Industry Dynamics with Social Learning: Evidence from Hydraulic Fracturing

Andrew Steck*

January 2022

Abstract

I model the interaction between dynamic decision making and social learning about new technologies in driving industry takeoff and productivity growth. Learning about the use of new technologies is an important factor in economic growth, but I demonstrate that anticipated social learning can lead to a free-riding dynamic in scenarios with high enough uncertainty. I consider the empirical setting of hydraulic fracturing in North Dakota, where firms learn about the optimal use of fracturing technology, in part due to detailed data published by regulators. The cumulative value of this learning process is a ceteris paribus 40% increase in ex-ante expected profitability. I model the impact of learning externalities on agents' decisions to drill shale oil wells, an optimal stopping problem. My estimates suggest that the social learning externality is too small to affect investment in later stages of industry development, after uncertainty is reduced. Conversely, I demonstrate that under conditions of higher uncertainty, anticipated social learning can lead to significantly lower industry investment and slower rates of industry learning. Under this scenario, I also demonstrate the potential for public tests of the technology to enhance welfare by leading to more investment and a higher learning rate.

*Department of Management, University of Toronto Mississauga and Rotman School of Management, University of Toronto; andrew.steck@utoronto.ca. Special thanks is owed to my dissertation committee: Allan Collard-Wexler, Chris Timmins, Jimmy Roberts, Steven Sexton, and Daniel Yi Xu. I have also benefited from feedback from numerous seminar and conference participants and discussants, especially Mark Agerton, Wesley Blundell, Thom Covert, Chuck Mason, and Peter Thompson. I am also grateful for conversations with Steve Slawson of Slawson Exploration and Rob Jacobs of Caird Energy on the industry. In addition, I thank Enverus (formerly DrillingInfo) for providing some of the data used in this work. I appreciatively acknowledge funding from the Duke University Graduate School and the Ottis Green Foundation, and computing resources from Compute Canada. Any errors are my own.

> *As a general rule, it seems likely that in the past 150 years the majority of important inventions, from steel converters to chemotherapy, from food canning to aspartame, have been used long before people understood* why *they worked, and systematic research in these areas was thus limited to ordered trial-and-error operations.* – Mokyr (1992)

1 Introduction

Much technological change occurs not in the laboratory, but on the shop floor and in the field, through a process of tinkering and exchange of experience with a new technology. Mokyr's characterization is particularly apt for one of the most significant recent technologies to be developed for oil extraction: hydraulic fracturing. Improvements in the application of this technology have led to stunning increases in productivity. In the North Dakota Bakken shale, the median well in 2008 began producing at a rate of 419 barrels of oil per day; by 2015 the figure had increased more than 200% to 1,265. Engineers' continual experiments, rather than scientific breakthroughs, are responsible for this growth, and experimentation and productivity improvement both continue (Gold, 2014). This process of adoption, perfection, and dispersion of new technologies is understood to be a key component of economic growth (see, e.g. Lucas (1993)), but questions remain: how does the possibility of social learning affect investment decisions, and in turn, how does investment affect learning?

I study the process of technological change in North Dakota's hydraulic fracturing industry by building and estimating a dynamic structural model. The model features current knowledge as a dynamic state and generates endogenous investment decisions and learning behavior. The model explicitly accounts for a firm's incentive to wait and potentially benefit from its rivals' experimentation under social learning. I propose a rational expectations equilibrium, and use approximate dynamic programming techniques to estimate the equilibrium in an empirical application. Model estimates suggest that anticipated social learning does not act as a drag on investment as the industry's knowledge matures; however, in states with higher uncertainty, a free-riding dynamic can arise that leads to lower investment and slower learning.

My empirical application focuses on the shale oil industry in North Dakota's Bakken formation. The industry has witnessed rapid productivity growth in recent years as firms have improved their use of inputs. The process has had dramatic effects on global energy markets: U.S. domestic oil production has returned to peak levels not seen since the 1980s, prompting falls in crude oil prices and political unrest in oil-rich nations. It is also an industry that features social learning: competing firms sometimes cooperate on shared wells or use the same sub-contractors, and regulators often collect and publish detailed production information. I estimate the model using data on oil production and hydraulic fracturing inputs available through the North Dakota Industrial Commission (hereafter, NDIC). The Bakken is well-suited to this study: over 10,000 wells have been drilled over the last decade; large swings in oil prices provide identifying variation; the NDIC's data is available to industry participants, allowing for social learning; and detailed cost estimates are available for a subset of wells, allowing for the construction of a reasonable model of expected profits.

My results first confirm that learning has occurred, as firms changed inputs to improve profits. This process has lead to an increase in expected profits of 40% over 10 years. I then construct and estimate a dynamic model that includes industry knowledge as a state. The estimates from this model suggest that anticipated social learning has a negligible impact on investment decisions. That is, in the estimated information states, the option value of anticipated learning from rivals on a single well is small, despite the overaching impact of industry learning on profits. Counterfactual simulations then study the free-riding effect that anticipated learning has in less-certain information states – such as those that might corresond to firms possessing limited attention or the beginning stages of learning about a new shale formation, or about a new technology in another industry. In these states, lower investment can lead to slower learning in a feedback loop. Finally, I consider the possible role for public test wells in jump-starting the industry's learning; these results demonstrate

that public tests of a new technology can be welfare enhancing by increasing learning and investment.

The remainder of this paper proceeds as follows. Below I summarize this paper's contribution to the literature, and then turn in Section 2 to an overview of the industry and relevant institutional details. Section 3 describes and summarizes my data sources, then provides empirical evidence of learning. I outline the structural model in Section 4, and Section 5 details empirical choices and estimation results. I conduct counterfactual analyses in Section 6, and conclude in Section 7.

Related Literature

This paper contributes to the literature that models learning in strategic and social environments, as well as that which studies the economics of the oil and gas industry.

The importance of learning to economic growth has long been recognized. As an example, Mokyr (1992) describes the gains of new knowledge or technologies as a true "free lunch", and Lucas (1993) calls growth, attributable in large part to learning, "a miracle". One strand of the economics literature, exemplified in Benkard (2000), has focused on "learning by doing" or the ability to reduce costs and inputs over repeated instances of production: through trial and error, the marginal cost of producing the thousandth widget is some fraction of that for the first widget. Complementary studies such as Griliches (1957), Foster and Rosenzweig (1995) and Conley and Udry (2010) have considered the adoption of new technology and social learning in agricultural settings. This paper is closer to the latter strand, modeling social learning about a new technology, though my empirical setting is oil drilling and hydraulic fracturing.

Oil firms, which I also refer to as operators following industry convention, face a complex dynamic problem: drilling a well in the current period will yield uncertain resources in future periods, which can then be sold for uncertain prices.[1] This com-

[1]Drilling and fracturing technically refer to different stages of the production process. As I focus only on fractured Bakken wells (which were all drilled prior to fracturing), I use the terms

bination of irretrievable up-front investment and uncertain future payoffs gives rise to real option value. Kellogg (2014) demonstrated the behavior of Texas oil drillers in the 1990s in response to changes in oil price volatility was consistent with a real options framework. His setting featured a constant (distribution of) underlying productivity, so firms' option value was due to possible changes in the price of oil or cost of drilling.

In the more modern setting of the hydraulic fracturing boom, operators have enjoyed rapid productivity growth as they have learned about the optimal use of the new technology. Covert (2015) documents this process in the North Dakota Bakken shale formation, providing evidence of increasing productivity over time. He also finds that operators place greater weight on information from their own wells compared to that published by a regulator on others' wells. In related work, Fetter, Steck, Timmins and Wrenn (2017) study learning about hydraulic fracturing fluids, using a regulatory change in chemical disclosure laws; their results suggest that firms exploit disclosure to learn from competitors, and that the knowledge is economically valuable.[2]

Both Covert and Fetter et al. focus on social learning's effect on the firm's decision of *how* to fracture a well, taking the decision to drill a well as given. That is, they do not model the extensive margin of *whether* to drill and fracture a new well. A rich literature in empirical industrial organization, e.g. Ryan (2012) and Collard-Wexler (2013), has demonstrated the importance of accounting for these investment and entry choices on industry and welfare outcomes. This paper bridges that gap by jointly modeling social learning and drilling decisions.

A related work, Hodgson (2021) models investment decisions and information externalities in the historical context of drilling for oil on the UK's continental shelf. Investigating a very similar economic issue to that in this paper, his model captures the information externality from *where* operators choose to drill in a conventional

interchangably in this paper.

[2]Other forms of learning have been studied in this industry, such as geographic learning in Hendricks and Porter (1996) and Levitt (2009), and inter-firm relationships in Kellogg (2011). This paper follows Covert (2015) in focusing on learning about the production function.

resource play when the drilling technology is not changing, whereas I focus on the information externality from *how* operators choose to fracture and so model technological progress explicitly. In a similar vein, Agerton (2020) studies drilling in Louisiana's Haynesville shale and finds in his context that the choice of *where* firms drill is more important for explaining productivity improvements than than *how* firms fracture.

This study is also related to two other recent works on strategic models of learning. Doraszelski, Lewis and Pakes (2018) model agents learning about competitors' play and equilibrium strategies in a repeated game. Jeon (2020) studies implications on shipping investment of having agents learn about demand shock parameters. Her model is similar in spirit to a strand of macroeconomic literature that models either agents or policymakers as having Bayesian beliefs about exogenous macroeconomic fundamentals, such as Cogley and Sargent (2005) or Orlik and Veldkamp (2014). In contrast to all of these studies, *the arrival of new information in my model is endogenous to current knowledge, as investment decisions depend on the current information state.* This significantly increases the difficulty of estimation in the present study, but represents a unique contribution.

2 Industry Overview

In order to inform modeling choices and data requirements, this section provides some institutional background on oil production, hydraulic fracturing, and the Bakken shale formation.

Producing oil or gas from drilling a well is different than producing widgets in a factory. Firms pay large sunk costs to drill each well (on the order of $10m to drill and fracture), and negligible marginal costs of maintenance and pumping. With low marginal costs, economic production can last for many years, but decreases exponentially over time from the rate of initial production (hereafter, IP). Further, even with the best seismic imaging technologies, there is significant variation in the realized productivity of wells - the largest and most experienced companies drill "dry holes" along with "gushers". As discussed in Anderson et al. (2014), firms also face physical

constraints that limit the ability to control production from an active well. The implication is that an operator's key decision in determining future production is *when* to drill new wells.

Hydraulic fracturing refers to to the process of injecting water and chemicals (fracturing fluid) and proppant, at high pressures, into shale or other low-porosity rock formations, in order to access trapped hydrocarbon molecules. The technology has been around in some form since the 1940s, but enjoyed rapid growth and development beginning in Texas the late 1990s. It has since spread widely, revitalizing the U.S. oil and gas industry, and upending global energy markets. Its use has a few key features that differentiate it from conventional oil and gas drilling. First, there is significantly less geological uncertainty - for example in the Bakken virtually every well finds and produces some oil. Second, as fracturing is used in rock formations with very low porosity, there is essentially no common pool problem. Third, how the well is fractured plays a large role in determining productivity.[3] The fracture is key to unlocking the valuable hydrocarbons, but done incorrectly can also damage the resevoir. Therefore a second key decision in oil production from a hydraulicly fractured well is *how* to fracture the well. As the technology in its modern form is relatively new, firms are still experimenting and learning about its optimal use.

In fact, the development of hydraulic fracturing as it is known today is a result of just such experimentation. Employees at Mitchell Energy in the 1990s were experimenting with gas wells in Texas' Barnett Shale. They were trying to reduce costs by fracturing wells with a watered down fracture solution, when they discovered that fracturing with mostly inexpensive water (slickwater) worked *better* than the expensive gels (Zuckerman, 2013) that predominated at the time. Gold (2014) describes how even before Mitchell's experiments, similar learning in Texas sandstones occurred

[3]When drilling and fracturing a well, an operator chooses the configuration of key input variables including: the lateral wellbore length, the number of fracture stages, the amount and type of proppant, the amount and composition of the fracturing fluid, and the injection pressure and rate. All of these choices combine to determine the quality of the fracture. The fracture then interacts with local geology to determine oil production.

by accident at another firm; engineers injected more water than intended due to a broken gauge, and the wells turned out to produce surprisingly well. The engineers published a paper and served as inspiration for Mitchell's later efforts. Gold also describes the attitude of some of those early innovators:[4]

> "Why it works is still generally unknown," Walker wrote. Not that this mattered to Walker. Engineers are problem solvers. If the wells were cheaper and gas production better, problem solved. A later generation of geologists and engineers could worry about why. They were making better wells and improving their company's bottom line.

This experimentation has continued through the present, as operators have tried wells with longer horizontal wellbores, denser fracture stages, varying amounts and types of proppant, and varying amounts and compositions of fluid. As some changes have proven profitable, the average well has used higher amounts of fluid and proppant per foot of wellbore, and more fracture stages (EIA, 2016). The production function is high-dimensional and complex enough that the learning process is ongoing at the time of this writing.

This paper's empirical application focuses on drilling in North Dakota's Bakken shale. The Bakken has seen more than 10,000 wells drilled over the last decade, produces predominantly oil, and has been a big contributor to a rapid rise in U.S. crude oil production. Hydrocarbon production in the Bakken is regulated by the North Dakota Industrial Commission, hereafter NDIC. The NDIC collects and publishes well-level data on inputs and outputs, enabling firms to observe results from competitors' wells when deciding how to fracture their own.

3 Data and Summary

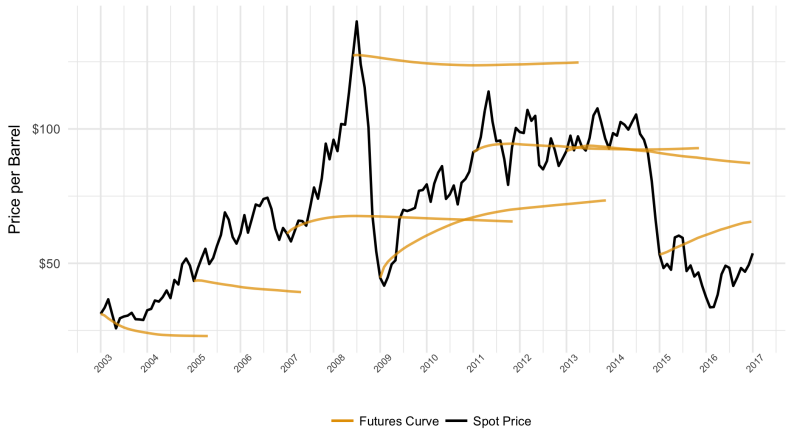
In this section I describe the data used in this study and discuss trends in key variables. To estimate my model, I require: oil prices and futures to determine revenue expectations; well-level production, characteristics, and fracturing inputs to estimate a decline curve and production function; and drilling and fracturing costs to estimate a function relating inputs to costs.

[4] The attitude described is reminiscent of many examples given in Mokyr (1992) where scientific understanding follows technological innovation, rather than vice-versa as is commonly presumed.

3.1 Oil Prices

Together with monthly production data, I use monthly oil prices to calculate operator revenue. I gather prices from the U.S. industry's benchmark futures curve, known as West Texas Intermediate (WTI) via Bloomberg. I collect prices along the futures curve up to 60 months out to capture the industry's expectations of medium-term price movements.[5] Figure 1 plots the price of crude and some illustrative curves from 2003 through 2016. The figure displays a few features of interest. First, the oil market has experienced two drastic price collapses in recent years, the first of which was coincident with the financial crisis. Second, market expectations of future prices can run the gamut from falling to stable to rising, but tend to reflect the recent past.

Figure 1: Monthly Oil Prices and Futures



Notes: Prices shown are for West Texas Intermediate (WTI) crude oil, and are taken from Bloomberg. The futures curves are drawn using copntemporaneous futures contracts, up to 60 months out.

3.2 Hydraulic Fracturing Inputs

To estimate learning over the production function, it is necessary to see operators' input choices. The NDIC publicly provides data on fracturing inputs, but housed in pdf documents of scanned images that are not machine-readable. I therefore supplement these pdfs with a pull from the NDIC's data server.[6] Because these data are

[5] Conversations with industry members suggest this is reasonable: firms use futures to inform expectations, and oftentimes to hedge price risk.

[6] An earlier version of this study instead used similar data provided by what was then known as DrillingInfo's Engineering Feed. Today the company is known as Enverus and I believe it has a different name for this data product.

sometimes missing or incorrectly transcribed, I supplement them with manual entries from the NDIC's pdf files, and from FracFocus, where available.[7]

Table 1 provides details on how those input variables have changed over time. From the first three columns, we see that fractures have used significantly larger quantities of fracturing fluid per foot over the last decade, and that the variation in these inputs has likewise increased. The second three columns show similar patterns in the use of proppant per foot.

Table 1: Fracturing Inputs by Year

	Fluid (gallons / ft)			Proppant (lbs / ft)		
	Median	Mean	Std. Dev.	Median	Mean	Std. Dev.
2005	83	80	41	55	60	48
2006	87	109	102	51	94	202
2007	96	122	109	72	84	82
2008	132	185	143	93	110	60
2009	175	210	153	96	134	92
2010	276	292	131	197	207	95
2011	289	293	114	220	229	122
2012	293	300	121	236	262	188
2013	317	350	229	245	294	203
2014	382	446	264	328	415	278
2015	414	504	289	384	478	292
2016	626	737	472	574	669	393

Data is taken from the NDIC's internal database on well stimulations. Missing and inaccurate data was corrected by hand using the NDIC's Completion Reports, housed in the NDIC's Wellfile pdfs.

I also collect administrative data on completions, monthly production and other well characteristics from the NDIC. Further information on this data and a table of summary statistics are available in Appendix A.

3.3 Well Costs

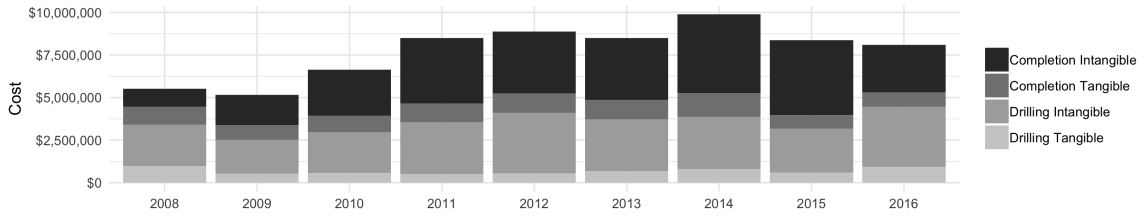
Operators are motivated by profit rather than production per se, so it is essential that my model include costs. I am able to obtain detailed ex ante cost estimates known as Authorization for Expenditures (AFEs) for some wells. These estimates

[7] In some instances, wells are re-fractured after a period of production, or if there was an issue in the original stimulation. I am able to observe re-fractures by matching well API numbers, and do not include them in my analysis.

are generated by operator engineers, and represent the operators' expected expenses for drilling and completing a given well. I am able to gather AFEs only in particular circumstances, which occur for roughly 400 of the wells in my sample, and some of these wells cannot be matched to the drilling, production and input datasets by lease and wellname. Appendix E provides more details.

Figure 2 shows the evolution of average well costs from the AFE dataset from 2008 to 2016, with costs broken down into four components: tangible drilling, intangible drilling, tangible completion, and intangible completion. The figure shows that costs have generally been increasing, before decreasing in 2015. Looking more closely at the subcomponents, the chart shows that most of the change has come from completion costs. Completion costs have become a larger fraction of total well costs, rising from roughly 40% to 60% between 2008 and 2014.

Figure 2: Estimated Well Costs



Notes: Average well costs per year are shown, for wells in the AFE subsample (see text for details).

3.4 Evidence of Learning

Section 2 argued that learning is an important feature of the oil shale industry. In this section, I provide evidence for, and an overview of, recent learning in the industry.

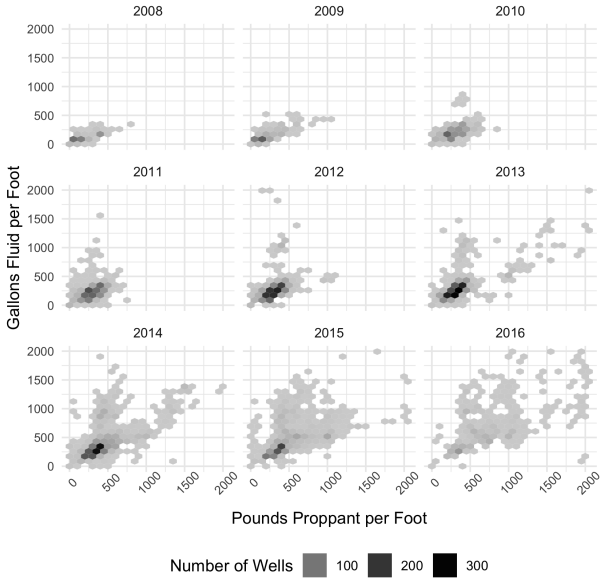
I provide three pieces of empirical evidence that learning is indeed occurring: first, I show the evolution of well configurations and input choices; I then discuss how this learning is consistent with profit maximization; third, I use regressions to suggest that operators learn from individual wells.

As the first evidence for operator learning, I discuss how the industry's well configurations and input choices have changed. Two key inputs in a hydraulic fracture are the amounts of proppant and fluid. Figure 3 plots yearly proppant and fluid use over the nine years ending in 2016. Both input variables are normalized into per-foot

terms, by dividing by the length of the horizontal wellbore. Three facts are visible from this figure: first, there is wide variation in the use of these inputs; second, the number of wells increased over the measured timespan, with a falloff in 2015 (as oil prices fell); third, there is a clear upward and rightward trend, indicating that operators were pursuing "bigger" fractures with more proppant and fluid. For example, a modal fracture in 2014 would have been one of the biggest fractures in 2008.

As operators began implementing larger fracture jobs, these new wells became important datapoints for operators trying to understand the fracturing production function. This can be seen from Figure 3, where the observed frontier of fracture intensity shifts up and to the right over time. Operators in later years had data on and experience with larger fracture jobs that they had lacked in earlier years.

Figure 3: Proppant and Fluid Use, by Year

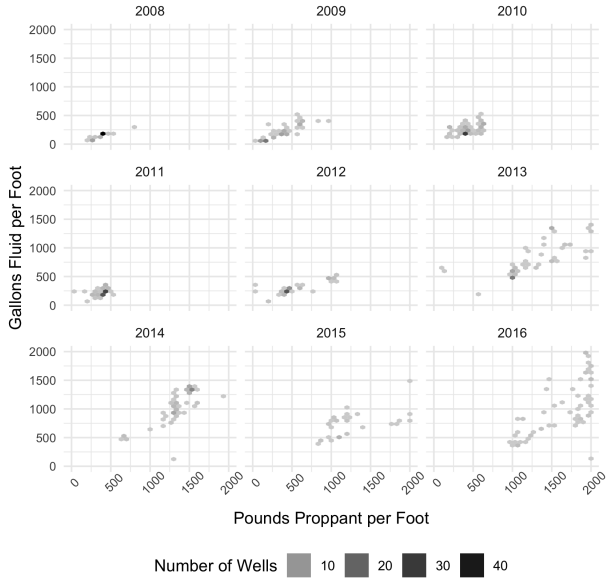


Notes: Each figure represents a heatmap, with a darker shade indicating more wells. Both input variables are transformed by dividing by the length of the horizontal wellbore. Values are truncated at 2000.

This industry-wide pattern is similar to that found at the individual operator level. As an example, consider Figure 4, which charts the use of inputs by EOG Resources, one of the most prolific operators in the Bakken. The figure illustrates that EOG fractured its wells more or less predictably until 2012, when they began to incorporate

some larger fractures. Over the next two years they tried different configurations, but the trend was to use more proppant and more fluid. EOG's smallest fracture in 2016 is larger than anything from 2012 or before. This pattern suggests that EOG was learning about (and experimenting with) the optimal fracturing configuration.

Figure 4: EOG Resources Proppant and Fluid Use, by Year

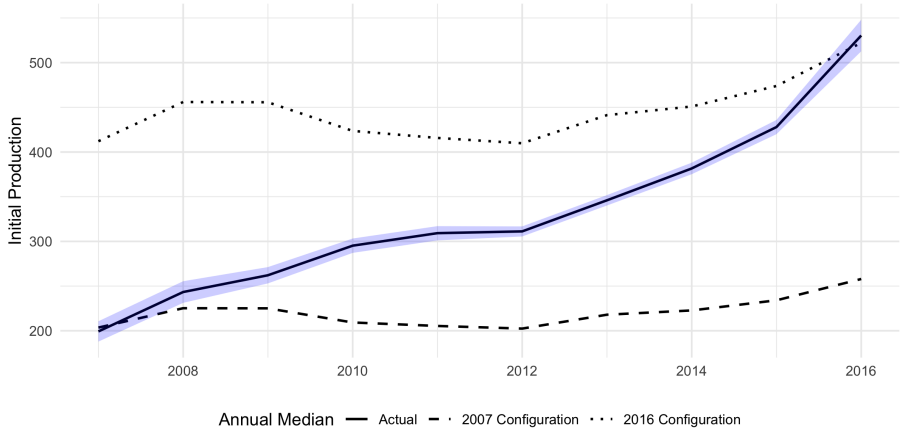


Notes: Each figure represents a heatmap, with a darker shade indicating more wells. Both input variables are transformed by dividing by the length of the horizontal wellbore. Values are truncated at 2000.

Next, I show that the changes discussed above are consistent with learning and profit-maximization. Rows 3-6 of Table 9 show that 6 Month IP per well increased significantly from 2005 to 2016. Figure 5 examines this evolution more closely, using a regression and counterfactual predicted values. The figure shows predicted median initial production by year for three scenarios. The first uses actual inputs and estimated location fixed effects, while the other two use median inputs for 2007 and 2016 along with estimated location fixed effects. Thus, the counterfactuals are created by predicting 6 Month IP if each well had been drilled with 2007 and 2016 median configurations, but in the same location and at the same time as it was drilled in reality. Details on the process of computing the counterfactual can be found in Appendix D. The chart shows a sizable gap between predicted production from the 2007 and 2016

configurations, which implies that operators accumulated valuable knowledge over those years. It also shows that most of the increase in production has been due to changes in well configurations, rather than changes in geology, i.e. firms are drilling wells in a more productive manner, not simply drilling in more productive locations.

Figure 5: 6 Month IP, Actual and Counterfactual



Notes: The solid line plots annual median 6 Month IP in barrels of oil per day; the dashed and dotted lines show median counterfactual 6 Month IPs if the wells were instead drilled with the median 2007 and 2016 configurations. See text and Appendix D for details.

The discussion above shows that operators have improved their output, but does not consider costs and profitability. As a simple measure of profitability I consider the ratio of cost to 6 Month IP.[8,9] Figure 6 shows the evolution of predicted cost/IP with two counterfactuals. The solid line shows the median of actual predicted cost/IP by year, while the two dotted lines plot median predicted cost/predicted IP if the median 2007 and 2016 configurations had been used. Calculation details can be found in Appendix D. The figure shows that the 2016 predicted cost/IP ratio is better than that of 2007, and that the industry has generally been improving its estimated cost/IP ratio over that timeframe.[10] This demonstrates that operators have made

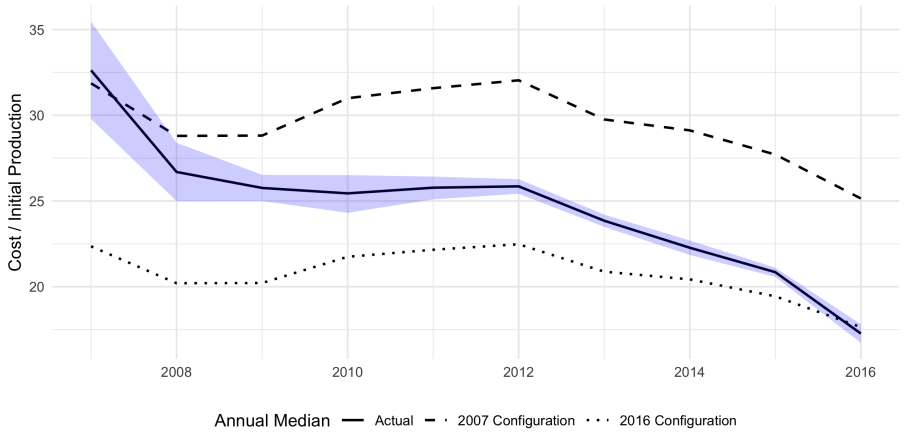
[8]Operators cannot control the commodity price of oil (and therefore revenues), but do have some control over their costs and levels of initial production.

[9]An alternative measure of profitability is the break-even price of oil; this approach requires many modeling assumptions and so I do not pursue it here. However, the decrease found in my measure is similar to the decrease in break-even prices found in industry analyses such as that at: http://digital.ogfj.com/ogfj/201706?pg=17pg17.

[10]Figures 5 and 6 are consistent with the results shown in Figure 9 of Covert (2015), where he

improvements on the profit margins they can control, and suggests that the industry's configuration changes are aimed at increasing profits, not simply production.

Figure 6: Cost/6 Month IP, Actual and Counterfactual



Notes: The solid line plots annual median 6 Month IP / predicted costs in barrels of oil per day per $ thousand; the dashed and dotted lines show median counterfactual 6 Month IPs / predicted costs if the wells were instead drilled with the median 2007 and 2016 configurations. See text and Appendix D for details.

As a final piece of evidence of operator learning, I estimate a simple regression designed to test whether individual wells have a measurable effect on the industry's learning. The procedure is as follows. For each well i in my dataset, I calculate the median inputs and outputs of all wells completed in the 12 months before well i's completion, and separately, of all wells in 6 to 18 month window after well i's completion.[11] I next create an indicator denoting whether well i is an "outlier" well. I define an outlier for cutoff k as a well with both input and output levels greater than k standard deviations above the pre-distribution medians. I next calculate the difference in post- and pre- median inputs, and then regress this difference on the outlier indicator and an operator fixed-effect.

The results of this exercise, using gallons of fracturing fluid per foot, are shown the second and third columns of Table 2, with standard errors clustered at the operator

performs a more flexible calculation in a similar spirit.

[11] I begin the 'after' window after a 180 day lag, as the majority of wells choose to have their data kept confidential for a 180-day period. Repeating this exercise for the 12 months immediately following each well's completion yields similar results, with coefficients of a slightly lower magnitude.

level. The table indicates that larger outliers predict greater increases in input usage. The coefficients at higher cutoff levels are also estimated to be statistically significant from zero. While these regressions should not be interpreted as causal, the pattern is suggestive: wells that make use of bigger fractures and produce more output than those in the recent past tend to be followed by more wells with bigger fractures.

One potential concern about this result is that the regression is just picking up a general time-trend in the industry of a conducting bigger fractures over time. As a robustness check of this result, I conduct a set of placebo regressions: I randomly assign an indicator to each well and treat it as the independent variable (i.e. placebo outlier)in a set of similar regressions to that above. I ensure that the number of placebo outliers matches the number of true outliers. I repeat this exercise for 1,000 bootstrap iterations, where in each iteration the independent variable is separately randomly assigned. The results of this exercise are shown in the last two columns of Table 2. The coefficients for the placebo test do not show a pattern, and are not statistically distinguishable from zero. This result argues against the possibility that the coefficients estimated in the second column are only picking up a time trend.

Table 2: Fluid per Foot Outlier Regressions

	Outlier		Placebo	
k	Coefficient	p	Coefficient	p
0.0	13.110	0.160	−0.032	0.984
0.5	21.808	0.034	−0.112	0.959
1.0	23.176	0.006	0.127	0.968
1.5	29.450	0.016	−0.086	0.986
2.0	37.321	0.000	−0.107	0.988
2.5	43.721	0.000	−0.273	0.975

Notes: each coefficient represents a separate regression, where the dependent variable is the change in 12 month pre- and post- fluid per foot medians, and the independent variables are an indicator for outlier wells and operator fixed effects. The "Placebo" specification represents a boostrapped regression, where in each iteration the indicators are randomly assigned, with the same number of positives as in the actual "Outlier" case.

In Appendix B, I argue against changes in $q(x)$, $c(x)$, or p explaining the observed trend in $x^*(p, \Gamma)$; I conclude that the most plausible explanation is a change in indus-

try knowledge Γ, or learning. This finding motivates the model of operator learning over optimal input use that follows.

4 Model

This section provides an overview of the industry model of social learning. First, I discuss the profit function from drilling a well, and the role of information in determining input choices and profits. I then outline a Bayesian learning process for incorporating new information, and the firm's dynamic problem. Finally, I turn to the industry-wide dynamics that arise from the firm-level model and specify the equilibrium concept that will be used.

I begin with a brief overview: each well i is a firm that faces a two-part decision.[12] The first decision is an optimal stopping problem of when to drill the well: making the investment of drilling a well enables oil to be produced and sold; however, there is also option value in waiting, as oil prices might rise or new knowledge might arrive. The second decision is how to drill and stimulate the well given the choice to drill has been made – this is referred to as the "static" problem below.[13] These two decisions, along with shocks, determine IP; subsequent production follows a deterministic decline curve. Both decisions are made with the goal of maximizing expected discounted profits: profits are determined from production via the prevailing oil price and drilling costs, both taken as given.

The key feature of the model is that firms are uncertain about the optimal method of drilling wells. Information on drilled wells is published by the regulator, allowing

[12] Modeling each well as an independent firm follows Kellogg (2014), but is potentially less realistic in my context, where firms are actively learning about how to increase productivity. It amounts to an assumption that firms with multiple potential wells treat them independently. As I demonstrate in Appendix C, learning in the industry appears to be operating primarily at the inter-firm rather than intra-firm level, which suggests this assumption is reasonable.

[13] As discussed above, Anderson et al. (2014) reformulated the benchmark model of Hotelling (1931), demonstrating that due to technological limitations on the production function, an oil firm's key production decision is *when* to drill a new well rather than *how much* oil to produce from an existing well. That is, the operator's problem is best construed as when to "tap another keg" instead of how to adjust the flow from already-tapped kegs. This is confirmed empirically in Newell et al. (2016) and Newell and Prest (2017), who find that new drilling is the primary margin of response to increased prices in natural gas and oil production, respectively.

firms to learn in a Bayesian manner from other wells. The industry thus has common knowledge in every period, knowledge that affects both stages of the two-stage decision described above. Having decided to drill, more knowledge might change how the firm drills and its expected profits. In turn, the firm anticipates the possibility of learning, and so may have additional option value from waiting. Modeling details on knowledge and learning follow a discussion of the static problem below.

4.1 Static Overview

I begin with the static problem of a firm i that has decided to drill in month t. The quantity of oil produced by well i is a linear function of a vector of inputs x and an iid shock:

$$q_i = \beta_0 + x_i'\beta_x + \epsilon_i^q \,. \tag{1}$$

Firms are price-takers in the global market, and receive price p_t per barrel of oil produced. The cost function is also linear in inputs:

$$c_i = \omega_0 + x_i'\omega_x + \epsilon_i^c \,, \tag{2}$$

where ϵ^c is an iid cost shock.[14]

With these functions defined, I then write profits:

$$\pi(x_i, p_t, \phi_i) = (p_t q_i(x_i) - c_i(x_i))\mathbb{1}\{\phi_i > g(x_i)\} \,. \tag{3}$$

Profits take the usual form, multiplied by a binary term governing feasible well scale. That is, if the realization ϕ_i is less than $g(x_i)$, then the well is a failure and no profits are realized.[15]

Going forward, I limit x_i to be a scalar: gallons of fracturing fluid per foot of wellbore.[16] I set $g(x)$ as the identity function. I also define the vector $X_i \equiv [1, x_i]$ for

[14] My cost data is not dense enough to reliably estimate changes in drilling and completion costs over time, so I treat $c(x)$ as fixed. An alternate possibility would be to model drilling costs as a function of oil prices (and consequently drilling demand).

[15] The well-scale shock could take alternate forms, but I use the simplest modeling choice that produces the desired shape in the expected profit curve: the important thing is that there is an input cutoff above which firms expect to realize lower profits. This shape could also be achieved with a penalty function that added extra costs, or reduced output or revenues.

[16] The fracturing design has many dimensions, and in principal I could account for more than one, such as pounds of proppant, number of stages, pressure, etc. In practice though, I quickly run up against the "curse of dimensionality" in the dynamic game, as the dynamic state has two dimensions for every input (including a constant). I select fluid intensity, but in my sample it has a

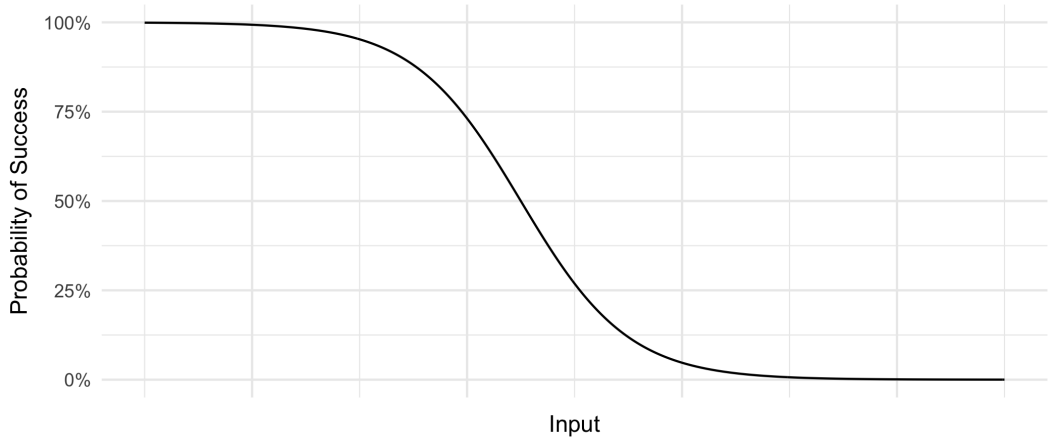
notational convenience.

I model the scale shock ϕ as following a logit distribution with parameters γ^*:

$$\Pr(\phi > x_i) = \frac{1}{1 + \exp(-X_i'\gamma^*)}\,. \tag{4}$$

An example of the shape of $\Pr(\phi > x)$ is shown in Figure 7. Further, ϕ is a latent variable in that it is not directly observable. Rather, it is only observed through the binary outcome $A_i \equiv \mathbb{1}\{\phi_i > x_i\}$. So each well is observed as either a 'success' ($A_i = 1$) or 'failure' ($A_i = 0$).

Figure 7: Example of Expected Probability of Success



Notes: plot of an example curve from Equation 4, depicting the anticipated probability of realizing $\phi > x$.

This construction can be interpreted as a model of reservoir integrity: a more intense stimulation can yield greater output, but could also cause reservoir damage and reduce well productivity. For models of fluid damage to reservoirs, see e.g., Bahrami et al. (2012), Putthaworapoom et al. (2012), and Eveline et al. (2017) from the petroleum engineering literature. This simplified model fits sensibly with the dynamic of progress in hydraulic fracturing: the basic technology has been around since the 1960s, but the recent revolution has observed much higher fracture intensities as a result of firm experimentation. The model is also consistent with the finite fracture intensities observed in my sample: even when the marginal costs of additional fluid are low relative to expected revenue of additional oil produced, the additional

65.8% correlation with proppant intensity.

fluid increases the risk of reservoir damage.

Firms are uncertain about (and will learn about) the parameters γ governing the distribution of ϕ. I model their uncertainty as Bayesian, with multivariate normal priors: $\gamma \sim \mathcal{N}(\mu, \Sigma)$. For convenience, I will refer to the firm's prior or information state as $\Gamma \equiv (\mu, \Sigma)$. The dynamic state governing drilling decisions will be the combination of information and oil prices: $S \equiv (p, \Gamma)$.

A firm that has decided to drill in state S, chooses inputs x so as to maximize expected profits:

$$x^*(S) = \arg\max_x E\left[(pq(x) - c(x))\mathbb{1}\{\phi > x\}\right] \tag{5}$$

$$= \arg\max_x \left(pq(x) - c(x)\right)\left(\frac{1}{1+\exp(-X'\mu)}\right). \tag{6}$$

From these equations, linearity in $q(\cdot)$ and $c(\cdot)$, and concavity in $\Pr(\phi > \cdot)$, we can see that each state S maps to an optimal input choice $x^*(S)$. For convenience I will also define expected profits in a given state, assuming optimal input use, as:

$$\pi(S) = \left(pq(x^*(S)) - c(x^*(S))\right)\left(\frac{1}{1+\exp(-X^*(S)'\mu)}\right) \tag{7}$$

4.2 Learning

Firms are statistically savvy, and change their estimates of γ to incorporate new information. Their Bayesian updating process is more challenging than most that have been used in empirical economics because they only ever observe A_i and x_i and never ϕ_i directly. This complication leads to non-conjugate posterior probabilities, which quickly limits empirical tractibility. To simplify the analysis I assume that firms update from a normal prior to a normal posterior using 'moment-matching': essentially approximating a non-normal posterior with a normal posterior. This method, under various names, has been used and explored in a number of statistical fields (Powell and Ryzhov, 2012); the specific form developed here follows derivations in Jaakkola and Jordan (2000).

This method yields convenient formulas for updating prior beliefs after observing n wells, $\Gamma^n = (\mu^n, \Sigma^n)$ to posterior beliefs Γ^{n+1} upon observing information for well

$n+1$, (X_{n+1}, A_{n+1}):

$$\Sigma^{n+1} = \left[(\Sigma^n)^{-1} + 2\lambda(\eta^{n+1}) X_{n+1} X'_{n+1}\right]^{-1}, \tag{8}$$

$$\mu^{n+1} = \Sigma^{n+1}\left[(\Sigma^n)^{-1}\mu^n + \left(A_{n+1} - \frac{1}{2}\right) X_{n+1}\right], \tag{9}$$

where the quantity η^n is

$$\eta^n \equiv \sqrt{X'_n \Sigma^n X_n + X'_n \mu^n}\,, \tag{10}$$

and the function λ is defined

$$\lambda(\bullet) \equiv \frac{\tanh\left(\frac{\bullet}{2}\right)}{4\bullet}\,. \tag{11}$$

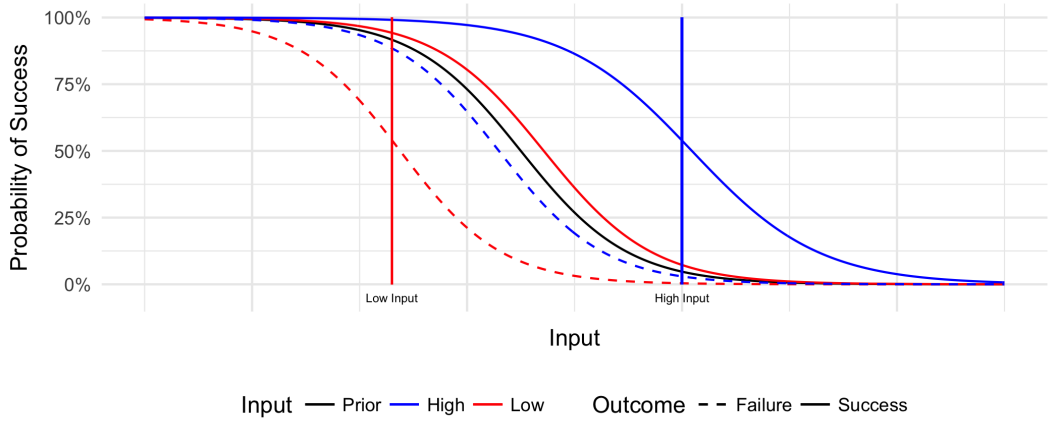
Several graphical illustrations of Equations 8 and 9 are shown in Figure 8. The prior expected probability of success is depicted as a black solid line, and four calculated posteriors are considered: whether the observed well is a success (solid line) or failure (dashed line); and whether the observed input is high (blue) or low (red). For example, when a high level of input is used, the prior expected probability of success is quite low. Thus if a failure is observed, the posterior is only shifted slightly to the left. Conversely, an observed success is quite surprising, and causes a large move in the posterior to the right. The opposite pattern is true for a low input well, when the prior expected probability of success is high: a success does little to change the posterior, but a failure moves the posterior significantly to the left.

This asymmetric learning is a useful feature of this model: it allows the value of new information to depend on its relative novelty. This matches the industry dynamic of paying attention (demonstrated in Section 3.4) to the success of the largest new stimulations, while learning little from the thousandth well stimulated in a tried and true fashion far below the frontier. Importantly, it also provides a nice model of how the frontier moves over time, as firms push the limits of feasibility incrementally; this will allow it to be useful in explaining the trends shown in Table 1 and Figure 3.

4.3 Dynamic Overview

Having outlined the firms' static decision and learning process, I turn to the dynamic problem of when to drill. This is modeled as an optimal stopping problem: in each

Figure 8: Examples of Learning



Notes: graphical example prior and posteriors; all curves depict the anticipated probability of realizing $\phi > x$ from Equation 4. Posteriors are calculated using Equations 8 - 9. The solid black curve is the prior. Red and blue curves correspond to posteriors from low and high input observations, respectively; solid curves indicate an observed success $A_{n+1} = 1$, and dashed curves indicate observed failure.

period, the firm can drill its well (entering the static problem described above, realizing shocks and profits), or wait until the next period when it will face the same decision, potentially in a more profitable state. Recall that the dynamic state governing expected profits is $S_t \equiv (p_t, \Gamma_t)$. State S in period t is made up of the current price of oil p and the current information state Γ, so the firm may have option value from changing oil prices and changing information. Once a well is drilled, it realizes shocks determining costs, revenues, and profits, and exits the game. I assume that potential wells are infinitely-lived, that drilled wells are replaced in the potential entrant pool each period (i.e. a constant number of potential wells each period), and that each potential well observes an iid normal profit shock ξ before deciding whether to drill.[17,18]

I recall the definition of expected profits in a state from Equation (7), drop time subscripts, and rewrite the firm's drilling problem in a Bellman setup:

$$V(S, \xi; \theta) = \max \{\pi(S) + \xi,\ \delta E[V(S', \xi; \theta)|S, \theta]\} \ . \tag{12}$$

[17]This assumption removes the issue of leasing decisions and lease expirations, which are potentially important. In practice, estimated drilling probabilities are high enough to suggest wells would be drilled before the expiration of standard leases.

[18]The shock ξ functions primarily as an empirical device to help the model rationalize observed drilling decisions. See discussion in Kellogg (2014), where stochasticity in $q(\cdot)$ plays a similar role.

$V(S)$ gives the value of an undrilled well in state S, and $\delta \in (0,1)$ is the shared discount factor. Normal parameters $\theta = (\mu_\xi, \sigma_\xi)$ govern the distribution of ξ, and as I show below, affect the expected value of waiting. Given this, a well is drilled if and only if current expected profits, $\pi(S) + \xi$, are greater than or equal to the discounted value of waiting, $\delta E[V(S', \xi; \theta)|S, \theta]$. Under some regularity conditions, this rule can be re-expressed in terms of a cut-off value of ξ: for each state S, there exists a cutoff $\xi^*(S)$ that governs drilling decisions. Wells with $\xi \geq \xi^*(S)$ drill, otherwise they wait.[19]

The final piece of the dynamic model is the expected value of waiting,

$$E[V(S', \xi; \theta)|S, \theta] \equiv \int_{S'} V(S', \xi; \theta) \Pr(S'|S, \theta)\,, \tag{13}$$

where $\Pr(S'|S, \theta)$ denotes the true probability of moving from state S to state S' in the next period. The state transition is composed of two parts: $\Pr(p'|p)$, and $\Pr(\Gamma'|\Gamma, p)$. Prices change exogenously, and Γ changes as wells are drilled, create new information, and that information is incorporated according to the Bayesian updating rules (8) - (9). A firm that waits may have a new expected profit in the next period – not only because of a change in price, but also because of new information. Written in this way, the challenge of an endogenous equilibrium becomes clear. The decision of whether to drill a well this period or not depends on the value of waiting $E[V(S', \xi; \theta)|S, \theta]$. In turn, the value of waiting depends in part on the possibility of learning new information (in S'), which depends on others' decisions to drill. I discuss this expectation term in further detail in Section 5.4, where I also describe how it is calculated empirically.

Given these components, I define a dynamic social learning equilibrium:

Definition. *An equilibrium is a state space $\mathcal{S}$, policy functions $\xi^*(S)$, and transition beliefs $\tilde{\Pr}(S'|S)$ such that $\forall S, S' \in \mathcal{S}$:*

1. *Drilling decisions follow the policy function $\xi^*(S)$, defined implicitly by*

$$\pi(S) + \xi^* = \delta E[V(S', \xi^*; \theta)],$$

[19] From Dixit and Pindyck (1994), the conditions are that the value of waiting for one period is monotonic in ξ for any S, and that the distribution of $\pi(S') + \xi$ has positive persistence.

so that a well is drilled if and only if $\xi > \xi^*(S)$; this cutoff represents the optimal drilling policy given beliefs $\tilde{\Pr}(S'|S)$. $\pi(S)$ is given by Equation (7) and $E[V(\cdot)]$ is defined according to Equation (13).

2. $\Pr(S'|S)$ *are defined by exogenous price transitions, and information states that evolve as new wells are drilled and observed. Information is incorporated according to Bayesian updating rules (8) - (9).*

3. *Expectations are rational, i.e. optimal behavior ensures that state transition beliefs are self-fulfilling, so* $\tilde{\Pr}(S'|S) = \Pr(S'|S)$.

Solving for this equilibrium empirically is computationally costly. I describe my approach in the following section.

5 Empirical Model and Estimates

This section describes how I take the above model to the data and presents estimates.

5.1 Static Profits

Wells produce oil over many periods. I follow industry practitioners in modeling production with a deterministic decline curve (known as an Arps model), so that IP is a sufficient statistic for production over the life of the well.

In addition to knowing future production, firms have expectations of future oil prices informed by the futures curve. Using the common industry annual discount factor $\delta = 0.9$, I can then write the revenue function as:

$$rev(S) = \psi \sum_{\tau=0}^{\bar{\tau}} \delta^{\tau} p_{\tau}^{f} q_{\tau}(x^*(S)) \,, \tag{14}$$

where $\psi = 0.7$ is the share of oil revenue accruing to the leaseholder, p_{τ}^{f} is the expected per-barrel price of oil in period τ and q_{τ} is production in period τ.[20] I set $\bar{\tau} = 240$ months, shorter than the industry's expectation of 540 months. However, the decline curve and discount rate are steep enough that total expected revenue is not very sensitive to increasing $\bar{\tau}$.

[20]The selected discount rate corresponds to an interest rate of 11.11%. The share of oil revenue not accruing to the well operator represents a typical royalty rate to the mineral rights lessor of 16.5%, state taxes of 11.5%, and a small marginal cost of 2% (Covert, 2015).

Because decline curves are deterministic, the expected revenue function can be expressed even more simply in terms of initial production.

$$rev(S) = \psi * 30 * p * h * IP(x^*(S)) * \sum_{\tau=0}^{\bar{\tau}} \delta^\tau \zeta_\tau \,, \tag{15}$$

where $IP(x^*(S))$ is initial production per foot of horizontal lateral, expressed in barrels / day, h is the length of the lateral, in feet, 30 is the number of days per month, and ζ_τ is per-day production in month τ as a fraction of IP.[21] Well lengths are taken to be exogenous, for now the later-year standard of $10,000$ feet. Finally, p is the "flattened" price incorporating information from futures contracts out to 60 months:

$$p = \frac{\sum_{\tau=0}^{\bar{\tau}} \delta^\tau p_\tau^f \zeta_\tau}{\sum_{\tau=0}^{\bar{\tau}} \delta^\tau \zeta_\tau} \,, \tag{16}$$

where p_τ^f is the futures price τ months in the future. For months after 60, I assume constant oil prices, i.e. $p_\tau^f = p_{60}^f$ for $\tau > 60$.[22]

So the only non-deterministic component of $rev(S)$ is $IP(x^*(S))$. Initial production per foot is given by a relationship of the following form:

$$IP(x_i) = \beta_0 + \beta_1 x_i + \epsilon_i^q \,, \tag{17}$$

where x_i is gallons of fracturing fluid used per foot.[23] Drilling and stimulation costs are given by:

$$c(x_i) = \omega_0 + \omega_1 x_i + \epsilon_i^c \,. \tag{18}$$

The simple production and cost coefficients $\hat{\beta}$ and $\hat{\omega}$ that will be used in Equation 7 are estimated by OLS and reported in Table 3.[24] The production estimates

[21] The decline curve ζ_τ is estimated non-parametrically in a first stage, using the well-month level production data: $\hat{\zeta}_\tau$ is found as the empirical mean of month τ production per day as a fraction of reported IP.

[22] The first two months in my sample (January and February 2006), only have futures prices out to 28 months. For those months I assume constant prices after 28 months.

[23] In principal, x could be a vector containing other stimulation choices such proppant, pressure, injection rate, stages, chemicals, etc.; in practice, however, the computational feasibility of the learning model will be limited by the dimensionality of x. I select gallons of fracturing fluid as the single input closest to a sufficient statistic for the fracture intensity.

[24] One might worry that x_i is endogenous if firms have any knowledge of ϵ_i^q. I argue that this is not a major concern in this setting. Fractures involve major logistical challenges, and are designed far in advance of drilling. Conversations with industry participants suggest that any last minute changes to inputs tend to be marginal.

demonstrate an increasing relationship between fracturing fluid intensity and initial production. Cost estimates are calculated on the subset of wells that can be matched to an AFE, and suggest an average well cost of $8.0 million.

Table 3: Production and Cost Regressions, for Dynamic Model

	6 month IP per Thousand Feet	Cost, $
Intercept	34.780^{***}	$7,551.934^{***}$
	(0.406)	(152.634)
Gallons Fluid per Foot	0.033^{***}	2.162^{***}
	(0.001)	(0.338)
Observations	11,328	421
Adjusted R^2	0.108	0.087

Notes: Production regression includes all wells in the sample with non-missing IP and input data. The cost regression includes wells in the AFE sample that could be matched to the input dataset, see Section 3.3 for details. *, **, and *** indicate statistical significance at the 10, 5, and 1 percent levels.

5.2 Price Transitions

I estimate $\Pr(p'|p)$ under the assumption that flattened prices are a martingale process with normal errors:

$$p' = p + \epsilon^p\,, \quad \epsilon^p \sim \mathcal{N}(0, \sigma^p)\,, \tag{19}$$

and recover a point estimate of $\sigma^p = 5.428$ with maximum likelihood.[25] For my dynamic application, I then form a matrix $\Pr(p'|p)$ by simulating 10,000 draws from each point on an evenly spaced grid with 15 nodes from $30 to $120.

5.3 Well Inputs and Learning

As in the abstract model, a firm that drills in state S calculates its optimal expected input $x^*(S)$ using Equation (5). However, as can be seen in Table 1 and Figure 3, there is significant variation in observed inputs, even in the same time period. To match this input dispersion, I assume that the actual input used is stochastic: $x_i(S) = x^*(S)\epsilon_i^x$, where ϵ_i^x is distributed $\epsilon_i^x \sim \text{LogNormal}(0, \sigma_x)$.[26] I estimate σ_x as

[25] Alquist and Kilian (2010) show that a non-change forecast for the spot price of oil outperforms many other forecast methods, including the use of the futures curve.

[26] Different wells drilled in similar areas at similar times, often even by the same firm, can have a wide range of stimulation configurations. Conversations with industry participants confirm that this variation is attributable to experimentation and iteration. This is outside the model in Section

the mean across months of observed within-month standard deviations of the log of observed inputs, and recover a value of $\sigma_x = 0.577$.

The industry updates priors Γ in the approximate Bayesian fashion described in Equations (8) - (9) to incorporate information from new wells. A new well observation is a pair (x_i, A_i). Input x_i is per-foot gallons of fracturing fluid, and outcome A_i is an indicator for whether the treatment was successfully completed in a timely fashion.[27]

Next I describe the process of estimating the information states $\{\Gamma\}_t$ using observed well data and the model. It is worth emphasizing that these information states are estimated outside the dynamic model; they are identified by the observed sequence of well inputs and outcomes, the optimal input rule (5), and updating equations (8) - (9). Intuitively, the level of inputs used and Equation (5) identify prior means $\{\mu\}$, and the rate of change in inputs and updating equations identify prior variances $\{\Sigma\}$. I provide more detail and results below.

First, recall that an information state is made up of a prior mean and covariance, $\Gamma = (\mu, \Sigma)$. Next, note that the Bayesian updating Equations (8) and (9) describe the calculation of Γ^{n+1} given Γ^n, x_{n+1}, and A_{n+1}. In other words, they form a function mapping a prior information state and information on a new well to a posterior information state, $\Gamma \times (x, A) \mapsto \Gamma'$

Now consider the sequence of observed well data $\{x_i, A_i\}_I$.[28] The mapping described above can be applied iteratively, so that an initial prior Γ^0 combines with a sequence of well data $\{x_i, A_i\}_I$ to yield a sequence of posteriors $\{\Gamma_i\}_I$. I use dates of well completions to translate this sequence of posteriors into the time periods of the model, $\{\Gamma^t(\Gamma^0)\}$, where my notation makes explicit the dependency on the original prior, and dependency on well data is implicit.

4.

[27]Specifically, $A_i = 1 \iff$ the treatment was completed in less than 75 days per fracture stage; this represents the 95th percentile of treatment days per stage, which has a heavily skewed distribution. See Appendix I for details.

[28]I observe A_i directly for all wells that are present in the FracFocus database (roughly 2012 and later). I calculate a fitted value of $\hat{A}_i$ using observed x_i and the econometrician's estimate of γ for earlier wells.

The next step is to observe that Equation (5) is a model-derived mapping from an information state Γ and price p to an optimal input x^*. I use this second mapping to generate a sequence of optimal inputs given initial priors (explicit), and observed well information and prices (implicit): $\{x_t^*(\Gamma^0)\}$.

Finally, I proceed to estimate priors by minimizing the sum squared distance from predicted inputs to observed inputs:

$$\hat{\Gamma}^0 = \arg\min_{\Gamma^0} \sum_t \left(\bar{x}_t - x_t^*(\Gamma^0)\right)^2 , \tag{20}$$

where $\bar{x}_t$ is the average input use in month t.[29] I perform this minimization using a grid search in Γ^0, and report estimates in Table 4.[30] The second and third columns of Table 4 present the 5th and 95th percentiles of $\hat{\Gamma^0}$ estimates calculated from 5,000 bootstrapped samples. Each bootstrap sample is constructed with a block bootstrap procedure, with wells drawn separately for each month, with replacement. The final column shows the corresponding information state in the final period of the data.

Table 4: Information State Estimates

	t = March 2006			t = March 2017
		Bootstrapped		
	Point Estimate	5th Percentile	95th Percentile	
$\hat{\mu}^t$	$\begin{bmatrix} 81.7 \\ -1.04 \end{bmatrix}$	$\begin{bmatrix} 38.4 \\ -1.04 \end{bmatrix}$	$\begin{bmatrix} 84.7 \\ -0.606 \end{bmatrix}$	$\begin{bmatrix} 13.9 \\ -0.015 \end{bmatrix}$
$diag(\hat{\Sigma}^t)$	$\begin{bmatrix} 0.249 \\ 4.2e^{-6} \end{bmatrix}$	$\begin{bmatrix} 0.028 \\ 3.3e^{-6} \end{bmatrix}$	$\begin{bmatrix} 4.61 \\ 2.67e^{-5} \end{bmatrix}$	$\begin{bmatrix} 3.5e^{-3} \\ 1.52e^{-8} \end{bmatrix}$

Notes: estimates of information states Γ^t, estimated via grid search. See text for details.

5.3.1 Value of Information

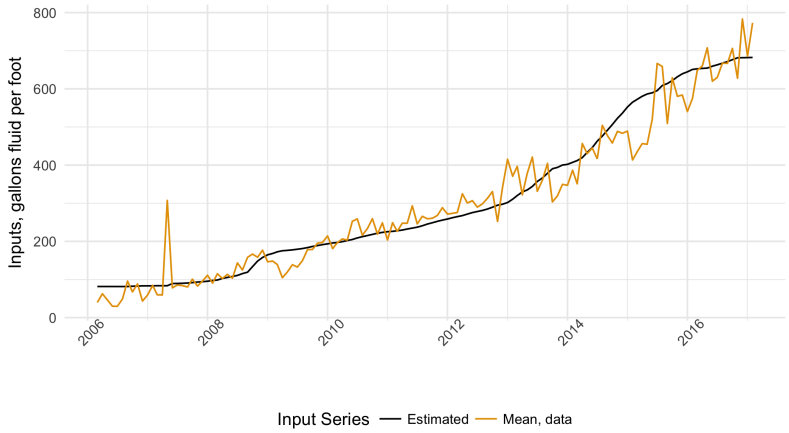
Figure 9 plots monthly mean inputs observed in my data, as well as the model's prediction using $\{\hat{x}_t^*(\hat{\Gamma}^0)\}$. The figure illustrates that the four estimated parameters in $\hat{\Gamma}^0$, combined with observed wells $\{x_i, A_i\}_I$, $\hat{\beta}$, $\hat{\omega}$, and Equations (8) and (9) produce

[29] I estimate priors trying to fit monthly averages because the number of wells drilled in each month varies significantly across the sample; an alternative objective function like $\arg\min \sum_i \left(x_i - x_i^*(\Gamma^0)\right)^2$ would underweight the low-input wells observed in the early months and the high input wells observed in the late months, and thus understate the learning process.

[30] I restrict Σ to have zeros in the off-diagonal in order to reduce the dimensionality of my dynamic state space. In practice this appears to be innocuous: when I allow non-zero off-diagonal elements, they are always estimated to be orders of magnitude smaller than Σ_{22}.

a coherent explanation of increasing fracture intensities.

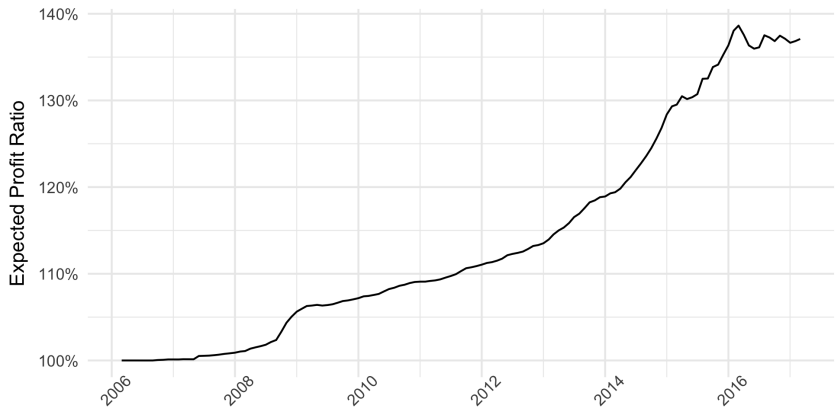
Figure 9: Model Predicted and Observed Inputs



Notes: observed means are calculated at the monthly level. Estimates plot the sequence of optimal inputs given estimated information states, $\{\hat{x}_t^*(\hat{\Gamma}^0)\}$. See text for details.

With an estimate of the industry's priors, I can also quantify the contribution of information to expected profits as per the model. Figure 10 illustrates the results of this calculation, where the graph is plotting the ratio $\dfrac{\pi(p_t, \hat{\Gamma}^t)}{\pi(p_t, \hat{\Gamma}^0)}$, i.e. the modeled expected profit in a given month divided by the modeled expected profit if no learning were to have taken place. The emphasizes that learning has value – the bigger wells that operators are learning to drill are associated with higher expected profits.

Figure 10: Ratio of Expected Profits: Learning to No Learning

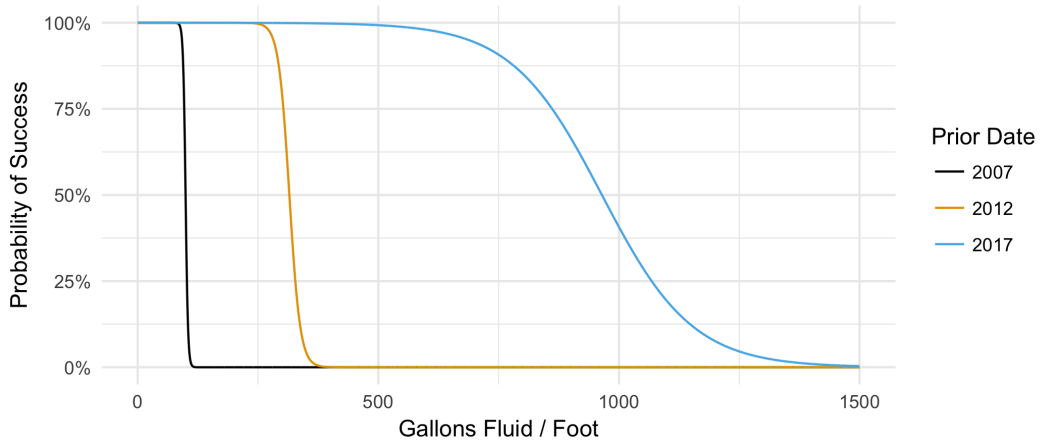


Notes: plot of $\pi(p_t, \hat{\Gamma}^t)/\pi(p_t, \hat{\Gamma}^0)$ over time, illustrating the gains from $\hat{\Gamma}^t$ relative to $\hat{\Gamma}^0$.

A graphic illustration of success probabilities with $\hat{\Gamma}^t$, for $t \in \{$ January 2007,

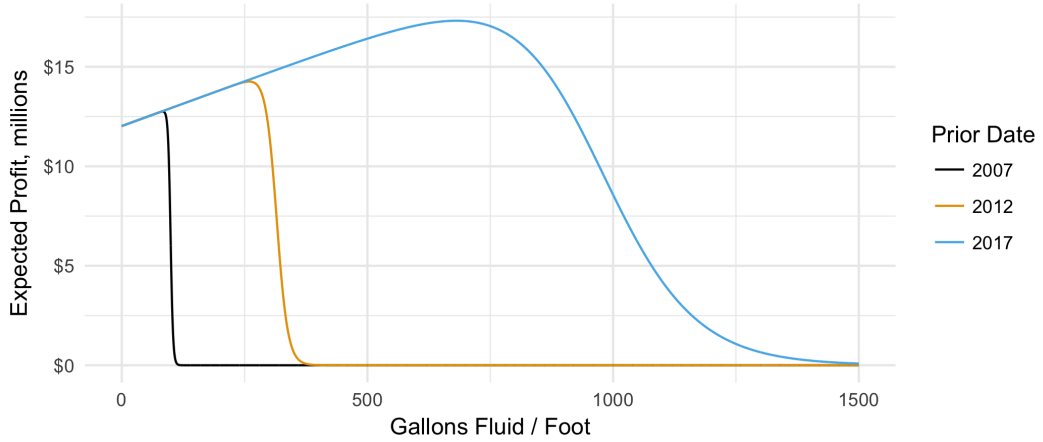
January 2012, January 2017 }, and expected profits using a $60 price of oil are shown in Figures 11 and 12. Figure 11 demonstrates how the industry has incorporated new information and significantly updated the expected probability of well success at midrange input levels. Figure 12 translates this change in prior to a change in expected profit, assuming a $60 price of oil: firms opt for higher inputs and expect higher profits in 2017 than they did in 2007.

Figure 11: Modeled Success Probabilities



Notes: anticipated probabilities of realizing $\phi > x$ under three different information states $\hat{\Gamma}^t$: January 2007, January 2012, and January 2017.

Figure 12: Modeled Expected Profits



Notes: NPV of expected profits for various input levels using a $60 price of oil, under three different information states $\hat{\Gamma}^t$: January 2007, January 2012, and January 2017.

For comparison, I also present the econometrician's estimate of γ^* in Table 5, estimated via logit regression on the subsample of wells in the FracFocus dataset.

Comparing Table 4 and Table 5, it can be seen that the industry's estimates of $\hat{\mu}^t$ is much closer to the econometrician's estimate in March 2017 than in March 2006.

Table 5: Econometrician's Estimate of ϕ Parameters

	γ
Intercept	3.702^{***}
	(0.082)
Gallons Fluid per Thousand Feet	-1.423^{***}
	(0.098)
Observations	7,645

Notes: Logit regression results for failure parameters γ, on the sample of wells included in FracFocus dataset. Outcome variable A_i is an indicator of timely well completion (treatment less than 75 days per stage), and x_i is the gallons of fracturing fluid per foot. *, **, and *** indicate statistical significance at the 10, 5, and 1 percent levels.

5.4 Calculating Value Functions

With all of the components for static profits and learning, I turn to solving for the equilibrium value functions. To begin, recall the form of expected values:

$$E[V(S', \xi; \theta)|S, \theta] \equiv \int_{S'} V(S', \xi; \theta) \Pr(S'|S, \theta) \,. \tag{21}$$

If I knew the transition probabilities $\Pr(S'|S, \theta)$, I could calculate the value functions V with a simple contraction mapping, as in Rust (1987). In my setting I cannot reasonably infer $\Pr(S'|S, \theta)$ from the data in a first stage, as the transition probabilities are endogenous with the value functions.

Instead, I use a monte carlo simulation and interpolation to derive $E[V(S', \xi; \theta)|S, \theta]$. Ultimately I require that the values $V(S, \xi; \theta)$ and expected future values $E[V(S', \xi; \theta)|S, \theta]$ are mutually consistent, and consistent with the model primitives as outlined in the Equilibrium Definition of Section 4.3. I describe the solution method in more detail below, and note now that it is similar in spirit to that used in Krusell and Smith (1998), or could be termed "Monte Carlo Value and Policy Iteration" in the approximate dynamic programming language of Powell (2007).

As global oil prices are exogenous to the state of drilling knowledge in North

Dakota, I can re-express the expected values as:

$$E[V(\Gamma', p', \xi; \theta)|\Gamma, p, \theta] = \sum_{p'} \int_{\Gamma'} V(\Gamma', p', \xi; \theta) \Pr(\Gamma'|\Gamma, p, \theta) \Pr(p'|p) \,, \tag{22}$$

where $\Pr(p'|p)$ is discretized and estimated as described in Section 5.2.

Next I turn to the more challenging dimension of the expectation, $\Pr(\Gamma'|\Gamma, p, \theta)$. I begin by rewriting it as:

$$\Pr(\Gamma'|S, \theta) = \sum_{k=1}^{K} \Pr(k \text{ wells drilled}|S, \theta) \Pr(\Gamma'|k \text{ wells drilled}, S, \theta) \,. \tag{23}$$

I set K, the maximum number of possible wells per month as 500, more than twice the maximum observed in the sample. For the purposes of simulation, the first quantity is known up to the distribution θ and cutoff $\xi^*(S; \theta)$: a well is drilled with probability $\rho(S; \theta) \equiv \Pr(\xi > \xi^*(S)|\theta)$, so the probability that k wells are drilled follows a binomial distribution. The remaining unknown quantity is $\Pr(\Gamma'|k \text{ wells drilled}, S, \theta)$. This does not have a closed-form solution and must be simulated.

So I proceed with a monte carlo approach, outlined in Algorithm 1. I begin with the first state-shock combination S, ξ on my grid.[31] First, I calculate the cutoff $\xi^*(S; \theta)$ and drilling probability $\rho(S; \theta)$. Next, I proceed to the monte carlo portion. For each of R sub-iterations, I draw a scale coefficient $\tilde{\gamma}^r$ from Γ and k^r from a Binomial(K, $\rho(S; \theta)$), which I use to simulate k^r new wells. For each simulated well, I simulate inputs $\tilde{x}$ by scaling optimal inputs $x^*(S)$ with drawn input shocks ϵ^x; I then simulate well success and failure $\tilde{A}$ using $\tilde{\gamma}^r$. After simulating k wells, I use Equations (8) - (9) to calculate $\tilde{\Gamma'}^r$ using $\{\tilde{x}\}_k$ and $\{\tilde{A}\}_k$. With $\tilde{\Gamma'}^r$ in hand, I interpolate to get $\tilde{V}(\tilde{\Gamma'}^r, \xi; \theta) = \sum_{p'} \tilde{V}(\tilde{\Gamma'}^r, p', \xi; \theta) \Pr(p'|p)$. I repeat this procedure for 2,500 monte carlo draws, and average over the results to get $E[V(S', \xi; \theta)|S, \theta]$. This entire procedure is then repeated for all states S and shocks ξ.

The above simulation of $E[V(S', \xi; \theta)|S, \theta]$ is nested within a fixed point itera-

[31] As mentioned previously, p is gridded into 15 points. The Σ dimensions are each gridded into 6 points, the μ dimensions are each gridded into 12 points, and ξ is gridded into 25 points, for a total of 1.9 million S, ξ combinations considered. The endpoints of the Σ and μ dimensions encompass estimated information states $\{\hat{\Gamma}\}$, and the endpoints of ξ are plus / minus 3 standard deviations from the mean (given by θ).

tion to find the equilibrium $V(\theta)$, as outlined in Algorithm 1. The process begins with value functions calculated from a myopic model, and is repeated until convergence.[32,33] Convergence is checked in the sup-norm of $E[V(S')|S]$, and I am using a tolerance of $\$15,000$.[34]

Begin with model primitives, and θ ;
Solve myopic problem for starting values $V^0(S, \xi; \theta)$;
Initialize $dist > tol$, $q = 1$, $R = 2,500$, $K = 500$;
while $dist > tol$ **do**
  foreach S, ξ **do**
    calculate cutoff $\xi^*(S)$;
    calculate drilling probability $\rho(S; \theta)$;
    for $r = 1 : R$ **do**
      draw 'truth' $\tilde{\gamma}^r \sim \Gamma(S)$;
      draw number of new wells k^r from Binomial(K, $\rho(S; \theta)$) ;
      simulate k^r wells using optimal input $x^*(S)$, input shocks $\epsilon^x_{k^r}$, and 'truth' $\tilde{\gamma}^r$;
      compute posterior $\tilde{\Gamma'}^r$ using Equations (8) - (9) ;
      interpolate $\tilde{V}(\tilde{\Gamma'}^r, p', \xi; \theta)$ using V^{q-1} and $\Pr(p'|p)$;
    end
    set $E[V^{q\prime}(S', \xi; \theta)|S, \theta] \equiv \frac{1}{R} \sum \tilde{V}(\tilde{\Gamma'}^r, p', \xi; \theta)$;
  end
  update value functions $V^q(S, \xi; \theta) = \max\{\pi(S) + \xi, \delta E[V^q(S', \xi; \theta)|S, \theta]\}$;
  $dist = \sup |V^q - V^{q-1}|$;
  $q = q + 1$;
end

Algorithm 1: Monte-Carlo Value Function Iteration

5.5 Estimation of Dynamic Model

Now I turn to estimating the dynamic parameters in the model, $\theta \equiv \{\mu_\xi, \sigma_\xi\}$, along with the values $V(S)$. This estimation procedure requires solving the nested fixed point detailed in Algorithm 1 for every candidate value of θ. As it is computationally

[32] I use temporal difference learning to smooth the convergence process. I use a burn-in period of 25 iterations, after which I use an updating weight of $\frac{1}{\text{iter } \#-25}$. This approach meets the criteria laid out for convergence outlined in Chapter 6 of Powell (2007).

[33] In the myopic model, firms assume that they will not learn ($\Pr(\Gamma'|\Gamma) = 1$), so I can use a standard fixed point contraction mapping.

[34] Robustness checks around the number of forward draws, convergence criterion with simulated versions have so far shown that these selections are robust.

costly to solve this fixed point, I estimate μ_ξ, σ_ξ with a grid search.[35] Estimates below are presented for the best $\hat{\theta}$ found from this search.

I evaluate the fit at each candidate θ as follows. First, I solve the nested fixed point and retrieve $V(S;\theta)$, and $E[V(S')|S;\theta]$. Then, to determine model fit, I calculate a non-linear least squares objective:

$$Q(\theta) = \sum_t \sum_{i \in t} \left(\mathbb{1}\{drilled_{it}\} - \rho(S_t, \theta) \right)^2 , \tag{24}$$

where t denote months in the data, $i \in t$ denote potential wells in month t, $\mathbb{1}\{drilled_{it}\}$ is an indicator that well i was drilled in month t, $\rho(S_t, \theta)$ is the probability of drilling given the state S_t and the calculated value and policy functions.

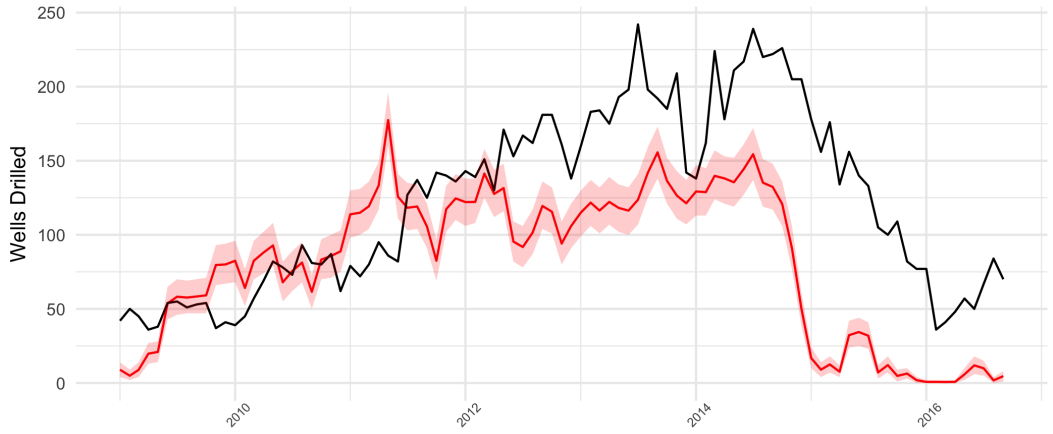
The critical assumption for this fit process is on the number of potential wells in each month. While my data contains the number of wells drilled every month, it is impossible to say with certainty the number of *potential* wells in each month that were not drilled. I proceed with the assumption that there are $K = 500$ potential wells in each month of my data, and maintain this assumption in the next section when I simulate counterfactual industry outcomes. This is significantly higher than the maximum number of wells, 240, observed in a single month. I restrict the estimation sample for the dynamic parameters to January 2009 onwards. This represents a point before the boom accelerated, but after the Bakken began to see a significant number of monthly completions.

This procedure yields estimates of $\mu_\xi = -3e^7$, and $\sigma_\xi = 2.8e^7$. Figure 13 gives an idea of how the estimated model fits in terms of predicted drilling patterns over time. The model captures the acceleration of drilling as oil prices rise and operators become more knowledgeable, as well as the deceleration with the collapse of prices in late 2015.

I next consider the effect of the social learning externality on drilling decisions.

[35] For a sense of the computational cost, it takes 250 cores on the Duke Compute Cluster on the order of 2 hours to solve the dynamic problem for a single θ to a tolerance of \$15,000 when the state space is gridded into roughly 2 million points. Code is written in Julia.

Figure 13: Actual vs. Model-Predicted Drilling



Notes: The black line plots ND Bakken completions by month. The red line plots mean model simulated drilling by month, given observed prices, estimated information states $\{\hat{\Gamma}^t\}$, and estimated policy functions and dynamic parameters; shaded area represents the middle 95% of outcomes from 1,000 simulations.

For each month in my data, I compare the equilibrium drilling cutoff $\xi^*(S_t)$, and the cutoff under a myopic model, where agents assume $\Pr(\Gamma' = \Gamma) = 1$, labeled $\xi^*_{MY}(S_t)$.[36] The first row of Table 6 presents some quantiles of these differences - it can be seen that in most states, the cutoff rule is higher under rational expectations, indicating that the possibility of learning is adding to the value of waiting. However, the second row shows that when these differences are translated into drilling probabilities, the magnitudes are negligible.

Table 6: Dynamic Impact of Rational Expectations

	Quantile				
	5%	25%	50%	75%	95%
$\xi^* - \xi^*_{MY}$	-675	238	1,148	6,228	24,646
$\rho - \rho_{MY}$	-3.1e^{-4}	-6.9e^{-5}	-1.0e^{-5}	-1.2e^{-6}	2.3e^{-6}

Notes: Quantiles of estimated differences between the rational expectations and myopic models. The quantities are calculated for all empirical states observed, using $\hat{\theta}$.

I conclude that although learning has played an integral role in the industry's development, the value of anticipated social learning in the states on the estimated information path is not large enough to appreciably change investment decisions.

[36] Calculation of the myopic model is a much simpler fixed point contraction, as transition probabilities are now known and exogenous.

Restated, firms have low enough uncertainty that they do not place much value on the possibility of moving to new information states. In the next section, I consider how this assessment changes when agents are estimated to have higher uncertainty - this could be a model for the industry in an earlier stage, learning about a new formation, or as a result of limited attention.

6 Counterfactuals

This section re-estimates industry priors under the assumption that only a limited number of wells are observed each month. The new information states lead to a similar learning path, but with much higher prior variances. The higher variances lead to significant differences between equilibrium behavior under rational expectations, and the alternative where drillers are myopic with respect to learning. I show that the expectation of social learning leads to a free-riding dynamic with lower levels of drilling and learning. Finally, I demonstrate how this dynamic can be overcome through subsidies and/or public test wells.

6.1 Alternative Prior Estimates, Limited Attention

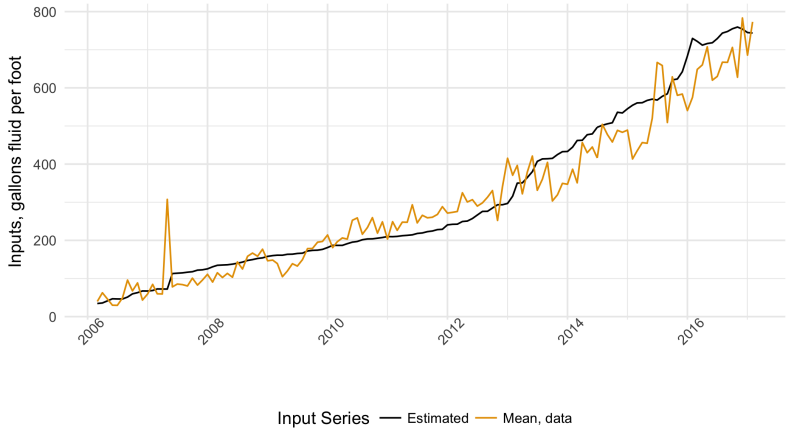
I begin by re-estimating industry priors, but using only a subsample of wells. The subsample is constructed by randomly selecting ten wells in each month, without replacement (if the month has ten or fewer wells, I select them all). Figure 14 demonstrates that the estimates using this subsample are still able to fit the industry's learning curve. However, a comparison between Table 7 and Table 4 shows that the estimated priors are much looser. This is sensible, as the means μ undergo a similar change despite a far smaller set of information.

Table 7: Information State Estimates, Limited Attention Alternative

	$t =$ March 2006	$t =$ March 2017
$\hat{\mu}^t$	$\begin{bmatrix} 21.0 \\ -0.43 \end{bmatrix}$	$\begin{bmatrix} 4.7 \\ -3.6e^{-3} \end{bmatrix}$
$\hat{\Sigma}^t$	$\begin{bmatrix} 12.4 & 0 \\ 0 & 2.1e^{-4} \end{bmatrix}$	$\begin{bmatrix} 1.4e^{-2} & 0 \\ 0 & 5.5e^{-8} \end{bmatrix}$

Notes: estimates of information states Γ^t, found via grid search, using only a random subsample of wells, with up to 10 wells for each month.

Figure 14: Model Predicted and Observed Inputs, Limited Attention Alternative



Notes: observed means are calculated at the monthly level; information states are estimated using a random subsample of wells, with 10 wells from each month.

6.2 Industry Progression under Looser Priors

I now present results showing that expected social learning can result in a free-riding dynamic, slowing drilling, and thus the learning rate of the industry. I do this by using simulating 1,000 counterfactual industries, using my estimates of θ from Section 6.1 above, and re-solving the value functions $V(S, \xi; \theta)$ on a grid of the new region of S.[37] I simulate three different scenarios: first, the baseline case with rational expectations; second, a case where wells are myopic about the possibility of learning, but incorporate new information when it arrives; third, a case where 25 public test wells are drilled at the beginning of the game to provide additional information.

For all simulations, I begin with the same starting information state of

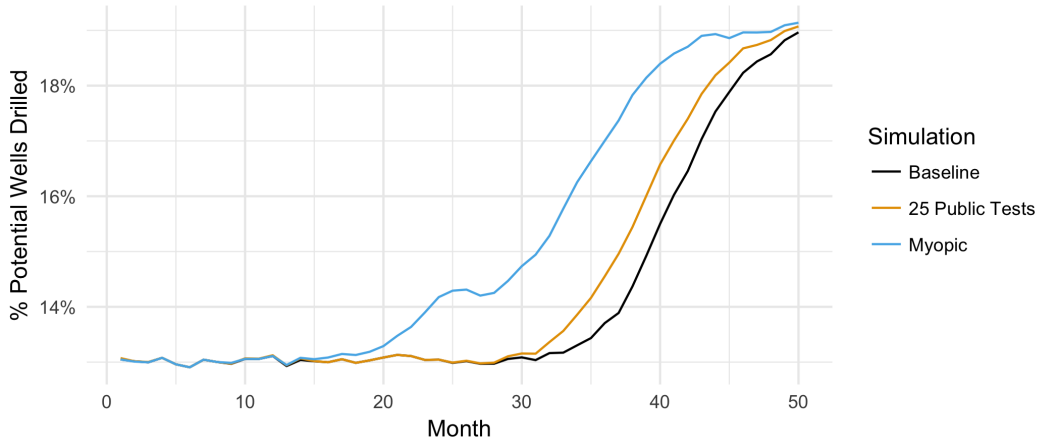
$$\mu = \begin{bmatrix} 11.09 \\ -0.04 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 7.5 & 0 \\ 0 & 5.0e^{-8} \end{bmatrix},$$

where the relatively loose priors mimic early uncertainty. I simulate with a constant \$75 price of oil to focus on the effects of social learning (though agents are still expecting the possibility of oil price swings). Finally, I hold the potential entrant assumption constant at 500 wells per period, and make sure that each of the three scenarios shares identical shocks across simulations.

[37] The new state-space grid has the same grid points in the p and ξ dimensions, but new points in the Γ dimensions, as informed by the alternative estimates of $\{\hat{\Gamma}\}$. The number of points in each dimension is unchanged from the baseline case.

The results are shown in Figures 15 - 16. Figure 15 plots mean drilling levels across the 1,000 simulation runs. It demonstrates that the rational expectation of social learning is an important factor when priors have higher uncertainty: fewer wells are drilled in the baseline case as the extra value of anticipated social learning causes more wells to wait before drilling. The figure also shows that the extra information provided by the test wells leads to more drilling under that simulation.

Figure 15: Simulated Drilling

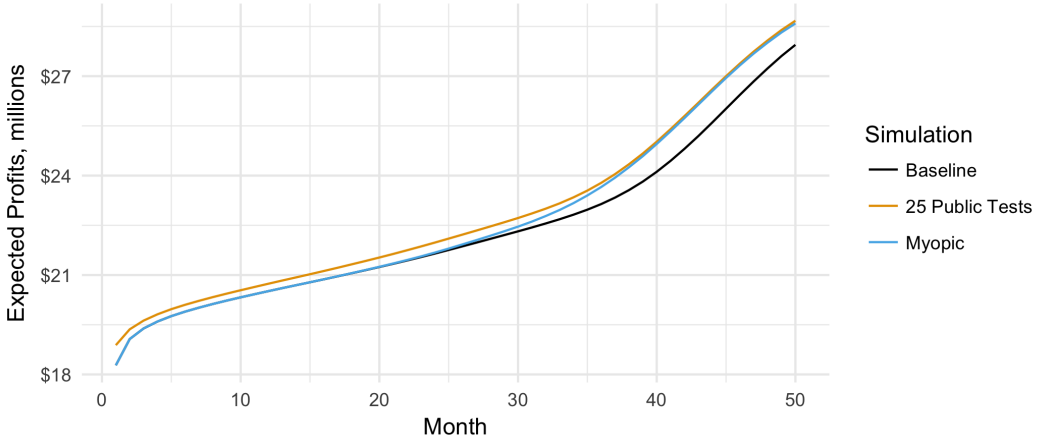


Notes: plotted lines show means across 1,000 simulations.

Figure 16 plots expected profits $\pi(S)$ in each period across simulations. The first thing to note is that drilling is higher under the myopic than the baseline simulation, even while expected profits are quite similar, highlighting that the social learning effect is operating through the $\delta E[V(S', \xi; \theta)$ channel. Moreover, the higher levels of early drilling lead to higher expected profits in later periods, demonstrating the possibility of positive feedback loops between investment and learning in this model. The figure also shows that the test wells provide valuable information in terms of expected profits; however, drilling levels are still lower than under the myopic simulations, as the expectation of social learning continues to matter.

Finally, I consider a back of the envelope cost-benefit analysis of the test well program. The 25 test wells have a simulated cost of $189.9m. In terms of benefits (and disregarding any revenue from the test wells), I calculate that the better infor-

Figure 16: Simulated Expected Profits



Notes: plotted lines show means across 1,000 simulations.

mation leads to an incremental NPV of $173.0m to the state, and $248.1m to private leaseholders, assuming a tax rate of 11.5% and a royalty rate of 16.5%. Finally, I calculate an incremental NPV of $1.7b in profits to the operators drilling the wells.

Table 8 shows similar welfare caulculations (with benefits aggregated) across different assumptions for the prevailing oil price and number of public test wells drilled. These figures demonstrate that public tests of a new technology can be welfare enhancing by accelerating the joint processes of investment and learning. It is worth highlighting that the "test" wells I have simulated here are not necessarily even intended to push the frontier, they simply represent the best design under the current industry knowledge. Much more impressive results could be obtained under an alternative assumption that these tests are designed to optimally increase available information.

Table 8: Public Test Welfare, Alternative Assumptions

Wells	Cost	NPV Aggregate Benefit				
		$55	$65	$75	$85	$95
5	-38.0	123	415	2,065	2,939	3,138
25	-189.9	95	330	2,130	3,548	4,747
50	-379.7	24	230	1,806	3,325	4,682

Notes: Figures are in millions of dollars. Assumptions include a constant price of oil and 50 month simulation run. Calculated values are incremental to simulated baselines, and discounted to NPVs. In all instances, the test wells are assumed to provide no monetary benefits. Tax rate: 11.5%; royalty rate: 16.5%.

7 Discussion

In this paper, I have developed a dynamic model of industry investment with endogenous social learning. I estimated the model using data from the shale oil industry in North Dakota's Bakken formation, where I find that learning has played a significant role in the growth and increasing profitability of the industry. However, my dynamic estimates reveal that anticipated social learning has had little to no effect on the industry's recent development: the estimated prior variances are too low to allow for much value from new information.

In the final section, I considered a counterfactual scenario of industry development under less less certain priors, as might correspond to the earliest stages of the industry, learning about a new formation, or limited firm attention. Here I found that the possibility of social learning can lead to a free-riding dynamic, where firms delay investment in order to learn from others. I show that this dynamic can lead to sizable differences in the industry's progress up the learning curve: less drilling yields less information, and lower future investment. Finally, I demonstrate that a policy designed to increase early information can have a significant effect in overcoming this free-riding: public test wells provide early information, leading subsequently to a higher level of investment and a higher rate of learning.

Taken together, the results of this paper suggest that policymakers interested in fostering industry development of new technologies face tradeoffs when designing disclosure policies. In the earlier stages of learning about a new technology, when

firms may expect to actively learn from their peers, disclosure can lead to a free-riding dynamic. In contrast, as the industry's understanding of the technology matures and prior variances fall, disclosure may increase welfare by facilitating social learning without risk of limiting investment.

References

Agerton, Mark, "Learning Where to Drill: Drilling Decisions and Geological Quality in the Haynesville Shale," Technical Report, Working Paper 2020.

Alquist, Ron and Lutz Kilian, "What do we learn from the price of crude oil futures?," *Journal of Applied Econometrics*, 2010, *25* (4), 539–573.

Anderson, Soren T, Ryan Kellogg, and Stephen W Salant, "Hotelling under pressure," Technical Report, National Bureau of Economic Research 2014.

Bahrami, Hassan, Reza Rezaee, and Ben Clennell, "Water blocking damage in hydraulically fractured tight sand gas reservoirs: An example from Perth Basin, Western Australia," *Journal of Petroleum Science and Engineering*, 2012, *88-89* (Supplement C), 100 – 106. Unconventional hydrocarbons exploration and production Challenges.

Benkard, C Lanier, "Learning and Forgetting: The Dynamics of Aircraft Production," *American Economic Review*, 2000, pp. 1034–1054.

Cogley, Timothy and Thomas J Sargent, "The conquest of US inflation: Learning and robustness to model uncertainty," *Review of Economic dynamics*, 2005, *8* (2), 528–563.

Collard-Wexler, Allan, "Demand Fluctuations in the Ready-Mix Concrete Industry," *Econometrica*, 2013, *81* (3), 1003–1037.

Conley, Timothy G and Christopher R Udry, "Learning about a new technology: Pineapple in Ghana," *The American Economic Review*, 2010, pp. 35–69.

Covert, Thomas, "Experiential and social learning in firms: the case of hydraulic fracturing in the Bakken Shale," Technical Report, Working Paper 2015.

Dixit, Avinash K and Robert S Pindyck, *Investment under uncertainty*, Princeton university press, 1994.

Doraszelski, Ulrich, Gregory Lewis, and Ariel Pakes, "Just starting out: Learning and equilibrium in a new market," *American Economic Review*, 2018, *108* (3), 565–615.

EIA, "Trends in U.S. Oil and Natural Gas Upstream Costs," Technical Report, U.S. Energy Information Administration 2016.

Eveline, Vena. F., I. Yucel Akkutlu, and George J. Moridis, "Numerical Simulation of Hydraulic Fracturing Water Effects on Shale Gas Permeability Alteration," *Transport in Porous Media*, Jan 2017, *116* (2), 727–752.

Fetter, T. Robert, Andrew L. Steck, Christopher Timmins, and Douglas H. Wrenn, "Learning by Viewing? Social Learning, Regulatory Disclosure, and Firm Productivity in Shale Gas," 2017.

Foster, Andrew D and Mark R Rosenzweig, "Learning by doing and learning from others: Human capital and technical change in agriculture," *Journal of political Economy*, 1995, pp. 1176–1209.

Gold, Russell, *The boom: how fracking ignited the American energy revolution and changed the world*, Simon and Schuster, 2014.

Griliches, Zvi, "Hybrid corn: An exploration in the economics of technological change," *Econometrica, Journal of the Econometric Society*, 1957, pp. 501–522.

Hendricks, Kenneth and Robert H Porter, "The timing and incidence of exploratory drilling on offshore wildcat tracts," *The American Economic Review*, 1996, pp. 388–407.

Herrnstadt, Evan M, Ryan Kellogg, and Eric Lewis, "The Economics of Time-Limited Development Options: The Case of Oil and Gas Leases," Technical Report, National Bureau of Economic Research 2020.

Hodgson, Charles, "Information Externalities, Free Riding, and Optimal Exploration in the UK Oil Industry," Technical Report, Working Paper 2021.

Hotelling, Harold, "The economics of exhaustible resources," *The journal of political economy*, 1931, pp. 137–175.

Jaakkola, Tommi S. and Michael I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, January 2000, *10* (1), 25–37.

Jeon, Jihye, "Learning and Investment under Demand Uncertainty in Container Shipping," Technical Report, Working Paper 2020.

Kellogg, Ryan, "Learning by drilling: Interfirm learning and relationship persistence in the Texas oilpatch," *The Quarterly Journal of Economics*, 2011.

_ , "The effect of uncertainty on investment: evidence from Texas Oil Drilling," *The American Economic Review*, 2014, *104* (6), 1698–1734.

Krusell, Per and Jr. Smith Anthony A., "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, October 1998, *106* (5), 867–896.

Lade, Gabriel E and Ivan Rudik, "Costs of inefficient regulation: Evidence from the Bakken," *Journal of Environmental Economics and Management*, 2020, *102*, 102336.

Levitt, Clinton J, "Learning through oil and gas exploration," *University of Iowa manuscript*, 2009.

Lucas, Robert E, "Making a miracle," *Econometrica: Journal of the Econometric Society*, 1993, pp. 251–272.

McCarthy, Kevin, Katherine Rojas, Martin Niemann, Daniel Palmowski, Kenneth Peters, and Artur Stankiewicz, "Basic petroleum geochemistry for source rock evaluation," *Oilfield Review*, 2011, *23* (2), 32–43.

Mokyr, Joel, *The lever of riches: Technological creativity and economic progress*, Oxford University Press, 1992.

Newell, Richard G and Brian C Prest, "Is the US the New Swing Producer? The Price-Responsiveness of Tight Oil," Technical Report, Resources for the Future 2017.

_ , _ , and Ashley Vissing, "Trophy Hunting vs. Manufacturing Energy: The Price-Responsiveness of Shale Gas," Technical Report, National Bureau of Economic Research 2016.

Orlik, Anna and Laura Veldkamp, "Understanding uncertainty shocks and the role of black swans," Technical Report, National Bureau of Economic Research 2014.

Powell, Warren B., *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, John Wiley & Sons, October 2007. Google-Books-ID: WWWDkd65TdYC.

_ and Ilya O. Ryzhov, *Optimal Learning*, John Wiley & Sons, April 2012. Google-Books-ID: hnsVMbx5HOAC.

Putthaworapoom, Natthapon, Jennifer Lynne Miskimins, and Hossein Kazemi, "Sensitivity Analysis of Hydraulic Fracturing Damage Factors: Reservoir Properties and Operation Variables," in "SPE-151060-MS" Society of Petroleum Engineers SPE January 2012.

Rust, John, "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 1987, pp. 999–1033.

Ryan, Stephen P, "The costs of environmental regulation in a concentrated industry," *Econometrica*, 2012, *80* (3), 1019–1061.

Zuckerman, Gregory, *The frackers: The outrageous inside story of the new billionaire wildcatters*, Penguin, 2013.

ONLINE APPENDIX

A Well Characteristics and Production

I collect administrative data on completions, production and other well characteristics from the NDIC for wells completed in the years between 2005 and 2016. One dataset of static well-level characteristics includes IP rates, drilling and completion dates, well locations, target formations, and for horizontal wells, the length of the horizontal wellbore. As different geological formations may have different learning processes, I use the formation variable to restrict my sample for this paper to wells targeting the Bakken play.[38] A separate NDIC dataset contains monthly well-level production details: quantities of oil, water and gas produced, days in production, quantities of oil and gas sold, and quantity of gas flared. Bakken wells derive almost all of their value from crude oil rather than natural gas (which is often flared instead of harvested, see e.g. Lade and Rudik (2020)), so I focus attention below and in the model on oil production and sales.

Table 9 shows the evolution of some of these variables over the timeframe of my study. The first two rows show that the number of active operators and completed wells rose steeply before falling off in 2015. The third through sixth rows show that the distribution of IP has also dramatically increased: a 25th percentile well in 2015 has an IP almost twice as high as a median well in 2008. This is partly explained by the increase in average length of a well, shown in the seventh row. The middle section of Table 9 shows annual aggregate production from all Bakken wells, and from Bakken wells completed that year. While aggregate production from Bakken wells has continued to increase, the production from new wells has fallen along with completions.

[38] Conversations with industry participants indicate that learning processes across formations are separate, as geological differences between formations cause them to respond differently to treatments.

The NDIC also publishes results from geological surveys, guiding estimates of the production potential of the Bakken shale in various locations. Higher measures of total organic content (TOC) and thicker shale layers indicate the possibility of more oil. The hydrogen index (HI) and S2-TMAX (S2) are both measures of thermal maturity. A higher hydrogen index indicates the presence of more hydrocarbons, and values as low as 200 can be considered mature. Ideal maturity is also indicated by S2-TMAX values between 435 and 460 degress celsius (McCarthy et al., 2011). The bottom section of Table 9 shows that wells are progressively drilled where the Bakken formation is thicker, with a higher TOC, a slightly lower HI, and roughly the same S2. The geological survey data is thus ambiguous as to whether operators are drilling "sweeter spots" over time.

Table 9: NDIC Summary Stats, Drilling and Production

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Wells Completed	22	66	153	412	475	766	1211	1761	1991	2175	1478	737
Operators	10	24	25	33	38	51	54	51	52	47	39	26
IP 25th perc.	47	48	78	149	175	210	223	216	241	277	326	378
IP 50th perc.	96	97	137	237	266	315	325	315	341	390	458	542
IP 75th perc.	154	203	266	403	387	439	436	432	456	528	600	726
IP 95th perc.	368	432	646	820	625	703	645	638	735	802	892	1026
Mean Length (feet)	5287	5849	7015	8654	8949	9220	9476	9498	9561	9592	9631	9655
Median IP/ft $*$ 100	1.9	1.6	1.9	3.1	3.7	3.8	3.6	3.5	3.7	4.2	4.8	5.8
Total Production	1.1	2.3	7.5	27.3	49.9	86.0	128.8	219.4	290.3	373.6	410.9	361.2
New Production	0.3	0.9	3.8	17.1	19.2	37.1	51.9	86.0	104.8	121.4	101.4	52.1
Mean Formation Width	37.2	34.5	39.6	44.0	46.4	45.5	43.7	43.0	44.9	45.2	45.7	46.8
Mean TOC	12.7	12.4	13.4	14.0	13.9	13.8	13.7	13.7	13.5	13.5	13.5	13.9
Mean HI	298	282	311	383	391	349	313	301	289	262	241	246
Mean S2	436	437	436	434	434	435	436	436	437	438	439	438

IP is reported initial production, in barrels of oil per day. Total and New Production are both reported in millions of barrels. Formation width is reported in feet and is the sum of the upper and lower Bakken thicknesses. The Total Organic Content is the percentage of organic content found in samples of the rock. The Hydrogen Index is unitless, and calculated from S2 and TOC measures. The S2 is measured in mg/g.

B Alternative Explanations

In this appendix, I argue against alternative explanations for the patterns in the data, and conclude that learning about the production function is the most plausible explanation.

Consider the illustrative model, where a well's expected profit can be written as a function of oil prices p, quantity q, costs c, as well as their determinants: well inputs x and operator knowledge Γ:

$$\pi(p, x; \Gamma) = pq(x^*(p, \Gamma)) - c(x^*(p, \Gamma)) \,. \tag{25}$$

I write $x^*(p, \Gamma)$ to denote that the operator is choosing optimal inputs given p, Γ, and the functions $q(x)$ and $c(x)$. Table 1 charts a monotonic increase in observed $x^*(p, \Gamma)$, even as Figure 1 shows that prices p behaved very non-monotonically. This phenomenon has a few potential explanations.

The first alternative explanation is that unit costs $c'_t(x)$ have fallen, where t indexes time. The direct costs for proppant and/or hydraulic fracturing fluid and/or the costs of recovering and disposing well flow back might have fallen.[39] Standard economic theory predicts that this would lead to operators increasing their uses of these inputs until marginal benefits are again equated to marginal costs. However, the evidence does not support this alternative. The EIA released a study in 2016 on industry costs.[40] They first find that proppant, fluid, and flowback costs make up only 26% of well costs on average.[41] The study then found that proppant and fluid costs were relatively stable from 2006 to 2015: fluids were slightly more expensive from 2010 to 2013, and proppant costs were actually increasing over the period. The EIA study also traces flowback costs: while it does show a fall in flowback costs from 2012 to 2014, it shows an increase from 2006 to 2012 (EIA, 2016). These small and ambiguous changes in marginal costs cannot explain the monotonic increase of proppant and fluid

[39] After the well is stimulated, some of the fracturing fluid flows back up the well; it must be carefully collected and recycled or disposed of due to the hazardous chemicals it contains.

[40] See Figure 2-5 on page 12 of the IHS report (EIA, 2016).

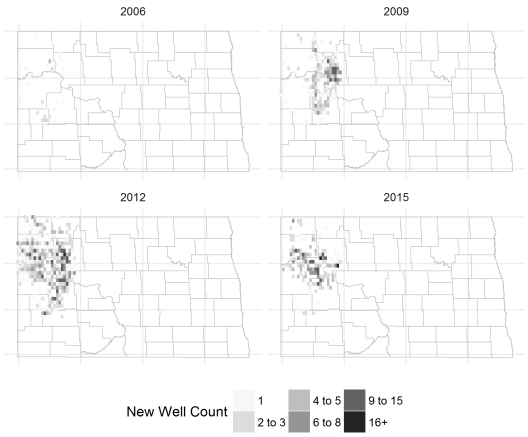
[41] This is in line with what I can calculate from my AFE dataset.

use in the Bakken.

The second possible explanation is that technology improved over the timeframe, so that $q(x)$ should be written $q_t(x)$. Such a shift in technology could expand operators' choice set over input configurations and change x^*. Again, the evidence rules out this possibility. The same EIA report shows in Figure 2-16 on page 19 that the average Bakken well has lagged behind average wells in other formations in proppant use per foot (EIA, 2016). The technology for more intense fractures clearly existed for some time before operators decided to take advantage of it in the Bakken.[42] So changes in technology q are not a satisfactory explanation for the observed increase in x.

A third possibility is that operators' configurations have changed because they are drilling in different places, or that the production technology is a highly location-specific $q_{loc}(x)$. Figure 17, which shows locations of wells drilled in 2006, 2009, 2012, and 2015 suggests that this is not the case. While the early years saw operators explore new areas, by later years wells tended to be drilled in areas that had been explored previously.

Figure 17: Drilling Locations Over Time



Notes: Each sub-figure represents a heatmap of new wells drilled in that year.

[42]This was confirmed by my conversations with an industry participant; he noted in early 2017 that his firm had been experimenting with their currently-used frac configurations as early as 2011.

I further test this last alternative through a series of regressions, regressing well inputs on fixed effects for operators, dates, and locations. Table 10 show the results of this exercise when dates and locations are binned into 25 binary variables. Focusing on the first row, the adjusted R^2 values illustrate that of the three sets of fixed effects, operators are the most important in determining how much fluid is used, followed by date; columns 4-6 show that the location fixed effects have little explanatory power when operators are taken into account. The second and third row show that similar patterns exist for pounds of proppant per foot and the number of frac stages.[43] I conclude that *who* fractures a well and *when* the well is fractured do much more to determine the well's inputs than *where* the well is fractured.

Table 10: Inputs by Operator, Date and Location; 25 bins

	Adjusted R^2						
Fluid per Foot	0.27	0.27	0.08	0.51	0.31	0.35	0.53
Proppant per Foot	0.40	0.29	0.14	0.61	0.42	0.39	0.62
Stages	0.17	0.22	0.02	0.33	0.18	0.24	0.34
Operator FE	✓			✓	✓		✓
Date FE		✓		✓		✓	✓
Location FE			✓		✓	✓	✓

Notes: Each entry represents a separate regression. Dependent variables are listed in the first column. Included independent variables are denoted by checkmarks, see text for description; location and date fixed effects are included as 25 bins.

Having argued against changes in $q(x)$, $c(x)$, or p explaining the observed trend in $x^*(p, \Gamma)$, I conclude that the most plausible explanation is a change in industry knowledge Γ, or learning. This finding suggests the model of operator learning over optimal input use that follows.

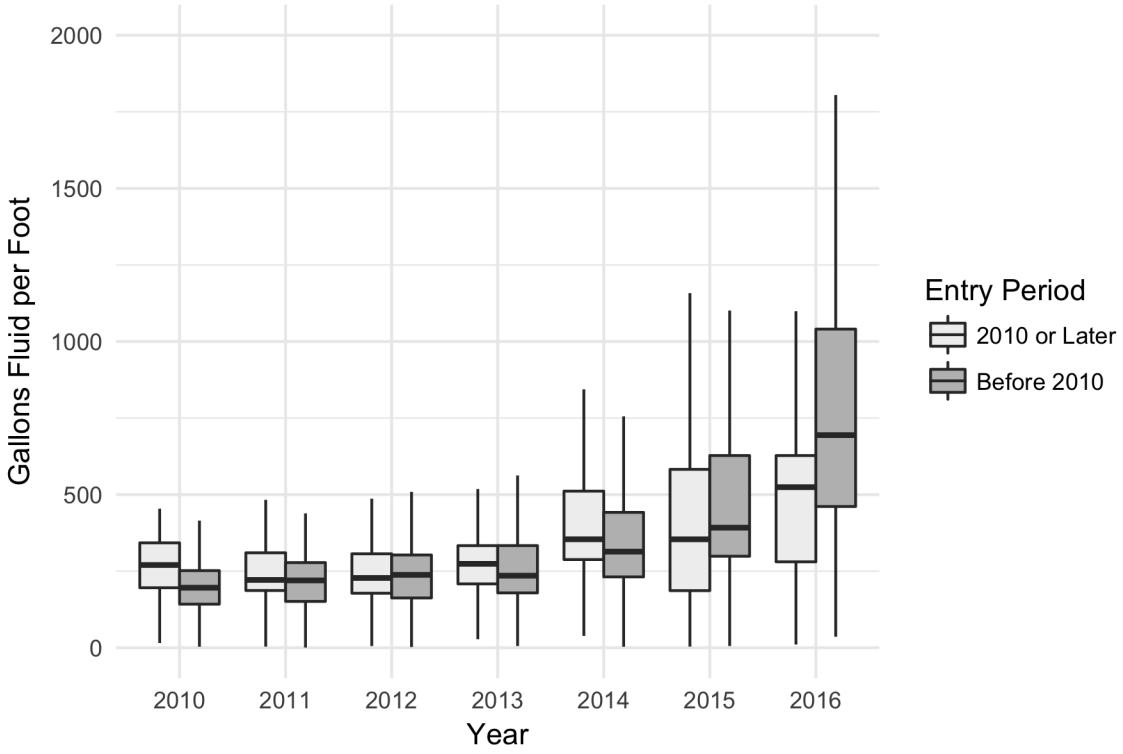
[43] Appendix F shows the results using 100 bins per variable; the results are similar.

C Inter-firm Learning

In defense of the assumption to treat each potential well as its own firm, this appendix argues that the learning in the data takes place predominantly at the inter-firm level, rather than the intra-firm level. The lack of demonstrable intra-firm learning suggests that firms are as likely to learn from each others' wells as their own, and aids the plausibility of the assumption of treating each well as its own firm.

Figure 18 graphs the distributions of gallons of fracturing fluid per foot usedi n wells from 2010 - 2016, with the sample split into two groups: those operated by early and late entrants. I define early entrants as those firms who were first active in the the Bakken prior to 2010. As Figure 18 shows, the input configurations do not seem to vary systematically between these two groups (with the exception of 2016), suggesting that the pattern of changing operational choices is common across the industry rather than intra-firm.

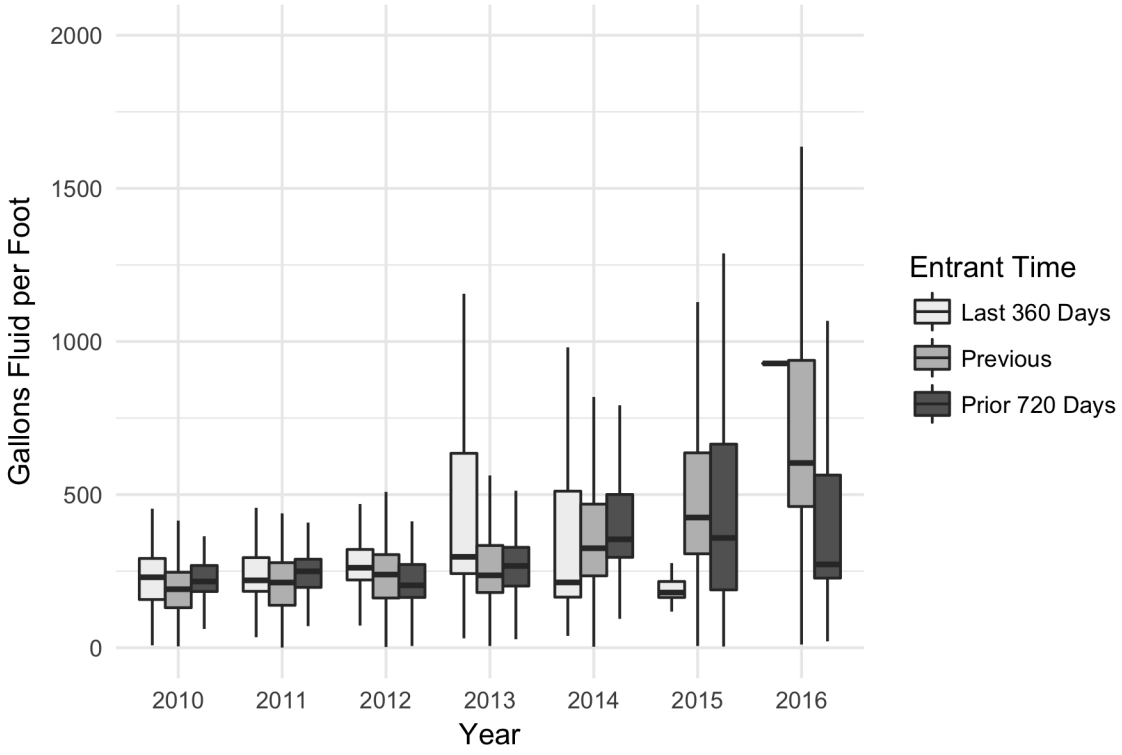
Figure 18: Gallons Fracturing Fluid per Foot by Entrant Date



This conclusion is reinforced by Figure 19; this figure performs a similar exercise, but divides the sample dynamically. For each year of wells, the sample of well-operators is split into three: those who entered the Bakken recently, in the last 360 days; those who entered in the last 361 to 1080 days; and those who entered more than 1080 days previously. From this figure it can be seen that there are no discernible patterns between entry date and input choices. In 2010 - 2012 for example, the distributions are quite similar. In 2015, it looks like the length of time the firm has been operating in the Bakken is positively correlated with gallons of fracturing fluid, but this conclusion is flipped in 2016, when it appears that the newest firms are using the most fluid, followed by the oldest firms, followed by the intermediate.

Analagous graphs with a different cutoff years, and different input choices (pounds of proppant per foot, fracture stages, maximum injection pressure) similar patterns and are available from the author upon request.

Figure 19: Gallons Fracturing Fluid per Foot by Entrant Date



D IP and Cost Predictions

This appendix describes the calculations performed to create Figures 5 and 6.

The first step is a regression of 6 month IP on well configuration and location:

$$q_w = \beta_0^p + \beta_h^p h_w + \beta_s^p s_w + \beta_f^p f_w + \alpha_w^p + \varepsilon_w^p, \tag{26}$$

where q_w is the log of 6 month initial production, h_w is the log of the well's horizontal length in feet, s_w is the log of pounds proppant, f_w is the log of gallons fluid, α_w represents a township fixed effect, and ε_w^p is the regression error. The results of this regression are shown in Table 11.

Table 11: Counterfactual Production Regression Results

	Log 6 Month IP (BBL/day)
Log Feet Length	0.292***
	(0.022)
Log Pounds Proppant	0.151***
	(0.009)
Log Gallons Fluid	0.134***
	(0.009)
Location FE	✓
Observations	11,460
R^2	0.511
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

I then take the estimated coefficients $(\hat{\beta}^p, \hat{\alpha}^p)$, and use them to generate predicted values for each well:

$$\hat{q}_w^{2007} \equiv \hat{\beta}_0^p + \hat{\beta}_h^p h^{2007} + \hat{\beta}_s^p s^{2007} + \hat{\beta}_f^p f^{2007} + \hat{\alpha}_w^p \tag{27}$$

$$\hat{q}_w^{actual} \equiv \hat{\beta}_0^p + \hat{\beta}_h^p h^{actual} + \hat{\beta}_s^p s^{actual} + \hat{\beta}_f^p f^{actual} + \hat{\alpha}_w^p \tag{28}$$

$$\hat{q}_w^{2016} \equiv \hat{\beta}_0^p + \hat{\beta}_h^p h^{2016} + \hat{\beta}_s^p s^{2016} + \hat{\beta}_f^p f^{2016} + \hat{\alpha}_w^p, \tag{29}$$

where h^{2007}, s^{2007}, and f^{2007} denote the median horizontal length, sand, and fluid amounts used by wells in 2007. Similarly, h^{actual} denotes the actual horizontal length used seen in the data.[44]. Thus the fitted values $\hat{q}_w$ represent estimates of initial production given the actual geography and counterfactual configurations. Figure 5 plots the medians by year of actual completion of these fitted values, $\hat{q}_w^{2007}$ and $\hat{q}_w^{2012016}$.

The routine for costs is similar. The regression in this case is:

$$c_w = \beta_0^c + \beta_h^c h_w + \beta_s^c s_w + \beta_f^c f_w + \varepsilon_w^c, \tag{30}$$

where c_w is the log of well costs in thousands of dollars and the other variables are defined as before. Note that I assume that location does not affect drilling and completion costs conditional on configuration choices. The results of this regression are shown in Table 12.

Table 12: Counterfactual Cost Regression Results

	Log Cost ($ 1000s)
Log Feet Length	0.429***
	(0.045)
Log Pounds Proppant	0.022
	(0.018)
Log Gallons Fluid	0.080***
	(0.018)
Observations	421
R^2	0.320

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

I then use the estimated coefficients $(\hat{\beta}^p, \hat{\omega}^p)$ to generate predicted costs for each

[44]I project actual inputs on the estimated location effects so that the only differences between the lines in Figure 5 are inputs used

well for which I do not have a cost estimate:

$$\hat{c}_w \equiv \hat{\beta}_0^c + \hat{\beta}_h^c h_w + \hat{\beta}_s^c s_w + \hat{\beta}_f^c f_w + \hat{\omega}_w^c, \tag{31}$$

as well as predicted costs for all wells if they were drilled in the mean 2007 and 2016 configurations:

$$\hat{c}_w^{2007} \equiv \hat{\beta}_0^c + \hat{\beta}_h^c h_w^{2007} + \hat{\beta}_s^c s_w^{2007} + \hat{\beta}_f^c f_w^{2007} + \hat{\omega}^{2007}, \tag{32}$$

$$\hat{c}_w^{2016} \equiv \hat{\beta}_0^c + \hat{\beta}_h^c h_w^{2016} + \hat{\beta}_s^c s_w^{2016} + \hat{\beta}_f^c f_w^{2016} + \hat{\omega}^{2016}. \tag{33}$$

Finally, I transform the predicted values to levels:

$$\hat{Q}_w^{2007} = \exp(\hat{q}_w^{2007}) + \frac{\hat{\sigma}_p^2}{2},$$

$$\hat{Q}_w^{2016} = \exp(\hat{q}_w^{2016}) + \frac{\hat{\sigma}_p^2}{2},$$

$$\hat{C}_w = \exp(\hat{c}_w) + \frac{\hat{\sigma}_c^2}{2},$$

$$\hat{C}_w^{2007} \exp(\hat{c}_w^{2007}) + \frac{\hat{\sigma}_c^2}{2},$$

$$\hat{C}_w^{2016} \exp(\hat{c}_w^{2016}) + \frac{\hat{\sigma}_c^2}{2}.$$

I then calculate three quantities for each well:

- $\hat{C}_w/Q_w$ - the estimated (or actual where available) cost over actual 6 month IP;
- $\hat{C}_w^{2007}/Q_w^{2007}$ - the 2007-configuration estimated cost over 6 month IP;
- $\hat{C}_w^{2016}/Q_w^{2016}$ - the 2016-configuration estimated cost over 6 month IP.

Figure 6 plots the median values of each of these three distributions by year of actual completion.

E AFE Sample Selection

Table 13 shows the 10th, 50th, and 90th percentiles of selected variables for two samples: those wells which can be matched to an AFE, and those that cannot. The final column displays p-values from a t-test for equality between means. The table shows that on average, AFE wells are drilled slightly earlier. Consistent with this, they tend to be slightly smaller in horizontal length, to use slightly less proppant and fracturing fluid, and to produce slightly less oil.

Table 13: AFE Sample Selection

		AFE	Non-AFE	t-test
N		421	11,309	
Completion Date		2012-09-13	2013-04-27	0.00
6 month IP Oil	10%	183	250	
	50%	712	879	0.00
	90%	1,779	2,342	
Horizontal Length	10%	5505	5750	
	50%	9,473	9,509	0.04
	90%	10,056	10,191	
Proppant (million lbs)	10%	1.51	1.29	
	50%	2.85	3.09	0.00
	90%	4.09	6.84	
Fluid (million gallons)	10%	1.11	0.87	
	50%	2.27	2.44	0.16
	90%	8.00	6.80	

Values shown are distribution percentiles.
The final column displays p-values for t-tests of mean equality across groups.

F Inputs regressed on FEs, 100 bins

Table 14 shows an alternative version of Table 10, when the geographic and date variables are divided into 100 rather than 25 bins. The conclusions from the table are similar: it appears that when the well is drilled, and who drilled the well do much more to determine input use than where the well was drilled. This suggests that the pattern of increasing input use in the data cannot be well explained by a specific-geography argument.

Table 14: Inputs by Operator, Date and Location; 100 bins

	Adjusted R^2						
Fluid per Foot	0.27	0.27	0.18	0.51	0.34	0.42	0.55
Proppant per Foot	0.40	0.29	0.25	0.61	0.44	0.47	0.64
Stages	0.17	0.22	0.07	0.34	0.19	0.27	0.35
Operator FE	✓			✓	✓		✓
Date FE		✓		✓		✓	✓
Location FE			✓		✓	✓	✓

Note that each entry represents a separate regression.
Dependent variables are listed in the first column.
Included independent variables are denoted by checkmarks, see text of Section B for description; location and date fixed effects are included as 100 bins.

G Industry Concentration

Table 15 shows some measures of industry activity and concentration over time. As has been shown elsewhere, the number of active operators in a given year has increased along with the number of wells completed. The fourth through eighth columns demonstrate the heterogeneity in operator size within a given year. For example, in 2012, the 95th percentile operator completes 186 wells, compared to the median operator's 25 wells, and the 5th percentile operator's 2 wells. The final column shows the Herfindahl-Hirschman Index compute by year in the share of completed wells; the values suggest that while there are a few big players, the industry is not concentrated in an absolute sense.

Table 15: Measures of Industry Concentration

			Wells per Operator					
Year	Wells Completed	Active Operators	p5	p25	p50	p75	p95	HHI
2006	45	12	1	1	2	4	10	0.178
2007	161	16	1	3	4	12	34	0.155
2008	1566	28	1	4	10	66	212	0.151
2009	1016	33	1	2	8	46	85	0.122
2010	1237	38	1	3	10	34	115	0.126
2011	1634	41	1	3	14	49	152	0.094
2012	2183	41	2	7	25	59	186	0.073
2013	2523	42	2	7	25	73	214	0.075
2014	2566	42	2	7	28	75	259	0.063
2015	1655	35	1	8	20	58	200	0.078
2016	693	29	1	6	16	39	67	0.067

As an example, Wells per Operator p50 denotes the median number of wells completed by a single operator in a given year.
HHI denotes the Herfindahl-Hirschman Index calculated by the share of completed wells by an operator in a given year.

H Drilling and Lease Expiration

In this section, I examine the timing of well drilling. While the model assumes that the firm's problem has an infinite horizon, in fact the mineral rights leases that grant operators the right to drill for oil have expiration dates. Herrnstadt et al. (2020) does an excellent job studying how this affects the timing of an operator's drilling decision and considering the wider economics of similar contracts. If a well is drilled after a lease has elapsed (and not been extended), the share of revenue accuring to the mineral rights holder will revert to 100% from the average royalty rate of less than 20% (though the mineral rights holder will become responsible for his share of drilling and production costs in this case).

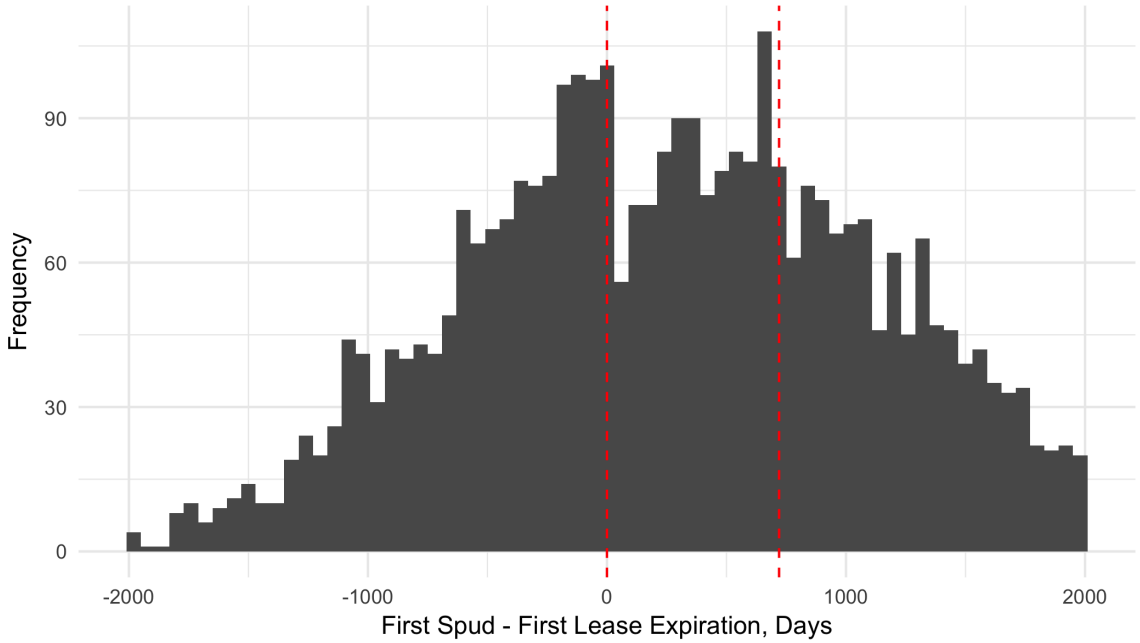
Using lease data available from Enverus (formerly DrillingInfo), I investigate the relationship between lease expiration and drilling in my dataset. I use a simple algorithm to match wells to leases, by restricting attention to "regular" leases that are made up of whole sections, quarter-sections, or sixteenths of sections. This choice allows me to quickly assign wells to leases using data from the NDIC's "scout ticket" dataset, and captures 7,805 wells, more than 60% of the wells in my sample. While it is not exhaustive, it suffices for the investigation undertaken in this section. I do not attempt to match the "irregular" leases: it would be a significant undertaking and not add anything to the primary analysis of this paper.

Figure 20 is a histogram of the difference in days, for each spacing unit, between when the drilling begins on the first well, and the earliest lease expiration date. Vertical lines are drawn at the earliest lease expiration, and two years afterward (which is the length of the typical extension clause). There are a few noteworthy takeaways. First, lease expiration and two years following lease expiration appear to be important dates that affect the timing of drilling decisions. Second, while operators are more likely to drill immediately before than immediately after the first lease expiration, they frequently drill at many other times than simply right before

the first lease expires. Third, the first spud on many spacing units takes place well after even the extension of the lease. It is worth highlighting here that his is in stark contrast to the findings of Herrnstadt et al. (2020) in the Haynesville shale.

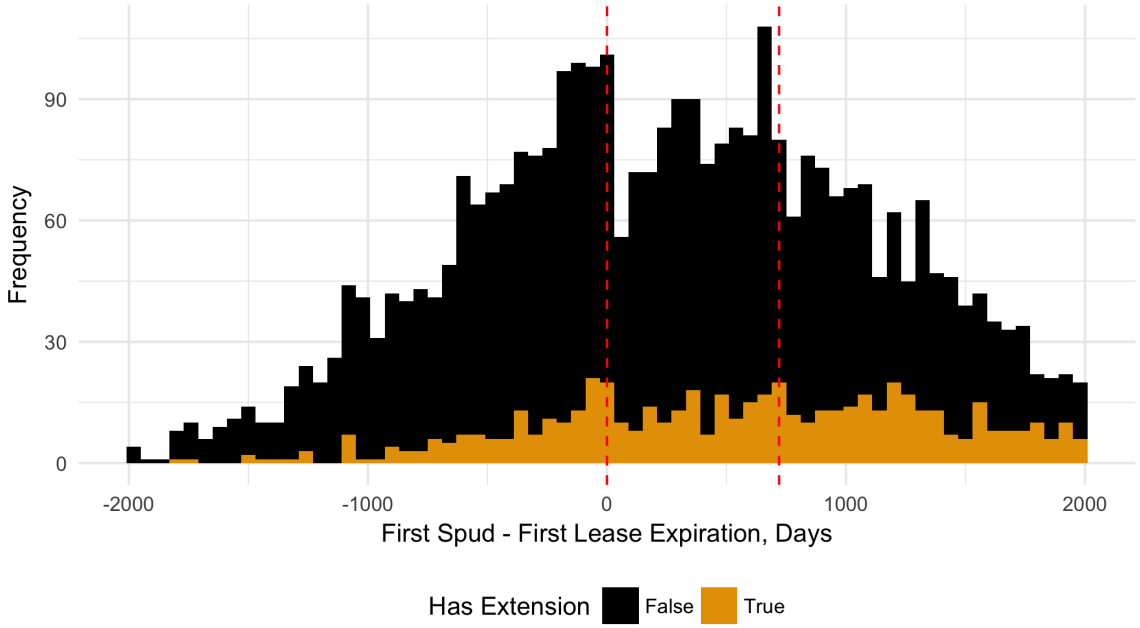
Figure 21 is a similar plot, but with the sample broken up into those spacing units whose earliest expiring lease does and does not have an extension clause. While relatively few leases feature such an extension clause, the patterns in drilling time are strikingly similar across the two groups. Again, this is in contrast to the findings presented in Herrnstadt et al. (2020).

Figure 20: Timing of Drilling and Lease Expiration



Notes: Positive numbers indicate the spud took place *after* the first lease expiration. The sample is made up of those wells which are assigned to 'regular' spacing units, see text for details.

Figure 21: Timing of Drilling and Lease Expiration, with and without Extensions



Notes: Positive numbers indicate the spud took place *after* the first lease expiration. The sample is made up of those wells which are assigned to 'regular' spacing units, see text for details.

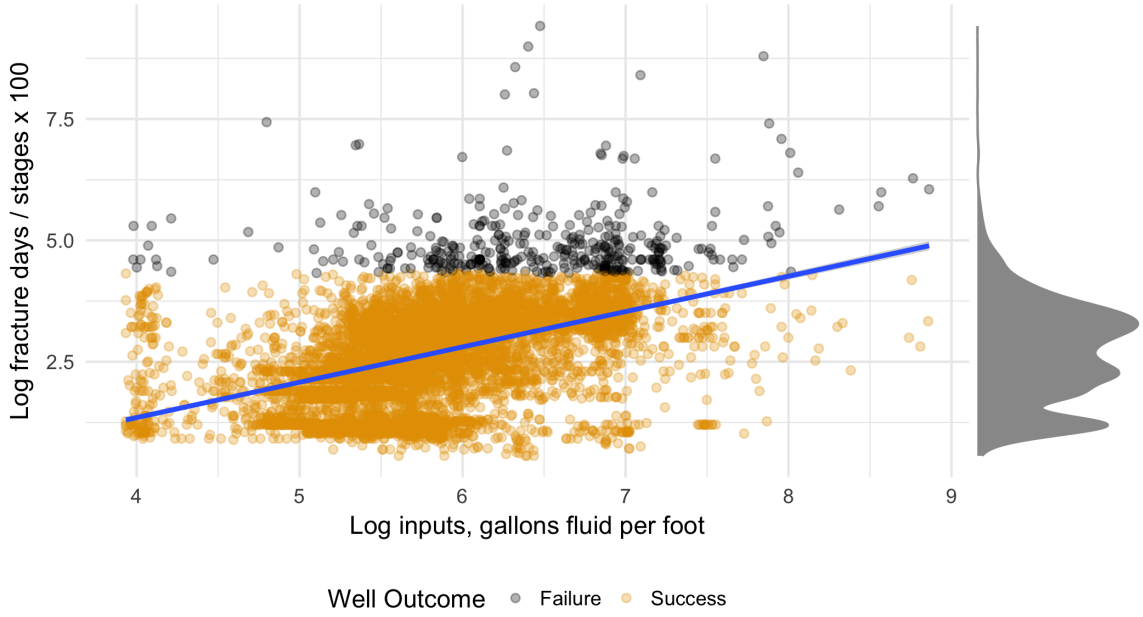
I Defining Well Success and Failure

This appendix describes the empirical choice made to define binary well success and failure, the data used in that definition, and how the chosen definition relates to when the well is drilled and operator size.

The learning model described in Section 4 requires a binary outcome for well i, A_i. I use fracture days, reported on FracFocus, normalized by the number of fracture stages, reported by the NDIC, as a measure of the success in the fracture job. The right panel of Figure 22 plots the density of this measure (with a log transformation of fracture days), and demonstrates that it is a very skewed measure empirically. I select the 95th percentile of this measure as a cutoff to define well success or failure. The main panel of Figure 22 is a scatterplot of this measure related to input choices; in the figure one can clearly see the relationship between the two variables.

Figure 23 is a companion graph, plotting the distributions of well input choices

Figure 22: Fracture Days and Inputs

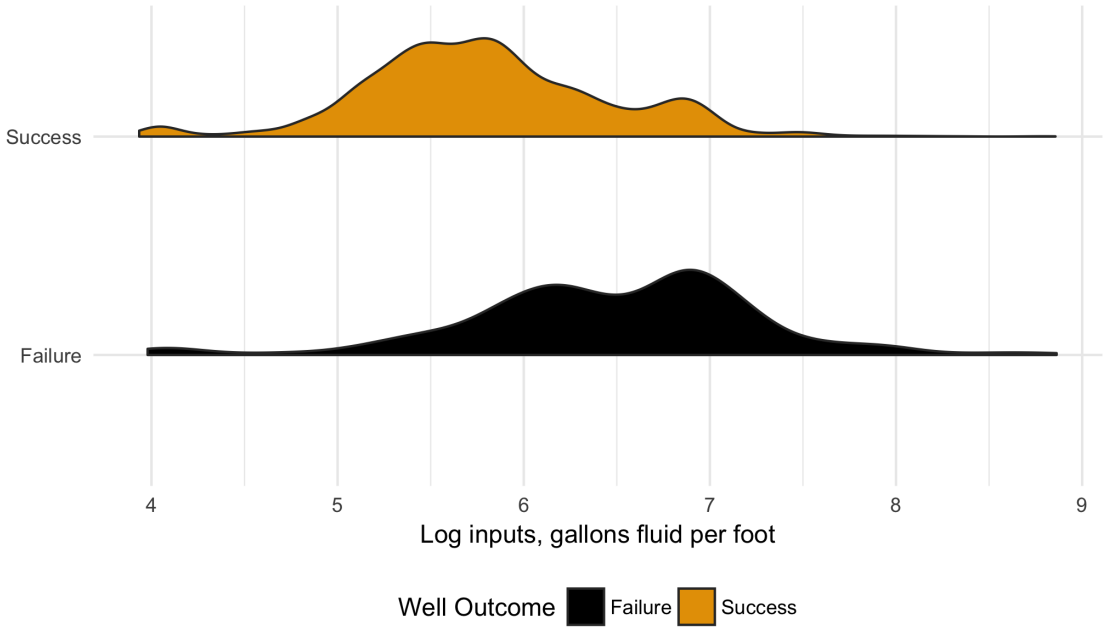


Notes:

by well success and failure. The figure makes clear the fact that successful and unsuccessful wells can both occur at any level of input choice, but also that unsuccesful wells are much more likely to have had a high level of inputs than successful wells.

Finally, I examine how input choices and well success vary with time and firm size in Table 16. The fifth column reports the percentage of observed well successes, as defined above using the FracFocus data on fracture days. The fourth column extends the sample to those wells that are excluded from FracFocus, by using fitted values from the econometrician's estimate of $\hat{\gamma}$ (reported in Table 5) to predict success. The top half of the table considers the evolution over time, and shows observed and perceived well sucesses falling as wells are fractured with higher levels of inputs. The bottom half of the table shows that the relationship of input choices and well outcomes to firm size: the largest quartile of firms evidentally use a higher average level of input, but the relationship among the the other three quartiles is less clear.

Figure 23: Well Outcomes and Inputs



Notes:

Table 16: Inputs and Outcomes by Year and Firm Size

Category	Count	Input	Percent Successful Inferred	Percent Successful Observed
2006-08	1772	181	96.9	–
2009-11	3887	228.5	96.5	–
2012-14	7272	398.9	94.6	94.3
2015-16	2348	568.7	93.2	93.2
1st Quartile Firm Size	4582	349.2	95.5	94.7
2nd Quartile Firm Size	2873	370.9	92.8	89.8
3rd Quartile Firm Size	5542	331.7	96.2	96
4th Quartile Firm Size	2323	419.9	94.9	92.2

Quartiles of firm size are determined by the number of wells observed in the sample.
Input refers to the mean observed input for a given category.
Percent Successful refers to the percentage of wells that are successful by my definition (see text for details); Observed values are taken directly from the data, while inferred values are augmented by predictions from the econometrician's estimate of $\hat{\gamma}$ for those wells without FracFocus data.