SUPPLEMENTARY MATERIALS TO:

The State of American Entrepreneurship:

New Estimates of the Quantity and Quality of Entrepreneurship for 32 US States, 1988-2014

Jorge Guzman, Columbia University
Scott Stern, MIT and NBER

Online Appendix

Table A1. Full Model Specification of Models in Table 3.
Dependent Variable: 1[IPO or Acquisition in six years or less]

	Nowcasting (up to real-time)	Full (2 year lag)
	(1)	(2)
Corporate Governance Measures		
Corporation	3.276***	3.061***
	(0.0788)	(0.0739)
Delaware	18.22***	
	(0.363)	
Name-Based Measures		
Short Name	2.487***	2.263***
	(0.0491)	(0.0456)
Eponymous	0.296***	0.315***
	(0.0168)	(0.0179)
Intellectual Property Measures		
Trademark		3.964***
		(0.219)
Patent - Delaware Interaction		
Patent Only		22.77***
		(1.059)
Delaware Only		15.18***
		(0.335)
Patent and Delaware		84.08***
		(3.320)
US CMP Cluster Dummies		
Local	0.418***	0.434***
	(0.0175)	(0.0182)
Traded Resource Intensive	0.876***	0.868***
	(0.0242)	(0.0243)
Traded	0.997	1.048*
	(0.0199)	(0.0212)
US CMP High-Tech Cluster Dummies		
Biotechnology	2.845***	2.173***
	(0.186)	(0.155)
E-Commerce	1.443***	1.348***
	(0.0466)	(0.0446)
IT	2.468***	2.175***
	(0.0857)	(0.0779)
Medical Devices	1.535***	1.301***
	(0.0587)	(0.0515)
Semiconductors	2.329***	1.590***
	(0.304)	(0.224)
N	18,764,856	18,764,856
R-squared	0.163	0.187

Table A2. Robustness Models. State Fixed Effects and State-Specific Time Trends
Dependent Variable: 1[IPO or Acquisition in six years or less]

	(1)	(2)	(3)
Corporate Governance Measures			
Corporation	2.499***	2.657***	2.572***
	(0.0630)	(0.0675)	(0.0651)
Name-Based Measures			
Short Name	2.280***	2.287***	2.286***
	(0.0460)	(0.0461)	(0.0461)
Eponymous	0.315***	0.313***	0.313***
	(0.0179)	(0.0178)	(0.0178)
Intellectual Property Measures			
Trademark	4.102***	4.017***	4.055***
	(0.230)	(0.223)	(0.228)
Patent - Delaware Interaction			
Delaware Only	15.22***	15.58***	15.40***
	(0.336)	(0.343)	(0.340)
Patent Only	21.78***	22.10***	21.58***
	(1.018)	(1.034)	(1.011)
Patent and Delaware	90.91***	92.35***	92.55***
	(3.613)	(3.661)	(3.696)
US CMP Cluster Dummies			
Local	0.439***	0.438***	0.440***
	(0.0185)	(0.0184)	(0.0185)
Traded Resource Intensive	0.858***	0.853***	0.859***
	(0.0241)	(0.0240)	(0.0242)
Traded	1.038	1.042*	1.036
	(0.0210)	(0.0211)	(0.0210)
US CMP High-Tech Cluster Dummies			
Biotechnology	2.278***	2.239***	2.276***
	(0.164)	(0.160)	(0.164)
E-Commerce	1.311***	1.302***	1.305***
	(0.0436)	(0.0432)	(0.0434)
IT	2.127***	2.135***	2.115***
	(0.0760)	(0.0762)	(0.0756)
Medical Devices	1.303***	1.302***	1.301***
	(0.0516)	(0.0515)	(0.0515)
Semiconductors	1.546**	1.572**	1.537**
	(0.221)	(0.222)	(0.220)
Year FE	Yes	No	Yes
State FE	Yes	Yes	Yes
State Trends	No	Yes	Yes
N	18,764,856	18,764,856	18,764,856
pseudo R-sq	0.192	0.190	0.193

We repeat the main regression model of Table 3 but include year fixed effects, and state-specific time-trends, to evaluate the robustness of our findings. We perform other tests on the performance of our predictive model in our appendix. Robust standard errors in parenthesis. * p < .05 , ** p < .01, *** p < .001

Table A3. Entrepreneurial Quality Models with High Employment Growth Outcomes

	(1)	(2)	(3)
Dependent Variable	Equity Growth (IPO or Acquisition)	Employment > 500	Employment > 1000
Corporate Governance Measures			
Corporation	3.008***	1.542***	1.378***
	(0.0860)	(0.0681)	(0.103)
Name-Based Measures			
Short Name	2.248***	1.568***	1.279***
	(0.0514)	(0.0635)	(0.0883)
Eponymous	0.304***	0.675***	0.781
	(0.0197)	(0.0595)	(0.112)
Intellectual Property Measures			
Trademark	3.984***	7.194***	6.243***
	(0.268)	(0.750)	(1.053)
Delaware Only	14.01***	12.61***	13.43***
	(0.354)	(0.626)	(1.149)
Patent Only	20.83***	26.52***	46.79***
	(1.101)	(2.607)	(6.684)
Patent and Delaware	80.56***	95.87***	131.4***
	(3.645)	(9.064)	(19.86)
US CMP Cluster Dummies			
Local	0.418***	0.954	0.960
	(0.0202)	(0.0563)	(0.0962)
Traded Resource Intensive	0.831***	1.357***	1.297**
	(0.0268)	(0.0700)	(0.112)
Traded	0.418***	0.954	0.960
	(0.0202)	(0.0563)	(0.0962)
US CMP High-Tech Cluster Dummies			
Biotechnology	2.114***	0.828	0.339
	(0.181)	(0.194)	(0.198)
E-Commerce	1.316***	1.213**	0.972
	(0.0496)	(0.0858)	(0.123)
IT	2.185***	1.008	0.922
	(0.0872)	(0.0935)	(0.146)
Medical Devices	1.231***	1.214*	1.181
	(0.0556)	(0.109)	(0.183)
Semiconductors	1.585**	3.342***	2.639*
	(0.245)	(0.825)	(1.164)
N	12842817	12842817	12708349
pseudo R-sq	0.184	0.103	0.100

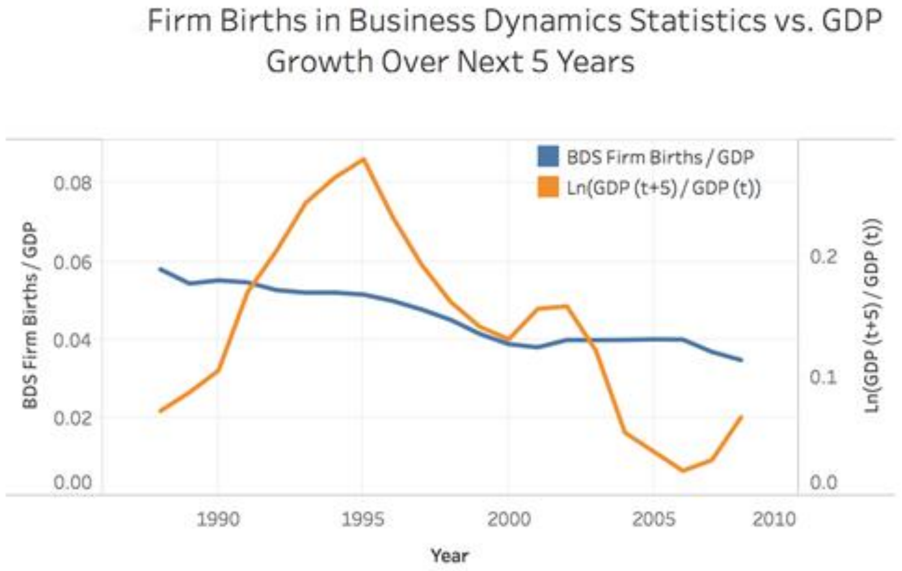
We develop models with the same regressor as our full information entrepreneurial quality model (Table 3, Column 5) but substitute high equity growth outcomes for high employment growth outcomes. Our outcome variable is 1 if a firm has high employment six years after founding and zero otherwise, at different thresholds. Employment measures are taken from the Infogroup USA panel data. We have a long-term project with the US Census to develop entrepreneurial quality estimates using continuous employment outcomes. Robust standard errors in parenthesis. * p < .05 , ** p < .01, *** p < .001

Table A4. Vector Autoregression Models (VAR) on the impact of changes in GDP Growth to Entrepreneurship

	VAR		SVAR		SVAR: US Recession	
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Ln(RECPI/GDP)	Ln(N/GDP)	Ln(RECPI/GDP)	Ln(N/GDP)	Ln(RECPI/GDP)	Ln(N/GDP)
Intersection ΔLn(GDP)(t)			0.024	0.005		
ΔLn(GDP)(t-1)	1.166	0.140	3.48**	-1.184		
	(0.80)	(0.561)	(1.08)	(0.62)		
ΔLn(GDP)(t-2)			1.08	-0.59		
			(1.07)	(0.59)		
ΔLn(GDP)(t-3)			-1.05	0.30		
			(0.85)	(0.50)		
Intersection 1[US Recession](t)					-0.0453	-0.011
1[US Recession](t-1)					-0.106*	0.059**
					(.06)	(.023)
1[US Recession](t-2)					-0.03	0.044*
					(.06)	(.025)
1[US Recession](t-3)					-0.01	0.033*
					(.038)	(.019)
Ln(RECPI/GDP)(t-1)	0.791***		0.208		0.477*	
	(0.105)		(0.22)		(.29)	
Ln(RECPI/GDP)(t-2)			-0.115		0.093	
			(0.241)		(.35)	
Ln(RECPI/GDP)(t-3)			0.678**		0.233	
			(0.229)		(.27)	
Ln(N/GDP)(t-1)		0.975***		1.27***		1.38***
		(0.0375)		(0.19)		(.20)
Ln(N/GDP)(t-2)				0.062		-0.029*
				(0.33)		(0.35)
Ln(N/GDP)(t-3)				-0.436		-0.42**
				(0.219)		(.211)

Notes: All models are run on a 27 observation time series representing each year observed in the data, from 1988 to 2014 (inclusive). VAR models are estimated through simultaneous equations, only the equation with GDP as a dependent variable is presented in the table. Three lag structure chosen as the optimal one using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). US Recession is a dummy variable equal to 1 if the year is 1992, 2001, 2008, or 2009. All regressions also pass VAR stability tests—the eigenvalues of all models lie within the unit circle. Standard errors in parenthesis. * p < .1 ** p < .05 *** p < .01

Figure A1



Business Dynamics Statistics downloaded from the US Census website (United States Census, 2017). GDP from Federal Reserve Economic Data (FRED) (St Louis Fed, 2017).

APPENDIX B[1]

Modeling Entrepreneurial Quality Through Governance Choices.

We begin our framework by developing a simple model to map early firm choices observable in business registration records to the underlying quality and potential of the firm. Our goal with this model is suggestive: its purpose is to provide clarity on the intuition through which we can use ex-ante firm choices and ex-post growth outcomes to measure underlying firm quality[2].

Suppose a firm has positive quality at birth, $q \in \mathbb{R}^+$. This quality creates firm value $V(q)$, a measure of the net-present value of its opportunities, which is also positive and increasing in q (i.e. $\frac{\partial V}{\partial q} > 0$). Both quality and value are unobservable to the analyst.

At birth, the firm must choose whether to use each of N independent binary governance options. These governance options reflect early choices that must be done around the birth of a firm such as whether to register as a corporation, whether to register locally or in Delaware, or the name of the firm[3]. The firm thus much choose a set $H = \{h_1, \ldots, h_N\}\,, h_i \in \{0,1\}\; \forall h_i$.

Each option offers benefit $b(q,h)$. The benefit is increasing in h, and the marginal benefit is also increasing in q. The option also has constant cost $c(h)$ plus an idiosyncratic component that is uncorrelated with quality and specific to this firm and option. This idiosyncratic component represents the different costs entrepreneurs could face due to heterogeneous preferences, institutional variation across corporate registries, local institutions (e.g. available financing), and firm characteristics (e.g. industry). Therefore:

$$\text{Benefit of h is } b(q,h)$$

[1] Appendix A is a set of 3 tables included in the main document.

[2] We model the governance decisions of firms in a sophisticated model in Guzman and Stern (mimeo).

[3] In this model, we focus on corporate governance options only, but the model naturally applies to other firm choices such as patenting, registering trademarks, and any other observable at birth.

$$\frac{\partial b}{\partial h} \geq 0\,, \frac{\partial^2 b}{\partial h \partial q} \geq 0$$

Cost of h is $C(q, h) = C(h) = c(h) + \epsilon$

$$E[\epsilon] = 0, \;\; E[\epsilon q] = 0$$

The Entrepreneur's Problem. The entrepreneur maximizes the value of the firm given the firm's quality, the available choices, and the idiosyncratic components:

$$H^* = \underset{H=\{h_1,\ldots,h_N\}}{\operatorname{argmax}} V(q) + \sum_{i=1}^{N} [b(q, h_i) - c(h_i) - \epsilon_i]$$

Since these choices are binary, the entrepreneur takes option h_i if $b(q, h_i) \geq c(h_i) + \epsilon_i$.

In this problem, for a given q and a given menu of governance choices, different values of H^* will occur. Since alx`l firms face the same set of options by assumption, the values of H^* will differ only due to q. Our goal is to understand what can be learned about true entrepreneurial quality q by looking at these choices.

Our first proposition studies how the value of H^* changes as q changes.

Proposition 1: $E[H^*]$ *is weakly increasing in* q.

Proof. First, note that the term $V(q)$ does not matter in the entrepreneur's problem, as it is constant given an original value of q. Therefore, the entrepreneur only maximizes $\sum_{i=1}^{N} [b(q, h_i) - c(h_i) - \epsilon_i]$, where the only terms that depend on q are $b(q, h_i)$. Since the marginal return to each h_i is increasing in q (i.e. $\frac{\partial^2 b}{\partial h \partial q} \geq 0$), then, for any two values $q'' > q'$, $P[b(q'', h_i) \geq c(h_i) + \epsilon_i] \geq P[b(q', h_i) \geq c(h_i) + \epsilon_i]$ which implies $E[H^*|q''] \geq E[H^*|q']$. QED

The relationship between H^* and q, in which the early entrepreneurial choices are determined in part by the firm quality, is the key insight on which we build our empirical approach.

Entrepreneurs must make choices early on, and they do so given their own potential and intentions for firm growth (their quality) as well as some idiosyncrasies. These choices, in turn, are observable in public records such as corporate registries, patent databases, or media, to name a few, and observing them for a firm can allow us to separate firms into different quality groups. To learn how we can do that we add more structure.

Firm growth outcomes. While the analyst cannot observe firm quality or value, we assume she is able to observe a growth outcome g, such as employment, IPO, or revenue with a lag. This growth outcome is more likely at higher values of $V(q)$, such that $E[g|q]$ is increasing in q, exhibiting first order stochastic dominance.

Since $E[H^*|q]$ is also increasing in q and is first order stochastic dominant, it follows from the transitivity of first order stochastic dominance (see Hadar and Russell, 1971) that $E[g|H^*]$ is also exhibits first order stochastic dominance in q.

Lemma 1 ($\mathbf{E[g|H^*]}$ is an increasing function of q)*: For any two $q'' > q'$, if $H^*(q)$ is a solution to the Entrepreneur's Problem, $E[g|H^*(q'')] \geq E[g|H^*(q')]$*

Proof. See above.

Now, consider a mapping f^{-1} which estimates the expected value of growth given H^*, $f^{-1}(g,H) \rightarrow \theta$. Then, if θ is the expected value of g given H^*, then $\hat{\theta}$ identifies a monotonic function of q.

Proposition 2 (Mapping g and H to Quality): *If a mapping $f^{-1}(g,H) \rightarrow \hat{\theta}$ is an estimate of $E[g|H^*]$, and H^* is a solution to the Entrepreneur's Problem, then $\hat{\theta}$ is a monotonic function of q.*

Proof: The proof is simple, since Lemma 1 shows that all mappings $E[g|H^*(q)]$ are monotonic in q, then if the value we use of H^* is a solution to the entrepreneur's maximization problem, then the values from function f^{-1} also need to be monotonic in q.

APPENDIX C

DATA APPENDIX

I. Overview of Data Appendix

This data appendix to the paper The State of American Entrepreneurship, by Jorge Guzman and Scott Stern, outlines in detail the use of business registration records in the United States, the steps and decisions we took when converting those records into measures for analysis, and robustness tests we ran to validate the potential for bias both due to specific assumptions about each measure as well as heterogeneity in our sample across geography and time. It serves the dual purpose of serving as an introduction for future users of business registration data while also providing detailed robustness verification and explaining the logic of specific decisions on many aspects of our data.

Section II of this appendix explains the development of our measures and dataset, including how we matched multiple datasets for analysis, how we built our measures using the merged dataset, and the economic rationale for the production of each one. Section III explains the differences between business registration records across the United States, their ease of access, and variation in the data they provide. It also highlights the potential for bias given the time when different data is observed (i.e. whether we observe the most recent value of a business or the original one) and performs numerous robustness tests to rule out the potential for bias driving our results given these differences. Section IV analyzes the potential for bias in our aggregate RECPI with a focus on guaranteeing that the predictive value of our indexes is high across geographies and time, and is not driven by a particularly large startup period (e.g. the dot-com bubble) nor driven by a particular area with many growth startups (e.g. Silicon Valley).

II. Using Business Registration Records to Find Signals of Quality

Our data set is drawn from the complete set of business registrants in thirty two states from 1988 to 2014. Our analysis draws on the complete population of firms satisfying one of the following conditions: (i) a for-profit firm whose jurisdiction is in the source state or (ii) a for-profit firm whose jurisdiction is in Delaware but whose principal office address is in the state. The resulting data set is composed of 27,976,477 observations. For each observation, we construct variables related to (i) the growth outcome for the startup, (ii) measures based on business registration observables and (iii) measures based on external observables that can be linked to the startup.

Growth outcome. The growth outcome utilized in this paper, *Growth*, is a dummy variable equal to 1 if the startup achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration. Both outcomes, IPO and acquisitions, are drawn from Thomson Reuters SDC Platinum[4]. Although the coverage of IPOs is likely to be nearly comprehensive, the SDC data set excludes some acquisitions. SDC captures their list of acquisitions by using over 200 news sources, SEC filings, trade publications, wires, and proprietary sources of investment banks, law firms, and other advisors (Churchwell, 2016). Barnes, Harp, and Oler (2014) compare the quality of the SDC data to acquisitions by public firms and find a 95% accuracy (Netter, Stegemoller, and Wintoki (2011), also perform a similar review). While we know this data not to be perfect, we believe it to have relatively good coverage of 'high value' acquisitions. We also note that none of the cited studies found significant false positives,

[4] Thomson Reuters's SDC Platinum is a commonly used database of financial information transferred to Refinitiv in 2018. More details are available at https://www.refinitiv.com .

suggesting that the only effect of the acquisitions we do not track will be an attenuation of our estimated coefficients.

We observe 13,406 positive growth outcomes for the 1988–2008 start-up cohorts), yielding a mean for *Growth* of 0.0007. In our main results, we assign acquisitions with an unrecorded acquisitions price as a positive growth outcome, because an evaluation of those deals suggests that most reported acquisitions were likely in excess of $5 million. We perform a series of robustness tests on different outcomes in the next section of this data appendix.

Start-up characteristics. The core of the empirical approach is to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures: (i) measures based on business registration observables and (ii) measures based on external indicators of start-up quality that are observable at or near the time of business registration. We review each of these in turn.

Measures based on business registration observables. We construct six measures of start-up quality based on information directly observable from the business registration record. First, we create binary measures related to how the firm is registered, including *corporation,* whether the firm is a corporation (rather than partnership or LLC) and *Delaware jurisdiction*, whether the firm is incorporated in Delaware. C*orporation* is an indicator equal to 1 if the firm is registered as a corporation and 0 if it is registered either as an LLC or partnership.[5] In the period of 1988 to 2008**,** 0.10% of corporations achieve a growth outcome versus only 0.03% of noncorporations. *Delaware jurisdiction* is equal to 1 if the firm is registered under Delaware, but has its main office in the source state (all other foreign firms are dropped before analysis). Delaware jurisdiction is favorable for firms which, due to more complex operations, require more certainty in corporate

[5] Previous research highlights performance differences between incorporated and unincorporated entrepreneurs (Levine and Rubinstein, 2013).

law, but it is associated with extra costs and time to establish and maintain two registrations. Between 1988 and 2008, 2.4% of the sample registers in Delaware; 37% of firms achieving a growth outcome do so.

Second, we create four measures that are based on the name of the firm, including a measure associated with whether the firm name is eponymous (named after the founder), is short or long, is associated with local industries (rather than traded), or is associated with a set of high-technology industry clusters.

Drawing on the recent work of Belenzon, Chatterji, and Daley (2017) (BCD), we use the firm and top manager name to establish whether the firm name is eponymous (i.e., named after one or more of the president, CEO, chairman, or managers (in the case of LLCs and partnerships)). *Eponymy* is equal to 1 if the first, middle, or last name of the top managers is part of the name of the firm itself.[6] We require names be at least four characters to reduce the likelihood of making errors from short names. Our results are robust to variations of the precise calculation of eponymy (e.g., names with a higher or lower number of minimum letters). We have also undertaken numerous checks to assess the robustness of our name matching algorithm. Not all states include the name of top managers[7]. Within those that do, 7.7% of the firms in our training sample are eponymous [an incidence rate similar to BCD], though only 2.4% for whom *Growth* equals one. It is useful to note that, while we draw on BCD to develop the role of eponymy as a useful start-up characteristic, our hypothesis is somewhat different than BCD: we hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Whereas BCD evaluates whether serial entrepreneurs are more likely to invest and grow companies which they name after

[6] For corporations, we consider top managers only the current president, for partnerships and LLCs, we allow for any of the two listed managers. The corporation president and two top partnership managers are listed in the business registration records themselves.

[7] These, and other, institutional differences are taken care of in our specifications through the inclusion of state fixed effects.in

themselves, we focus on the cross-sectional difference between firms with broad aspirations for growth (and so likely avoid naming the firm after the founders) versus less ambitious enterprises, such as family-owned "lifestyle" businesses.

Our second measure relates to the length of the firm name. Based on our review of naming patterns of growth-oriented start-ups versus the full business registration database, a striking feature of growth-oriented firms is that the vast majority of their names are at most two words (plus perhaps one additional word to capture organizational form (e.g., "Inc."). Companies such as Google or Spotify have sharp and distinctive names, whereas more traditional businesses often have long and descriptive names (e.g., "Green Valley Home Health Care & Hospice, Inc."). We define *short name* to be equal to one if the entire firm name has three of less words, and zero otherwise. 46% of firms within the 1988-2008 period have a short name, but the incidence rate among growth firms is more than 73%. We have also investigated a number of other variants (allowing more or less words, evaluating whether the name is "distinctive" (in the sense of being both non-eponymous and also not an English word). While these are promising areas for future research, we found that the three-word binary variable provides a useful measure for distinguishing entrepreneurial quality.

We then create four measures based on how the firm name reflects the industry or sector that the firm within which the firm is operating. To do so, we take advantage of two features of the US Cluster Mapping Project (Delgado, Porter, and Stern, 2016), which categorizes industries into (a) whether that industry is primarily local (demand is primarily within the region) versus traded (demand is across regions) and (b) among traded industries, a set of 51 traded clusters of industries that share complementarities and linkages. We augment the classification scheme from the US Cluster Mapping Project with the complete list of firm names and industry classifications contained in Reference USA, a business directory containing more than 10 million firm names and industry codes for companies across the United States. Using a random sample of 1.5 million Reference USA records, we create two indices for every word ever used in a firm name. The first of these indices measures the degree of localness, and is defined as the relative incidence of that word in firm names that are in local versus non-local industries (i.e., $\mathrm{\rho_i} = \frac{\sum_{\mathrm{j=\{local\ firms\}}} 1[\mathrm{w_i} \subseteq \mathrm{name_j}]}{\sum_{\mathrm{j=\{non-local\ firms\}}} 1[\mathrm{w_i} \subseteq \mathrm{name_j}]}$). We then define a list of Top Local Words, defined as those words that are (a) within the top quartile of ρ_i and (b) have an overall rate of incidence greater than 0.01% within the population of firms in local industries (see Guzman and Stern, (2015, Table S10) for the complete list). Finally, we define local to be equal to one for firms that have at least one of the Top Local Words in their name, and zero otherwise. We then undertake a similar exercise for the degree to which a firm name is associated with a traded name. It is important to note that there are firms which we cannot associate either with traded or local and thus leave out as a third category. Just more than 19% of firms have local names, though only 5% of firms for whom growth equals one, and while 54% of firms are associated with the traded sector, 59% of firms for whom growth equals one do.

We additionally examine the type of traded cluster a firm is associated with, focusing in particular on whether the firm is in a high-technology cluster or a cluster associated with resource intensive industries. For our high technology cluster group (Traded High Technology), we draw on firm names from industries include in ten USCMP clusters: Aerospace Vehicles, Analytical Instruments, Biopharmaceuticals, Downstream Chemical, Information Technology, Medical Devices, Metalworking Technology, Plastics, Production Technology and Heavy Machinery, and Upstream Chemical. From 1988 to 2008, while only 5% firms are associated with high technology, this rate increases to 16% within firms that achieve our growth outcome. For our resource intensive cluster group, we draw on firms names from fourteen USCMP clusters: Agricultural Inputs and Services, Coal Mining, Downstream Metal Products, Electric Power Generation and Transmission, Fishing and Fishing Products, Food Processing and Manufacturing, Jewelry and Precious Metals, Lighting and Electrical Equipment, Livestock Processing, Metal Mining, Nonmetal Mining, Oil and Gas Production and Transportation, Tobacco, Upstream Metal Manufacturing. While 14% of firms are associated with resource intensive industries, and 13% amongst growth firms.

Finally, we also repeat the same procedure to find firms associated with more narrow sets of clusters that have a closer linkage to growth entrepreneurship in the United States. We specifically focus on firms associated to Biotechnology, E-Commerce, Information Technology, Medical Devices and Semiconductors. It is important to note that these definitions are not exclusive and our algorithm could associate firms with more than one industry group. For Biotechnology (Biotechnology Sector), we use firm names associated with the US CMP Biopharmaceuticals cluster. While only 0.19% of firms are associated with Biotechnology, this number increases to 2.2% amongst growth firms. For E-commerce (E-Commerce Sector) we focus on firms associated with the Electronic and Catalog Shopping sub-cluster within the Distribution and Electronic Commerce cluster. And while 5% of all firms are associated with e-commerce, the rate is 9.3% for growth firms. For Information Technology (IT Sector), we focus on firms related to the USCMP cluster Information Technology and Analytical Instruments. 2.4% of all firms in our sample are associated with IT, and 12% of all growth firms are identified as IT-related. For Medical Devices (Medical Dev. Sector), we focus on firms associated with the Medical Devices cluster. We find that while 3% of all firms are in medical devices, this number increases to 9.6% within growth firms. Finally, for Semiconductors (Semiconductor Sector), we focus on the sub-cluster of Semiconductors within the Information Technology and Analytical Instruments cluster. Though only 0.04% of all firms are associated with semiconductors, 0.5% of growth firms are.

Measures based on external observables. We construct two measures related to start-up quality based on information in intellectual property data sources. Although this paper only measures external observables related to intellectual property, our approach can be utilized to measure other externally observable characteristics that may be related to entrepreneurial quality (e.g., measures related to the quality of the founding team listed in the business registration, or measures of early investments in scale (e.g., a Web presence).

Building on prior research matching business names to intellectual property (Balasubramanian and Sivadasan, 2010; Kerr and Fu, 2008), we rely on a name-matching algorithm connecting the firms in the business registration data to external data sources. Importantly, because we match only on firms located in California, and because firms names legally must be "unique" within each state's company registrar, we are able to have a reasonable level of confidence that any "exact match" by a matching procedure has indeed matched the same firm across two databases. In addition, our main results use "exact name matching" rather than "fuzzy matching"; in small-scale tests using a fuzzy matching approach [the Levenshtein edit distance (Levenshtein, 1965)], we found that fuzzy matching yielded a high rate of false positives due to the prevalence of similarly named but distinct firms (e.g., Capital Bank v. Capitol Bank, Pacificorp Inc v. Pacificare Inc.).

Our matching algorithm works in three steps.

First, we clean the firm name by:

- expanding eight common abbreviations (“Ctr.”, “Svc.”, “Co.”, “Inc.”, “Corp.”, “Univ.”, “Dept.”, “LLC.”) in a consistent way (e.g., “Corp.” to “Corporation”)
- removing the word “the” from all names
- replacing “associates” for “associate”
- deleting the following special characters from the name: . | ’ ” - @ _

Second, we create measures of the firm name with and without the organization type, and with and without spaces. We then match each external data source to each of these measures of the firm name. The online appendix contains all of the data and annotated code for this procedure.

This procedure yields two variables. Our first measure of intellectual property captures whether the firm is in the process of acquiring patent protection during its first year of activity. *Patent* is equal to 1 if the firm holds a patent application in the first year. All patent applications and patent application assignments are drawn from the Google U.S. Patent and Trademark Office (USPTO) Bulk Download archive. We use patent applications, rather than granted patents, because patents are granted with a lag and only applications are observable close to the data of founding. Note that we include both patent applications that were initially filed by another entity (e.g., an inventor or another firm), as well as patent applications filed by the newly founded firm. While only 0.2% of the firms in 1988–2008 have a first-year patent, 14% of growth firms do.

Our second intellectual property measure captures whether a firm registers a trademark during its first year of business activity. *Trademark* is equal to 1 if a firm applied for a trademark within the first year, and 0 otherwise. We build this measure from the Stata-ready trademark DTA file developed by the USPTO Office of Chief Economist (Graham et al, 2013). Between 1988 and 2008, 0.11% of all firms register a trademark, while 4.7% of growth firms do.

III. Observing Entrepreneurship Across States using Business Registration Records

III.A Business Registration Records State by State

While the act of registering a business is essentially the same across the United States, and carries basically the same benefits, corporation registries do vary in their internal operation across jurisdictions. While we have high confidence that firms register at the same point in their lifespan independent of state, the exact information we are able to get from each state is more nuanced. Business registration records vary in accessibility of the data, fields available, the exact definition and information within each field, and ease of use of data files. Each of these creates considerations in our use of business registration files, and has shaped the definition of our final sample.

Though business registration records are a public record, access to full datasets of registration records varies substantially in availability, cost and operational procedures required to get the files. In one end of the spectrum, we found several states that posted bulk data files publicly and allowed anonymous download of such files (Alaska, Florida, Washington, Wyoming, and Vermont). There was also another set of states for which access to these files required interfacing directly with the corporations office and filing some forms, but the procedure to access the data was relatively straightforward, and the costs where reasonable and appeared in line with a principle of trying to simply recuperate the costs of an administrative task (California, Massachusetts, Ohio, and others). There were other states that charged costs that we found higher than what would appear to be the appropriate to cover an administrative cost, and while we decided to pay for some of those in the low end (e.g. $1,250 for Texas) we avoided others that where substantially higher (e.g. $59,773.42 for New Jersey). Finally some states appeared to be outright evasive on fulfilling requests for data that is supposed to be public record, and suggested that either providing such data was impossible for them (e.g. Wisconsin) or deflected multiple attempts to contact individuals in their corporations division, through both phone and email, to ask for the records (e.g. Pennsylvania). In selecting our sample states, we tried to balance ease of access with economic importance, spending extra effort to get the top 5 by GDP (California, Texas, New York, Florida, and Illinous). We do note, however, that there did not appear to be any discernible pattern as to which states fell under different access regimes for their registration data. In prior work (Guzman and Stern, 2016) we have called on business registration offices to open access to such data.

The state corporations offices also vary in the fields that they provide or that can be generated from the information in their records. There were a number of fields which we were only able to get for a small number of states, such as date the firm becomes inactive (though most states record it, many where do not do consistently), firm industry, and stated mission of the firm, and as such decided not to use these fields in our national analysis even though their ability to explain growth seemed promising. There were also states that did not have fields that are important in our analysis and had to be dropped. In two cases (North Carolina and Ohio) we received the data from the corporations office but found they did not record the jurisdiction of foreign firms (firms registered in a different state), and we were unable to know which firms were from Delaware and which were from other states. We decided to drop these two states from our analysis. For two other states (New York and Washington) we found many firms had a missing address or had the address of their registered agent rather than the firm. We were able to keep these states for our national indexes, but unable to do any micro-geography analysis for them and included a caveat in our national map (note that state-level indexes are not affected by this issue since we do record the firm in the state correctly). Finally, not all states provided the current manager or president of the firm, and as such we were unable to estimate eponymy for all states and did not include it in the main prediction model.

The state corporation offices also differ in the exact specification of each field and only provided exactly equivalent fields for jurisdiction and registration date in all states. States vary, for example, in the specific set of corporate types that they allow. Specifically, only some states include an extra type of corporation or LLC for trade services (e.g. plumbing, law, etc) called a "Professional Corporation" or "Professional LLC". While a promising category, we are unable to take advantage of this extra categorization since it doesn't exist in all states, and instead only split into corporation and non-corporation firms in our analysis. Within corporations, the share of firms that registers a corporation changed through time due to the introduction of the LLC. LLC as a legal form was introduced at different times in different states, and in some states the introduction occurs within our sample years (for example, it was introduced in Massachusetts in 1995). As such, the role of corporations varies across years with the main effect being adverse selection of low-quality firms that would have registered as LLC but are instead corporations in the early years. We view this as a bias that only works against our results and do not control for it. We are also unable to differentiate between S-Corporations and C-Corporations since those are tax statuses rather than legal forms, and corporations can change from one to the other year to year. Finally, while non-profit status is also a tax status (e.g. as a 501(c) organization), all states also allow firms to registered specifically as a non-profit corporation and we are hence able to drop these firms (and the related benefit corporations, cemetery corporations, religious corporations, and trusts) directly through registration data before our analysis.

States also vary in the firm name information they provide. Only some states provided the list of all names an entity has had (e.g. Massachusetts and Texas). For those states, we are able to recover the original name of the firm and use such name when matching to intellectual property records and when creating our name-based measures. In cases where we did not have the original name, we used instead the current (provided) name. Only one state (Massachusetts) provides information to recover the original address of firms, and only for a subsample, while all other states only provide the current firm address. We investigate the possibility of any bias that could incur in our analysis by using the current address and firm name, rather than original ones, in the next section. Furthermore, states only provide the name of the current president or manager, and not the original firm founding, an issue we also evaluate in the next section.

Finally, states also vary in the ease of use of the data they provide, and no two states provide the data in the same format. Some states provide simple comma-delimited files that are easy to import in Stata, or fixed-length fields that can be imported through a Stata dictionary, while other states provide lists of transaction records that then need to be pre-processed through scripts that then produce the files that can be added to Stata.

III.B Estimating Potential Biases from Changes in Firm Location.

A main concern in our analysis is the potential of bias from changes in firm location. The data we receive from business registries holds the *current* location of the firm, but our goal in understanding entrepreneurial quality geography is to understand the *initial* location of the firm. (Importantly this does not impact our firm-level quality estimates, and hence we can analyze variation across different unbiased ex-ante quality levels of firms.) Firms are likely to move for many reasons. Ex-ante better firms might be more likely to start close to the center of an entrepreneurial cluster as it might have more value for the local externalities and move out of high potential clusters if unsuccessful, while ex-post successful firms (with lower quality ex-ante) might be more likely to move into such clusters. The potential direction and effect of this bias is in principle unclear.

While we are unable to study the extent of this bias in all states, we are able to perform a sub-sample study in Massachusetts. Using Massachusetts offers several important benefits that support the robustness of any forthcoming conclusions. First, our samples are beneficial: We are able to obtain two samples in Massachusetts that are almost exactly two years apart (one from January 06, 2013 (Commonwealth of Massachusetts, 2013), and one from November 24, 2014 (Commonwealth of Massachusetts, 2014)); furthermore, a sample from January 2013 provides the earliest possible snapshot that includes all 2012 firms (the most recent firms for which we estimate our full quality model, and the data we use for our full US snapshot), and hence includes the address in the firm's actual registration. Second, Massachusetts requires firms to update their address (among other things) in a yearly annual report guaranteeing we observe the new address for all firms that move. In other states, such annual report is not necessary. If a firm doesn't report its new address, we would continue to observe the original business address even after it moves, and our analysis will hold no bias. And third, the period we consider is a period in which there is considerable geographic migration of high-quality firms within Massachusetts, from Route 128 to the Cambridge and Boston area (see Guzman and Stern, 2017 for further details). Each of these details guarantees that our estimate is most likely to be an upper bound, and the extent of bias identified in this analysis is, if anything, likely to be lower in our national sample.

For this analysis, given that the ZIP Code is the smallest unit of geographic measurement that we use in this paper, we focus all of our analysis in ZIP Code level variation[8]. First, for each firm, we keep their 2013 ZIP Code (observed in January 06, 2013) their 2015 ZIP Code (observed

[8] This also helps protect from noise that could occur from "fuzzy" address matching approaches rather than exact ZIP Code matching.

in November 24, 2014). We also geocode each ZIP Code to assess the distance of any geographic move and remove all firms that have an invalid ZIP Code (e.g. due to typos)[9]. Finally, we estimate the leave-self-out quality of each ZIP Code for each firm using the average quality of all firms from 1988-2012 in our sample period.

We begin by documenting the extent to which a firm changes location at all. Table B3 presents the rates of change in ZIP Code for each 2-year group in our data. The first column indicates the age of the firm in 2013, when we first observe it, and the second column the share of firms that stay in the same ZIP Code in the next two years for the group. These estimates are not conditional on survival, and thus capture the share of total firms that will change from one category to the next in the total sample (i.e. it controls for changes in survival probability), the quantity we are interested on. Firms under 4 years or less (at 2013) are most likely to change address, with a probability of change between 2.9% and 3.6%. This probability then drops quickly, and in the 26-year-old cohort the probability of change is only 0.3%. Because our measure implicitly also includes likelihood of survival at different cohorts, we can estimate the overall likelihood that a firm record will have a different address after N years by simply doing the running product of the probability of same ZIP Code (under the assumption the migration dynamics have been the same historically). Column 3 includes this result. For the cohort of 10-year-old firms, we estimate 95% of the records to still contain the original ZIP Code, and for 26-year-old firms we estimate this share at 88%. We repeat this exercise with only the top 10% of quality firms in the distribution. While the likelihood of change of ZIP Code for a high-quality firm is higher, even within this group, we estimate 89% of records still contain the original ZIP Code by 10 years and 81% by 26

[9] We consider all ZIP Codes we cannot geocode through the Google API to be invalid.

years. In unreported tests, we find the share of firms that move in the top 1% is not meaningfully higher than the top 10%.

In our paper, most of our micro-geography results are done based on spatial visualizations. We therefore would also like to know *how far* are the firms moving. If firms are moving to contiguous ZIP Codes around the same high-quality cluster, perhaps due to small relocations or even ZIP Code redistricting, then the impact of those moves on our maps is small. On the contrary, if they move over large distances, then the impact is large. Using geocodings for each ZIP Code we estimate the distance of each ZIP code to another. We find 25% of all firms move less than 4 miles (25th percentile is 3.5), 50% of all firm moves are on less than 8 miles (50th percentile is 7.2), and 90% of all moves are 30 miles or less (90th percentile is 28.7).

Finally, any firm movement across ZIP Codes can only bias our results if it is systematic. If the moves are instead random, then average ZIP Code quality (our measure) would be constant even after there is firm migration. We estimate the difference in ZIP Code quality before and after a firm move (ZIP Code quality is estimated using all firms in that ZIP Code in November 24, 2014, without the moving firm included in either the source of destination ZIP Codes), and present a histogram of this measure in Figure B1. This difference in ZIP Code quality has a mean and median both basically centered at zero, therefore suggesting these moves are unbiased.

As a final test, we investigate whether this difference can vary by firm quality or age – i.e. if firms of higher or lower quality (or age) can systematically move to higher or lower average quality ZIP Codes. To do so, we run an OLS regression of firm quality on difference in ZIP Code (both in natural logs to account for the substantial skewness in entrepreneurial quality measures and be able to interpret this as an elasticity). The coefficient is -0.007 with a p-value of .61 using robust standard errors and an R^2 of .0001. This effect is (basically) indistinguishable from zero.

We also regress log-age on difference in ZIP Code quality to get a coefficient of -.0014 with a p-value of .94 and R^2 of .000.

III.C Analyzing Other Potential Sources of Bias in the Use of Business Registration Records

We now turn to analyzing the potential for bias in our estimates due to the specific nature of our sample. We specifically comment on six specific areas where there exists the possibility of bias: the impact of unobserved name changes, the role of re-incorporations on our data, the impact of spin-offs vs new firms, changes of ownership, changes in firm location, and the role of subsidiaries as separate corporate entities. We review each one in turn.

Name changes. As mentioned in section I of this appendix, we receive the original name for only some states in our dataset and only the current name in the rest of the states. While changes in name that correlate to growth could bias the relationship between our name-based measures and growth, it is unlikely to bias our most important measures. Specifically, changes in name cannot impact firm legal type (corporations vs non-corporations) or firm jurisdiction (Delaware). Our name-matching algorithm to match patents and trademarks uses firm names and assumes that the name we use is the same name as in the patent. While this can result in bias, it is only a bias that would work against our results – since we look for patents around the registration date, we can have false negatives for firms where we are looking for the wrong (new) name in the patent record but the firm had a previous name, but false positives are much less likely. These governance and intellectual property measures are, in fact, the most important in our study, and we find the fact that they cannot be affected by name changes assuring. Perhaps a risk in using only original names in some states is that the rate of false negatives will change depending on states. In unreported robustness tests, we have found the variation in results from using always the final name for all states (and hence implicitly having the same bias for all states) to be immaterial for our results.

Change of Ownership. Our dataset differs from other datasets in what is a firm and how it changes depending on ownership. The Longitudinal Business Database is built using tax records from corporate entities. As such, establishments that change ownership might bias the sample in different way and users of this data take substantial care to make sure changes in ownership do not drive their results (e.g. see the data appendix of Decker, Haltiwanger, Jarmin, and Miranda, 2014). Our data is different. Changes in ownership do not affect the registered firm and, unless the firm is closed down and re-incorporated, changes in ownership do not change anything in registration records.

The potential for re-incorporations. We argue in our analysis that we identify the extent to which firms are born with different quality, which is observed to the entrepreneur. An alternative hypothesis would be that entrepreneurs change their firm type once they observe their potential, at which point they re-incorporate the firm differently (e.g. as a Delaware corporation). To study the possibility of this bias we take advantage of institutional details of the process through which firms re-incorporate to observe the instances when it occurs. When a low potential firm (e.g. a Massachusetts LLC) re-incorporates as a high-quality firm (e.g. a Delaware corporation), it is done in two steps. First, a new firm is registered under the high quality regime; then, the old firm is merged into the new firm so that the new firm holds the old firm's assets and other matters (note that it is not possible to just "convert" the firm among firm types without creating a new target firm).

Once again, we use our Massachusetts data, which also includes a list of all mergers that have occurred among registered firms and the date of each merger. Obviously, firms can merge for many reasons and re-incorporation is only one of them. We create a measure *Re-registration,* which is equal to 1 only when the target firm was registered close to the merger date (90 days window). The facts we identify are included in Table B4. We review each in turn.

We identify a total of 6,767 mergers where the target firm is in Massachusetts (we drop all other firms earlier in our data, including firms registered before 1988 and firms with domicile outside Massachusetts). Of those, 3,041 firms (44.94%) are re-registrations, which are 2,847 new firms (sometimes multiple firms merge into one), while the rest are not. This total is low relative to the total firms in our sample for Massachusetts, 518,921 firms, suggesting that at most 0.55% of firms can potentially have a bias. We identify 1,905 cases in which both the source and target are in our dataset, with the rest likely being firms either registered before 1988 or with a foreign domicile.

We now proceed by studying our five most significant variables in this transition: patent, trademark, Delaware Jurisdiction, Corporation). Our main goal is to understand the extent to which founders of low-quality firms might later on re-register as high quality firms. To do so, we estimate the number firms that "gain" each of these observables, where a "gain" means the source firm did not have the observable, but the new firm does (e.g. the source firm is not a Corporation but the new firm is). We also compare this number with the total number of firms with this measure equal to 1 in our Massachusetts sample. As can been seen in Table B5, in all cases, the share of firms that gain a positive observable is always less than 3%. In Delaware, the observable which might hold the most bias, only 0.84% of all Delaware firms are re-registrations of firms changing corporate form, while the other 99.2% is not.

III.D Robustness Tests on Variations of Growth Outcome

In this section, we document a number of robustness tests done on our main predictive model and variations of our growth outcome variable. Our goal in these tests is to guarantee our sample is not sensitive to specific sub-sample issues in our definition of growth, such that small variation in the growth criteria would lead to widely different results, and to validate that spurious correlations are not driving our estimates. Given our focus on predictive value of our early stage measures rather than causal inference, we will look at the difference in coefficient magnitudes when comparing other coefficients to this baseline model, rather than statistical significance. That is, we seek to know whether changing our definition of growth would lead to different spatial and time-based indexes of EQI, RECPI and REAI rather than understanding if the magnitude itself is equal to one another in a statistical sense. We present all regressions in Table B4, with column 1 presenting our baseline model, columns 2-5 presenting alternate robustness models, and columns 6-9 presenting the absolute percentage difference between the coefficients of the baseline model and the alternative model.

Model (1) is our existing full information model presented in Table (5), with growth defined as an IPO or acquisition within six years, which we include here as a baseline model.

In Models 2 and 3 we focus on increasing the threshold of growth for which we measure a firm as having achieved growth. In Model 2, we investigate whether our results could be driven by a large number of low-value exits that are sold at a loss for stockholders. We use a different growth measure that is equal to 1 only for IPOs and acquisitions with a recorded firm valuation of over $100 million dollars. The number of growth firms drops significantly from 13,406 growth firms to 1,378, a drop of 90%. Delaware Only and Patent and Delaware have the highest percentage difference, with the Delaware Only coefficient being 2.5 times higher than the baseline model and the Patent and Delaware coefficient being 3.5 times higher. Importantly, we highlight that our use of SDC Platinum as a source of acquisitions is likely to lead to a positive selection in our sample: SDC Platinum is already more likely to include transactions that are significant in value and less likely to represent mergers that are only a sell of small assets of a firm.

Model 3 increases our threshold of quality further and includes only IPOs. IPO outcomes represent the top-end of growth successes in our sample, and understanding if our dynamics hold in this set might prove a particularly useful regularity. The number of growth firms drops substantially to 1,477**,** a share that appears broadly in line with patterns of exit of venture backed events in Kaplan and Lerner (2010). We also drop our Corporation measure before running this regression since it is endogenous – all IPOs are necessarily corporations, as it is not possible for non-corporations to sell shares. Our coefficients exhibit more variation than those in Model 2, with the most notable differences in Patent measures and Delaware measures. Patent independently increases by 1.6 times, Trademark increase almost 1.1 times while the interaction term increases by 2.4 times. The importance of name based measures also increases, with firms with short names being 33% more likely to grow in the IPO model than the baseline model, as well as some sector measures, particularly an association to Traded industries, increases the likelihood of IPO by 24%, an association to Local industries (already a negative correlation to growth), which increases the likelihood of IPO by 52% relative to the baseline model, and being a biotechnology firm, which is 1.2 times more likely to grow relative to the baseline. Assuming IPO measures are a higher value version of our growth outcome, it would appear that the effect of our measures is even starker in this high value growth outcome compared to our main growth measure. This further supports our view that our measures relate to real outcomes where, if anything, we could have even larger variation in quality when selecting stricter growth measures.

Models 4 and 5 test for biases that could relate to the window of growth in 6 years rather than a longer number of years. Changing the number of years allows us to investigate potential differences in dynamics of firms depending on their observables and industry sector and investigate to what extent this could bias our results. In Model 4, we define growth as an IPO or acquisition within 9 years instead of 6 years. Given that the time-window is three years longer, we drop the last three years (2006-2008) in our training sample from this regression, since the full growth window will not have elapsed for those years. The number of growth firms in these years increases by 50% from 11,500 to 17,248 after excluding these extra years. This might appear to be lower than would be expected since the average years to IPO or close to six, but we note that growth outcomes are skewed and the median is much lower than six years. The largest variation in relative magnitude is for firm with Delaware Only measure, which are 21% more likely to grow than in the 6 year window, and for firms to be corporations which are 21% more likely to grow relative to baseline.

Finally, in Model 5 we use an unbounded IPO outcome that is equal to 1 if a firm ever has an IPO. We run this regression on our 1988-2008 sample, implicitly allowing the most recent firms at least 9 years to achieve such outcome. As in Model 3, we find looking at IPO growth basically makes our estimates starker and highlights the ability of our measures to correlate significantly to growth outcomes at the very top end.

Evaluating Entrepreneurial Quality Estimates

Even if our model has strong predictive capacity, another potential source of concern could be heterogeneity within subsamples. Specifically, if one state (California) holds a disproportionate number of growth outcomes, or if growth outcomes occur disproportionately on a small number of years (the late 1990s), it is possible that our model is mostly fitting that region or time-period but does not have the external validity to work outside of the training years and states. If so, our prediction of quality in future years would be poor even if such predictions are good in the sample years.

We begin testing the accuracy across states in Table B1. We perform three different tests. In Column 3, we estimate the share of state growth firms in the top 5% of the state quality distribution using our 30% training sample. All states appear to separate growth firms in a within a small percentage at the top of the distribution[10]. The share of firms in the top 5% is highest in Massachusetts (67%), and Colorado (60%) and lowest in North Carolina (21%); California (53%) is only around the median, and there does not appear to be a discernible relationship between this statistic and the distribution of venture capital or high technology clusters. Our second test evaluates to what extent do our observables characterize the growth process in a region. To do so, we re-run our full information model (Model 1 of Table 4) separately for each state and calculate the pseudo-R^2 of each model. Once again, variation in this measure appears to be stable, with our measures having important relationship to growth outcomes in all states. Finally, we measure the relationship between entrepreneurial quality estimated from these states' specific models to our global quality measure. In column 5 we report the correlation between the two. [11] All correlation

[10] We are unable to estimate this measure for Alaska, Vermont and Wyoming due to the low number of growth firms that the states have.

[11] Another potential approach to test the difference in predictive measures between quality estimated with a state and national model would be to look at the distribution of the difference between these two measures ($d_i = \theta_{i,state} - \theta_i$ and test for H_0 : $d_i = 0$. However, because the state model implicitly includes a state fixed effect this would counfound

measures are high, with the highest one being in Texas (.973) and the lowest in North Carolina (.528), all other states are between .598 and .970. In conclusion, while there is variation in state performance each of these three tests, we find our estimate of quality with a national index to hold good predictive capacity at the state level.

We repeat the same three tests for each year in Table B2. The robustness of our model across years appears to be even higher than the robustness across states. The share of top 5% varies from 40% to 60%. Interestingly both the best predictive accuracy (share in top 5%) and the best fit between our observables and growth do not occur in the late 1990s but in the years 2005 to 2008. Both the stability across a long period of time and the fact that this accuracy appears to be improving gives us confidence in the quality of our predictions in the years following 2008, where growth is unobserved.

quality and ecosystem effects.

APPENDIX D
INFOGROUP SAMPLE

Our paper, though mostly focused on efforts to study the relationship of entrepreneurship to economic growth, also considers a section on the possibility of achieving High Growth employment outcomes. To study this question, we use data from Infogroup USA to estimate predicted employment six years after founding.[12] Infogroup USA is a database of local businesses which is sold for marketing and research purposes (similar to Dunn and Bradstreet). The dataset was originally created by collecting all firms who advertised in the Yellow Pages, but quickly moved to include other ways of capturing firms. The data is a list of over 10 million establishments and includes the name of the establishment, the address of the establishment, the parent establishment (if any), the industry code, and the estimated employment and sales (inclusive of children establishment for parents), as estimated by Infogroup.

We received annual snapshots of the Infogroup USA database, purchased by MIT Libraries, for the years 1997 to 2014. We deleted all child establishments and kept only 'headquarter' locations. Using the name-based matching algorithm that we used to match all other datasets, we matched each of our firms to the sample of firms in the file six years ahead and in the same state. If a firm is found, and their employment level is above 500, then we record a variable *Employment Growth 500* as 1, else, we give it the value of 0. We repeat this exercise for the threshold of 1000 employees.

[12] This data was received through MIT Libraries and is available in the following link. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GW2P3G Non MIT affiliates can contact ReferenceUSA directly at http://resource.referenceusa.com/contact-us/

REFERENCES

N. Balasubramanian, J. Sivadasan, (2010) “NBER Patent Data-BR Bridge: User guide and technical documentation” *SSRN Working paper #1695013*

B. Barnes, N. Harp, D. Oler, (2014) “Evaluating the SDC mergers and acquisitions database” *The Financial Review*. 49(4): 93-822.

Belenzon, Sharon, Chatterji, Aaron and Brendan Daley. 2017. “Eponymous Entrepreneurs” American Economic Review 107(6):1638-55, June 2017

C. Churchwell. (2016). “Q. SDC: M&A Database”. *Baker Library – Fast Answers.* Url: http://asklib.library.hbs.edu/faq/47760. (Accessed on January 17, 2017.)

Commonwealth of Massachusetts. 2013. *Business Registration Database.* Commonwealth of Massachusetts, Corporations Division. https://www.sec.state.ma.us/cor/coridx.htm (Received: January 06, 2013).

Commonwealth of Massachusetts. 2014. *Business Registration Database.* Commonwealth of Massachusetts, Corporations Division. https://www.sec.state.ma.us/cor/coridx.htm (Received: November 24, 2014).

M. Delgado, M. Porter, S. Stern, (2016) “Defining clusters in related industries” *Journal of Economic Geography*. 16 (1): 1-38

S. Graham, G. Hancock, A. Marco, A. F. Myers, (2013) “The USPTO case files data set: Descriptions, lessons and insights” *SSRN Working Paper #2188621*

W. R. Kerr, Shihe Fu, (2008) "The Survey of Industrial R&D--Patent Database Link Project." *J. Technol. Transf.* 33, no. 2

V. I. Levenshtein, (1965) "Binary codes capable of correcting deletions, insertions, and reversals." *Doklady Akad. Nauk SSSR* 163(4): 845–848

R. Levine, Y. Rubinstein, (2013) “Smart and illicit: Who becomes an entrepreneur and does it pay?” *NBER Working Paper #19276*

J. Netter, M. Stegemoller, and M. B. Wintoki. (2011) “Implications of Data Screens on Merger and Acquisition Analysis: A Large Sample Study of Mergers and Acquisitions from 1992 to 2009” *The Review of Financial Studies.* 24 (7): 2316–2357.

St. Louis Fed. 2017. “Real Gross Domestic Product” *Federal Reserve Economic Data (FRED).* https://fred.stlouisfed.org/series/GDPCA (Accessed on July of 2017)

ReferenceUSA. 2014. “ReferenceUSA Business Historical Data Files.” Harvard Dataverse. https://doi.org/10.7910/DVN/GW2P3G (Accessed on the summer of 2016)

United States Census. 2017. “Legacy Firm Characteristics Tables 1977-2014”. *Business Dynamics Statistics*. https://www.census.gov/programs-surveys/bds.html. United States Census Bureau (Accessed on July of 2017).

TABLE B1

Goodness of Fit Measures of Entrepreneurial Quality Model Across States

State	Total Growth Events	(1) Median of: share in top 5%	(2) Median of: share in top 10%	(3) Correlation of State Model and National Model
Alaska	1	-	-	-
Arkansas	45	42.9%	50.0%	63.8%
Arizona	95	50.0%	62.5%	87.8%
California	4166	52.6%	58.1%	94.7%
Colorado	49	60.0%	71.4%	65.9%
Florida	1148	30.7%	37.1%	91.5%
Georgia	475	38.8%	44.2%	96.0%
Iowa	60	50.0%	50.0%	79.6%
Idaho	43	33.3%	50.0%	91.6%
Illinois	332	54.5%	57.6%	88.6%
Kentucky	111	45.5%	45.5%	82.3%
Massachusetts	1069	67.3%	73.8%	94.9%
Maine	22	-	-	-
Michigan	176	22.7%	29.4%	84.9%
Minnesota	298	51.5%	53.3%	94.7%
Missouri	134	38.5%	50.0%	84.5%
North Carolina	185	21.4%	26.9%	52.8%
New Jersey	431	53.7%	63.6%	94.6%
New Mexico	29	29.2%	36.7%	84.1%
New York	883	57.3%	59.3%	97.0%
Ohio	344	38.7%	45.5%	93.6%
Oklahoma	109	42.9%	42.9%	85.2%
Oregon	218	42.1%	50.0%	94.5%
Rhode Island	3	-	-	-
South Carolina	103	38.5%	50.0%	59.8%
Tennessee	177	42.9%	47.6%	91.6%
Texas	1785	42.3%	51.3%	97.3%
Utah	220	45.0%	55.6%	94.9%
Virginia	212	47.8%	61.9%	93.5%
Vermont	17	-	-	-
Washington	326	47.5%	53.8%	91.5%
Wisconsin	140	30.0%	45.5%	78.2%
Average		43.5%	50.8%	86.0%

This table performs two goodness-of-fit estimates for entrepreneurial quality measures across states. Columns 1 and 2 repeat our out of sample 10-fold cross validation process (Figure 2) across each state. Specifically, it estimates the share of out of sample growth firms who are in the top 5% and 10% of the state's entrepreneurial quality distribution, for 10 different random out of sample, samples. The median of this estimate is reported. Column 3 reports the correlation between the quality measures and a second quality estimate, built only with data of each state independently. States with 10 growth events are not included.

TABLE B2

Quality of Predictive Algorithm By Cohort (70% Test Sample)

Cohort Year	(1) Total Growth Firms in Test Sample	(2) Share of Growth Firms Top 10% of Sample	(3) Share of Growth Firms Top 5% of Test Sample	(4) Share of Growth Firms Top 1% of Test Sample	(5) Correlation with Single Year Quality
1988	239	55%	42%	30%	0.87
1989	225	61%	47%	32%	0.94
1990	273	52%	43%	26%	0.94
1991	320	55%	44%	24%	0.89
1992	373	56%	43%	28%	0.95
1993	404	54%	40%	26%	0.94
1994	442	54%	45%	25%	0.93
1995	557	53%	40%	22%	0.94
1996	612	63%	50%	26%	0.93
1997	528	65%	52%	32%	0.93
1998	534	67%	56%	36%	0.96
1999	648	68%	58%	37%	0.89
2000	627	70%	60%	46%	0.96
2001	451	60%	49%	38%	0.96
2002	442	61%	49%	35%	0.94
2003	493	60%	50%	37%	0.93
2004	452	64%	53%	42%	0.95
2005	428	65%	56%	41%	0.94
2006	493	64%	56%	44%	0.93
2007	409	65%	59%	47%	0.95
2008	433	64%	56%	46%	0.97

We run our main Full Information model on a random 30% of our data, and predict the other 70%. The results above reflect the distribution of the growth firms in this 70% test sample when sorted by predicted quality.

TABLE B3

Test of changes of address using a Massachusetts subsample

P(Address Change) by Age

	All Firms		*Top 10% of Quality*	
Lifespan	P(Address Change) in Two Years	Lifetime Probability	P(Address Change) in Two Years	Lifetime Probability
0-2	3.6%	96.4%	6.2%	93.8%
2-4	2.9%	89.5%	5.2%	83.5%
4-6	2.0%	86.3%	3.7%	77.6%
6-8	1.5%	84.8%	2.2%	75.4%
8-10	1.2%	83.6%	1.8%	73.6%
10-12	1.0%	95.4%	1.3%	92.4%
12-14	1.0%	94.5%	1.4%	91.1%
14-16	0.8%	93.7%	1.0%	90.0%
16-18	0.8%	92.9%	0.7%	89.4%
18-20	0.6%	92.3%	0.6%	88.7%
20-22	0.5%	89.0%	0.6%	82.9%
22-24	0.4%	88.6%	0.9%	82.0%
24-26	0.3%	88.3%	0.7%	81.3%

Cohort of Age 0 is the 2012 Cohort

Lifetime probability of address change is the implied probability of changing address for a firm

TABLE B4
Other implementations of our model

We estimate a logit model with *Growth* as the dependent variable, under different definitions of Growth. These models are estimated with an earlier sample of 32 US states Incidence ratios reported; Robust standard errors in parenthesis.

	Models					*Share Difference with Baseline*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Growth (Only				Growth	IPO	Growth	
	Original	Acq >=	IPO in 6	Growth in 9	IPO	(Only Acq	in 6	in 9	IPO
	Regression	100M)	Years	Years	(Ever)	>= 100M)	Years	Years	(Ever)
Short Name	2.263***	2.666***	3.018***	2.089***	3.286***	18%	33%	8%	45%
	(0.0456)	(0.190)	(0.195)	(0.0361)	(0.171)				
Eponymous	0.315***	0.182***	0.276***	0.339***	0.261***	42%	12%	8%	17%
	(0.0179)	(0.0531)	(0.0560)	(0.0160)	(0.0420)				
Corporation	3.061***	8.525***		2.453***		179%		20%	
	(0.0739)	(0.942)		(0.0529)					
Trademark	3.964***	3.749***	3.322***	4.246***	3.489***	5%	16%	7%	12%
	(0.219)	(0.406)	(0.407)	(0.245)	(0.351)				
Patent Only	22.77***	71.39***	59.71***	18.65***	51.96***	214%	162%	18%	128%
	(1.059)	(9.009)	(6.879)	(0.801)	(4.627)				
Delaware Only	15.18***	52.53***	31.65***	12.04***	25.32***	246%	108%	21%	67%
	(0.335)	(3.900)	(2.147)	(0.243)	(1.324)				
Patent and Delaware	84.08***	381.6***	284.8***	71.83***	216.1***	354%	239%	15%	157%
	(3.320)	(38.08)	(27.95)	(2.796)	(16.81)				
Local	0.434***	0.549***	0.660***	0.441***	0.589***	26%	52%	2%	36%
	(0.0182)	(0.0814)	(0.0824)	(0.0156)	(0.0593)				
Traded Resource Intensive	0.868***	0.935	1.319***	0.857***	1.312***	8%	52%	1%	51%
	(0.0243)	(0.0808)	(0.0974)	(0.0208)	(0.0774)				
Traded	1.048*	1.020	1.304***	1.141***	1.215***	3%	24%	9%	16%
	(0.0212)	(0.0653)	(0.0804)	(0.0204)	(0.0585)				
Biotech Sector	2.173***	2.429***	4.697***	2.269***	5.595***	12%	116%	4%	157%
	(0.155)	(0.368)	(0.604)	(0.157)	(0.565)				
Ecommerce Sector	1.348***	1.134	1.158	1.265***	1.272**	16%	14%	6%	6%
	(0.0446)	(0.113)	(0.110)	(0.0368)	(0.0949)				
IT Sector	2.175***	1.535***	1.714***	2.035***	1.723***	29%	21%	6%	21%
	(0.0779)	(0.166)	(0.178)	(0.0642)	(0.143)				
Medical Dev. Sector	1.301***	0.975	1.123	1.271***	1.126	25%	14%	2%	13%
	(0.0515)	(0.119)	(0.124)	(0.0449)	(0.0982)				
Semiconductor Sector	1.590***	2.248**	1.195	1.862***	2.450***	41%	25%	17%	54%
	(0.224)	(0.614)	(0.445)	(0.238)	(0.583)				
State Fixed Effects	Yes	Yes	Yes	Yes	Yes				
Observations	18764856	18613648	18681641	14214629	18707865				
Pseudo R-squared	0.187	0.306	0.240	0.155	0.235				

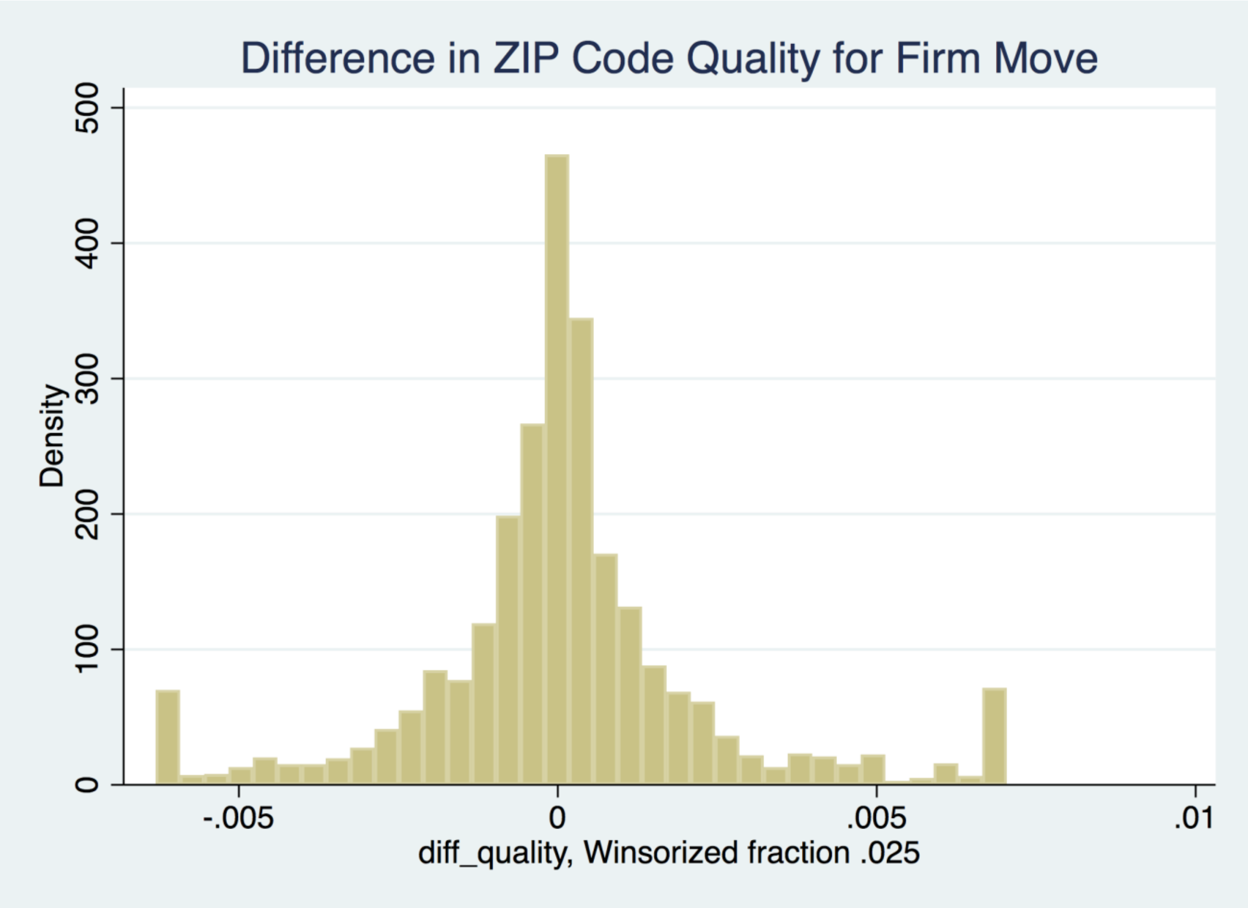
TABLE B5

Re-Registrations in Massachusetts

General Statistics	
Total Massachusetts Firms in Sample	518,921
Firms founded through a re-registration	2,847
Share of Firms Founded through re-registration	0.55%
Re-incorporations with source and destination firm in sample	1,905
Corporations	
Firms that Gain Corporation = 1	573
Total Corporations in Sample	310,061
Share	0.18%
Delaware Jurisdiction	
Firms that Gain Delaware = 1	259
Total Delaware Firms in Sample	30,781
Share	0.84%
Patents	
Firms that Gain Patent = 1	51
Total Patent Firms in Sample	2,670
Share	1.91%
Trademark	
Firms that Gain Trademark = 1	36
Total Trademark Firms in Sample	1,463
Share	2.46%
Short Name	
Firms that Gain Short Name = 1	234
Total Short Name Firms in Sample	250,212
Share	0.09%

A firm is coded as gaining an observable if the source firm of the re-registration did not have such observable at birth but the new firm does.

FIGURE B1



Difference in ZIP Code Quality for Firm Move
Density
0
100
200
300
400
500
-.005
0
.005
.01
diff_quality, Winsorized fraction .025