The mean and variance of SMB are determined by the current state, and the state variable S_t is not dependent on the past information beyond one period.

SMB_t under each state is assumed to follow the normal distribution and the parameters of the distribution function are only contingent on the state k, so

$$f(y_t | S_t = k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-(y_t - \mu_k)^2}{2\sigma_k^2}\right) \tag{6}$$

for $k = 1, 2$. The log-likelihood function is

$$\ln \mathscr{L}(y_1, y_2, \ldots, y_T; \theta) = \sum_{t=1}^{T} \ln[\Pr(S_t = 1) f(y_t | S_t = 1) + \Pr(S_t = 2) f(y_t | S_t = 2)] \tag{7}$$

and the regime probability $\Pr(S_t = k)$ can be estimated with the following recursive representation proposed by Gray (1996):

$$\Pr(S_t = 1) = (1-q)\left[\frac{f(y_{t-1} | S_{t-1} = 2)\Pr(S_{t-1} = 2)}{f(y_{t-1} | S_{t-1} = 1)\Pr(S_{t-1} = 1) + f(y_{t-1} | S_{t-1} = 2)\Pr(S_{t-1} = 2)}\right]$$

$$+p\left[\frac{f(y_{t-1}|S_{t-1}=1)\Pr(S_{t-1}=1)}{f(y_{t-1}|S_{t-1}=1)\Pr(S_{t-1}=1)+f(y_{t-1}|S_{t-1}=2)\Pr(S_{t-1}=2)}\right] \quad (8)$$

where the lowercase p and q are the transition probabilities defined in equation (5) and $\Pr(S_t=2)=1-\Pr(S_t=1)$.

Table 3 presents the estimation results of the above Markov-switching model along with an unconditional normal distribution model as its comparison. The sample period is from July 1940 to December 2008 instead of starting from July 1926 because it has to be trimmed short in the following sections to accommodate the portfolio positions with longer holding periods. According to the log-likelihood values, AIC, and BIC statistics of these two models, the Markov-switching model fits the sample better than the model with the assumption that SMB follows an unconditional normal distribution. The expected mean of the low SMB state is insignificantly different from zero, which explains why SMB can disappear over an extended period. The average annualized returns under two different states are -2.67% and 44.97%.

[Insert Table 3 here.]

Panel B of Figure 1 displays the smoothed probability in state 2 (high SMB state). Table 3 also shows the transition probabilities p and q, which are 0.9579 and 0.8090, respectively. These results imply that the low SMB regime is more persistent than the high SMB regime. On average the high SMB regime lasts for 5.2 months, and the low SMB regime keeps at the same state for 23.8 months. If the true data generating process of SMB follows the description of this Markov-switching model, it is no surprise that the small stock premium could disappear over a long period during the 1980s and most of the 1990s then resurfaces in recent years.

From Figure 1 we can also see that SMB is persistently low from 1946 to 1963, which indicates that the experience from the 1980s and 90s indeed has a predecessor. Repeat the same exercise done in Table 1 for this period, we can find that portfolio 1 has an average size premium at -1.77% per annum, while portfolio 10 has a slightly positive 0.42% average size premium. The average of SMB from 1946 to 1963 is -0.74%, which mostly stems from the low size premium of small stocks instead of the difference between their respective CAPM projections.[10] These results show that the temporary disappearance of the size effect is a recurring event. However, when we look at a longer time span, the small stock premium could still hold true at least on average.

[10]CAPM beta is still negatively related to firm size during this period, but the slope of the security market line calculated with returns on size portfolios and their respective betas is smaller than it is calculated with the full sample.

3 Size as a Genetic Code or a Short-Lived Characteristic?

If the size premium ceases to exist like Horowitz, Loughran, and Savin (2000b) assert, or its magnitude has no relation to firm size, there is no need to give a "premium" to a small firm when estimating its cost of equity capital. In fact, given what we see in Panel C of Table □ we might have to give small-cap firms a discount if the negative size premium of portfolio 1 remains. The data from the last 10 years seem to restore the order of the size premium and the necessity to add it to small firms, but I will show in this section that it still remains to be proved whether a small-cap firm should require this size premium in its cost-of-equity estimation.

3.1 Design of the $t+j$ Portfolio

Fama and French (2007) find that the return differential between small and big firms is mainly driven by small-cap firms moving up the size rank to become large-cap firms. This perspective changes the assumption of the size premium a small firm should get in the long run. The logic is simple: a small firm becomes a big firm because its market capitalization increases faster than its peer, which usually results from its fast growing price. However, small firms cannot keep the higher average return of old once they become big firms, otherwise the small stock premium will turn into a big stock premium. Although this is mainly an explanation of the small stock premium instead of the size premium, the discussion in the previous section shows that these two premiums are related.

Since the Fama-French size portfolios are constructed in each June and are held for a whole year until they are rebalanced in June next year, their finding implies that some firms are likely to switch to different size groups sooner than a year, especially for the small firms to become big firms. The usual practice of the size premium estimation is to calculate it with annually rebalanced size portfolios,[11] then we add this number to a firm's cost of equity for the following years to discount its future cash flows to the present value. We know this is probably a proper assessment of the discount factor for the first year, but is it still proper if an originally small firm becomes a big firm from the second year on and does not warrant such a premium hereafter?

To investigate whether the size premium is changing over time and how it evolves, I design the following $t+j$ size portfolio approach. In the traditional size portfolio formation, securities are assigned to each portfolio in June and the portfolios are held from July to June next year under a buy-and-hold strategy. In the $t+j$ size portfolio approach I also choose to sort securities in June of each year t, but instead of holding the portfolios for the following year, I also look at the monthly returns for an one-year holding period from July of year $t+j-1$ to June of year $t+j$, where $j = 2,\ldots,15$.[12] All the firms are identified and tracked by their CRSP permanent number. If a firm goes bankrupt or is merged by another firm in the following years, then it is taken out of the portfolio once it is off the CRSP tape. Otherwise it keeps in the same $t+j$ size portfolio as assigned in the initial sorting date no matter how big or how small its market capitalization becomes.

[11] For getting the size premium estimation, some practitioners rebalance the size portfolios more frequently. For example, Ibbotson Associates sorts and assigns all eligible companies to different size portfolios with the closing price and shares outstanding data for the last trading day of March, June, September and December instead of June each year.

[12] This approach reduces to the traditional size portfolio formation when $j = 1$.

For example, the firms in t+2 portfolios from July 1989 to June 1990 were sorted and assigned to different size portfolios in June 1988; the same composition of firms is used in t+1 portfolios from July 1988 to June 1989, which are 12 months immediately after the sorting date. The t+3 portfolios in July 1990 also consist of the same firms, except for those were delisted during the first two years. There is also another set of t+2 portfolios from July 1988 to June 1989, each consists firms sorted by their June 1987 size. We can string together all the t+2 portfolios to see how firms perform a year after its original sorting date for a whole year. The same process is done for all t+j size portfolios. This approach allows us to follow the average performance of firms j years after they were assigned to a specific size group.

If a firm's size behaves as a characteristic and this attribute follows the firm for an extended period of time, return patterns among different t+j size portfolios should not change much for different j. On the other hand, if a small firm deserves a lower size premium after it becomes a bigger firm, the size premium in the following years will decrease accordingly. By tracking the historical performance of firms sorted by size, we can get a better idea on how the size premium of a firm behaves and whether it is a good indicator of an extra risk source.

3.2 Size Premium is Changing Over Time

Practitioners usually consider a fixed size premium for a firm for subsequent years, which implies that either firms will not migrate to other size groups, or they will still demand the same size premium even after they switch to different size groups. To make a valid comparison between different t+j portfolios, I change the starting date of all portfolios from July 1926 to July 1940 to accommodate the t+15 portfolios,

which have companies being sorted in June 1926 but will not report the first return observation until July 1940.[13]

Table 4 presents the average size premiums of different $t+j$ size portfolios in reference to the respective CAPM projected returns on the traditional size portfolios. The "traditional" size portfolio means that firms are sorted and assigned to different size portfolios according to their June market capitalization, and the portfolios are held from July of the same year to June next year. The definition of the average size premium of a $t+j$ size portfolio is

$$\begin{aligned} SP_{i,t}^{t+j} &= R_{i,t}^{t+j} - \left(R_{f,t} + \beta_i (R_{m,t} - R_{f,t})\right), \text{ and} \\ SP_i^{t+j} &= \frac{1}{T}\sum_{t=1}^{T} SP_{i,t}^{t+j}, \end{aligned} \tag{9}$$

where $R_{i,t}^{t+j}$ represents the time t return on the $t+j$ portfolio of firms in the ith size group, and β_i is the same as in equation (1).

[Insert Table 4 here.]

The first decile size portfolio, which contains firms with the lowest market capitalizations among all listed firms on the sorting date, usually has a large and significant CAPM alpha and a beta too low to project the realized return. Table 1 shows that portfolio 1 has a size premium of 3.39% per annum with the sample period from July 1926 to December 2008. The corresponding number in Table 4 is the average size premium of the $t+1$ portfolio for portfolio 1. Although the benchmark is still calculated with the same beta, it drops to 1.49% because the sample period here does not start until July 1940. The difference reflects a large historical size premium for the

[13]The security return data on CRSP tape start from December 1925, so June 1926 becomes the first available sorting date.

small firms from 1926 to 1940. The premiums change a lot with different sample periods, but the pattern is nevertheless revealing. The smallest firms still get a bigger size premium, while the biggest firms even get a size discount.

If firms are supposed to be awarded a fixed size premium for years, we should see the numbers in Table 4 remain stable over different $t+j$ portfolios within each size group. The result is apparently contrary to this hypothesis. The size premium of portfolio 1 drops dramatically two years after the initial sorting date and becomes insignificantly different from zero in the third year. After that the small firms get a discount and such a discount gradually becomes significantly different from zero. On the other hand, portfolio 10 sees its size premium going up from the negative value in the first two years to a positive but insignificant number for the most part of the following eight years. Most of the size portfolios have a declining size premium after the sorting date except for portfolio 10, which reflects the fact that returns on different size portfolios tend to converge to the same number over years. Table 5 shows that the difference in average returns on different size portfolios gradually becomes insignificant as sorting dates pass by.

[Insert Table 5 here.]

If history can be any guide to the future performance, we are likely to over-estimate the cost of equity capital of small firms and under-estimate the cost of equity of big firms by the current treatment of the size premium.

3.3 Robustness Check

We have seen in Table 1 that the historical averages of both the size premium and the small stock premium are sensitive to the choice of the sample period, but the

pattern remains unchanged if given a long enough horizon. Here I will verify that the findings in this section are not sensitive to different breakpoints of size groups.

Fama and French (2007) divide firms into two groups in terms of size to explain the cause of the Fama-French *SMB* factor, so I also divide all the acting firms into two groups according to the NYSE median market-cap breakpoint in each June.

For better examining the relation between firm size and the corresponding return performance, I also rank firms according to their size each June and form three portfolios with firms of their size in the bottom 30%, middle 40%, and top 30% (S-30%, M-40% and B-30% hereafter) by the NYSE market-cap breakpoints.

The size premiums calculated with new breakpoints are displayed in Table 5. The big size portfolios (Big or B-30%) all have very small and insignificant size premiums like the size premium of portfolio 10 reported in Table 4. Please be noted that I still use the traditional size portfolio approach (it is equivalent to the t+1 portfolio here) with the new breakpoints and the sample period from 1926 to 2008 to estimate CAPM betas. The size premiums of "Small" and "S-30%" size portfolios are significant through t+1 to t+4 or t+5 portfolios, respectively, and they are also declining as j goes up. Ten or seven years after the initial sorting dates, these two small size portfolios even have a discount. These characteristics are all consistent with the pattern shown in portfolio 1 in Table 4.

[Insert Table 5 here.]

Comparing Table 5 to Table 4, it is apparent that the size premium for small stocks in the traditional sense does exist no matter how many size groups the stocks

are divided into, but it fades out gradually if the same composition of firms is held longer than a year.[14]

If a group of firms have the same stream of expected future cash flows, it is possible that the firm with a higher risk is going to be priced lower. Such a firm may end up having a higher return because it is more likely to have a higher dividend yield. However, small firms do not only gather higher returns through higher dividend yields, they usually have higher capital appreciation rates too. Fama and French (2007) explain that migration of stocks across size groups is the cause of the small stock premium.[15] Once a small firm's market capitalization increases and it is qualified as a big firm, a size premium should not apply anymore. According to Table 4 and 6, small firms did have higher size premiums when they were first assigned to the small size portfolio, but this effect does not persist. A firm which belongs to portfolio 1 sees its size premium turns into a discount after a few years if it is still expected to be compensated as a small stock. It is probably reasonable for a small firm to get a larger discount factor than the CAPM suggests because it bears higher risks than the model can explain for the time being, but the usual practice could very likely over-compensate the risks a small firm is bearing.

If the size effect has to be considered in the cost-of-equity estimation, we should search for the root of this short-lived premium and identify the risk source it represents. This is just as important as how much it is, if not more important.

[14]The small stock premium fades away until it is barely noticeable. However, the size premium for small stocks sometimes becomes a size discount if the same composition of stocks is held for a few years.

[15]In their article Fama and French use "size premium" to refer to the fact that small-cap firms have higher returns than big-cap firms without risk adjustment, which is equivalent to the "small stock premium" used in this paper. As shown earlier that these two premiums are related.

4 Size Premium under Different Economic Situations

Section 3 shows that a small firm can have a higher size premium only in the short run. Over a longer time span, a firm's size and even its sensitivity to risk are all subject to change, and its size premium changes accordingly.[16] In light of these results, I propose not to include a fixed size premium in the long-term cost-of-equity estimation. However, the size premium, no matter how short-lived it is, still appears to exist in the first few years for small firms. Take the popular discounted cash flow method as an example, the first few years matter the most if given a steady stream of future cash flows. By excluding the size premium from the cost-of-equity estimation, one might argue that we are also likely to understate the risk a small firm is taking.

The simplest way to resolve this conundrum seems to apply a time-varying cost of equity by adding different size premiums to the estimation according to the results in Table 4. The short-term size effect is thus accounted for, and the long-term size premium is also no longer permanent. However, Table 4 only displays the standard deviation of the average of the size premium, the variation of the annual size premium per se is much larger. If the size premium swings between high and low levels like the two-regime small stock premium model shown in section 2.3, adding an average size premium into the short-term cost-of-equity estimation may not help the matter. We could easily over-estimate the cost of equity of small firms in one period and suppress their value, while under-estimate the cost of equity in another period

[16] CAPM betas of all size groups are monotonically decreasing from t+1 through t+15 portfolios. These results are not shown in the tables, but they are available upon request. In this paper I use the traditional size portfolios with the full sample (July 1926 to December 2008) to estimate CAPM betas to get a consistent benchmark in all cases but ones in Table 1.

and bring the price to an un-deserving high level. In this section I will examine the likelihood of this scenario.

The concept of connecting financial distress to firm size has been discussed in the asset pricing literature to explain the anomalous cross-sectional pattern of stock returns. Queen and Roll (1987) find that a firm's unfavorable mortality rate is a decreasing function of its size, and Campbell, Hilscher, and Szilagyi (2008) further show that size has a negative relation with the excess return between safe and distress stocks. I will examine from a different angle to see whether economic distress has an effect on the size premiums.

I divide the sample period into several two-regime scenarios according to different macroeconomic variables related to distress and calculate the size effect under each regime. There are two reasons for this experiment: the first is that only the systematic risk should be taken into account when pricing a firm or an asset. If small firms are supposed to be awarded a higher premium sheerly because of their failure risk, then we should be able to distinguish different patterns of their size premium under different economic situations. Second, in light of the success of a simple Markov-switching model used on the small stock premium in section 2, it is natural to try a two-regime model on the size premium as well. However, the estimation of the size premium is highly contingent on the choice of the asset pricing model and the sample period, so I do not investigate the possible regime shifts of the size premium directly. Instead, I will try to explore the relation between the size premium and three different candidates of macroeconomic variables. If the size premium is at least partly driven by systematic risk sources, its magnitude should vary as the economic environment changes.

4.1 Identifying the States of Economy

The first state variable is an indicator variable which identifies the economic status during a business cycle: a dummy variable which equals 1 for months in the expansion period and 0 for months in the contraction period.[17] When in distress, smaller firms usually get hit harder because they have thinner cushion in common equity and their ability to raise capital via new debts, bank loans, or even government bailouts is also poorer than big firms. On the other hand, small firms which survive the storm can often see a sudden boom in their stock returns, as were evidenced by their bigger beta.[18] Whether the bigger volatility in the stock return for the small stock portfolio can translate to separate size premiums is the focus of the investigation. According to NBER's Business Cycle Dating Committee, there are 14 business cycles since 1926 to date with the shortest contraction period being 6 months and the shortest expansion period being 24 months.

The second indicator is the market trend, which is similar to the idea of the business cycle. I distinguish the bull and bear markets by a Markov-switching model on the CRSP value-weighted market portfolio return with the similar procedure laid

[17] NBER's Business Cycle Dating Committee publishes the U.S. business cycle peak and trough months on the NBER website. Their latest announcement on 12/01/2008 declares that the previous expansion period peaked in December 2007 and a recession soon followed. The conclusion of the current recession has not yet been determined as the writing of this paper. I assume all of year 2008 fell into the contraction period to make the sample period consistent with other state variables.

[18] Fama and French (1993) point out that small firms do not participate in the economic boom of the middle and late 1980s for an unknown reason. This finding is consistent with the argument of the disappearance of the size effect in the 1980s and 1990s. Indeed, the small stock premium was -10.4% per annum from December 1982 to July 1990, the expansion period right after the longest recession since the Great Depression. However, small firms greatly outperform big firms during the economic booms after the Great Depression or the recession caused by 1973 oil crisis, with average small stock premiums at 55.9% and 23.1%, respectively. It is probably premature to judge the experience in the 1980s as a new norm or just an anomaly. Nonetheless, the magnitude of SMB during the expansion periods in the middle 1930s and the late 1980s could counter the argument raised by Fama and French (1993).

out in section 2.3.[19] Regime 1 represents the state of the bear market with a lower mean return and higher volatility; regime 2 indicates the bull market with a higher mean return and lower volatility. An indicator variable is used to represent the bull market with its value being equal to 1 when the regime 2 smoothed inference of the month is greater than 0.5, and 0 otherwise. The reason to use a dummy to identify the market trend instead of the realized market return is to filter out noise. When we apply the size premium on the cost of equity capital estimation, we look for the long-term performance instead of the short-term disturbance. Looking too much into the day-to-day or month-to-month performance will mix up true trend and noise. For instance, even during the huge market downturn in the Great Depression, when the Dow Jones Industrial Average (DJIA) dropped from then historical high of 381.17 on 9/3/1929 to the following lowest point of 41.22 on 7/8/1932, we can still see the market posted double digit gains on return during the process. In February and June 1931, the monthly returns derived from the DJIA were 12.40% and 16.90%, respectively. These were great rallies even in any bull market, but they still cannot stop the free fall of the stock market and the investment environment would not be changed simply because of a sudden spark of life. Since the cost of equity capital and the size premium are all about the long term prospect of the firm, it is more fitting to examine the general market trend in this simple fashion.

The third indicator is the credit spread between AAA and BAA corporate bond rates. The data are obtained from the Federal Reserve Bank of St. Louis website. Although we cannot link a firm's size directly to its credit rating, large firms usually get better ratings and lower borrowing rates.[20] When there is abundant credit

[19]There is no consensus on the definition of bear or bull markets other than a general description. Here I adopt the market trend definition of the model 1 in Chen (2009).

[20]According to the summary statistics provided by Altman and Rijken (2004), firm's credit rating is negatively related to the market value of equity. I also compare the average market values between

floating in the market, the credit spread tends to narrow down because banks and funds compete against each other for an investment opportunity without thinking too much about the risk. This process will eventually drive the spread down. On the other hand, the credit spread increases when the credit market is in a dire condition and investors take default risks more seriously. Every banker will think twice before lending money out. When the credit spread is high, it is more likely that small firms endure a higher borrowing cost than big firms, therefore their failure risk induced by the poorer credit rating is also higher. I continue to apply the same technique previously used in the market trend indicator to separate the credit spreads into two different states, and then convert the smoothed inference into a dummy variable using the 0.50 threshold.

The transition probabilities of staying in the same state for the Markov-switching model of the market trend are 0.892 (bear market) and 0.963 (bull market); they are 0.987 (low credit spread) and 0.974 (high credit spread) for the credit spread. The common feature of these macroeconomic variables is that the states defined by them are all very persistent, so we can link these variables with the shift of the size premium over a longer span instead of the month-by-month movement. Once the state variable of the market trend shifts to the bull market state, it would stay put for 27 months on average, and a credit spread dummy remains in the state of a lower mean value for 78 months.

[Insert Figure 2 here.]

firms with investment grade ratings and with non-investment grade ratings over the past 15 years. The average size of firms with better credit is 9 to 10 times bigger than the size of poorer rating firms. The sample includes all firms in the Compustat database from 1994 to 2008.

Figure 2 illustrates three different dummy variables on the right-hand side and their original data on the left.[21] It has to be noted that these state variables are all asymmetrical. We see expansion periods more often than contraction periods, longer bull markets than bear markets, and more days with low credit spreads than days with high ones. Over the total 822 observations, there are 698 months identified as in the expansion period, 646 months in the bull market, and 552 months in the low credit spread regime.

4.2 The Size Premium under Different Economic Environments

These state variables do not highly coincide with each other, but they are all capable of separating the size premium of small stocks under different states. I also use the $t+j$ portfolio approach to see whether these states can identify the size effect of stocks over the long run. Table 7 and 8 present the size premiums of the first and the 10th size portfolios under different economic situations.

[Insert Table 7 here.]

[Insert Table 8 here.]

The first column of Table 7 or 8 shows the same average size premiums as the corresponding column in Table 4. Through the second column to the last, the average size premiums under different states of the same macroeconomic variable are paired with each other. The second and third columns are the average size premiums in the expansion or contraction state identified by the business cycle dummy; the fourth and fifth columns show the averages during bull or bear markets from the market

[21] I use the GDP growth rate for the business cycle dummy as its "original data". However, it is well known that the Business Cycle Dating Committee of the NBER does not determine the peaks and troughs by the GDP data alone.

trend dummy; and the last two columns are average size premiums in the high or low state of the credit spread dummy.

The last row of each table shows the number of observations in a specific state. These three dummy variables post asymmetric states as earlier mentioned, but the credit spread dummy is significantly different from the others because the state brings the higher average returns has a lot less observations than the state brings the higher return for the other two dummy variables.[22]

Small stocks usually have a high and significant size premium, and this premium is even more pronounced in the expansion period or the high credit spread period, and interestingly, during the bear market. Portfolio 1 has a positive premium for most of the $t+j$ portfolios during the market downturn because the market trend dummy successfully identifies the low return period of the market, which in turn drives the benchmark even lower than the drop of the realized return on small stocks. The time series dynamics of the size premium revealed by the $t+j$ portfolio approach present a different scenario for the business cycle dummy. It is indecisive whether a small firm has a greater size premium during the expansion or contraction period.

Table 8 displays the size premium, or more precisely, the size discount of portfolio 10. Large firms usually can be explained well by the CAPM or other asset pricing models, so the common practice does not require a size premium on them. Even under different states, the size premiums are still small in magnitude comparing to the corresponding statistics of portfolio 1. If we focus on the first few $t+j$ portfolios, the business cycle does not seem to play an important role. The average size premi-

[22]The state generates the higher average return does not necessarily have the higher size premium. The latter also depends on the sensitivity to the market risk and the market return under this "unfavorable" state.

ums under different regimes of the market trends or credit spreads are much more different, but they are still not as pronounced as their counterparts in portfolio 1.

A one-sided t test on unequal sized variables is also applied here to compare the difference between average size premiums under different economic states. The size premiums in Table 7 and 8 are shown in **boldface fonts** if the difference is significant at the 10 percent level. We cannot reject the null hypothesis that none of the size premium pairs of portfolio 1 or 10 are significantly different during different periods of business cycles. The same test for different market trends shows the similar result for the first nine years for portfolio 1 and the first two years for portfolio 10. The state variable derived from the credit spread data is the most successful of all. The difference of the average size premiums of $t+j$ portfolios is significant at 10 percent level for most of the cases for portfolio 1, and it is also significant for the first 6 years for portfolio 10.

The size premium a small firm should demand for bearing higher risks is limited only in the first few years and its magnitude is difficult to predict. The empirical results imply that we should be very careful to identify the risks a firm is bearing instead of taking it only by the firm's current size. If there are other systematic risks which is related to size, we should reconsider whether that is the cause of a firm being riskier than the others and assign the specific risk premium to it accordingly.

5 Conclusion

This study verifies the existence of the size effect of annually rebalanced size portfolios with a longer sample period, but suggests not to include the size premium in the cost-of-equity estimation of small firms because this effect is only short-lived.

The assertion of the disappearance of the size effect in the 1980s and 90s was just a result of sample selection. Similar events of temporary disappearance of the size effect from different periods were found but they have never been proved permanent. Suffice it to say that the size effect did not simply disappear because it was revealed by academics and exploited by practitioners. It is shown in section 2 that the small stock premium can be better captured by a two-state Markov-switching model rather than the usual stationary normal distribution assumption. This empirical evidence is consistent with the story of the temporary disappearance of the size effect in the 1980s and 1990s.

Using the $t+j$ portfolio approach designed for this study, I demonstrate that the small stock premium declines if we hold the size portfolio longer than the usual one-year holding period rule. This can be considered as evidence of Fama and French (2007)'s finding that the size premium stems from small firms moving up the size rank to become big firms. Since firms move between size groups, the size premium should not be considered as a constant and it has to reflect the new size group they are currently in. The popular perception of a fixed size premium used by practitioners in the cost-of-equity estimation is obviously mistaken. I track the size premiums of different size portfolios for the subsequent 15 years after their formation date and find that most of the premiums converge toward zero, so firms should not be awarded a size premium for a long-term estimation.

If the size premium of a firm is estimated with the assumption that a firm moves from one size group to another all the time, it should be time-varying as well. The average size premium of portfolio 1, which includes all NYSE, NASDAQ and AMEX firms with market capitalization less than the first decile market-cap breakpoint of all NYSE listed firms, is 1.49% for the first year after its creation for the past 68 years. The same composition of firms still merit an average of 1.02% premium in the following year, but it declines rapidly after that. Adding a fixed size premium according to a firm's current size could very well overstate the relation between a firm's size and the risk it is bearing.

Certain macroeconomic variables can help us to distinguish the possible regimes of the size premium. These variables include the business cycle, the market trend, and the credit spread. However, the decision to distinguish the size premium of a firm under the assumption of one specific state is very difficult to make given how highly volatile the monthly size premium is. Adding a naive size premium to a firm's cost of equity capital estimation still potentially introduces more errors no matter this size premium is fixed or time-varying.

References

Altman, Edward I., and Herbert A. Rijken, 2004, How Rating Agencies Achieve Rating Stability, *Journal of Banking and Finance* 28 11, 2679–2714.

Banz, Rolf W., 1981, The Relationship between Return and Market Value of Common Stocks, *Journal of Financial Economics* 9, 3–18.

Berk, Jonathan B., 1995, A Critique of Size-Related Anomalies, *Review of Financial Studies* 8, 275–86.

Black, Fischer, 1972, Capital Market Equilibrium with Restricted Borrowing, *Journal of Business* 45, 444–455.

Blume, Marshall E., and Robert F. Stambaugh, 1983, Biases in Computed Returns: An Application to the Size Effect, *Journal of Financial Economics* 12 3, 387–404.

Booth, David G., Donald B. Keim, and William T. Ziemba, 2000, Is There Still a January Effect?, in *Security Market Imperfections in Worldwide Equity Markets* (Cambridge University Press, Publications of the Newton Institute. Cambridge; New York and Melbourne).

Campbell, John Y., Jens Hilscher, and Jan Szilagyi, 2008, In Search of Distress Risk, *Journal of Finance* 63 6, 2899–2939.

Chen, Shiu-Sheng, 2009, Predicting the Bear Stock Market: Macroeconomic Variables as Leading Indicators, *Journal of Banking and Finance* 33 2, 211–23.

Cochrane, John H., 2005, *Asset pricing*. (Princeton University Press Revised Edition. Princeton and Oxford).

Fama, Eugene F., and Kenneth R. French, 1992, The Cross-Section of Expected Stock Returns, *Journal of Finance* 47, 427–65.

Fama, Eugene F., and Kenneth R. French, 1993, Common Risk Factors in the Returns on Stock and Bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 1995, Size and Book-to-Market Factors in Earnings and Returns, *Journal of Finance* 50, 131–55.

Fama, Eugene F., and Kenneth R. French, 2007, Migration, *Financial Analysts Journal* 63, 48–58.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, Return, and Equilibrium: Empirical Tests, *Journal of Political Economy* 81, 607–636.

Gray, Stephen F., 1996, Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process, *Journal of Financial Economics* 42, 27–62.

Horowitz, Joel L., Tim Loughran, and N. E. Savin, 2000a, The Disappearing Size Effect, *Research in Economics* 54 1, 83–100.

Horowitz, Joel L., Tim Loughran, and N. E. Savin, 2000b, Three Analyses of the Firm Size Premium, *Journal of Empirical Finance* 7 2, 143–53.

Keim, Donald B., 1983, Size-Related Anomalies and Stock Return Seasonality: Further Empirical Evidence, *Journal of Financial Economics* 12 1, 13–32.

Kothari, S. P., Jay Shanken, and Richard G. Sloan, 1995, Another Look at the Cross-Section of Expected Stock Returns, *Journal of Finance* 50, 185–224.

Lintner, John, 1965, The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *The Review of Economics and Statistics* 47, 13–37.

Pratt, Shannon P., and Roger J. Grabowski, 2008, *Cost of Capital: Applications and Examples*. (Wiley) 3 edn.

Queen, Maggie, and Richard Roll, 1987, Firm Mortality: Using Market Indicators to Predict Survival, *Financial Analysts Journal* 43, 9.

Reinganum, Marc R., 1981, Misspecification of Capital Asset Pricing: Empirical Anomalies Based on Earnings' Yields and Market Values, *Journal of Financial Economics* 9 1, 19–46.

Reinganum, Marc R., 1983, The Anomalous Stock Market Behavior of Small Firms in January: Empirical Tests for Tax-Loss Selling Effects, *Journal of Financial Economics* 12 1, 89–104.

Roll, Richard, 1983, On Computing Mean Returns and the Small Firm Premium, *Journal of Financial Economics* 12 3, 371–86.

Sharpe, William F., 1964, Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *The Journal of Finance* 19, 425–442.

Figure 1: The return difference between the first and the 10th decile size portfolios and the smoothed probability of the high small stock premium regime. Panel A shows the annual portfolio return difference between small and big stocks. It is apparent that big firms outperform small firms most of the time from the mid-1980s to late 1990s. This account for the "disappearance" of the size effect in that time span. Similar situation also happened in the 1950s and late 1960s to early 1970s. The smoothed inference of the high SMB regime is shown in Panel B.

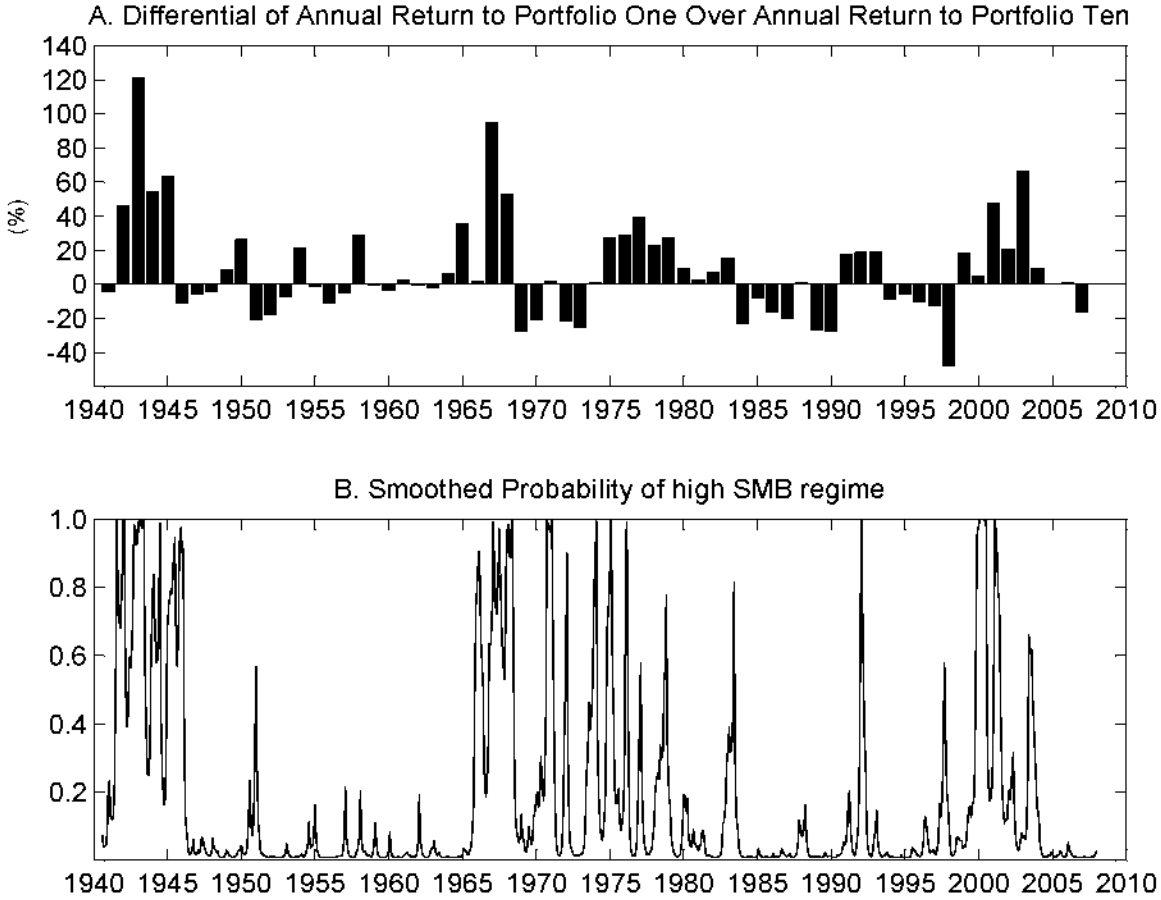


Figure 2: Three different dummy variables indicates three different economic environments. The first row includes the GDP growth rate of the U.S. and the business cycle dummy. The second row presents the CRSP monthly return and the market trend dummy variable derived from the smoothed probability of the bull market regime. The third row contains the credit spread and the high credit spread dummy also generated from the smoothed inference of a two-state Markov-switching model.

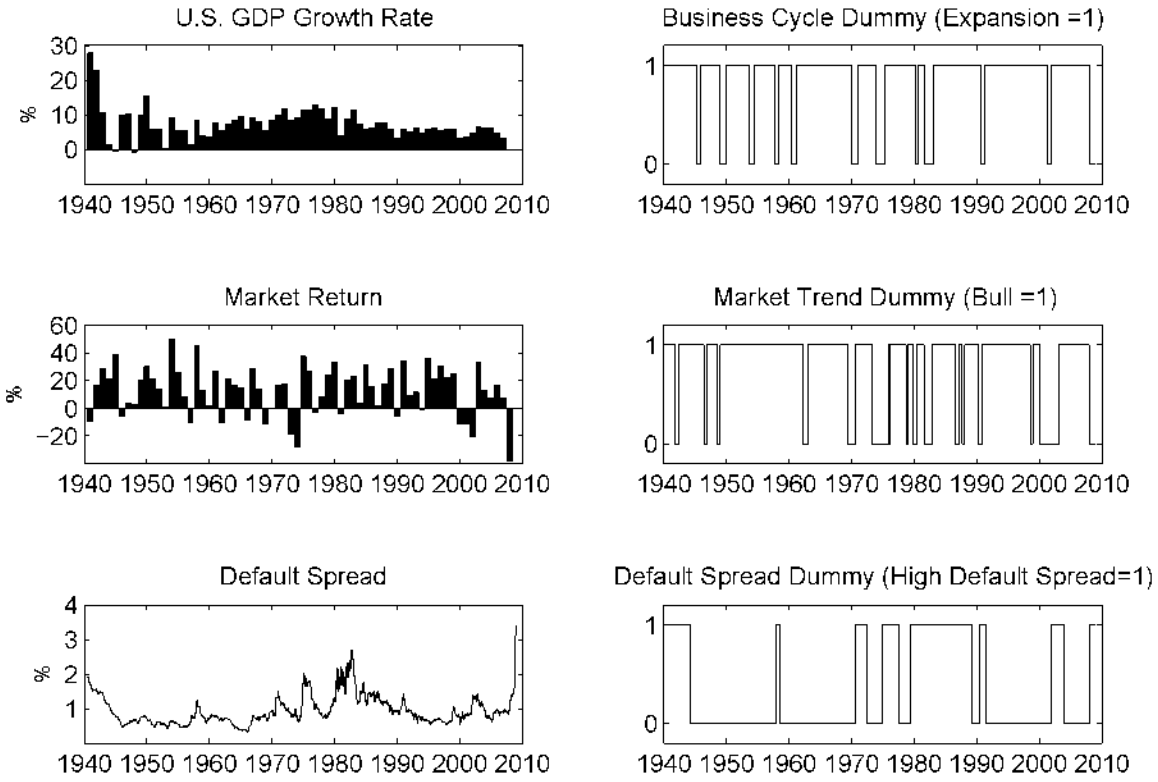


Table 1: **Returns on Size Portfolios and Size Premiums in Reference to CAPM**

Panel A. Full Sample (1926.7 to 2008.12)

	1 (Small)	2	3	4	5	6	7	8	9	10 (Big)
Mean Return	17.36	14.79	14.52	14.37	13.68	13.22	12.75	12.16	11.66	10.14
Standard Dev.	35.46	30.86	28.39	26.58	25.08	23.68	22.77	21.82	20.24	17.80
β	1.46	1.40	1.34	1.27	1.25	1.20	1.16	1.13	1.05	0.93
Size Premium	3.39	1.21	1.37	1.70	1.21	1.08	0.85	0.53	0.54	-0.10

Panel B. 1926.7 to 1980.6

	1 (Small)	2	3	4	5	6	7	8	9	10 (Big)
Mean Return	20.44	16.19	15.61	15.23	14.14	13.84	12.58	12.22	11.45	9.70
Standard Deviation	41.17	34.89	31.96	29.55	27.82	26.30	25.13	23.80	22.12	19.04
CAPM β	1.60	1.48	1.41	1.32	1.29	1.24	1.19	1.14	1.07	0.93
Size Premium	5.14	1.79	1.80	2.11	1.30	1.38	0.50	0.54	0.33	-0.29

Panel C. 1980.7 to 1998.6

	1 (Small)	2	3	4	5	6	7	8	9	10 (Big)
Mean Return	12.93	14.50	15.96	16.52	17.23	16.96	17.16	15.94	16.84	17.40
Standard Dev.	17.63	17.89	17.77	17.66	17.16	16.24	16.09	15.58	15.32	14.32
β	0.95	1.07	1.10	1.10	1.09	1.05	1.08	1.04	1.04	0.96
Size Premium	-2.99	-2.61	-1.40	-0.90	-0.08	0.01	-0.03	-0.93	0.01	1.31

Panel D. 1998.7 to 2008.12

	1 (Small)	2	3	4	5	6	7	8	9	10 (Big)
Mean Return	9.14	8.05	6.48	6.26	5.23	3.61	6.03	5.36	3.87	-0.03
Standard Dev.	25.11	26.08	23.24	22.94	21.33	19.83	19.57	20.24	17.13	16.10
β	1.06	1.21	1.15	1.13	1.13	1.08	1.08	1.14	0.98	0.92
Size Premium	7.47	6.59	4.95	4.68	3.66	1.97	4.38	3.80	2.07	-1.92

All securities in NYSE, AMEX and NASDAQ are sorted at the end of June of each year t and are assigned to ten different size portfolios according to NYSE breakpoints. The size portfolios are constructed with securities in each size group with their respective market cap as weights and are held from July of year t through June of year $t+1$.

β's are estimated with regression of monthly portfolio returns in excess of the Ibbotson Associates risk free rate on the CRSP value-weighted market returns in excess of the same risk free rate.

The size premium is calculated by subtracting the product of the CAPM beta and the equity premium from the size portfolio returns in excess of the risk free rate. All the equity risk premiums in different panels are estimated from their respective sample periods.

Returns, standard deviations and size premiums are all annualized and in percentage points.

Table 2: **Prices of Fama-French Risk Factors**

	1926.7-2007.12	1926.7-1980.6	1980.7-1998.6	1998.7-2007.12
$R_m - R_f$	0.64 (0.17)	0.70 (0.23)	0.84 (0.29)	-0.04 (0.44)
SMB	0.24 (0.11)	0.29 (0.14)	-0.04 (0.17)	0.47 (0.37)
HML	0.38 (0.12)	0.41 (0.15)	0.41 (0.18)	0.24 (0.35)

I calculate the price of risk of the Fama-French (1993) three factors with Fama and MacBeth (1973)'s two-pass regression approach. These data are retrieved from Professor French's website at Dartmouth. Test portfolios are obtained from 25 portfolios formed on size and book-to-market equity and 17 industry portfolios. Since there exist missing values in one of the 25 size/BM portfolio, it is taken out of the portfolio set. The returns on the remaining 41 test portfolios are named as R_{it}, $i = 1,2,\ldots,N, N = 41$.

First we find beta estimates from the time-series regressions,

$$R^e_{it} = \alpha_i + \beta_i R^e_{mt} + s_i SMB_t + h_i HML_t + \varepsilon_{it} \qquad t = 1,2,\ldots,T, \forall i.$$

where $R^e_{it} = R_{it} - R_{ft}$ and $R^e_{mt} = R_{mt} - R_{ft}$.

Then estimate the factor risk premiums λ from a cross-sectional regression,

$$E_T(R^e_i) = \beta_i \lambda_1 + s_i \lambda_2 + h_i \lambda_3 + a_i, \qquad i = 1,2,\ldots,N.$$

Since the pricing errors a_i are likely to be correlated, we follow Cochrane (2005)'s suggestion to run a GLS cross-sectional regression and the estimations of the price of risk are

$$\begin{aligned} \hat{\lambda} &= (\beta \Sigma^{-1} \beta)^{-1} \beta \Sigma^{-1} E_T(R^e), \text{and} \\ \sigma^2(\hat{\lambda}) &= \frac{1}{T}\left[(\beta \Sigma_f^{-1} \beta)^{-1} + \Sigma_f\right] \end{aligned}$$

where β is an N-by-3 matrix with $[\beta_i \; s_i \; h_i]$ in each row, $\lambda = [\lambda_1 \; \lambda_2 \; \lambda_3]$, f is a T-by-3 matrix of the risk factors, R^e_{mt}, SMB, HML.

The sample period is broken down like in Table 11. The parameter estimates in each subperiod use only observations from that subperiod. Standard deviations of λ estimates are reported in parentheses.

The insignificance of parameters in the subperiod from July 1996 to December 2007 probably results from sample selection and short sample period. The most interesting finding is on λ_2, the price of the risk factor SMB. During the sample period from July 1980 to June 1996, the price of this factor is not only insignificant but also much smaller in its value.

Table 3: **Regime Switching Model of the return difference between the 1st and 10th decile Size Portfolios**

	Regime Switching Model			Unconditional Normal Dist	
	Parameter	Standard Deviation		Parameter	Standard Deviation
μ_1	-0.002436	0.00189	μ	0.004590	0.001825
μ_2	0.036465	0.01184			
σ_1^2	0.001263	0.00013	σ^2	0.052284	0.000136
σ_2^2	0.008167	0.00179			
p	0.9579	0.01991			
q	0.8090	0.11592			
Log-Likelihood Value	1367.73901			1257.87773	
AIC	-2723.47802			-2511.75546	
BIC	-2695.20758			-2502.33198	

Table 4: **Size Premium of *t+j* Decile Size Portfolio**

	Small	2	3	4	5	6	7	8	9	Big
t+1	1.49	0.57	0.94	1.26	0.87	0.48	1.02	0.48	0.50	-0.19
	(0.56)	(0.42)	(0.34)	(0.31)	(0.26)	(0.22)	(0.18)	(0.16)	(0.12)	(0.11)
t+2	1.02	1.70	1.63	1.50	1.16	0.53	0.36	0.84	0.36	-0.14
	(0.52)	(0.40)	(0.33)	(0.29)	(0.25)	(0.21)	(0.18)	(0.15)	(0.13)	(0.11)
t+3	-0.67	1.33	1.51	0.77	1.46	0.47	0.34	0.52	0.17	0.03
	(0.48)	(0.39)	(0.32)	(0.29)	(0.25)	(0.22)	(0.18)	(0.15)	(0.13)	(0.12)
t+4	-1.60	1.96	0.79	1.69	0.82	-0.04	0.59	0.37	0.40	0.10
	(0.45)	(0.37)	(0.32)	(0.29)	(0.25)	(0.22)	(0.18)	(0.16)	(0.12)	(0.12)
t+5	-0.83	1.42	1.26	0.58	-0.44	0.73	0.88	0.53	0.27	0.10
	(0.44)	(0.37)	(0.31)	(0.27)	(0.24)	(0.20)	(0.19)	(0.15)	(0.12)	(0.12)
t+6	-0.18	0.43	0.91	0.38	0.29	0.90	0.49	0.77	0.18	0.14
	(0.44)	(0.36)	(0.30)	(0.27)	(0.23)	(0.21)	(0.19)	(0.14)	(0.13)	(0.12)
t+7	-1.57	0.51	0.43	0.27	0.66	0.89	-0.78	0.12	0.50	0.29
	(0.43)	(0.35)	(0.30)	(0.26)	(0.24)	(0.21)	(0.17)	(0.15)	(0.14)	(0.12)
t+8	-1.31	-0.54	0.86	0.99	0.19	0.12	0.34	0.27	0.64	0.11
	(0.42)	(0.33)	(0.30)	(0.25)	(0.23)	(0.20)	(0.18)	(0.14)	(0.13)	(0.13)
t+9	-1.38	-0.46	0.43	-0.02	0.98	0.01	1.27	-0.42	0.47	0.16
	(0.39)	(0.32)	(0.30)	(0.26)	(0.24)	(0.21)	(0.20)	(0.17)	(0.14)	(0.13)
t+10	-1.61	-0.72	-0.65	1.22	-0.08	0.33	-1.02	-0.26	0.76	0.20
	(0.38)	(0.31)	(0.30)	(0.25)	(0.23)	(0.21)	(0.20)	(0.19)	(0.13)	(0.14)
t+11	-1.30	-0.62	-0.76	0.05	0.12	0.18	-0.36	0.56	-0.12	0.31
	(0.39)	(0.31)	(0.28)	(0.26)	(0.24)	(0.20)	(0.21)	(0.17)	(0.13)	(0.14)
t+12	-1.62	-1.60	-0.83	1.11	0.12	0.37	0.14	-0.21	-0.17	0.33
	(0.39)	(0.30)	(0.30)	(0.26)	(0.23)	(0.21)	(0.20)	(0.16)	(0.14)	(0.14)
t+13	-1.40	-2.30	-0.20	0.72	0.36	-0.04	-0.62	-0.51	-0.26	0.35
	(0.38)	(0.31)	(0.30)	(0.26)	(0.25)	(0.21)	(0.19)	(0.18)	(0.15)	(0.14)
t+14	-2.64	-1.08	-1.22	0.90	-0.45	-1.08	-0.91	-0.84	-0.26	0.42
	(0.38)	(0.31)	(0.31)	(0.27)	(0.25)	(0.22)	(0.21)	(0.19)	(0.15)	(0.15)
t+15	-3.14	-0.86	-1.50	-0.01	-1.02	-1.29	-0.83	-0.81	-1.21	0.68
	(0.39)	(0.31)	(0.30)	(0.26)	(0.24)	(0.24)	(0.23)	(0.20)	(0.16)	(0.15)

Standard deviations of mean returns (or return differential in the last column) are in the parentheses.

CAPM betas used in this table are estimated with full sample period (July 1926 to December 2008) instead of the trimmed sample period (July 1940 to December 2008) for the $t+j$ portfolios. The size premium of the $t+1$ portfolios here and the size premium of the Panel A of Table 1 should be the same if given the same length of sample.

Table 5: **Average Returns on *t+j* Decile Size Portfolio and Decile 1- Decile 10 Return Difference**

	Small	2	3	4	5	6	7	8	9	Big	1-10
t+1	16.17	14.85	14.78	14.61	14.02	13.29	13.58	12.76	12.27	10.68	5.49
	(0.81)	(0.74)	(0.69)	(0.67)	(0.63)	(0.60)	(0.59)	(0.57)	(0.53)	(0.49)	(0.63)
t+2	15.71	15.98	15.47	14.84	14.30	13.33	12.92	13.13	12.13	10.73	4.97
	(0.80)	(0.74)	(0.69)	(0.67)	(0.63)	(0.60)	(0.60)	(0.57)	(0.54)	(0.48)	(0.60)
t+3	14.01	15.61	15.35	14.12	14.61	13.27	12.89	12.81	11.94	10.90	3.12
	(0.79)	(0.75)	(0.69)	(0.66)	(0.63)	(0.62)	(0.59)	(0.57)	(0.53)	(0.48)	(0.58)
t+4	13.08	16.23	14.64	15.03	13.97	12.77	13.14	12.66	12.17	10.97	2.12
	(0.78)	(0.73)	(0.69)	(0.66)	(0.65)	(0.61)	(0.59)	(0.55)	(0.53)	(0.48)	(0.56)
t+5	13.85	15.69	15.10	13.93	12.71	13.53	13.43	12.81	12.04	10.97	2.88
	(0.78)	(0.73)	(0.70)	(0.66)	(0.64)	(0.60)	(0.58)	(0.56)	(0.53)	(0.47)	(0.55)
t+6	14.50	14.71	14.76	13.72	13.44	13.71	13.04	13.06	11.95	11.01	3.49
	(0.78)	(0.74)	(0.69)	(0.65)	(0.62)	(0.60)	(0.59)	(0.56)	(0.53)	(0.47)	(0.55)
t+7	13.12	14.79	14.27	13.61	13.80	13.70	11.77	12.41	12.27	11.15	1.96
	(0.79)	(0.73)	(0.68)	(0.63)	(0.63)	(0.60)	(0.59)	(0.56)	(0.53)	(0.47)	(0.56)
t+8	13.38	13.73	14.70	14.34	13.34	12.92	12.89	12.55	12.41	10.98	2.40
	(0.78)	(0.72)	(0.68)	(0.64)	(0.63)	(0.61)	(0.58)	(0.55)	(0.52)	(0.47)	(0.55)
t+9	13.30	13.82	14.27	13.33	14.13	12.82	13.82	11.86	12.24	11.03	2.27
	(0.76)	(0.70)	(0.69)	(0.64)	(0.63)	(0.60)	(0.59)	(0.55)	(0.53)	(0.46)	(0.51)
t+10	13.08	13.56	13.20	14.57	13.07	13.13	11.54	12.03	12.53	11.07	2.00
	(0.75)	(0.69)	(0.69)	(0.64)	(0.63)	(0.59)	(0.59)	(0.55)	(0.53)	(0.46)	(0.50)
t+11	13.38	13.65	13.09	13.40	13.27	12.99	12.19	12.85	11.65	11.18	2.20
	(0.74)	(0.70)	(0.68)	(0.63)	(0.63)	(0.58)	(0.58)	(0.54)	(0.53)	(0.46)	(0.49)
t+12	13.06	12.68	13.02	14.46	13.27	13.18	12.69	12.08	11.60	11.20	1.87
	(0.74)	(0.68)	(0.69)	(0.63)	(0.63)	(0.59)	(0.56)	(0.55)	(0.53)	(0.46)	(0.50)
t+13	13.28	11.97	13.65	14.07	13.51	12.77	11.93	11.78	11.51	11.21	2.07
	(0.74)	(0.68)	(0.69)	(0.62)	(0.61)	(0.59)	(0.58)	(0.54)	(0.53)	(0.46)	(0.49)
t+14	12.04	13.19	12.62	14.25	12.70	11.72	11.65	11.45	11.51	11.28	0.76
	(0.73)	(0.67)	(0.67)	(0.62)	(0.62)	(0.59)	(0.59)	(0.55)	(0.52)	(0.46)	(0.48)
t+15	11.54	13.42	12.34	13.34	12.12	11.52	11.72	11.48	10.56	11.55	-0.01
	(0.74)	(0.66)	(0.66)	(0.63)	(0.60)	(0.59)	(0.58)	(0.53)	(0.52)	(0.46)	(0.50)

Standard deviations of mean returns (or return differential in the last column) are in the parentheses.

Table 6: **Robustness Check: Size Premium of Different Size Portfolios in Reference to CAPM Projected Return**

	Small	Big	S-30%	M-40%	B-30%
t+1	0.96	0.02	0.91	0.91	-0.05
	(0.32)	(0.05)	(0.40)	(0.21)	(0.06)
t+2	1.51	0.05	1.60	0.77	0.02
	(0.31)	(0.05)	(0.38)	(0.20)	(0.07)
t+3	1.09	0.11	0.94	0.70	0.08
	(0.30)	(0.06)	(0.36)	(0.19)	(0.08)
t+4	0.99	0.14	0.72	0.65	0.13
	(0.28)	(0.07)	(0.35)	(0.18)	(0.08)
t+5	0.44	0.20	0.95	0.46	0.15
	(0.26)	(0.07)	(0.34)	(0.17)	(0.08)
t+6	0.30	0.23	0.49	0.52	0.21
	(0.25)	(0.07)	(0.32)	(0.17)	(0.09)
t+7	0.03	0.24	-0.10	0.07	0.28
	(0.24)	(0.07)	(0.30)	(0.17)	(0.09)
t+8	0.17	0.20	-0.25	0.37	0.19
	(0.23)	(0.08)	(0.30)	(0.16)	(0.09)
t+9	0.10	0.21	-0.31	0.52	0.15
	(0.23)	(0.09)	(0.29)	(0.16)	(0.10)
t+10	-0.22	0.17	-1.05	-0.14	0.26
	(0.22)	(0.09)	(0.27)	(0.16)	(0.10)
t+11	-0.35	0.22	-1.04	-0.30	0.24
	(0.21)	(0.09)	(0.26)	(0.16)	(0.10)
t+12	-0.28	0.21	-1.30	0.23	0.18
	(0.21)	(0.10)	(0.27)	(0.16)	(0.11)
t+13	-0.28	0.13	-1.16	-0.02	0.16
	(0.21)	(0.10)	(0.26)	(0.16)	(0.11)
t+14	-0.50	0.07	-1.52	-0.55	0.21
	(0.21)	(0.11)	(0.26)	(0.16)	(0.12)
t+15	-0.97	0.10	-1.68	-0.87	0.22
	(0.20)	(0.12)	(0.26)	(0.17)	(0.12)

Standard deviations of mean returns (or return differential in the last column) are in the parentheses.

Table 7: **Average Size Premium of Portfolio 1 under Different Economic Environments**

	Total	Expansion	Contraction	Bull Mkt	Bear Mkt	High CS	Low CS
t+1	1.49	2.07	-1.78	0.65	4.57	**5.45**	**-0.45**
	(0.56)	(0.61)	(1.42)	(0.57)	(1.57)	**(1.15)**	**(0.62)**
t+2	1.02	1.36	-0.86	0.15	4.24	**4.57**	**-0.71**
	(0.52)	(0.56)	(1.35)	(0.53)	(1.47)	**(1.01)**	**(0.60)**
t+3	-0.67	-0.70	-0.47	-1.08	0.84	2.17	-2.06
	(0.48)	(0.52)	(1.30)	(0.50)	(1.32)	(0.90)	(0.57)
t+4	-1.60	-1.51	-2.09	-2.13	0.35	**2.62**	**-3.67**
	(0.45)	(0.48)	(1.30)	(0.47)	(1.23)	**(0.83)**	**(0.54)**
t+5	-0.83	-0.82	-0.87	-1.33	1.02	**3.34**	**-2.87**
	(0.44)	(0.48)	(1.19)	(0.45)	(1.24)	**(0.79)**	**(0.53)**
t+6	-0.18	-0.23	0.06	-0.72	1.80	**3.18**	**-1.83**
	(0.44)	(0.47)	(1.17)	(0.45)	(1.21)	**(0.75)**	**(0.54)**
t+7	-1.57	-1.67	-0.97	-1.26	-2.70	**2.56**	**-3.59**
	(0.43)	(0.46)	(1.16)	(0.43)	(1.24)	**(0.72)**	**(0.53)**
t+8	-1.31	-1.27	-1.51	-1.30	-1.32	**1.60**	**-2.73**
	(0.42)	(0.44)	(1.28)	(0.43)	(1.14)	**(0.72)**	**(0.51)**
t+9	-1.38	-1.25	-2.12	-1.93	0.64	**3.54**	**-3.79**
	(0.39)	(0.42)	(1.13)	(0.42)	(1.01)	**(0.68)**	**(0.48)**
t+10	-1.61	-1.47	-2.36	**-2.99**	**3.48**	**2.38**	**-3.56**
	(0.38)	(0.40)	(1.13)	**(0.40)**	**(1.03)**	**(0.65)**	**(0.47)**
t+11	-1.30	-1.21	-1.83	**-2.64**	**3.61**	**1.22**	**-2.54**
	(0.39)	(0.41)	(1.17)	**(0.40)**	**(1.03)**	**(0.65)**	**(0.48)**
t+12	-1.62	-1.80	-0.61	**-2.60**	**1.97**	**1.23**	**-3.01**
	(0.39)	(0.41)	(1.13)	**(0.41)**	**(1.06)**	**(0.69)**	**(0.47)**
t+13	-1.40	-1.22	-2.42	-2.20	1.55	0.35	-2.25
	(0.38)	(0.40)	(1.16)	(0.40)	(1.03)	(0.68)	(0.47)
t+14	-2.64	-2.33	-4.37	-3.39	0.11	**0.33**	**-4.09**
	(0.38)	(0.40)	(1.12)	(0.39)	(1.04)	**(0.67)**	**(0.46)**
t+15	-3.14	-3.20	-2.82	**-4.41**	**1.53**	**1.30**	**-5.32**
	(0.39)	(0.42)	(1.12)	**(0.39)**	**(1.12)**	**(0.74)**	**(0.45)**
Number of Observations	822	698	124	646	176	270	552

The standard deviation of the average size premium is in the parenthesis.

The first column shows the average size premium of the first decile size portfolio, which is the same as the first column of Table 4.

The number of observations in each state is in the last row of the table. The second and third columns are the expansion and contraction states; the fourth and fifth columns are the bull and bear market states; and the last two columns are the high and low credit spread states.

The size premiums are shown in **boldface fonts** if the difference is significant at the 10 percent level using a one-sided t test.

Table 8: **Average Size Premium of Portfolio 10 under Different Economic Environments**

	Total	Expansion	Contraction	Bull Mkt	Bear Mkt	High CS	Low CS
t+1	-0.19	-0.17	-0.27	-0.29	0.21	**-1.10**	**0.26**
	(0.11)	(0.12)	(0.29)	(0.11)	(0.32)	**(0.20)**	**(0.13)**
t+2	-0.14	-0.14	-0.12	-0.39	0.80	**-1.10**	**0.34**
	(0.11)	(0.12)	(0.29)	(0.11)	(0.34)	**(0.20)**	**(0.13)**
t+3	0.03	0.03	0.05	**-0.34**	**1.38**	**-0.87**	**0.47**
	(0.12)	(0.12)	(0.30)	**(0.11)**	**(0.35)**	**(0.20)**	**(0.14)**
t+4	0.10	0.04	0.43	**-0.33**	**1.66**	**-0.63**	**0.45**
	(0.12)	(0.13)	(0.31)	**(0.11)**	**(0.35)**	**(0.21)**	**(0.14)**
t+5	0.10	-0.03	0.85	**-0.42**	**2.02**	**-0.73**	**0.51**
	(0.12)	(0.13)	(0.32)	**(0.11)**	**(0.36)**	**(0.21)**	**(0.14)**
t+6	0.14	0.00	0.95	**-0.43**	**2.22**	**-0.59**	**0.50**
	(0.12)	(0.13)	(0.33)	**(0.11)**	**(0.38)**	**(0.21)**	**(0.15)**
t+7	0.29	0.11	1.29	**-0.37**	**2.68**	-0.29	0.57
	(0.12)	(0.13)	(0.34)	**(0.12)**	**(0.39)**	(0.22)	(0.15)
t+8	0.11	-0.08	1.17	**-0.49**	**2.30**	-0.55	0.43
	(0.13)	(0.14)	(0.33)	**(0.12)**	**(0.42)**	(0.22)	(0.16)
t+9	0.16	0.01	1.03	**-0.52**	**2.67**	-0.60	0.54
	(0.13)	(0.14)	(0.32)	**(0.12)**	**(0.44)**	(0.21)	(0.17)
t+10	0.20	0.03	1.16	**-0.45**	**2.60**	-0.51	0.55
	(0.14)	(0.15)	(0.34)	**(0.12)**	**(0.46)**	(0.22)	(0.17)
t+11	0.31	0.12	1.37	**-0.45**	**3.10**	-0.38	0.65
	(0.14)	(0.16)	(0.36)	**(0.12)**	**(0.49)**	(0.22)	(0.18)
t+12	0.33	0.20	1.08	**-0.43**	**3.11**	-0.37	0.67
	(0.14)	(0.16)	(0.37)	**(0.13)**	**(0.49)**	(0.23)	(0.18)
t+13	0.35	0.18	1.27	**-0.42**	**3.15**	-0.25	0.64
	(0.14)	(0.16)	(0.39)	**(0.13)**	**(0.48)**	(0.24)	(0.18)
t+14	0.42	0.21	1.55	**-0.28**	**2.96**	-0.14	0.68
	(0.15)	(0.16)	(0.38)	**(0.13)**	**(0.51)**	(0.24)	(0.19)
t+15	0.68	0.49	1.76	**-0.13**	**3.67**	-0.03	1.03
	(0.15)	(0.16)	(0.39)	**(0.13)**	**(0.53)**	(0.24)	(0.19)
Number of Observations	822	698	124	646	176	270	552

The standard deviation of the average size premium is in the parenthesis.

The first column shows the average size premium of the 10th decile size portfolio, which is the same as the last column of Table 4.

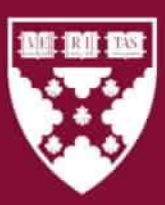
Column 2 to column 7 use the same dummy variables to separate different states as the corresponding columns in Table 7.

The size premiums are shown in **boldface fonts** if the difference is significant at the 10 percent level using a one-sided t test.

Working Paper 21-095

Deregulation, Market Power, and Prices: Evidence from the Electricity Sector

Alexander MacKay
Ignacia Mercadal



Harvard Business School

Deregulation, Market Power, and Prices: Evidence from the Electricity Sector

Alexander MacKay
Harvard Business School

Ignacia Mercadal
University of Florida

Working Paper 21-095

Funding for this research was provided in part by Harvard Business School.

Deregulation, Market Power, and Prices: Evidence from the Electricity Sector*

Alexander MacKay
Harvard University†

Ignacia Mercadal
University of Florida‡

December 12, 2022

Abstract

When deciding whether to introduce market competition in a regulated industry, a regulator faces an important tradeoff. Market-based prices can provide incentives to allocate resources more efficiently and reduce costs, but the presence of market power may lead to increased markups. We construct a novel dataset on electricity generation, wholesale transactions, and retail sales to investigate the impact of deregulation in the context of the U.S. electricity sector. We find that the higher markups charged by generation companies more than offset the efficiency gains, leading to higher wholesale prices. Downstream, incumbent utility retail prices rose one-for-one with the increase in variable costs of procurement, while the introduction of alternative retail suppliers generated modest retail markups for some customers. These results highlight the role of market power in deregulated markets, and show that consumers may prefer regulated prices to market-based prices when markets are not perfectly competitive.

Keywords: Deregulation, Market Power, Markups, Prices, Electricity
JEL Classification: L51, L94, D43, L13, L43, Q41

*An earlier version of this paper circulated under the title, "Shades of Integration: The Restructuring of the U.S. Electricity Markets." We thank Steve Cicala, Leemore Dafny, Tatyana Deryugina, Shane Greenstein, Akshaya Jha, Paul Joskow, C.-Y. Cynthia Lin Lawell, Bentley MacLeod, Nancy Rose, Marcelo Sant'Anna, David Sappington, and Richard Schmalensee for helpful comments. We thank seminar and conference participants the University of Florida, the IIOC, Rice, ITAM, the NBER Economics of Electricity Markets and Regulation Workshop, UChicago, the Northeast Workshop on Energy Policy and Environmental Economics, EARIE, the European Summer Meeting of the Econometric Society, MIT, Washington University in St. Louis, the University of Mannheim, and the ASSA Annual Meeting (TPUG). We are grateful for the research assistance of Tridevi Chakma, Laura Katsnelson, Gabriel Gonzalez Sutil, and Catrina Zhang.

†Harvard University, Harvard Business School. Email: amackay@hbs.edu.

‡University of Florida, Department of Economics. Email: imercadal@ufl.edu.

1 Introduction

Organizations that make large investments in broadly used infrastructure are often subject to special regulation. These organizations are typically characterized as natural monopolies, and regulation has been used to ensure the fair provision of services in the absence of competition. Industries such as electricity, airlines, telecommunications, and railroads have been subject to strict controls by governmental agencies, including the determination of prices. Over the past 50 years, technological progress and other factors changed how policymakers viewed many of these industries, leading to waves of deregulation. A common element of deregulation efforts has been the introduction of free entry and market-determined prices, with the goal of lowering prices to consumers.

However, the impact of deregulation on prices is theoretically ambiguous. Market-based prices provide incentives for profit-maximizing firms to reduce costs, but they also can provide firms with the ability to increase markups. When cost efficiencies are outweighed by the presence of market power, market-based prices can be higher than regulated rates. Thus, deregulation can lead to higher profits and lower consumer welfare.

We study the tradeoff between efficiencies in production and increased markups in the context of the restructuring of the U.S. electricity sector that started in the late 1990s. A significant objective of restructuring was to promote market-based—as opposed to regulated—prices in wholesale and retail markets. Toward this end, policymakers oversaw the divestment of generation facilities by regulated utilities and the introduction of alternative retail suppliers. Over 20 years later, we have yet to fully understand the consequences of these deregulation efforts (Bushnell et al., 2017). Previous studies have found that generation costs declined in deregulated markets (Fabrizio et al., 2007; Davis and Wolfram, 2012; Cicala, 2015, 2022), but the evidence on the impacts on prices is less conclusive (Borenstein and Bushnell, 2015; Bushnell et al., 2017). Contrary to the objectives of deregulation, we show that prices increased in deregulated markets, despite modest reductions in marginal and average variable costs. Markups increased substantially, indicating the widespread exercise of market power in U.S. electricity markets and highlighting the potential costs of deregulation.

To understand how markups changed in the electricity sector, we construct a novel dataset that covers the annual electricity flows from generation to final consumption for each electric utility territory from 1994 through 2016. Our dataset has the unique advantage of including purchases through bilateral contracts, in addition to purchases in the centralized wholesale markets run by independent system operators (ISOs).[1] From 2000 through 2016, the vast majority—over 85 percent—of wholesale electricity was sold with such contracts. Thus, a key contribution of our paper is to provide a more comprehensive view of prices in upstream and downstream markets, as that allows us to better understand the mechanisms behind higher

[1]The focus of the previous literature has been on the centralized wholesale markets. See, e.g., Borenstein et al. (2002); Puller (2007); Mercadal (2022).

prices and markups.

Using these data, we compare utilities that were subject to state-specific deregulation policies to similar utilities in other states that remained tightly regulated with a difference-in-differences matching approach (Deryugina et al., 2019). This approach has two important elements that allow us to measure the price effects of deregulation. First, policy variation at the state level allows us to observe both deregulated and regulated markets over the same time period. Second, our dataset allows us to match individual utilities based on generation technology, controlling not only for initial differences but also exposure to differential cost shocks in the future.[2] We then study how prices, costs, and markups have evolved across comparable utilities.

We find substantial price increases for consumers in deregulated states relative to consumers in regulated states. However, consistent with earlier findings, marginal costs declined in deregulated states, indicating that higher prices are driven by higher markups. Overall, we estimate that gross markups—retail prices minus the marginal cost of generation—increased by 15 dollars per MWh from 2000 to 2016. Relative to 1999 price levels, this change in markups corresponds to a 19 percent increase in prices over the period. Using our comprehensive data on wholesale markets, we find that wholesale prices increased despite declining generation costs. Thus, markups by generators increased by roughly 9 dollars per MWh, representing over 60 percent of the overall increase in gross markups. Thus, we find market power in the generation market to be the primary driver of price increases.

For a clearer picture of the mechanism behind prices increases, we focus in on the procurement costs for incumbent utilities. During the early years of deregulation, utilities faced higher procurement costs despite little change to generation costs and market prices. Because of the divestiture of generation assets, utilities were forced to obtain more electricity from purchases rather than own generation, and wholesale prices were higher than generation costs. The rates that incumbent utilities charged to their customers—which remained regulated to reimburse average variable costs—went up due to the introduction of this markup, which is analogous to double marginalization.

Several years later, around 2005, wholesale prices began to increase even though generation costs started to fall. Why? When states passed deregulation measures, they also adopted provisions to make the transition less sudden for consumers. Key provisions were price caps and long-term procurement contracts. When these expired (around 2005), utilities no longer had the bargaining power to insist on low wholesale prices. Generators could now sell to ISO markets or retail power marketers, and prices were no longer tied to price caps to downstream consumers. As a result, generators charged utilities more for their contracts, and wholesale prices increased. If there had been no market power, we would instead expect wholesale prices to fall along with the decline in generation costs.

[2]Fuel mix, for example, greatly determines how generators will be affected by shocks to fuel prices.

It is important to note that we measure market power using markups, the difference between price and marginal cost. In order to distinguish market power from competitive rents, which could arise in a competitive market in the presence of cost heterogeneity, we use a proxy for the marginal cost at the market level. In a competitive market, individual plants may have prices above marginal costs if a higher-cost plant determines the market price. However, the most expensive plants should not earn meaningful markups if the market is competitive. Consistent with market power, we find substantial increases in markups over the highest-cost plants.

Market power can exist even with competitive market mechanisms, such as auctions, when there are a limited number of potential suppliers. The previous literature has documented market power in electricity markets (e.g., Borenstein et al., 2002; Puller, 2007; Bushnell et al., 2008; Ito and Reguant, 2016; Mercadal, 2022), but its overall impact on consumers has not been studied. Several characteristics of electricity make these markets particularly prone to market power (Borenstein, 2002). Both demand and supply are inelastic, yet supply must meet demand at every moment since large amounts of electricity cannot be stored efficiently. Transportation is expensive, constraining the degree to which generators compete across local markets (Ryan, 2021; Mercadal, 2022). Entry is limited due to large sunk investments, long planning horizons, and high risk. As a result of these factors, only a few generators are typically competing to serve demand for a certain area at a particular moment, and the relative scarcity can give them substantial market power. Deregulation did not fundamentally change these factors.

We present several indirect tests of market power that point to market power at the wholesale level as the main driver of price increases. First, concentration among generators remained constant at high levels between 1995 and 2015. Higher markups did not attract significant entry, which is consistent with the presence of significant entry barriers. Second, states with lower potential competition saw bigger markup increases. Finally, we show that markups increased more in markets with more inelastic demand, as measured by the proportion of residential consumers.[3] Taken together, these findings support market power as the main driver behind our results.

We also show that the market restructuring intended by deregulation was delayed for several years. Despite the divestiture of generation assets, utilities maintained a high degree of vertical integration through contracts and umbrella ownership, where different companies are subsidiaries of the same parent/holding company. Thus, we distinguish between *apparent deregulation*—the share of a market supplied by companies other than the incumbent utility—and *effective deregulation*—the share of a market supplied by companies unaffiliated with the incumbent.[4] In wholesale markets, we find that the use of contracts delayed the onset of effective

[3]Residential customers tend to be less sensitive to prices than industrial and commercial customers. For instance, they are more likely to stay with the incumbent, even at higher prices, after retail competition is introduced.

[4]We use the term "affiliate" as a company belonging to the same parent company.

deregulation by many years, compared to apparent deregulation. In retail markets, caps on rates and other factors slowed the introduction of competitive retailers. Consistent with these delays, we observe a larger impact on prices once restructuring measures are fully in effect. Thus, distinguishing between apparent deregulation and effective deregulation can be important to accurately measure policy impacts.

We believe we are the first to document the extent to which electric deregulation in the U.S. yielded higher prices and to present evidence of an underlying mechanism: market power at the wholesale level that dominated cost efficiencies. Though there was early awareness of the potential for market power in deregulated markets,[5] the fact that the effects of market power could considerably exceed the savings from increased cost efficiency is surprising. Moreover, our analysis shows that contracts play a key role in market dynamics since they explain why we observe the effects of deregulation with a delay.

The existing literature on the consequences of deregulation is surprisingly scarce, given the importance of the electric sector for the economy and decarbonization efforts. The literature has documented gains in productive efficiency in several dimensions. Fabrizio et al. (2007) show that restructured plants reduced costs through better plant operation, spending less on labor and nonfuel costs for a given level of output. Davis and Wolfram (2012) also find better operational performance for deregulated nuclear plants, which increased output by 10 percent. Cicala (2015) shows that procurement costs decline in gas and coal plants after deregulation. Finally, Cicala (2022) shows that costs have also declined because of more efficient dispatch after ISOs were established to coordinate the usage of transmission and increase inter-utility trade. Our results on costs are consistent with this literature, since we also find moderate declines in fuel costs for power plants in restructured states.

However, the existing literature on restructuring has not yet determined whether these cost reductions have translated into lower prices for consumers. In a review of the literature, Bushnell et al. (2017) conclude that the effect is unclear. Findings differ across studies due to the differences in time periods, the use of different methods, switching focus between wholesale and retail prices, and the inclusion of other price determinants like stranded costs, among others (see, e.g., Joskow, 2005; Kwoka, 2008b; Su, 2015). Our dataset has the advantage of covering the whole industry, measuring flows from generation to retail, spanning a period of over 20 years, and capturing both costs and prices. This allows us to present a clear picture of the changes underwent by the industry, and using detailed firm-level data allows us to account for some of the confounding factors that are common concerns in the literature.[6]

[5]For example, Borenstein and Bushnell (2000) write "Market power among generators is likely to be a more serious and ongoing concern than has been anticipated by most observers," due to the combination of "inelastic short-run demand and supply (at peak times) with the real-time nature of the market."

[6]Although the deregulation process varied across countries, studies of the consequences of deregulation in other markets have found results that are consistent with ours. Newbery and Pollitt (1997) finds that costs went down after the restructuring of the electricity market in the UK in the 1990s, but prices barely decreased, leading to a substantial increase in profits. Bertram and Twaddle (2005) analyze the evolution of price-cost margins in New Zealand

Borenstein and Bushnell (2015) examine the consequences of restructuring between 1998 and 2012 and argue that the prices differences are primarily explained by differential responses to higher natural gas prices, which significantly affect marginal costs but not as much average costs. We consider this possibility, yet we find an increasing gap between prices (which increase) and marginal costs (which decrease) in markets after deregulation. In particular, natural gas prices fell in the latter half of our sample. Thus, changes in fuel costs do not seem to explain the rising prices observed in deregulated states. We conclude instead that increasing markups suggest the presence of market power.

The role of vertical integration in electricity markets has been discussed by Bushnell et al. (2008) and Mansur (2007), who show that spot wholesale electricity markets are more competitive when generators are vertically integrated because they have fewer incentives to increase prices. Our paper complements these finding by examining the market as a whole instead of focusing on the spot market, which, as of 2016, made up less than 25 percent of the entire wholesale market. We further add to the literature by examining the role of intermediate degrees of vertical integration. Previous studies in the transaction costs literature have identified the potential substitutability of long-term contracts and vertical integration (e.g., Coase, 1960; Joskow, 1987; MacKay, 2022). Here, we demonstrate how such alternative arrangements may be employed to side-step the intended effects of regulatory policies.

The paper proceeds as follows: Section 2 provides a background of deregulation efforts. Section 3 describes our dataset and key summary statistics. Section 4 details our empirical strategy and provides our main results for prices, costs, and markups, as well as a discussion on the mechanism. Section 5 presents supporting evidence for the role of market power in deregulated markets. In Section 6, we discuss the timing of the observed effects, explore the role of contracts in delaying deregulation effects, and provide a detailed case study on Illinois to illustrate how effective deregulation may be delayed. In Section 7, we explore several possible alternative mechanisms, and we conclude that our findings are most consistent with the exercise of market power. Section 8 concludes.

2 Background

2.1 Overview of Deregulation Efforts

In the 1970s and 1980s, a wave of deregulation encouraged entry and allowed market-based prices in many industries that had been considered natural monopolies, such as telecom, airlines, and surface freight.[7] Although the details of the deregulation process varied across indus-

after deregulation and show that cost decreased but prices increased in the decade following market restructuring. Our approach exploits detailed utility-level data in both deregulated and regulated markets during the same period, allowing to better control for other factors affecting costs and prices during the period under study.

[7]Market-based prices are those determined by demand and supply, as opposed to cost-based prices determined by a regulator as a function of cost.

tries, the principles motivating this process were the same: reduced entry barriers and market competition will increase efficiency and reduce prices. There was a consensus that some industries had undergone significant changes in their cost structures, allowing for beneficial effects of competition. In telecom, major changes in demand and technology had moved the sector away from a natural monopoly, making it an obvious candidate for deregulation.

Many of these deregulation efforts have been considered successful because prices have fallen, though in some cases at the cost of reduced quality (Borenstein and Rose, 2014; Viscusi et al., 2018; Joskow, 2005). However, even in successful cases, these industries remain highly concentrated, often appear in controversial merger cases, and engage in behavior that raises concerns about market power (Borenstein, 1989; Borenstein and Rose, 1994, 2014; Viscusi et al., 2018). For example, after the deregulation of airlines and the subsequent fall in prices, concentration increased (Kahn, 1988) and continued to increase afterwards. Telecom also remains highly concentrated, even after significant growth in demand and technological improvement (Viscusi et al., 2018). High levels of concentration suggest that market power may be an important concern in deregulated industries, where characteristics like high fixed costs or network economies that once led them to be regulated may make them prone to market power. For example, Rubinovitz (1993) finds that over 40 percent of the price increase after deregulation in cable markets in the United States was due to the exercise of market power.

The next section describes how competitive markets were introduced in the electricity sector and provides a brief background of the overall deregulation process.

2.2 Deregulation in U.S. Electricity Markets

Traditionally, electric utilities in the U.S. and the world were vertically integrated companies that included generation, transmission from power plants to towns and cities, distribution along power lines to final consumers, and retail sales to these consumers. Because electricity was considered a natural monopoly, a single utility served each local market, and electricity prices were regulated to avoid monopoly pricing. Utilities were reimbursed based on their average costs of generation. Following a wave of what was considered successful deregulation in other sectors, the electricity sector started its own process of deregulation in the 1990s.

The decision to implement competitive markets occurred at the state level and was determined by local politics (Borenstein and Bushnell, 2015).[8] Though specific implementation details varied across states, state-level deregulation typically involved two main components. The first was vertical separation: most states required utilities to divest some or all of their gen-

[8]On average, states that passed deregulation measures had higher pre-deregulation rates than those that remained regulated, but the decision to deregulate was not necessarily driven by price differences. For example, IOUs in deregulated states like Oregon and Texas had lower-than-average rates, while some states with higher rates like Vermont and Florida remained regulated. Within states, there is meaningful variation in rates offered by different utilities, resulting in a weaker relationship between deregulation and pre-deregulation prices at the utility level.

eration assets to encourage the creation of a competitive generation sector.[9] As we show in the paper, states varied in how strict the separation between utilities and generation was required to be, and in many cases utilities split themselves into generation and distribution subsidiaries under the same parent company. After deregulation, utilities and alternative retailers in deregulated states procured all electricity from wholesale markets, either through long-term contracts or in a centralized auction organized by transmission operators.[10]

The second major component of the process was the introduction of market-based prices. In restructured markets, prices were no longer dictated by the regulator based on costs, but instead determined by market forces. At the wholesale level, contract prices were determined by mutual agreement between buyer and seller, and centralized auctions cleared at the lowest price at which supply would meet demand. At the retail level, market-based prices included the introduction of competitive retailers who could sell energy at unregulated prices to final consumers. Partly because of uncertainty about whether deregulation would be effective and whether consumers would be protected from high prices, states differed in how they implemented retail competition. Twenty years later, a substantial share of industrial and commercial customers have switched to competitive retailers, but, in most states, the large majority of residential consumers still purchased from the incumbent utility.[11] Typically, incumbent utilities were still required to offer "bundled service," in which they provided electricity at regulated rates in addition to the delivery services that they also provided for competitive retailers.[12]

To ease the transition to deregulated markets, many states implemented caps that limited the rates utilities could charge for customers for several years. States that implemented these programs included Connecticut (expired in 2004), Delaware (2005), Illinois (2006), Maryland (expired between 2004 and 2008), Massachusetts (2004), and Virginia (2006). Along with the price caps, utilities typically signed long-term contracts with the newly divested generation facilities with terms that matched the rate caps. These contracts and price caps play an important role measuring the effects of deregulation, which we address in Section 6.

Deregulation was expected to bring increased efficiency by providing incentives to reduce costs, since under market-based prices lower costs translate into higher profits. Evidence indicates that in fact power plants are both operated more efficiently (Fabrizio et al., 2007; Cicala, 2015) and dispatched more efficiently (Cicala, 2022). Although previous research has found

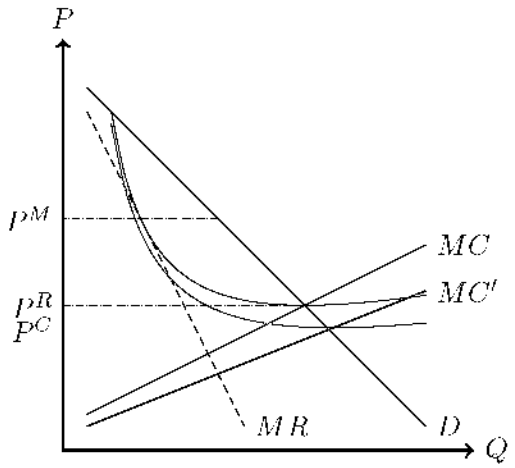
[9]Competitive generation was allowed in a limited fashion since 1978 (Public Utility Regulatory Policies Act, known as PURPA), but entry was limited due to the lack of incentives for utilities to purchase from new entrants or to share transmission assets with competing generation facilities.

[10]There were initially six centralized markets organized by independent system operators (ISOs), the entities in charge of coordinating the use of transmission assets. These are the California ISO (CAISO), Electric Reliability Council of Texas (ERCOT), the New York ISO (NYISO), the New England ISO (NEISO), the Midwest ISO (MISO), and the Pennsylvania-New Jersey-Maryland Interconnection (PJM). Prior to the implementation of ISOs, several markets operated power pools, which served a similar function.

[11]See Hortaçsu et al. (2017) for a discussion of the causes of this phenomenon.

[12]Competitive retailers were able to make use of the distribution grid to sell directly to end consumers. Their consumers paid a regulated distribution rate to the utility, in addition to paying for the electricity from the retailers.

Figure 1: Market-Based and Regulated Prices



Notes: Figure illustrates how market power could increase prices in deregulated markets, despite the presence of cost efficiencies. The thick black line labeled MC plots the marginal cost curve under a regulated regime. The regulated prices are set to reimburse average costs, which are plotted with the thin black curve (unlabeled). With efficient investment, average costs equal MC at the intersection with the demand curve, D, resulting in price P^R. Cost efficiencies from deregulation are illustrated with a downward shift in the marginal cost curve to the thick gray line MC'. In a competitive market, prices will equal $P^C < P^R$. With market power, firms could raise prices up to P^M, which is determined by the intersection of MC' and the marginal revenue curve, MR.

evidence of significant market power in deregulated electricity markets (Borenstein et al., 2002; Puller, 2007; Mansur, 2007; Ito and Reguant, 2016), the literature so far has paid less attention to the role that market power may have in translating this efficiency gains into lower prices for consumers. This paper helps to fill this gap.

Figure 1 illustrates how market power could increase prices in deregulated markets, despite the presence of cost efficiencies.[13] The market demand curve is plotted by the black line labeled D. The thick black line labeled MC plots the marginal cost curve under a regulated regime. The regulated prices are set to reimburse average costs in the market,[14] which are plotted with the thin black curve. With efficient investment, average costs equal MC at the intersection with the demand curve, resulting in regulated price P^R.

In competitive markets, profit incentives could lead firms to more efficiently allocate the supply of electricity. These potential cost efficiencies are illustrated with a downward shift in the marginal cost curve. The new marginal costs are plotted with the thick gray line MC'. In a competitive market, prices will be determined by the intersection of the demand curve with the marginal cost cure, resulting in price $P^C < P^R$. With market power, firms could raise prices up to P^M. P^M is the monopoly price and is determined by the intersection of MC' and the marginal revenue curve, MR. In this figure, deregulation could result in prices ranging from

[13]While this figure does not take into account cost heterogeneity, which is characteristic of electricity markets, the measure of costs used in our empirical analysis does.

[14]For the purposes of the figure, average costs include a fair rate-of-return on capital.

P^C to P^M, depending on the degree of market power.

Based on the motivation for deregulation efforts, the regulator's problem can be cast as a decision between regimes in order generate the lowest retail prices.[15] Overall, whether retail prices increase or decrease after restructuring is an empirical question, depending on the relative importance of efficiency gains and market power.

2.3 Market Power in Electricity Markets

Despite electricity being a homogeneous product, suppliers can have substantial market power. Transportation over long distances is expensive, which limits the effective size of geographic markets. Further, large amounts of electricity cannot be stored efficiently. Thus, supply and demand for a particular location at a particular point in time can be quite inelastic, providing individual suppliers with opportunities to exercise market power.

In centralized ISO auctions, market power is present when suppliers shade their bids upward above their true marginal costs. The degree to which suppliers can do so depends on the rival sources of generation that can provide to that particular market for that particular time window. Prices for bilateral contracts, which represent the vast majority of wholesale electricity transacted, may also reflect restrictions on procurement imposed by public utility commissions. While such restrictions may have a benefit (e.g., a greater share of renewable energy), they often serve to reduce potential competition for a contract and increase market power. For example, it is generally understood that one reason why prices rose sharply in Illinois at the beginning of deregulation was due to poor auction design.

Previous work in the literature has shown significant degrees of market power among generators (Puller, 2007; Hortaçsu et al., 2017; Borenstein et al., 2002; Mercadal, 2022). During the crisis in California at the beginning of its deregulation process, for example, all generators had market shares below 10 percent and still were able to charge markups of around 100 percent (Borenstein et al., 2002; Borenstein, 2002). For prices to fall, substantial efficiency gains would be required to compensate for markups of this magnitude.

Indeed, the example of California is illustrative because it happened at the beginning of the restructuring process, when utilities still retained significant market power. The restructuring process lead to changes in market structure that changed the balance of market power between buyers and sellers. For instance, the introduction of retail competition could allow generators to charge larger markups, as a greater number of buyers in the wholesale market can increase the relative bargaining power of generators. Section 5 documents that, in fact, concentration among buyers has decreased in deregulated markets, while concentration among sellers has remained constant.

While nationwide deregulation measures facilitated the exchange of electricity across geographic markets, local deregulation did not do much to increase within-market competition.

[15] We provide a simple formalization of this problem in Appendix B.

Utilities tended to sell of their entire portfolio of generation to a single new entity.[16] Further, there was limited entry of independent generators over time. Thus, generating facilities in deregulated markets did not realize a meaningful increase in local competition.

3 Data

3.1 Dataset Construction

To measure prices and markups, we use annual measures of generation, purchases, and retail sales within each utility's distribution territory. We obtain measures of quantities (MWh) and expenditures, allowing us to calculate average generation costs, average wholesale prices, and average retail prices. Our data accounts for the fact that, while the structure of the deregulated market changed, the geographical territories for distribution essentially remained unchanged, and the ultimate delivery of electricity to consumers continues to be the responsibility of the incumbent utilities.

We construct our unique dataset from several sources. Our main sources of data are reports provided by the Energy Information Administration (EIA) and the Federal Energy Regulatory Commission (FERC) from 1994 through 2016. These reports are publicly available, though they have not previously been combined at this level of detail. Utility-level aggregate data on generation, purchases, and sales is obtained from the operational data in form EIA-861. Form EIA-861 also provides more detailed measures of retail sales, which we use to construct state-specific measures of bundled service and delivery service for each utility. Bundled service refers to the provision of energy and its delivery using the utility's distribution grid; delivery service is the delivery of energy sold by a competitive retailer using the utility's grid. Form EIA-860 collects operational information on power plants, which we use to measure entry and exit of generation capacity.

Detailed data on purchases of electricity is obtained from FERC Form 1, which includes both purchases from centralized auctions and bilateral contracts. One of the key contributions of our data collection effort is to also incorporate bilateral contracts into the empirical study of electricity wholesale markets. These data are used by public utility commissions to set rates and are subject to audits. In addition, we augment the transaction-level data with information on firm ownership structure to construct an indicator of whether a purchase is made from an affiliated company. We use this measure to track what fraction of total sources obtained by a utility come from the same parent company versus independent suppliers.[17] The data on ownership structure was manually constructed from a combination of sources, including

[16] "The fact that these assets (power plants) were sold in large lots, sometimes entire power systems to a single buyer, demonstrates the greater concern regulators placed on vertical than horizontal market power." (Ishii and Yan, 2007)

[17] We are also able to use this data to measure the share of sources coming directly from the markets run by the Independent System Operators (ISOs).

current corporate structure from S&P Global, data on corporate structure, name changes, and mergers and acquisitions collected by the Edison Electric Institute (Edison Electric Institute, 2019), and manual Google search for confirmation.

Deregulation measures were implemented by 21 states in this period.[18] This definition includes measures that introduced market-based prices at the wholesale or wholesale and retail levels, and vertical separation measures including the strengthening of the wholesale market and free entry. Four states—Arizona, Arkansas, Nevada, and Montana—initially passed deregulation measures but later rescinded them. We remove them from our sample. We also remove Hawaii and Alaska, as the electricity infrastructure in these states is quite different from the rest of the United States. Finally, because Nebraska and Tennessee do not have investor-owned utilities with generation resources, they are not included in the sample. Thus, our sample of utilities covers 17 states that implemented deregulation measures and 25 states that did not. For additional details, see Appendix A.

3.2 Unit of Analysis and Key Variables

The unit of analysis in our study is the service area covered by investor-owned utilities (IOUs) in each state. Electric service in the United States is provided by three types of entities: IOUs, nonprofit cooperatives, and public utilities. IOUs were the primary target of deregulation measures—because they could make profits, were substantially larger than other types of utilities, and provided the vast majority of electricity service. In 1994, the 250 IOUs provided 75 percent of generation and 76 percent of retail service in the United States.[19] Since investor-owned utilities are subject to different regulations across states, we treat each utility with service areas in different states as separate utility-state entities. For some parts of our analysis, we will consider the state-wide electricity "market," as all utilities in that state are under the jurisdiction of the same state-specific regulatory commission.

Though deregulation measures ended generation and retail service for several utilities in our sample, these utilities continued to own and operate distribution lines and provide delivery service to retail customers. Because our focus is on the impact to consumers, we define our unit of analysis as each utility's service area. Service areas (i.e., the distribution infrastructure) are quite stable over time. For a visual representation of the geographic coverage of these areas, see Figure A1 in the Appendix. We also account for mergers of utilities throughout our sample period; if utilities merge at any point, we treat them as a single merged entity throughout our sample. For our analysis, we focus on utilities that had generation resources in 1994, at the

[18] Our sample of states that deregulated includes Rhode Island, New York, California, New Hampshire, Massachusetts, Pennsylvania, New Jersey, Delaware, Maryland, Connecticut, Illinois, Maine, Ohio, Texas, Virginia, Oregon, and Michigan.

[19] In 1994, 3,207 utilities reported to the EIA. The remaining 2,957 utilities that were not IOUs consisted of 2,194 municipal utilities and cooperatives, which tended to be much smaller, and 156 publicly run power authorities at the federal, state, or subdivision level.

beginning of our sample. Our final sample consists of 154 merged IOUs that provided over 70 percent of generation and over 70 percent of retail service in 1994.

The key outcomes of interest are retail prices, wholesale prices, and costs. For our primary measure of retail price, we use the "default" price available to residential, industrial, and commercial customers of a utility. We construct this measure by taking the average price for bundled service for each customer type and weighting these measures by the share of consumption by each customer type in the service area. Thus, we adjust for the fact that the composition of customers electing retail service from competitive sources changes over time. For Texas and Maine, several utilities no longer provide bundled service; for these utilities we instead use the average bundled price offered by all retailers in the state.[20]

For wholesale prices, we use the (weighted) average price for purchased electricity by each utility, which we obtain from the detailed transaction data in FERC Form 1. This measure has the advantage of reflecting demand and supply conditions that are local to each utility's service area. We also use these transaction data to capture the share of purchases that come from ISOs and affiliated companies.[21]

For generation costs, we use generator-specific fuel receipts data from EIA to construct a measure of marginal costs. For each utility, we sort its associated generation facilities by fuel costs. We then measure marginal costs as the average fuel cost for the 75th through 100th percentile of MWh generated.[22]This measure captures the marginal cost at the market level, i.e., the marginal cost of the marginal plant. We use the most expensive plants (instead of the average variable cost across all plants) because these plants are most likely to supply the marginal unit of electricity and their costs would determine prices in perfectly competitive markets. Thus, markups over this measure of marginal cost reflect market power and not competitive rents or profits. We use a range of costs (rather than, e.g., the 100th percentile) because the marginal unit varies over the course of the day and over the year. Our results are not sensitive to the lower-end percentile used in this calculation; we obtain similar results for changes in markups if we use the 60th-100th or 90th-100th percentiles instead. With the 75th-100th percentile, marginal costs are approximately equal to wholesale purchase prices in the pre-deregulation period, which we view as a reasonable starting point to test for market power after deregulation.

Our primary measure of costs uses, for each service area, all generators that were owned by the utility at the beginning of our sample (in 1994). That is, we ignore changes to ownership over time that may have been brought about as a result of deregulation. Thus, we preserve a

[20]Throughout, we consider annual quantity-weighted prices as our analysis focuses on price levels. Utilities differ in terms of how much electricity prices can vary month-to-month or with consumption. Existing evidence suggests that consumers are not particularly responsive to such variation (Ito, 2014; Deryugina et al., 2019).

[21]Our measure is somewhat conservative in that a utility may sell generation to a power marketer who then supplies electricity to a delivery customer of the utility. We cannot track this in the data, but if we could it would increase our measure of affiliated purchases.

[22]Before constructing the measure, we winsorize individual generator fuel costs at the 99th percentile.

proxy for generation costs that are specific to each utility's service area. The set of generators are reasonably stable over time; three-fourths of these generators appear in at least 20 years of our sample. To account for investment in new generation resources, we also calculate marginal costs at the state level using the 75th to 100th percentile of costs across all (current) utility and independent power producer generation facilities within the state. We consider retail markups, wholesale markups, and gross markups (retail prices minus generation costs) using these measures. For some analyses, we also consider average variable fuel costs across all generation units, which provides a more accurate measure of profits/rents.

When the unit cost of a given fuel at a specific power plant is not available, we impute it using the average unit cost for that fuel in the state and year; we then use plant-specific measures of fuel consumption and generation to calculate fuel cost per MWh. To the extent that within-state procurement costs for particular fuel types are correlated, this imputation will not affect our results. Due to reporting requirements, our measure of fuel costs in deregulated states comes disproportionately from smaller municipal utilities and coops,[23] which typically have higher procurement costs than the larger generation companies. Thus, our measure can be interpreted as an upper bound on costs. As we will see in the next section, our findings would only change if fuel costs for deregulated generators rose much faster relative to those for municipalities and coops, which we think is unlikely.

3.3 Summary Statistics

In this section, we provide some summary statistics of key variables in our sample. We identify similarities and differences between the treated and control utilities in our sample, where treated utilities are those in deregulated states. Some of the differences motivate our nearest-neighbor matching approach, which we describe in Section 4.

Table 1 shows the key variables for treated and control utilities in 1994. Column (1) reports the mean across the 78 IOUs in the deregulated states, and column (2) reports the mean across the 76 IOUs in the control states. Overall, utilities in deregulated and control states were similar in size in 1994, in terms of retail and generation output. There are some differences in generation mix across the two groups, in terms of the marginal generation units (75th-100th percentile by fuel cost). Markets in deregulated states were more likely to rely on oil (0.19 versus 0.07). This gives rise to a difference in marginal fuel costs, which are substantially larger in deregulated states in 1994. Despite this, the p-values of the difference in means for these variables, which are reported in column (3), are greater than 0.05, indicating no statistically significant differences. This finding, despite the economically meaningful differences in the share of oil and mean fuel costs, reflects the presence of a great deal of heterogeneity among utilities within each group.

[23] We do not directly observe fuel receipts for 60% of power plants in deregulated states and 17% in regulated states.

Table 1: Characteristics of Deregulated, Control, and Matched Control Utilities in 1994

	(1)	(2)	(3)	(4)	(5)
	Deregulated	Control		Matched Controls	
	Mean	Mean	p-value of Difference from (1)	Mean	p-value of Difference from (1)
ln(MWh Retail)	15.21	15.22	0.977	15.40	0.717
ln(MWh Generated)	14.70	14.60	0.857	14.59	0.891
Marginal Generation Share: Coal	0.50	0.54	0.705	0.53	0.817
Marginal Generation Share: Gas	0.12	0.15	0.639	0.12	0.943
Marginal Generation Share: Nuclear	0.02	0.02	0.763	0.01	0.575
Marginal Generation Share: Oil	0.19	0.07	0.078	0.16	0.735
Marginal Generation Share: Water	0.18	0.20	0.763	0.18	0.960
Marginal Fuel Costs	65.69	37.89	0.137	59.11	0.795
Retail Price	78.76	58.95	0.001	59.78	0.002
Number of Unique Utilities	78	76		72	

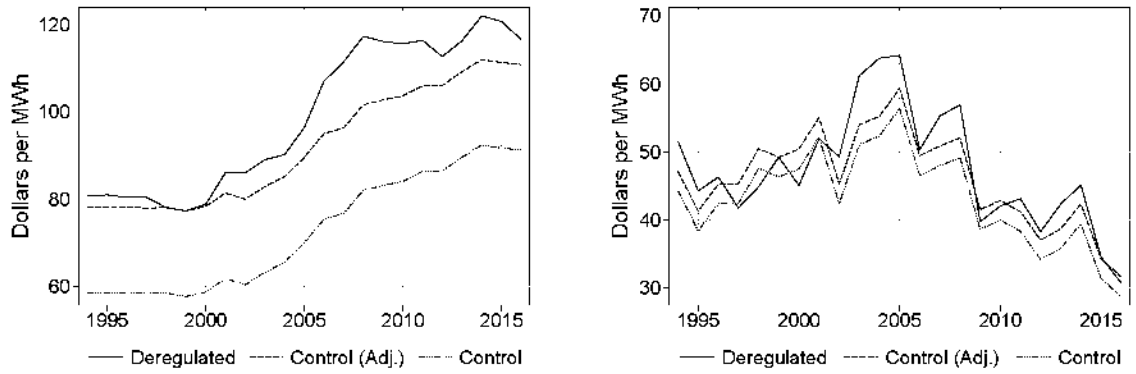
Notes: Table displays 1994 characteristics for 78 investor-owned utilities in states that later deregulated and 76 investor-owned utilities in states that did not deregulate. Columns (1) and (2) report the mean characteristics for each group, and column (3) reports the p-value of the difference in means. Column (4) reports the means for matched controls using a nearest-neighbor methodology, and column (5) reports the p-value of the difference in means between matched controls and the deregulated utilities. The first eight variables: (log) retail MWh, (log) generation MWh, marginal generation share by fuel type, and marginal fuel costs are used as matching variables.

Both of these features: mean differences across groups and heterogeneity within groups motivate our use of a matching procedure. By matching each deregulated utility to a set of similar controls, we can account for some of the heterogeneity in utility types. Specifically, we match utilities to three nearest neighbors based on 1994 values of (log) retail MWh, (log) generation MWh, marginal costs, and generation mix. Thus, we obtain a utility-specific control group that reflects both the type of generation and the size of the utility. We draw nearest neighbors from the pool of 76 control utilities. We provide additional details of our matching procedure in Section 4.2.

Column (4) in Table 1 reports the means for the nearest-neighbor controls, which are weighted by the number of times each utility is selected. Overall, the group becomes more similar to the deregulated utilities in terms of generation mix and fuel costs. For example, the difference in the oil share shrinks from 0.12 to 0.03. Marginal fuel costs for the matched control group increase to 59.5 dollars per MWh, which is close to the mean of 65.7 in the deregulated group. Correspondingly, the p-values for the matching variables tend to increase. The average p-value for the matching variables increases from 0.615 in column (3) to 0.787 in column (5). Note that the matching procedure only selects 72 out of the 76 possible control utilities as nearest neighbors.

Overall, utilities in deregulated states had higher prices than similar utilities in control states (79 versus 59 dollars per MWh). In 1994, implied gross markups are a small fraction of the

Figure 2: Aggregate Measures of Electricity Prices and Generation Costs



(a) Retail Prices

(b) Marginal Fuel Costs

Notes: Panel (a) plots the quantity-weighted default retail price for investor-owned utilities in deregulated states (solid line) and in control states (dotted light grey line). Panel (b) plots the average fuel costs of generation for all generating facilities that in 1994 belonged to utilities in deregulated states (solid black line) and control states (dotted line). The dashed line in both panels plots retail prices and fuel costs for control states after adjusting for level differences in 1999.

retail price. Thus, our measure of fuel costs can explain much of differences in prices across the two groups. In addition, the difference in prices between the two groups was stable before the onset of deregulation. In Figure 2, we present the time series of average prices for both groups, where we weight the average by retail MWh in each service territory. Panel (a) shows the mean retail price for deregulated states with a solid line and the mean for control states, after adjusting for level differences in 1999, with a dashed line. From 1994 to 1997, prices were stable in both groups. From 1998 to 2000, prices in deregulated states fell slightly, while prices in control states remained flat. Starting in 2001, prices in both states began to rise. Deregulated prices outpaced control prices until 2005, when the gap between the two widened further.

Likewise, panel (b) of Figure 2 shows marginal fuel costs for the two groups. As described above, we calculate the marginal costs based on the 75th through 100th percentiles of fuel cost for the generators that utilities in each group owned in 1994. After accounting for level differences, fuel costs for generation facilities in deregulated markets closely tracked fuel costs in control markets from 1994 through 2004. Starting in 2005, generation costs began to decline, and they declined more rapidly in deregulated markets. This pattern can largely be explained by the greater use of natural gas generators in deregulated states, as the price of natural gas fell significantly with the expansion of fracking.[24]

The general patterns we observe are not sensitive to the particular measure of costs. In

[24]Using only generators that appear in at least 20 years of our sample (three-fourths of the 1994 generation facilities), the time series of marginal fuel costs are almost identical, indicating that lower average costs in deregulated states were not driven by the retirement of expensive generation facilities.

Figure A4 of the Appendix, we show similar trends using average variable costs rather than our proxy for marginal costs. In Figure A5 of the Appendix, we present trends costs using statewide measures of marginal and average variable costs, rather than utility-specific measures. As in panel (b) of Figure 2, we find declining costs in both deregulated and control states in the latter half of our sample.

Thus, though retail prices rose substantially in deregulated states, there was no corresponding rise in fuel costs in these states. Using our localized measure of generation costs, we find that fuel costs in deregulated markets declined overall. This high-level finding is consistent with an increase in markups in deregulated states relative to control states, and motivates our more in-depth empirical analysis in Section 4.

4 Measuring the Effects of Deregulation

4.1 Empirical Strategy

The goal of our analysis is to evaluate the effect of electricity restructuring on markups and prices. For this, we compare utilities in restructured states to those that remained vertically integrated and regulated, and we examine the evolution of costs, wholesale prices, and retail prices over time. Specifically, we use a difference-in-differences matching approach, which we describe in greater detail in the next section.

By individually matching utilities based on their size and fuel costs prior to the onset of deregulation, we are able to nonparametrically control for changes in macroeconomic factors—such as fuel costs and demand for electricity—when measuring a number of outcome variables. Matching on fuel costs also allows us to control some relevant geographical variation, since plants in different locations may face different fuel costs.[25] Intuitively, we are using the data to provide an answer to the question, "What happened for similar utilities in states that did not deregulate?"

Because a state decision to restructure its electricity sector was not completely random, causal inference in this context is difficult.[26] A causal interpretation of our findings would require the assumption of parallel trends, which has several nuances in our context. First, it requires that there were no ongoing trends that differentiated the two groups outside of deregulation. Though comparable utilities in states that implemented deregulation measures initially had higher retail prices (Table 1), markups were similar, and costs and prices follow similar trends from 1994 through 1999 (Figure 2). This suggests that the parallel trends assumption

[25]For robustness, we include a specification where we also include whether or not the utilities are in the same geographic area (Census region) in the matching procedure. This does change the set of matched utilities but has little impact on our results. We report this alternative specification in Tables A5 and A6 in the Appendix.

[26]This is highlighted by how little we know about the consequences of restructuring 20 years later (Bushnell et al., 2017), in spite of the sector's importance and the urgency of market rules that can aid the transition to decarbonization.

may be reasonable before the onset of restructuring.

Second, the parallel trends assumption requires that shocks unrelated to deregulation did not differentially affect deregulated and control states after implementation. The primary concern on this front arises from changes in fuel costs and environmental regulation, which we control for using our matching approach since the effect of these shocks depends primarily on the fuel mix.

Third, the assumption requires that the effects of deregulation did not spill over into control states. Because of the ongoing integration of electricity markets across states, it is indeed plausible that deregulation could have affected retail prices in neighboring states. However, if we account for spillovers, the data suggest that our findings may be a conservative *lower bound* of the effects of deregulation, as we also observe large increases in retail prices and markups in control states (Figure 2).

A final consideration is whether other aspects of markets that affected market power and cost efficiency developed differently following deregulation. For example, we expect entry decisions to follow different dynamics in restructured and vertically integrated states. We do not want to control for all of these factors, as some endogenous responses are part of the effect we want to estimate. Keeping this distinction in mind, we examine alternative mechanisms that could potentially affect our findings in Section 7. Though we find some differences in policies affecting deregulated and control states, these differences do not provide a consistent alternative explanation for the changes in prices and markups we observe. Thus, despite the above caveats, we believe our empirical results provide a compelling narrative that suggest the widespread presence and practice of market power.

4.2 Difference-in-Differences Matching Estimator

To measure changes in outcomes for deregulated utilities, we match utilities in states that implemented market-based prices (the "deregulated" group) to utilities in states that did not (the "control" group) based on pre-deregulation retail MWh, generation MWh, and fuel costs, using our measure of marginal costs. We then apply a difference-in-differences adjustment to the bias-corrected matching estimator developed by Abadie and Imbens (2006, 2011). Our estimation procedure closely follows the approach of Deryugina et al. (2019). Though we use the term "control" and "counterfactual," it is important to note that the state-specific decision to deregulate was not purely random, as discussed in the previous section.

For each of our 78 deregulated utilities, we use 1994 outcomes to identify the three nearest neighbors from the pool of 76 control utilities in our sample. By matching based on 1994 values, we can observe how outcomes evolve prior to deregulation and assess the plausibility of the parallel trends assumption. We use match on log generation MWh, log retail MWh, marginal costs,[27] and the shares of (marginal) generated MWh coming from five fuel types:

[27]When matching, we transform marginal costs using the inverse hyperbolic sine, $f(z) = \ln\left(z + \sqrt{1+z^2}\right)$, which

coal, natural gas, oil, nuclear, and water. We use a least-squares metric to calculate distances between utilities, with equal weights across the three variables. We scale up the fuel type distance measures so that, across all potential matched pairs, roughly equal weight is put on fuel types as the combination of the other three variables.[28] We use this distance to select the three nearest neighbors for each deregulated utility, allowing control utilities to be matched to multiple deregulated utilities.

We use these nearest neighbors to construct counterfactual outcomes and employ standard difference-in-differences techniques to adjust for pre-period differences. Let Y_{it} denote an outcome of interest (e.g., retail prices) for utility i in period t, where $t-0$ corresponds to the year deregulation measures are implemented. Let $Y_{it}(1)$ indicate the outcome with deregulation and $\widehat{Y}_{it}(0)$ indicate estimated counterfactual without deregulation. Given $Y_{it}(1)$ and $\widehat{Y}_{it}(0)$, we can obtain a utility-specific estimate of the effect of deregulation on the outcome, $\widehat{\Delta Y}_{it}$:

$$\widehat{\Delta Y}_{it} - Y_{it}(1) - \widehat{Y}_{it}(0). \tag{1}$$

We observe the outcome $Y_{it}(1)$ for the deregulated utilities in our data. The counterfactual outcome, $\widehat{Y}_{it}(0)$, is unobserved and is calculated as follows. For each deregulated utility i, we select three nearest neighbors using the above procedure. We calculate the counterfactual outcome, $\widehat{Y}_{it}(0)$, as the average value of $Y_{it}(0)$ across the three matched control utilities plus the difference between deregulated and matched control outcomes in the period prior to deregulation. Thus, outcomes are indexed so that $Y_{i0}(1) - \widehat{Y}_{i0}(0)$. By indexing the levels to a baseline period, we obtain a utility-specific "difference-in-differences" estimate for any outcome of interest.

To quantify the average impact of deregulation across our utilities, we take the weighted average of the utility-specific treatment effects:

$$\widehat{\overline{\tau}}_t - \frac{\sum_i \omega_i \widehat{\tau}_{it}}{\sum_i \omega_i}. \tag{2}$$

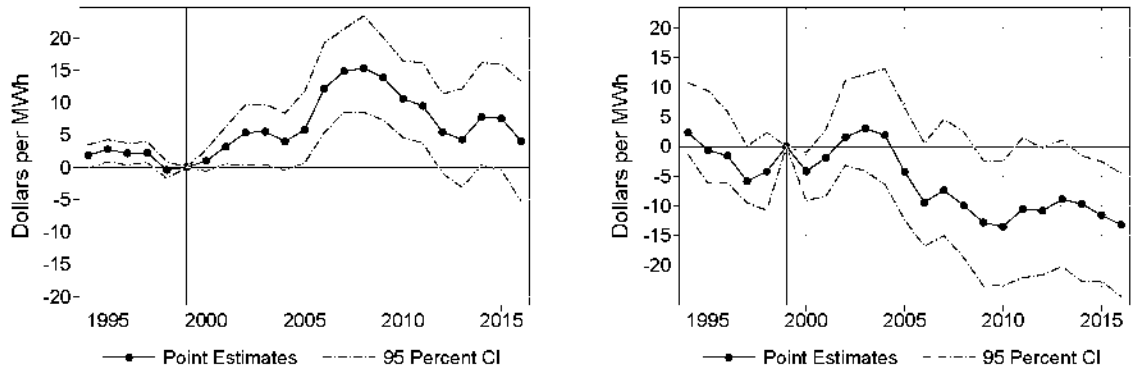
where ω_i is the retail MWh provided by the deregulated utility in 1994. Our weighting variable is chosen to capture the size of the utility with respect to consumption in its service area.

For our main analysis, we use 1999 as our baseline period across all states. Though there is some variation in terms of when deregulation measures legally came into effect across states, in practice, the restructuring effects all happened within a few years. This timing has little impact on the results we measure, which occur over 15 years after deregulation. Using a common baseline period has the advantage of making the empirical results more transparent, especially

is approximately the natural log function plus 0.7 for $z > 5$ and also has $f(0) = 0$.

[28] Specifically, we scale up the shares by $\sqrt{30}$, though we obtain similar point estimates with alternative scaling factors (i.e., 1 or $\sqrt{300}$). The procedure yields reasonable nearest-neighbor matches for individual utilities. For the matched pairs, the chosen weight prioritizes the fuel mix. We match over three-quarters of the utilities almost exactly based on fuel types.

Figure 3: Estimates of Changes in Prices and Costs After Deregulation



(a) Retail Prices

(b) Marginal Costs

Notes: Figure displays difference-in-differences matching estimates of changes in (a) retail prices and (b) fuel costs for deregulated utilities. Each deregulated utility is matched to a set of three control utilities based on 1994 characteristics. The estimated effects are indexed to 1999, which is the year prior to the first substantial deregulation measures. The dashed lines indicate 95 confidence intervals, which are constructed via subsampling.

for concerns about macroeconomic trends, such as changes in fuel prices. Our results are similar if we instead index treatment communities to their legal deregulation date.[29]

As in Deryugina et al. (2019), we employ a subsampling procedure to construct confidence intervals for our matching estimates.[30] Consider a parameter of interest, $\widehat{\theta}$. For each of $N_b = 500$ subsamples, we select without replacement $B_1 = R \cdot \sqrt{N_1}$ deregulated utilities and $B_0 = R \cdot \frac{N_0}{\sqrt{N_1}}$ control utilities, where R is a tuning parameter, N_1 is the number of deregulated utilities, and N_0 is the number of control utilities. For each subsample, we calculate $\widehat{\theta}_b$. The matching estimator converges at rate $\sqrt{N_1}$ (Abadie and Imbens, 2006, 2011), and the estimated CDF of $\widehat{\theta}$ is given by:

$$\widehat{F}(x) - \frac{1}{N_b} \sum_{b-1}^{N_b} 1 \left\{ \frac{\sqrt{B_1}}{\sqrt{N_1}} \left(\widehat{\theta}_b - \widehat{\theta} \right) + \widehat{\theta} < x \right\} \tag{3}$$

The lower and upper bounds of the confidence intervals can then be estimated as $\widehat{F}^{-1}(0.025)$ and $\widehat{F}^{-1}(0.975)$. We employ $R = 3$ ($B_1 = 26$) for the confidence intervals and standard errors reported in the paper.

4.3 Prices, Costs, and Markups

We first show that retail electricity prices increased for customers in deregulated states. Panel (a) of Figure 3 displays the average change in retail prices relative to matched controls. Leading up to the baseline year of 1999, there is little difference in price trends for deregulated and control utilities. From 2000 to 2005, deregulated utilities saw modest increases in retail prices, which an average difference of 3.9 dollars per MWh over that period. In 2006, deregulated utilities realized a sharp rise in retail prices, with an average difference of 12.6 dollars per MWh from 2006 to 2011 and an overall increase of 7.9 dollars per MWh from 2000 to 2016. The increases in the latter years are large in magnitude. The average retail price for deregulated utilities in 1999 was 78.0 dollars per MWh, so an increase of 12.6 dollars per MWh corresponds to a 16 percent increase in prices relative to the baseline. We reiterate that these changes are difference-in-differences effects, i.e., increases above and beyond the price trends occurring in control utilities.

A natural question is whether the price changes reflect underlying changes in costs. Panel (b) of Figure 3 plots the relative marginal generation costs for deregulated utilities. Relative to control utilities, deregulated utilities saw a *decrease* in generation costs in the post-deregulation period. From 2000 to 2016, fuel costs declined by 6.9 dollars per MWh in the deregulated utilities. Thus, despite declining costs, prices rose in deregulated states.

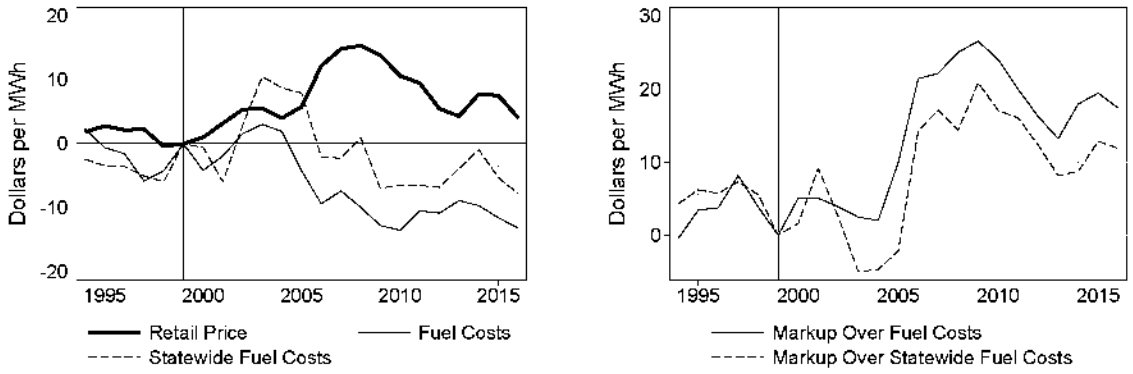
The combined effects of increasing prices and decreasing costs suggest that markups to consumers rose in deregulated states. To illustrate this, we combine the retail price effects and the generation costs on the same plot in panel (a) of Figure 4. The difference between the retail price (in thick solid black) and the fuel costs (in thin solid black) is the gross markups paid by end consumers above the generation costs of electricity. The gross markups are plotted in panel (b). The increase in gross markups was modest from 2000 until 2005. Markups spiked in 2006, with an increase of over 20 dollars per MWh from 2006 through 2011.

Our finding of increasing markups is robust to our measure of costs. As an alternative measure to the utility-specific generation costs, we calculate marginal costs from all utility and independent power producer generators within the same state. An argument for using this measure as opposed to the utility-specific measure is that, in a competitive market, consumers may obtain electricity from a lower-cost source that is nearby but outside of their service area. Additionally, this alternative measure accounts for entry of new plants. The dashed line in panel (a) plots the change in statewide fuel costs. Though the decline is not as quite large as the utility-specific measure, we find that statewide fuel costs decline in deregulated utilities relative to their controls. The dashed line in panel (b) plots the gross markup for retail prices using this alternative measure of costs. We still find large increases in gross markups to consumers using

[29]For a comparison, see Figures A2 and A3 in the Appendix.

[30]Matching estimators do not meet the regularity conditions required for bootstrapping (Abadie and Imbens, 2008), and subsampling provides great flexibility in terms of calculating treatment effects.

Figure 4: Prices, Costs, and Gross Markups



(a) Prices and Costs

(b) Gross Markups

Notes: Figure displays difference-in-differences matching estimates of changes in prices, costs, and gross markups for deregulated utilities. Panel (a) provides the point estimates for retail prices (thick line) and utility-specific fuel costs (thin solid line) from Figure 2 on the same plot. The dashed line on the plot represents an alternative measure of costs reflecting the average statewide fuel costs for all generators in each utility's state. Panel (b) displays the changes in the gross markups, which are defined as the retail price minus fuel costs, using both measures of costs from panel (a).

this alternative measure.

Table 2 summarizes the estimated difference-in-differences coefficients, as well as the baseline measures, for our key outcomes of interest.[31] The overall changes in retail prices and gross markups from 2000-2016 are large and highly significant. The changes in generation costs and wholesale markups we observe are economically meaningful and statistically significant at the 0.10 level. We find stronger effects for prices and generation costs starting around 2006. As discussed earlier, our findings are similar if we index each utility to state-specific implementation dates, rather than calendar time. Figure A2 in the Appendix shows that the share of own generation divested looks nearly identical using both measures of time. Appendix Figure A3 plots the corresponding effects on prices and costs, which are similar to the estimates in Figure 3 above.

As a robustness check, we estimate an alternative version of our matching procedure where we also weigh whether or not the control utility is in the same geographic area. For this procedure, we use Census regions (Northeast, Midwest, South, and West), and we choose a scaling factor that meaningfully changes the mix of matched control utilities. This has little impact on our results. We report the summary stats and outcomes with this specification in Tables A5 and A6 of the Appendix.

[31] The changes in markups in Table 2 do not always equal difference in changes between prices and costs because there are some periods where we do not observe wholesale prices for some utilities. In these cases, we do not calculate retail or wholesale markups.

Table 2: Relative Changes in Prices, Costs, and Markups

	(1) Retail Price	(2) Wholesale Price	(3) Generation Cost	(4) Retail Markup	(5) Wholesale Markup	(6) Gross Markup
1999 Values	78.06	42.81	48.89	34.95	−5.22	29.13
2000-2005	4.14	−0.42	−0.63	4.87	−0.26	4.74
	(1.74)	(2.97)	(2.88)	(2.45)	(4.52)	(3.59)
2006-2011	12.73	3.46	−10.59	9.38	12.30	23.18
	(2.95)	(3.49)	(4.82)	(3.84)	(6.03)	(5.72)
2012-2016	5.83	7.41	−10.83	2.63	16.40	16.80
	(3.83)	(4.18)	(4.92)	(4.10)	(6.56)	(6.01)
2000-2016	7.66	3.16	−7.11	5.62	8.82	14.71
	(2.30)	(2.99)	(3.59)	(2.73)	(4.68)	(4.40)

Notes: Table displays the estimated difference-in-differences matching cofficients for prices, costs, and markups between deregulated and control utilities in dollars per MWh. The first row provides the baseline values in 1999, and the remaining rows provide the average effect for the specified time period. Standard errors are displayed in parentheses.

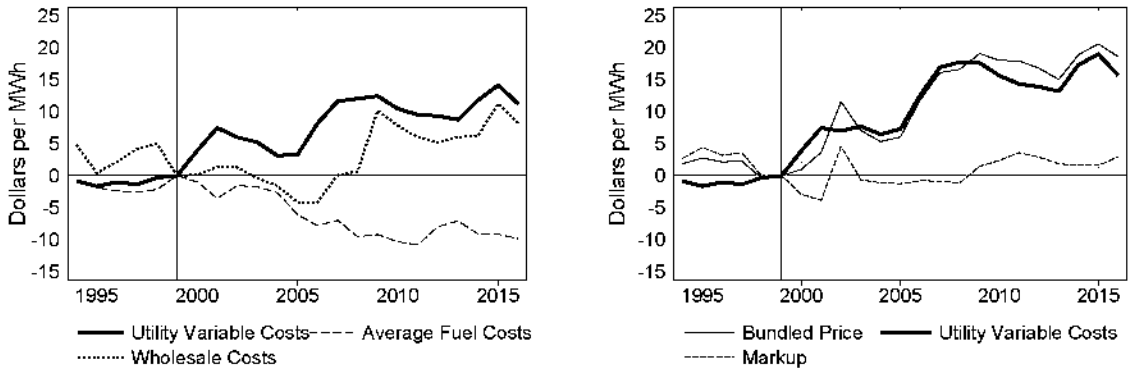
Section C in the Appendix discusses the variation in these effects across states. We estimate some heterogeneity across states. Most deregulated states realized meaningful price increases, with 9 states realizing price effects exceeding 5 percent. We estimate that consumers in some states did benefit from deregulation, with consumers in Virginia and Illinois realizing meaningful decreases in prices.

4.4 Where is the Increase in Markups Coming From?

Our above findings indicate an increase in gross markups paid by end consumers and higher prices. To unpack these changes, we now focus on incumbent utilities. In most states, even after deregulation, these utilities were required to continue to offer "bundled" service—i.e., providing retail electric service in addition to distribution—at regulated prices based on the procurement costs of electricity. At the same time, the utilities were required to switch from own generation to wholesale market purchases to supply these consumers. By studying how costs and prices moved for incumbent utilities, we illustrate the important role of generation markets and the underlying mechanisms that explain the estimated price changes.

Panel (a) of Figure 5 shows the impact of deregulation on the procurement costs for utilities using our difference-in-difference matching approach. The average variable costs for utilities (thick black line) increased shortly after the divestiture of generation facilities in 2000, and it remained 5 to 15 dollars per MWh higher throughout the sample period. The variable cost of electricity is the weighted average of the average fuel cost for generation by the utility (thin dashed line) and the average cost of electricity purchased from wholesale markets (dotted line).

Figure 5: Utility Costs and Markups



(a) Utility Variable Costs

(b) Utility Prices, Costs, and Markups

Notes: Figure displays difference-in-differences matching estimates of changes in costs, prices, and markups for regulated electric service in deregulated states. The thick black line in both panels shows the change in average variable costs for utilities. Each utility's average variable cost is calculated as the weighted average of generation fuel costs and wholesale purchase prices. Changes in these variables are shown in panel (a). Variable costs increase from 2000 through 2005 despite no increase in generation fuel costs (dashed line) and wholesale purchase prices (dotted line) because utilities procured a greater fraction of electricity from wholesale markets. Panel (b) plots the regulated bundled price (think solid line) and the utility markup (dashed line), defined as the bundled price minus the average variable cost.

Two factors contribute to the increase in average variable costs. The first is that, by separating from generation facilities, deregulated utilities had to procure a greater portion of the electricity sources from the wholesale market. For a utility, obtaining electricity from the wholesale market was more expensive than generation, as wholesale prices reflect a markup. In 1999, the mean wholesale markup over average variable generation costs was 17.1 dollars per MWh. Thus, despite the fact that wholesale prices and fuel costs both *declined* over the period 2000 to 2005, utility variable costs *increased* by 5.6 dollars per MWh. With deregulation, utilities effectively paid a market-based markup to generation facilities that they had previously owned.

The second factor that led to an increase in average variable costs for utilities was the increase in wholesale prices beginning in 2007. Though wholesale prices remained relatively flat in the initial years of deregulation, they eventually increased substantially, rising by 8 dollars per MWh from 2012 to 2016. The increase in wholesale prices, combined with the significant declines in fuel costs, indicate that wholesale markups for generators increased substantially in deregulated states. Our difference-in-differences estimate for the increase in wholesale markups is 8.9 dollars per MWh from 2000 to 2016, which is over 60 percent of the overall increase in gross markups.

For bundled service, incumbent utilities were required to charge prices equal to the variable costs for electricity. We should expect then, that, ceteris paribus, utility variable costs should

move one-for-one with prices for bundled electric service. Indeed, panel (b) of Figure 5 shows that the increase in utilities' average variable costs (thick solid line) fully explains the increase in regulated bundled prices (thin solid line). In other words, the increase in retail prices we observe did not arise from an increase in utility "markups"—i.e., additional charges to cover higher distribution costs, stranded costs payments, or other features. Utility markups moved similarly in deregulated and control states, as shown by the dashed line in the figure.

The changes documented in Figure 5 point to the role of two fundamental economic mechanisms in explaining price increases in deregulated states. First, the divestiture of generation facilities allowed for double marginalization, as generators were able to charge markups to downstream utilities. This mechanism corresponds with the price increases we observe before 2005, where utility variable costs increased despite declines in wholesale prices and fuel costs. Average generation markups did not increase, but markups were applied to a much larger share of generated electricity. Over this period, retail markups for incumbent utilities remained constant, though there were modest retail markups for alternative retail suppliers.

The second mechanism was an increase in the exercise of market power by generators, which corresponds to the rise of wholesale prices after 2005. Prior to this year, generators in many states were not able to raise prices due to the presence of long-term contracts and rate caps at the retail level. In Section 6, we examine the timing of this change in more detail.

Although some have viewed market power in wholesale electricity markets as a factor only during a few hours of peak demand, our findings indicate that market power is more pervasive than that. At an annual level, we find substantial markup increases even over the costs of the most expensive power plants, which typically determine prices on an hour-by-hour basis. Moreover, our data suggest that generators are signing longer-term (annual or longer) contracts at a markup over generation costs. We observe similar price increases in ISO markets and contract markets, as shown in Figure A7 in the Appendix.

5 Market Power in Wholesale Electricity Markets

In this section, we present evidence supporting the presence of market power in wholesale electricity markets. We first look at how concentration of buyers and sellers has evolved in wholesale markets. Deregulation did not substantially change seller concentration, and there was a notable lack of entry. Buyer concentration fell, potentially decreasing buyers' bargaining power and contributing to higher wholesale prices. Second, we show that there is correlation between measures of potential competition—i.e., features of market structure at the time of deregulation—and the change in wholesale markups. Third, we show that changes in prices are not positively correlated with changes in fuel costs. In fact, states with higher fuel costs realized greater declines in costs yet relatively higher prices. Fourth, we show that utilities with a more elastic demand, as measured by a higher share of industrial consumers, saw a higher

increase in markups. Finally, section 5.4 shows that the effects on rates for incumbent utilities did not significantly vary by customer type, despite different elasticities. These findings are consistent with market power being exercised at the wholesale level.

We run these analyses at the state level. While ISO markets have integrated markets across utility serving areas, creating, in some cases, larger market areas, we think it is reasonable to use historical regions due to the cost of transmission and because entry has been limited. Taken as a whole, these pieces of evidence support our earlier finding of generator market power as the main driver of price increases after deregulation.

5.1 Upstream and Downstream Concentration

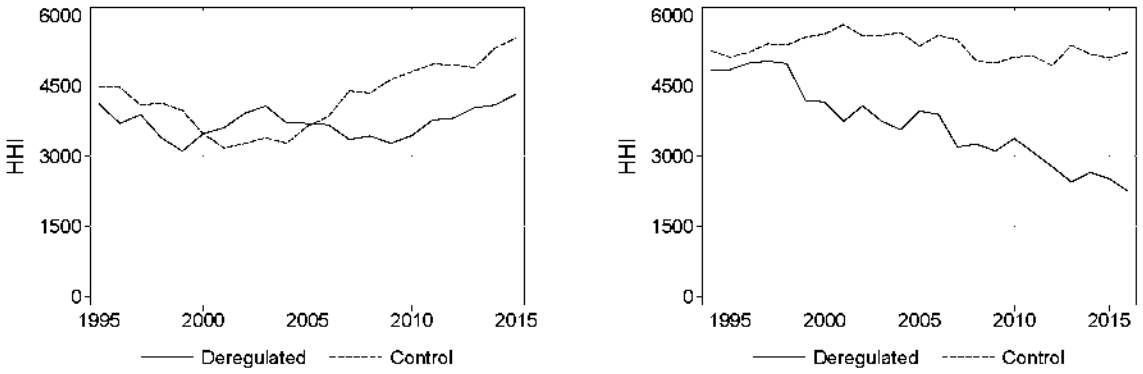
In this section, we use our detailed data, which provides a complete map of the corporate structure of the electricity industry, to accurately measure concentration at the wholesale and retail level over time.[32] Our findings indicate that concentration among wholesale sellers has remained high over the last two decades despite significant changes in market structure. Concentration among wholesale buyers has decreased over time, as expected with the introduction of retail competition, though it has remained high. While concentration is not necessarily an accurate measure of market power, these findings suggest that buyers have lost market power relative to sellers, which contributes to explain why utilities had to agree to higher prices when they sign contracts with new providers after their existing contracts expired.

We evaluate changes in concentration in upstream and downstream markets by calculating the Herfindahl-Hirschman Index (HHI) for restructured and control states. We find that concentration remained high in the upstream market for sellers. Though utilities were forced to divest their generation assets, this did not result in a substantial reduction in concentration. Often, a utility's entire generation portfolio was transferred to a single new entity, resulting in minimal changes to local competition. In the downstream market, we find that concentration decreased. Both forces—high concentration upstream and lower concentration downstream—could have increased wholesale prices (and markups) in restructured states. Decreasing concentration, or increased competition, in the retail market could increase wholesale prices through a reduction in buyer power. Initially, utilities were by far the largest buyers in their local markets. After vertical separation, utilities could purchase from several generation owners, some of which were affiliated companies. Over time, as retail competition increased, utilities' market share in the downstream market declined (see Figure 13 in Section 6). We think this change in the relative balance of bilateral market power may have contributed to the increase in markups in restructured states.

Panel (a) of Figure 6 shows the evolution of the mean HHI among firms that sell electricity to investor-owned utilities, as reported in FERC Form 1. Sellers have been aggregated to the

[32] We track ownership up until the ultimate parent company level.

Figure 6: Concentration Upstream and Downstream by Restructured Status



(a) Seller HHI

(b) Buyer HHI

Notes: The figure shows the evolution of the mean HHI over time, where the HHI is computed at the state level for both buyers and sellers. Buyers include investor-owned utilities and power marketers, as reported in EIA data. Sellers include all firms that sell to an investor-owned utility, as reported in FERC Form 1 data. For sellers, concentration is calculated at the parent company level.

parent company level, such that if a utility reports purchasing from a certain power plant, and the plant is owned by Exelon, for example, we consider that transaction as a purchase from Exelon. Both deregulated and control states were highly concentrated at the beginning of our sample and remained so, with average HHI levels consistently above 3,000.[33] Despite shifting an increasing share of energy to wholesale markets and encouraging independent generation, seller concentration did not decrease.[34]

Panel (b) of Figure 6 shows the evolution of the mean HHI among buyers for restructured and regulated states, where buyers include both investor-owned utilities and power marketers. Concentration remained roughly constant between 1995 and 2015 in regulated states. In restructured states, on the other hand, concentration started falling in the late 1990s, when the restructuring process started, and continued to do so through 2016. This pattern mirrors the increase in competition we observe in the retail sector. By the end of our sample, buyer HHI had crossed from the highly concentrated to the moderately concentrated range.

In summary, Figure 6 indicates that concentration among buyers decreased in restructured states, while seller concentration remained constant. This is consistent with sellers maintaining a high degree of market power and provides an explanation for the large markups we observe when prices are deregulated. In particular, we would expect buyers bargaining power to have decreased around 2005 when they had to sign new procurement contracts after the existing

[33]The US Department of Justice considers an HHI above 2,500 to be "highly concentrated," and an HHI between 1,500 and 2,500 to be "moderately concentrated."

[34]Regulated utilities generate most of their energy, so concentration measures for sellers in regulated states describe very small markets. After restructuring occurs in deregulated states, concentration measures are more representative because a much larger share of the market is traded.

ones expired. This correlation is not necessarily causal because market concentration is endogenous, but it is consistent with market power as the main explanation for our findings.

The above findings suggest that the entry of new generation plants did not not substantially affect upstream market concentration after deregulation. In a competitive market with free entry, we would expect high markups to attract new entrants, so we examine the entry of new generators over time. Persistently high markups are only possible if there are significant entry barriers, since otherwise new firms would enter the market to capture these high profits. Figure 7 shows the evolution of new capacity in the United States over time as a fraction of total capacity, net of retiring capacity. The figure shows an entry boom in the early 2000s, a period of optimism boosted by high capital availability and low gas prices (Kwoka, 2008a). These high levels of investment were rather an exception, since for most years entry of new capacity is relatively low (below 3 percent) for both deregulated and control states, though slightly lower in deregulated states.

Kwoka (2008a) documents the paucity of investment and lists several reasons, including large investment costs for new generators (e.g., $225 million for a gas generator of efficient size), long lead times for construction, the need for new transmission connections, the fact that incumbents already have plants in the best locations,[35] and time lags for regulatory approval ranging from 8 to 14 months. Further, unlike many other capital investments, investments in new generation plants are almost entirely sunk, as they plants cannot be repurposed for other uses. This, coupled with the long repayment period over decades, subjects any investor to a high degree of risk. In electricity markets, special risks include regulatory policy uncertainty, fuel cost uncertainty, environmental policy uncertainty, and technological uncertainty, all making investments in new generation more difficult.

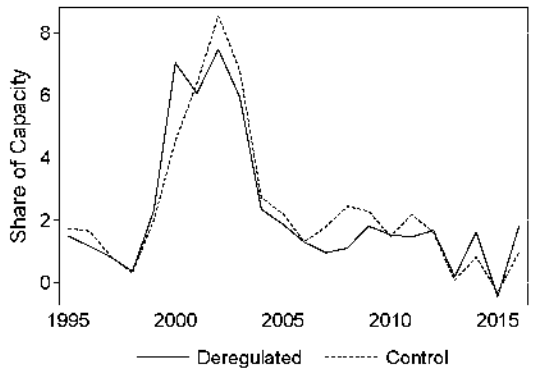
5.2 Supply-Side Factors

The previous section showed that upstream and downstream markets were highly concentrated and remained so after deregulation. Concentration levels, though suggestive, may not be a definitive indication for the presence of market power. In markets with homogeneous products, concentrated markets can still deliver close to marginal cost pricing when firms compete in prices, as in the classic Bertrand model.

To provide further evidence for the presence of market power, we examine heterogeneity across utilities. If electricity markets were characterized by near-perfect competition, then there would be no correlation between measures of market structure (such as concentration) and estimated changes to markups—any competition would be sufficient to drive prices down to marginal costs. On the other hand, if firms can exercise market power, then we might expect that variables correlated with competition will also correlate with changes in markups.

[35] Thermal plants need to be close to water and transmission. Renewable plants close to transmission and in an area with high wind or solar energy potential.

Figure 7: Net Entry of New Capacity



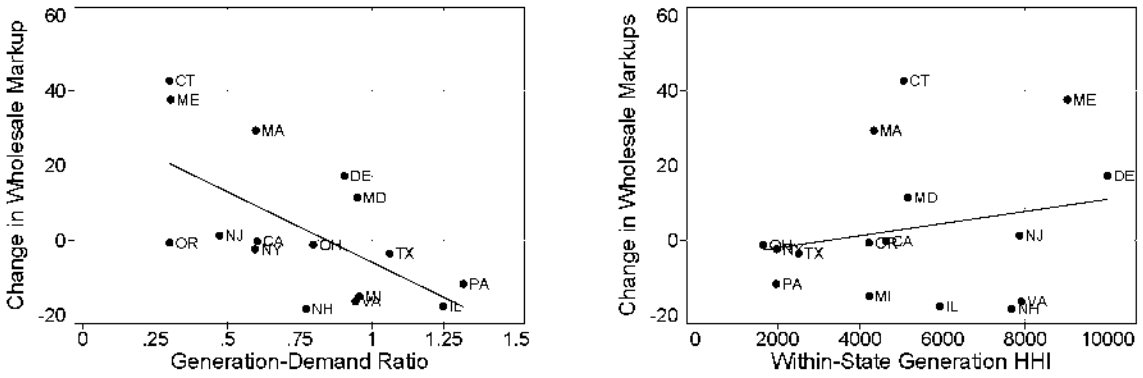
Notes: Figure displays the evolution of new nameplate capacity as a fraction of total capacity, net of retiring capacity, distinguishing between deregulated and regulated states. Only operating plants are included.

We consider two variables that would be expected to affect the intensity of competition in deregulated markets. We focus on upstream markets, as we estimate changes to be primarily driven by increases in wholesale markups. First, we consider a measure of potential *within-market* competition. For the pre-deregulation period, from 1994 to 1999, we calculate the average MWh generated and the average retail MWh demanded. We use the ratio of the two as a measure of the total potential within-market competition for generators. A lower value of this measure indicates that local generation is relatively scarce and imports of electricity from other service areas are more likely to be needed to cover demand. Since deregulation is state-specific, a higher value indicates that a greater share of production is subject to the effects of deregulation. If deregulation increases the role of competitive forces in the local market, then higher values should lead to less market power after deregulation. A ratio exceeding one indicates that local capacity exceeds demand, as the utility was a net exporter before deregulation.

Second, we consider a measure of *cross-market* competition. We exploit the fact that deregulated states varied in terms of the number of incumbent investor-owned utilities. In states with more utilities, after restructuring there are potentially more sellers to purchase electricity from in the newly created wholesale market. We capture the potential impact of competition from generators outside of the service area by measuring the within-state HHI of generation for each utility from 1994 to 1999. A lower concentration value would mean that the average buyer has more choices from the same state but outside the local service area after deregulation.

Figure 8 plots the impact on wholesale markups against our measures of competition. For this analysis, we measure costs and markups using our statewide measure of marginal costs. Impacts on markups are aggregated at the state level and across years 2000–2016, and the measures of competition are calculated relative to 1994–1999. We aggregate utilities to the state level, weighing each utility by retail MWh in 1994. We drop Rhode Island from our plots,

Figure 8: Potential Competition and Markups



(a) Within-Market Competition

(b) Cross-Market Competition

Notes: The figure shows the correlation between the estimated impact on wholesale markups, aggregated across years 2000-2016, and two measures of potential competition, aggregated across pre-deregulation years 1994-1999. Panel (a) presents correlation with generation-demand ratio, used as a measure of within-market potential for competition. Panel (b) presents correlation between estimated impact on wholesale markups and the within state generation HHI, interpreted as measure of cross-market potential for competition. We aggregate utilities to the state level, weighing each utility by retail MWh in 1994. We drop Rhode Island from our plots, as the generation plants for largest utility were very small and exited our sample after 1999, so we have no measure of wholesale markups for that utility.

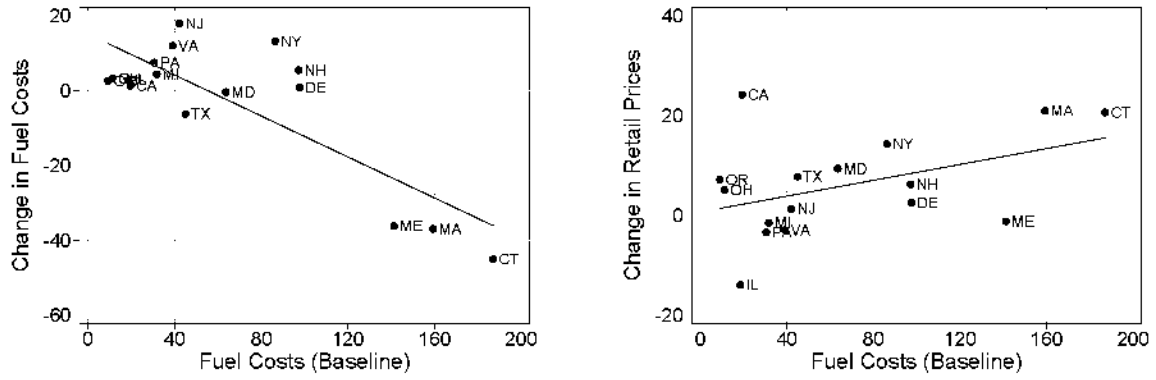
as the generation plants for largest utility were very small and exited our sample after 1999, so we have no measure of wholesale markups for that utility.

Panel (a) plots the change in wholesale markups versus the generation-demand ratio. Consistent with the presence of market power, lower potential within-market competition is associated with greater increases in wholesale markups. The correlation coefficient is -0.33. Panel (b) plots the change in wholesale markups against the within-state generation HHI. Consistent with the presence of market power, more concentrated markets have larger increases in wholesale markups. The correlation coefficient is 0.36.[36] These figures are in line with our explanation of increased markups in deregulated markets coming from market power.

To further investigate market power from the supply side, we analyze how the effects of restructuring varied across states according to pre-deregulation fuel costs. In a perfectly competitive market, we expect prices to be determined by marginal costs and therefore to move in proportion to costs. Therefore, states that see the largest declines in costs are expected to see commensurate effects on prices under competitive conditions. We examine whether this holds in Figure 9, which plots the relationship between pre-deregulation fuel costs, aggregated across pre-deregulation years 1994–1999, and impacts on both fuel costs and retail prices, aggregated across years 2000–2016. We aggregate utilities to the state level, weighing each utility by retail MWh in 1994. As before, we drop Rhode Island due the exit of its generation plants.

[36]The correlation coefficient for our two measures of potential competition is -0.20.

Figure 9: Changes in Fuel Costs and Prices Relative to Baseline Costs



(a) Fuel Costs

(b) Retail Prices

Notes: Panel (a) in the figure shows the correlation between pre-deregulation fuel costs and the estimated effect on fuel costs. Panel (b) shows the correlation between pre-deregulation fuel costs and the estimated effect on retail prices.

Panel (a) shows the correlation between pre-deregulation fuel costs and the estimated effect on fuel costs. States that had the highest costs initially saw the largest reductions, suggesting that inefficiencies explained the higher costs. The correlation coefficient is -0.78. Panel (b) plots pre-deregulation fuel costs against the estimated effect on prices. In a perfectly competitive world, both panels would look similar. By contrast, what we find is that states that had the highest pre-deregulation costs and highest cost declines also saw the largest price *increases*. The correlation coefficient between the average price impact and the baseline fuel costs is 0.49. The observation from these two figures is consistent with a market in which firms have market power, not a competitive one. Utilities might have been able to exert market power by inflating their costs in a regulated environment, and by charging higher markups in a deregulated market with market-based prices.

5.3 Elasticity of Demand

As an additional check to confirm that our findings are driven by firms' market power, we examine how the effects on markups vary with the elasticity of the demand. Although we do not directly estimate the elasticity of demand, we observe the share of industrial, commercial, and residential customers served by each utility, which is highly correlated to elasticity. Residential customers are typically less responsive to prices, while industrial customers have higher electricity bills and more flexibility over the timing of their consumption, which makes them more sensitive to prices (Fan and Hyndman, 2011; Burke and Abayasekara, 2018). In line with this categorization, retail competition has generally resulted in greater switching for industrial customers, while residential customers face significant switching and search costs and stay longer

Table 3: Markups and Demand Elasticity

	Gross Markup		Wholesale Markup	
	(1)	(2)	(3)	(4)
Share Residential 1994–1999	118.5***		146.9***	
	(32.83)		(40.20)	
Share Industrial 1994–1999		-87.39***		-122.0***
		(13.68)		(12.56)
Constant	-30.51**	37.74***	-49.87***	40.16***
	(14.84)	(11.88)	(17.01)	(8.123)
Year FE	Yes	Yes	Yes	Yes
Observations	733	733	603	603

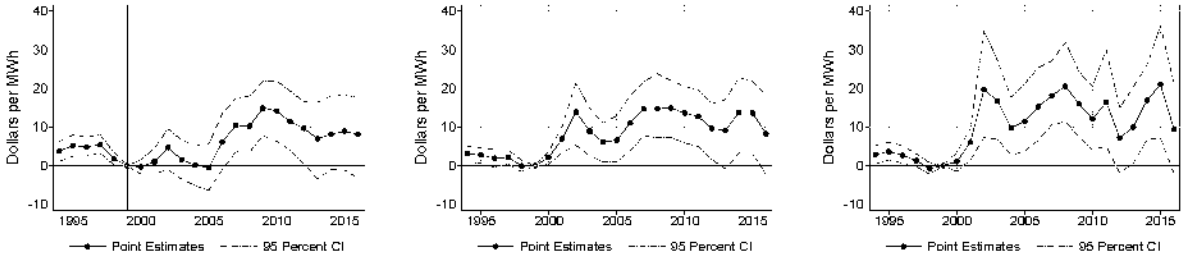
Notes: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. The dependent variable is the estimated effect on markups, which is regressed on the average share of residential and industrial customers from 1994 through 1999. Gross markup is retail price minus fuel cost. The sample contains observations at the utility level between 2006 and 2016. Coefficients are calculated using median regression with retail MWh sold in 1994 as weights.

with the incumbent provider (Hortaçsu et al., 2017). Importantly, the proportions of each group in a utility service area are arguably exogenous since for the majority of households and businesses electricity expenses are not significant enough to be a determinant factor in their location decisions. Consistent with this hypothesis, we find larger effects on markups for utilities that have a relatively higher share of residential customers or a lower share of industrial customers.

We examine the relationship between the estimated effect on markups and the share of residential or industrial customers in the area served by a given utility, which is strongly correlated with the elasticity of the demand faced by the utility. Table 3 presents results from regressing the estimated effect on markups on the share of residential or industrial customers in a utility's area, on average, from 1994 through 1999, using outcomes between 2006 and 2016. The sample is restricted to this period because this is when markups changed and we are interested in the mechanism behind this change. We use the shares from 1994 through 1999 because they are not affected by the prices charged by the utility in subsequent years. This provides a relatively clean proxy for the elasticity of the demand in that market. We analyze the relationship between markups and demand elasticity using both wholesale markups and gross markups, which are retail prices minus fuel costs, and find similar results for both measures. To mitigate the impact of outliers, we drop five utilities that do not have any residential customers, and we use median regressions.

Results in Table 3 indicate that utilities with a higher share of residential customers from 1994 to 1999, which is our proxy for more inelastic demand, had larger increases in markups. We also find that the share of industrial customers has a negative relationship with changes in markups, which would be expected when industrial customers exhibit more elastic demand.

Figure 10: Effects on Utility Rates by Customer Type



(a) Residential (b) Commercial (c) Industrial

Notes: Figure displays difference-in-differences matching estimates of changes in bundled service retail prices for deregulated utilities. These prices are determined by procurement costs for the utilities. Each deregulated utility is matched to a set of three control utilities based on 1994 characteristics. The estimated effects are indexed to 1999, which is the year prior to the first substantial deregulation measures. The dashed lines indicate 95 confidence intervals, which are constructed via subsampling.

These findings are consistent with deregulated firms exerting market power, charging higher markups in markets with more residential consumers and less elastic demand.

5.4 Heterogeneity in Effects by Customer Type

To further investigate the potential role of market power, we examine the effects of deregulation on different types of customers. We consider the three primary classes of electricity customers: residential, commercial, and industrial. To isolate the effect arising from the upstream market, we focus on bundled service rates available from local utilities. Though deregulation allowed for market-based prices, utilities that continued to operate in these retail markets were required to offer prices based on average variable costs. In effect, these utilities offered a price equal to the cost of procurement from the wholesale market, plus additional fees to cover distribution costs.

Observing similar changes in these rates across different classes of customers would be consistent with the exercise of market power in the wholesale market. Upstream generation facilities have little ability to price discriminate across different types of customers when selling to a utility, which bundles demand across customer types. If we observed instead that, for example, residential customers saw much greater increases in prices, we might infer that greater market power is exercised in downstream markets, where retailers can easily distinguish among types of customers. Alternatively, differential changes by customer type may also indicate special fees or subsidies provided as a result of deregulation to specific types of customers.

Figure 10 plots the difference-in-difference matching estimates of changes in utility retail prices by customer type. Overall, we find similar effects across different types of customers. All three types observe statistically significant increases in prices, with an average effect between 10 and 15 dollars per MWh from 2009 through 2016. Consistent with cost-based regulation

of these prices, these effects are very similar to the change in utility variable costs we report in panel (a) of Figure 5, which also average between 10 and 15 dollars per MWh over the same period. Overall, the fact that we observe similar increases in cost-based prices across customer types further suggests the important role upstream market power to increase prices in deregulated markets.[37]

One notable difference is that commercial and industrial customers realized price increases as early as 2001, whereas residential prices did not begin to increase until 2006. This is consistent with practice of implementing rate freezes along with deregulation, which fixed rates at pre-deregulation levels. Rate freezes were disproportionately targeted toward residential and small commercial customers. Thus, in many states, large commercial and industrial customers were immediately subject to the changes in variable costs realized by utilities in the aftermath of deregulation. We discuss the increase in utility variable costs and the rate freezes in more detail in Sections 4.3 and 6, respectively.

Consistent with our findings above, industrial and commercial customers are much more likely to switch away from the regulated utility rates. This transition was gradual, in contrast with the sudden increase in prices we observe.[38] See Figure A6 in the Appendix for estimated effects on the consumption of bundled service from the incumbent utility by customer type.

6 Delayed Effects of Deregulation

Price effects that result from deregulation may not be realized until many years after deregulation measures are enacted. Though many utilities were forced to legally separate from generation facilities abruptly, other measures were put in place that delayed actual changes to the structure of the market. For example, many utilities signed long-term procurement agreements with now independently operated generation facilities. These contracts effectively postponed the implementation of a competitive wholesale market, as much of the generation capacity was under long-term contracts. The possibility of delayed *effective deregulation* can explain why we observe larger price increases after some time.

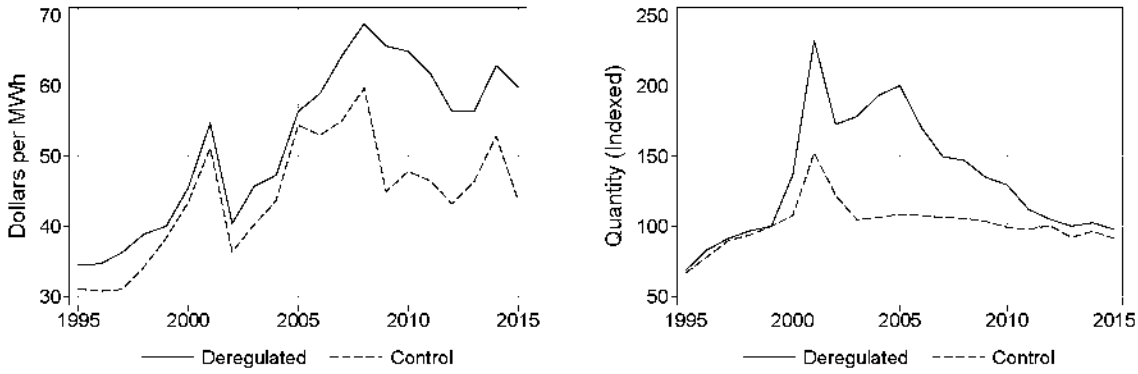
6.1 Long-Term Contracts

When deregulation measures were passed, most states imposed rate freezes or rate caps to guarantee low prices for consumers during the initial post-deregulation adjustment period. At the same time, utilities were vertically separated and signed long-term contracts with generators. The rates of these contracts were low because utilities were in good bargaining positions: there were no other significant buyers in the area and generators knew that their retail rates

[37] These results further suggest that the significant differences in markups across utilities shown in Table 3 are due to differential upstream behavior, as opposed to downstream price discrimination to different customer types.

[38] With the exception of Texas and Maine, which fully eliminated regulated rates for some utilities.

Figure 11: Contract Purchases



(a) Average Price: Largest Seller

(b) Average Quantity: Largest Seller

Notes: Figure plots mean characteristics for the largest buyer-seller relationships for each utility. We identify the largest seller to each utility by looking at aggregate MWh transacted for each seller-utility pair in each year. Panel (a) of the figure displays the average price paid to the largest sellers, and panel (b) displays the average quantity sold for that buyer-seller relationship. Quantities are based on MWh and are indexed to 100 for 1999 values. Values are plotted separately for utilities in deregulated states (solid lines) and control states (dashed lines).

were capped, so utilities could not pay more without incurring in losses. This situation changed around 2005, when both rate caps and contracts expired.[39] Two changes decreased utilities' bargaining position. First, utilities could pay more since they were allowed to increase rates if costs increased. Second, generators could sell to other buyers besides the utility, since wholesale centralized markets were starting to pick up (see Figure 14) and retail electricity providers had gained some market share.[40]

We examine the use and expiration of large long-term contracts in our data. Although we do not observe the exact expiration date of procurement contracts, we have annual data on transactions by seller for every utility, which allows us to explore how contracts evolved. Figure 11 presents characteristics of the contracts with the largest seller for each utility each year, separately by deregulated and control states. In panel (a), we see that initially prices in both groups moved roughly together, with utilities in restructured states paying only slightly more for energy. After 2005, the two series diverge, increasing substantially more in restructured states. Panel (b) on the right shows how the quantities purchased from the largest seller have evolved. The values are indexed to 100 in 1999. There is an early spike after 2000, when utilities purchased more energy after divesting a significant share of their power plants. The

[39] See the discussion of the case of Illinois in Section 6.2 for an illustration. Several states had similar timelines. For example, Maryland's rate freezes and rate caps began to expire in 2004, Delaware's price cap expired in 2006, Massachussetts' in 2004, Connecticut mandated a 10% reduction below 1996 rates for the period 2000-2003, and Virginia had price caps for the first six years after deregulation (expiring in 2006). All these states saw wholesale prices increasing around 2005.

[40] Section 5.1 shows how seller concentration remained fairly constant in the wholesale market during the last two decades, while buyer concentration decreased as retail competition became stronger.