NBER WORKING PAPER SERIES

# THE STATE OF AMERICAN ENTREPRENEURSHIP: NEW ESTIMATES OF THE QUALITY AND QUANTITY OF ENTREPRENEURSHIP FOR 32 US STATES, 1988-2014

Jorge Guzman
Scott Stern

Working Paper 22095
http://www.nber.org/papers/w22095

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2016, Revised July 2019

Previously circulated as "The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 15 US States, 1988-2014." We are thankful for comments and suggestions by Marianne Bertrand, Erik Brynjolffson, Ankur Chavda, Catherine Fazio, Joshua Gans, John Haltiwanger, Bill Kerr, Fiona Murray, Abhishek Nagaraj, Roberto Rigobon, David Robinson, and Hal Varian, as well as seminar and conference participants at Duke University, Harvard Business School, University of Toronto, the University of Virginia, the Kauffman Foundation New Entrepreneurial Growth Conference, and the NBER Pre-Conference on Entrepreneurship and Economic Growth, and to four anonymous referees. We also thank Open Corporates for providing data for New York and Michigan, and RJ Andrews for his development of the visualization approach. Sarah Andries, Jintao Chen, Ji Seok Kim, and Yupeng Liu provided excellent research assistance. Finally, we acknowledge and thank the Jean Hammond (1986) and Michael Krasner (1974) Entrepreneurship Fund and the Edward B. Roberts (1957) Entrepreneurship Fund at MIT, and the Kauffman Foundation for financial support. All errors and omissions are of course our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w22095.ack

The State of American Entrepreneurship: New Estimates of the Quality and Quantity of Entrepreneurship for 32 US States, 1988-2014
Jorge Guzman and Scott Stern
NBER Working Paper No. 22095
March 2016, Revised July 2019
JEL No. C53,L26,O51

## ABSTRACT

Assessing the state of American entrepreneurship requires not simply counting the quantity but also the initial quality of new ventures. Combining comprehensive business registries and predictive analytics, we present estimates of entrepreneurial quantity and quality from 1988-2014. Rather than a secular pattern of declining business dynamism, our quality-adjusted measures follow a cyclical pattern sensitive to economic and capital market conditions. Consistent with the role of investment cycles as a driver of high-growth entrepreneurship, our results highlight the role of economic and institutional conditions as a driver of both initial entrepreneurial quality and the scaling of new ventures over time.

Jorge Guzman
Columbia Business School
Uris Hall, 711
116th St & Broadway
New York, NY 10027
jag2367@gsb.columbia.edu

Scott Stern
MIT Sloan School of Management
100 Main Street, E62-476
Cambridge, MA 02142
and NBER
sstern@mit.edu

Over the past two decades, economists have made significant progress in advancing the measurement of entrepreneurship. The pioneering studies of Haltiwanger and co-authors (Davis and Haltiwanger, 1992; Davis et al, 1996; Haltiwanger et al, 2013; Decker et al, 2014) moved attention away from simply counting the density of *small and medium sized firms* towards the measurement and growth dynamics of *young firms*. These studies established that a disproportionate share of new job creation is associated with new firms, and economic growth is grounded in business dynamics. A separate stream of research focusing on selective samples of high-performance entrepreneurial ventures and the institutions that surround them reinforce this perspective. For example, Kortum and Lerner (2000) find that venture capital is associated with higher levels of innovation, and Samila and Sorenson (2011) find that venture capital has a positive impact on aggregate income, employment, and new establishment formation.

Notwithstanding these advances, there is increasing recognition that the relationship between entrepreneurship and economic growth depends not simply on the quantity but also on the underlying *quality* of new firms (Schoar, 2010; Hurst and Pugsley, 2011). While systematic population-level indices of the quantity of entrepreneurial activity (such as the Business Dynamics Statistics database, hereafter BDS) document a secular decline in the rate of business dynamism and the "aging" of US private sector establishments (Hathaway and Litan, 2014a, 2014b, 2014c), researchers focused on venture capital and high-growth firms have documented a sizeable increase after the Great Recession in the funding of growth-oriented entrepreneurial ventures (Gornall and Strebulaev, 2015).

To put these differences in perspective, consider the gap between the rate (relative to GDP) of firm births per year as measured by the Business Dynamics Statistics versus the rate (relative to GDP) of successful growth firms founded in a particular year (i.e., the number of firms founded in a given year that achieved an IPO or significant acquisition within six years of initial business registration), for the 32 states which will form the basis of our analysis. While the BDS shows a slow and steady decline of approximately 40% (consistent with Hathaway and Litan (2014a)), the realization of growth experienced a much sharper up-and-down cycle, with 1996 representing the most successful start-up cohort in US history, followed by a relatively stable level from 2001 to 2008.[2] Moreover, while it has long been known that the growth consequences of start-up activity

[2] This divergence is reinforced by comparing BDS firm births and economic growth. While the BDS has little cyclical variation (and is on a downward decline), GDP growth is far more variable with a sharp upward trend through the 1990s and a downward decline over the subsequent

are concentrated in the outcomes of a very small fraction of the most successful firms (Kerr, Nanda, and Rhodes-Kropf, 2014), prior attempts to use population-level data to characterize the rate of entrepreneurship have largely abstracted away from initial differences across firms in the ambitions of their founders or their inherent growth potential.[3]

Simply put, alternative definitions of entrepreneurship suggest different assessments of the state of American entrepreneurship.

Not simply a matter of measurement, characterizing entrepreneurial quality by cohort allows for the empirical assessment of important economic and policy questions. For example, in line with the debt deflation theory suggested by Fisher (1933), Bernanke and Gertler (1989) suggest that high-quality entrepreneurship may be reduced during a recession due to structural financing constraints (while low-quality entrepreneurship may be unaffected). And, this reduction in growth-oriented entrepreneurship can exacerbate a downturn through reduced business dynamism. Assessing this theoretical claim empirically requires the development of consistent measures for the quantity and quality of entrepreneurship at founding and observing how these measures change at different points in the business cycle. From a policy perspective, fostering growth-oriented entrepreneurship may be fundamentally different than focusing on policies that enhance the environment for "Main Street" businesses (Aulet and Murray, 2011; Mills and McCarthy, 2014; Chatterji, 2018), and so being able to differentiate between new ventures in terms of their growth potential can offer more targeted and effective entrepreneurship policy.

Building on Guzman and Stern (2015; 2017), this paper develops and implements a novel approach to the measurement of both the quantity and quality of entrepreneurship which we then use to provide substantive insight into both theoretical and policy questions.[4] Our approach to measuring entrepreneurial quality combines three interrelated insights. First, a practical requirement for any growth-oriented entrepreneur is business registration (as a corporation,

period. In recent work, Decker et al (2016) show that high growth in the High Tech sector (as defined by a collection of NAICS codes) has followed a more cyclical pattern than the Retail and Services sectors, which account for the bulk of the decline in new firms.

[3] The challenge is fundamentally a measurement problem: "The problem is that it is very difficult, if not impossible, to know at the time of founding whether or not firms are likely to survive and/or grow." (Hathaway and Litan, 2014b).

[4] In our earlier work, we undertook preliminary explorations of the approach that we develop in this paper. In Guzman and Stern (2015), we introduced the overall methodology in an exploratory way by examining regional clusters of entrepreneurship such as Silicon Valley at a given point in time. We then focused on a single US state (Massachusetts) to see if it was feasible to estimate entrepreneurial quality over time on a near real-time basis (Guzman and Stern, 2017). This paper builds on these earlier exercises to develop an analysis for 32 "representative" US states (comprising more than 80% of overall GDP) over a 27-year period, introduce new economic statistics that allow for the characterization of entrepreneurial quantity and quality over time and place, consider the relationship between alternative metrics of entrepreneurship and measures of economic performance, and consider the changing nature of regional entrepreneurship for selected metropolitan areas. Passages of text describing our methodology and approach, as well as the Data Appendix, draw upon these earlier papers (with significant revision for clarity and concision as appropriate).

partnership, or limited liability company). These public documents allow us to observe a "population" sample of entrepreneurs observed at a similar (and foundational) stage of the entrepreneurial process. Second, moving beyond simple counts of business registrants (Klapper, Amit, and Guillen, 2010), we are able to measure characteristics related to entrepreneurial quality *at or close to the time of registration.* These characteristics include how the firm is organized, how it is named, and how the idea behind the business is protected. These start-up characteristics may reflect choices by founders who perceive their venture to have high potential. In other words, though observed start-up characteristics are not causal drivers of start-up performance, they may nonetheless represent early-stage "digital signatures" of high-quality ventures. Third, we leverage the fact that, though rare, we observe meaningful growth outcomes for some firms, and are therefore able to estimate the relationship between these growth outcomes and start-up characteristics. This mapping allows us to form an estimate of entrepreneurial quality for any business registrant within our sample, even those in recent cohorts where a growth outcome (or lack thereof) has not yet had time to be observed.

We use this predictive analytics approach to propose three new statistics for the measurement of growth entrepreneurship: the Entrepreneurship Quality Index (EQI), the Regional Entrepreneurship Cohort Potential Index (RECPI), and the Regional Entrepreneurial Acceleration Index (REAI). EQI is a measure of *average quality* within any given group of firms, and allows for the calculation of the probability of a growth outcome for a firm within a specified population of start-ups. RECPI multiples EQI and the number of start-ups within a given geographical region (e.g., from a zip code or town to the entire five-state coverage of our sample), and so is a measure of the quality-adjusted quantity of entrepreneurship. Whereas EQI compares entrepreneurial quality across different groups, RECPI allows the direct calculation of the expected number of growth outcomes from a given start-up cohort within a given regional boundary. REAI, on the other hand, measures the ratio between the realized number of growth events for a given start-up cohort and the expected number of growth events for that cohort (i.e.., RECPI). REAI offers a measure of whether the "ecosystem" in which a start-up grows is conducive to growth (or not), and allows variation in ecosystem performance across time and at an arbitrary level of geographic granularity.

We calculate these measures on an annual basis for 32 U.S. states for the period 1988-2014. We document several key findings. First, in contrast to the secular and steady decline observed in

the BDS, RECPI / GDP has followed a cyclical pattern that seems sensitive to the capital market environment and overall economic conditions. Second, while the peak value of RECPI / GDP is recorded in 2000, the overall level during the first decade of the 2000s is actually *higher* than the level observed between 1990 and 1995, with an additional upward swing beginning in 2010.[5] Even after controlling for change in the overall size of the economy, the third highest level of entrepreneurial growth potential is registered in 2014. Finally, there is striking variation over time in the likelihood of start-up firms at a given quality level to realize their potential (REAI): REAI declined sharply in the late 1990s, and did not recover through 2008. While we focus on estimates of entrepreneurial quality based on a predictive model of equity growth outcomes (the achievement of an IPO or significant acquisition within six years of founding), these broad patterns of results also hold if one focuses on alternative definitions of equity growth (e.g., only focusing on IPOs) or alternative growth measures such as the realization of more than 500 employees within the first six years after founding.[6]

We use these measures to assess the long-standing theoretical debate regarding the relationship between entrepreneurship and the business cycle. While a long line of theoretical work emphasizes the potential for economic growth to stimulate and nurture the founding of new growth-oriented ventures, either by improving the balance sheet of investors (Fisher, 1933; Bernanke and Gertler, 1989; Carlstrom and Fuerst, 1997; Rampini, 2004), or the investor expectations of follow-on capital availability (Caballero, Farhi, and Hammour, 2004; Nanda and Rhodes-Kropf, 2013), others have emphasized the potential for a "cleansing effect" of recessions, whereby downturns instead enhance the potential of the firms that are founded ventures at that time (Schumpeter, 1939; Cabarello and Hammour, 1994). A key insight of this theoretical literature is that economic conditions shape the aggregate entrepreneurial potential of start-up cohorts and the overall incidence of high growth start-ups, but not the total number of start-ups founded for each cohort. We use our measurement of both quantity and initial entrepreneurial quality to test this hypothesis directly. Using a structural vector autoregression (SVAR) model,

[5] We use a "nowcasting" index for the most recent cohorts which only use start-up characteristics available within the business registration data, and compare that index to an "enriched" index which captures events that might occur early within the life of a start-up such as the initial receipt of intellectual property.

[6] Our employment growth results are based on a private sector data source, Infogroup USA. While we highlight the robustness of our core findings to this potentially noisy measure of employment growth in this paper, we do not undertake a systematic assessment of employment-oriented growth outcomes which could more naturally be conducted with a comprehensive administrative dataset such as the LBD.

we provide the first evidence that quality-adjusted quantity (RECPI) is procyclical while quantity is unrelated to economic conditions.

Our approach of course comes with important limitations and caveats. First, and most importantly, we strongly caution against a causal interpretation of the regressors we employ for our predictive analytics—while factors such as eponymy and the form of business registration are a "digital signature" that allows us to differentiate among firms in the aggregate, these are not meant to be interpreted as causal factors that lead to growth per se (i.e., simply registering their firm in Delaware is not going to directly enhance an individual firm's underlying growth potential). And, while we are encouraged by the robustness of our core approach across multiple states and time periods, we can easily imagine (and are actively working on identifying) additional firm-level measures which might allow for even more differentiation in quality, or account directly for changing patterns over time and space in the "drivers" of growth. Finally, though we show some robustness of our findings to the use of employment-oriented growth outcomes, a more complete assessment of the differences between equity growth outcomes and employment-oriented outcomes remains outstanding.

Keeping in mind these caveats, our findings nonetheless do offer a new perspective on the state of American entrepreneurship. Most importantly, our results highlight that the recent shift in attention towards young firms (pioneered by Haltiwanger and co-authors) is enriched by directly accounting for initial heterogeneity among new firms. Even within the same industry, there is significant heterogeneity among new firms in their ambition and inherent potential for growth.[7] Policies that implicitly treat all firms as equally likely candidates for growth are likely to expect "too much" from the vast majority of firms with relatively low growth potential. Second, the striking decline in REAI after the boom period of the 1990s is the first independent evidence for an often-cited concern of practitioners: even as the number of new ideas and potential for innovation is increasing, there seems to be a reduction in the ability of companies to scale in a meaningful and systematic way.

Our approach holds promise for multiple areas of economics research and policy. For example, a predictive analytics approach to entrepreneurial quality allows for the assessment of the relative importance of passive versus proactive growth firms in the overall process of firm

[7] In recent work, Decker et al (2016) use business dynamics estimates to also document important variation in rates of dynamisms across industries.

growth, and the specific role of venture capital in enabling the growth of firms with high initial quality (Catalini et al, 2019). As well, it is possible to use our approach to examine the role of gender differences among founders in the process of attracting venture capital and overall firm growth (Guzman and Kacperczyk, 2019). And, it is possible to use this approach (which uses firm choices regardless of location) to assess the role of location (e.g., whether the firm is in Silicon Valley) in facilitating firm growth (Guzman and Stern, 2015), shaping the incentives to migrate between locations (Guzman, 2019), and in assessing the role of local institutions (such as universities) and policies (such as tax) in shaping the process of firm founding and growth (Stern and Tartari, 2018; Fazio, et al, 2019).

The rest of this paper is organized as follows. Section I provides an overview of entrepreneurial quality in economics and briefly outlines our theoretical intuition. Section II explains our methodology, and Section III our dataset and the estimation of entrepreneurial quality for our sample. Sections IV and V describe the variation in our key statistics across geography and time. Section VI compares the relationship of our index to an alternative measure of economic growth using employment outcomes. In section VII, we empirically test the relationship between GDP growth and changes in national entrepreneurship. Section VIII concludes.

## I. Entrepreneurial Quality: Do Initial Differences Matter?

Economists have long sought to understand the role of firm-specific characteristics in industry dynamics. Gibrat (1931) provides the foundational benchmark in this area: Gibrat's Law proposes that the growth rate of firms (and the variance in that growth rate) are independent of firm size (Sutton, 1997). Despite broad patterns consistent with Gibrat's Law, a large literature beginning with Mansfield (1962) instead emphasizes deviations from proportional growth. This literature first emphasized that smaller firms have both higher growth rates and lower probabilities of survival (Mansfield, 1962; Acs and Audretsch, 1988, among others), but over time additional research suggested that younger firms also had high average growth rates and lower survival probabilities (Evans, 1987; Dunne, Roberts, and Samuelson, 1988).[8]

[8] Not simply a set of empirical regularities, these findings formed the foundations for important theoretical work, notably Jovanovic (1982) and subsequent formal models of firm and industry dynamics (Ericson and Pakes, 1995; Klepper, 1996; Hopenhayn, 1992; Klette and Kortum, 2004).

Davis and Haltiwanger (1992) clarified this empirical debate by simultaneously considering the role of size and age, developing systematic evidence that virtually all net job creation was in fact due to younger firms (which are small because they are young) rather than smaller firms per se (Davis, Haltiwanger, and Shuh, 1996; Jarmin, Haltiwanger, and Miranda, 2013; Akcigit and Kerr, 2018). Building on these studies, Decker et al (2014) extend this approach to document an overall decline in the rate of new firms that have at least one employee, which the authors characterize as a reduction in the rate of business dynamism, with meaningful variation across industry groups (Decker et al, 2016).

However, the role of young firms in shaping job creation is not homogenous across the population of new firms. The vast majority of new firms are associated with no net new job growth, and consequently a very small fraction of new firms is disproportionately responsible for net new job growth. Using surveys and aggregate economic comparisons, some have suggested that these differences in growth are accounted for by underlying differences in the firms themselves (Hurst and Pugsley, 2011; Kaplan and Lerner, 2010; Schoar, 2010). Yet, beyond broad industry effects, systematic studies of firm dynamics have yet to incorporate such ex ante differences in a way that ties closely to the highly skewed ex post distribution of firm growth.

Accounting for this skew requires confronting a measurement quandary: at the time that a company is founded, one cannot observe whether that particular firm will experience a skewed growth outcome (or not). On the one hand, this challenge is fundamental, since entrepreneurship involves a high level of uncertainty and luck. And, some outsized successes certainly result from unlikely origins. Ben & Jerry's, for example, was founded with the intention to be a one-store, home-made ice-cream shop. With that said, many of the most successful firms in the economy were founded with a strong growth orientation. For example, Jeff Bezos founded Amazon with the intention of first creating the "world's biggest bookstore" in order to take advantage of the nascent potential of electronic commerce (Stone, 2013). To the extent that the new firms that ultimately contribute to the skew are disproportionately drawn from firms with significant growth ambitions and underlying potential at their time of founding, identifying these growth-oriented firms can contribute significantly to the understanding of firm dynamics.

Our key insight is that while it may be difficult to identify the potential for growth based on traditional economic metrics (e.g., profits during the first of operations), the founders themselves likely have information about the underlying quality of their idea and their personal level of

ambition, and make choices at the time of founding consistent with their objectives and potential for growth. Specifically, we can take advantage of the fact that entrepreneurs who assess the underlying quality of their venture to be higher are more likely to make choices that result in "digital signatures" associated with growth-oriented start-up firms, and that firms with these resulting digital signatures are themselves more likely to grow. In other words, we can map the realized performance of start-ups to the early-stage choices of founders. By mapping the relationship between growth outcomes and these founder choices, we are able to form an estimate of entrepreneurial quality at founding.

To understand this intuition, consider a simple model where all new firms have an underlying quality level $q$ (e.g., the underlying value of the idea and the ambition and capabilities of the founder) that is observable to the entrepreneur but not to the econometrician. Firms with a higher level of $q$ are more likely to realize a meaningful growth outcome $g$. In addition, all entrepreneurs face a set of binary corporate governance and strategy choices $H = \{h_1, \ldots, h_N\}$, such as how to register the firm (e.g., as an LLC or corporation), what to name the firm (e.g., whether to name the firm after the founders) and how to protect their underlying idea (e.g., whether to apply for either a patent or trademark). Suppose further that while the cost of each corporate governance choice $h$ is independent of the quality of the idea (but might vary idiosyncratically across entrepreneurs), the expected value of each of these choices is increasing in underlying quality (i.e., firms with a higher $q$ receive a higher marginal return to each element of $H$). Finally, suppose that while the econometrician cannot observe underlying quality, she is able to observe both the corporate governance choice bundle $H^*$ as well as growth outcomes $g$. The proof in Appendix B demonstrates that the mapping between $g$ and $H$ allows us to form a consistent estimate of the underlying probability of growth conditional on initial conditions $H$ (we refer to this estimate as $\theta$), and importantly that this mapping is a monotonically increasing function of $q$.

## II. The Measurement of Entrepreneurial Quality and Performance

Building on this discussion, we now develop our empirical strategy. Our goal is to estimate the relationship between a growth outcome, $g$, and early firm choices, $H^*$, in order to form an estimate of the probability of growth $\hat{\theta}$ for all firms at their time of founding. This approach (and our discussion) builds directly on Guzman and Stern (2015; 2017).

We combine three interrelated insights. First, as the challenges to reach a growth outcome as a sole proprietorship are formidable, a practical requirement for any entrepreneur to achieve growth is business registration (as a corporation, partnership, or limited liability company). This practical requirement creates a public record of all registered companies, allowing us to form a population sample of entrepreneurs "at risk" of growth at a similar (and foundational) stage. Second, we are able to potentially distinguish among business registrants through the measurement of characteristics related to entrepreneurial quality observable at or close to the time of registration. For example, we can measure start-up characteristics such as whether the founders name the firm after themselves (eponymy), whether the firm is organized in order to facilitate equity financing (e.g., registering as a corporation or in Delaware), or whether the firm seeks intellectual property protection (e.g., a patent or trademark). Third, we leverage the fact that, though rare, we observe meaningful growth outcomes for some firms (e.g., those that achieve an IPO or high-value acquisition within six years of founding). Combining these insights, we measure entrepreneurial quality by estimating the relationship between observed growth outcomes and start-up characteristics using the population of at-risk firms. For firm *i* born in region *r* at time *t* with start-up characteristics $H_{i,r,t}$ and growth outcome $g_{i,r,t+s}$, we estimate:

$$\theta_{i,r,t} = P\left(g_{i,r,t+s} \middle| H_{i,r,t}\right) = f\left(\alpha + \beta H_{i,r,t}\right) \quad (1)$$

This model allows us to *predict* quality as the probability of achieving a growth outcome given start-up characteristics at founding, and so estimate entrepreneurial quality as $\hat{\theta}_{i,r,t}$. As long as the process by which start-up characteristics map to growth remains stable over time (an assumption which is itself testable), this mapping allows us to form an estimate of entrepreneurial quality for any business registrant within our sample (even those in recent cohorts where a growth outcome or not has not yet had time to be observed).

We use these estimates to propose three new entrepreneurship statistics capturing the level of entrepreneurial quality for a given population of start-ups, the potential for growth entrepreneurship within a given region and start-up cohort, and the performance over time of a regional entrepreneurial ecosystem in realizing the potential performance of firms founded within a given location and time period.

### *A. The* Entrepreneurial *Quality Index (EQI)*

To create an index of entrepreneurial quality for any group of firms (e.g., all the firms within a particular cohort or a group of firms satisfying a particular condition), we simply take the *average* quality within that group. Specifically, in our regional analysis, we define the *Entrepreneurial Quality Index* (EQI) as an aggregate of quality at the region-year level by simply estimating the average of $\theta_{i,r,t}$ over that region:

$$EQI_{r,t} = \frac{1}{N_{r,t}} \sum_{i \in \{I_{r,t}\}} \theta_{i,r,t} \quad (2)$$

where $\{I_{r,t}\}$ represents the set of all firms in region *r* and year *t*, and $N_{r,t}$ represents the number of firms in that region-year. To ensure that our estimate of entrepreneurial quality for region *r* reflects the quality of start-ups in that location rather than simply assuming that start-ups from a given location are associated with a given level of quality, we exclude any location-specific measures $H_{r,t}$ from the vector of observable start-up characteristics.

### *B. The Regional Entrepreneurship Cohort Potential Index (RECPI).*

From the perspective of a given region, the overall inherent potential for a cohort of start-ups combines both the quality of entrepreneurship in a region and the number of firms in such region (a measure of quantity). To do so, we define *RECPI* as simply $EQI_{r,t}$ multiplied by the number of firms in that region-year:

$$RECPI_{r,t} = EQI_{r,t} \times N_{r,t} \quad (3)$$

Since our index multiplies the *average* probability of a firm in a region-year to achieve growth (quality) by the number of firms, it is, by definition, the expected number of growth events from a region-year given the start-up characteristics of a cohort at birth. This measure of course abstracts away from the ability of a region to realize the performance of start-ups founded within a given cohort (i.e., its ecosystem performance), and instead can be interpreted as a measure of the "potential" of a region given the "intrinsic" quality of firms at birth, which can then be affected by the impact of the entrepreneurial ecosystem, or shocks to the economy and the cohort between the time of founding and a growth outcome.

### *C. The Regional Ecosystem Acceleration Index (REAI).*

While RECPI estimates the *expected* number of growth events for a given group of firms, over time we can observe the *realized* number of growth events from that cohort. This difference can be interpreted as the relative ability of firms within a given region to grow, conditional on their initial entrepreneurial quality. Variation in ecosystem performance could result from differences across regional ecosystems in their ability to nurture the growth of start-up firms, or changes over time due to financing cycles or economic conditions. We define REAI as the ratio of realized growth events to expected growth events:

$$(4)\ REAI_{r,t} = \frac{\sum g_{i,r,t}}{RECPI_{r,t}}$$

A value of REAI above one indicates a region-cohort that realizes a greater than expected number of growth events (and a value below one indicates under-performance relative to expectations). REAI is a measure of a regional performance premium: the rate at which the regional business ecosystem supports high potential firms in the process of becoming growth firms.

Together, EQI, RECPI, and REAI offer researchers and regional stakeholders the ability to undertake detailed evaluations (over time, and at different levels of geographic and sectorial granularity) of entrepreneurial quality and ecosystem performance.

## III. Data and Entrepreneurial Quality Estimation

Our analysis leverages business registration records, a potentially rich and systematic data for the study of entrepreneurship. Business registration records are public records created when an individual registers a new business as a corporation, LLC or partnership. Appendix C of the Supplementary Materials in this paper provides a rich and detailed overview of this data set, as do the data appendixes in our prior work (Guzman and Stern, 2015; 2017).

We focus on 32 US states from 1988-2014 (see Appendix C for a list). While it is possible to found a new business without business registration (e.g., a sole proprietorship), the benefits of registration are substantial, and include limited liability, various tax benefits, the ability to issue and trade ownership shares, and credibility with potential customers. Furthermore, all corporations, partnerships, and limited liability companies must register with a Secretary of State

(or equivalent) in order to take advantage of these benefits: the act of registering the firm triggers the legal creation of the company. As such, these records reflect the population of businesses that take a form that is a practical prerequisite for growth.

Concretely, our analysis draws on the complete population of firms satisfying one of the following conditions: (a) a for-profit firm in the local jurisdiction or (b) a for-profit firm whose jurisdiction is in Delaware but whose principal office address is in the local state. In other words, our analysis excludes non-profit organizations as well as companies whose primary location is not in the state. The resulting dataset contains 27,976,477 observations.[9] For each observation we construct variables related to: (a) a growth outcome for each start-up; (b) start-up characteristics based on business registration observables; and (c) start-up characteristics based on external observables that can be linked directly to the start-up. We briefly review each one in turn and provide a more detailed summary in our data appendix.

*Growth*. The growth outcome utilized in this paper, Growth, is a dummy variable equal to 1 if the start-up achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration,[10] as reported in the Thomson Reuters SDC database.[11] During the period of 1988 to 2008, we identify 13,406 firms that achieve growth, of which 1,378 are IPOs and 12,028 are acquisitions, representing 0.07% of the total sample of firms in that period.

*Start-Up Characteristics*. At the center of our analysis is an empirical approach to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures of start-up characteristics: (a) measures based on business registration data observable in the registration record itself, and (b) measures based on external indicators of start-up quality that are observable at or near the time of business registration.

[9] The number of firms founded in our sample is substantially higher than the US Census Longitudinal Business Database (LBD), done from tax records. For example, for Massachusetts in the period 2003-2012, the LBD records an average of 9,450 new firms per year and we record an average of 24,066 firm registrations. We have yet to explore the reasons for this difference. However, we expect that it may be explained, in part by: (i) partnerships and LLCs that do not have income during the year do not file a tax return and are thus not included in the LBD, and (ii) firms that have zero employees and thus are not included in the LBD.

[10] In our Data Appendix (Section III, Table A4) we investigate changes in this measure both in the threshold of growth (e.g. only IPOs) as well as the time to grow; all results are robust to these variations

[11] 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) 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. Further, none of the cited studies found significant false positives, suggesting that the only effect of the acquisitions we do not track will be simply an attenuation of our estimated coefficients.

### *A. Measures Based on Business Registration Observables*

We construct twelve measures based on information observable in business registration records. We first create two binary measures that relate to how the firm is registered: *Corporation*, whether the firm is a corporation rather than an LLC or partnership, and *Delaware Jurisdiction*, whether the firm is registered in Delaware. We then create two additional measures based directly on the name of the firm. *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.[12] We hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Our second measure relates to the structure 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."). We define *Short Name* to be equal to one if the entire firm name has three or fewer words, and zero otherwise.[13]

We then create several measures based on how the firm name reflects the industry or sector within which the firm is operating, taking advantage of the industry categorization of the US Cluster Mapping Project ("US CMP") (Delgado, Porter, and Stern, 2016) and a text analysis approach. We develop eight such measures. The first three are associated with broad industry sectors and include whether a firm can be identified as local (*Local*), or traded (*Traded*), or traded within resource intensive industries (*Traded Resource Intensive*). The other five industry groups are narrowly defined high technology industries that could be expected to have high growth, including whether the firm is associated with biotechnology (*Biotech Sector*), e-commerce (*E-Commerce*), other information technology (*IT Sector*), medical devices (*Medical Dev. Sector*) or semiconductors (*Semiconductor Sector*).

### *B. Measures based on External Observables*

We construct two measures related to start-up quality based on intellectual property data sources from the U.S. Patent and Trademark Office. *Patent* is equal to 1 if a firm holds a patent

[12] Belenzon et al (2014; 2017) perform a more detailed analysis of the interaction between eponymy and firm performance, highlighting name as a signal chosen by entrepreneurs given differences in growth intention.

[13] Companies such as Akamai or Biogen have sharp and distinctive names, whereas more traditional businesses often have long and descriptive names (e.g., "New England Commercial Realty Advisors, Inc.").

application within the first year and 0 otherwise. We include patents that are filed by the firm within the first year of registration and patents that are assigned to the firm within the first year from another entity (e.g., an inventor or another firm). Our second measure, *Trademark*, is equal to 1 if a firm applies for a trademark within the first year of registration.

Table 1 reports summary statistics and sources. A detailed description of all variables as well as the specific set of US CMP clusters used to develop each industry classification are provided in the Data Appendix (Appendix C).

### *C. Estimation of Entrepreneurial Quality*

To estimate entrepreneurial quality for each firm in our sample, we regress *Growth* on the set of start-up characteristics observable either directly through the business registration records or otherwise related to the early-stage activities of growth-oriented start-ups. In Table 2, we present a series of univariate logit regressions of *Growth* on each of these start-up characteristics. All regressions are run on the full sample of firms from 1988 to 2008. To facilitate the interpretation of our results, we present the results in terms of the odds-ratio coefficient and include the McFadden pseudo $R^2$.[14]

Our univariate results are suggestive, and highlight a relationship between early firm choices and later growth. Measures based on the firm name are statistically significant and inform variation in entrepreneurial outcomes. Having a short name is associated with a 3 times increase in the probability of growth, and having an eponymous name with a 70% *lower* probability of growth. Corporate form measures are also significant. Corporations are 3.4 times more likely to grow and firms registered under Delaware jurisdiction (instead of the local jurisdiction) are 24 times more likely to grow. These magnitudes are economically important and have strong explanatory power—the pseudo-$R^2$ of a Delaware binary measure alone is 0.09—indicating a potential role of firm governance choices as a screening mechanism for entrepreneurial quality. Intellectual property measures have the highest magnitude of all groups. Firms with a patent close to their birth

[14] In all our models, we use logit rather than OLS for our predictions for two reasons. First, a large literature documents firm sizes and growth rates as much closer to log-normal than linear (Gibrat, 1931; Axtell, 2001). While we stress that entrepreneurial quality is a distinct measure from firm size, it is still more natural to use a functional form that best fits the known regularities of the data. While OLS is known to perform better than logit in estimating marginal effects (see Angrist and Pischke, 2008), logit performs better than OLS in prediction of binary outcomes (Pohlman and Leitner, 2003), consistent with the objective of this paper. We have also undertaken exploratory work investigating a non-parametric approach involving unstructured interactions of start-up characteristics. The results from such an exercise result in an even more skewed distribution of estimated entrepreneurial quality.

are 90 times more likely to grow, while firms with a trademark are 45 times more likely to grow. Finally, the set of US CMP Cluster Dummies, implied from firm name, are also informative. For example, firms whose name is associated with local industries (e.g. "Taqueria") are 79% less likely to grow, while firms whose name is associated with the biotechnology sector are 12 times more likely to grow. These coefficients highlight the value of early firm name choices as indicators of firm intentions and signals of a firm's relationship to an industry.

It is of course important to caution against causal interpretations of these findings (and our subsequent regression estimates). If a firm with low growth potential changes its legal jurisdiction to Delaware, this decision need not have any impact on its overall growth prospects.[15] Instead, Delaware registration is an informative signal—based on the fact that external investors often prefer to invest in firms governed under Delaware law—of the ambition and potential of the start-up at the time of business registration.

In Table 3, we turn to a more systematic regression analysis to evaluate these relationships. We begin in the first three specifications by evaluating the joint role of related groups of measures.[16] (3-1) investigates the core corporate governance measures, indicating that corporations are 4.1 times more likely to grow and Delaware firms are 23 times more likely to grow.[17] Interestingly, both of these coefficients are actually larger than the odds ratio in the univariate analysis. In (3-2), we focus on two measures based on firm name: firms with a short name are 3 times more likely to grow while eponymous firms are 78% *less* likely to grow. Finally, in (3-3), we study the relationship of intellectual property measures to *Growth*. Firms with a patent are 50 times more likely to grow and firms with a trademark are 8 times more likely to grow.

We then estimate our core predictive analytics models in (3-4) and (3-5) by combining these measures alongside industry and sector controls (i.e., firm names indicating whether the firm is in a local versus traded industry, or associated with a particular industry cluster). Our first

[15] While it is possible the firms might "game" the algorithm by selecting into signals of high-quality (e.g., changing their name), this incentive is bounded by the objectives of the founders. For example, it is unlikely that a founder with no intention to grow would incur the yearly expense (around $1000) to maintain a registration in Delaware. As well, firms using names to signal that they serve a local customer base (e.g. "Taqueria") are unlikely to change their names in ways that affect their ability to attract customers. Finally, if firms with low underlying quality did choose to invest in signals associated with high-quality, that would undermine the empirical correlation between start-up characteristics and firm growth.

[16] We include state fixed effects to account for idiosyncratic differences across states in corporate registration policies and fees. Though differences across states likely influence the "marginal" registrant (and would be of independent interest), it is unlikely that firms with significant growth potential would be deterred from registration depending on the state in which they were founded. All of our core findings are robust to the inclusion or exclusion of state fixed effects.

[17] Since these are incidence-rate ratios (odds-ratios), the joint coefficients can be interpreted multiplicatively: Delaware corporations are 94.3 times more likely to grow ($23 \times 4.1 = 94.3$).

specification (3-4) uses only business registration observables. The coefficients associated with each of the business registration measures is roughly equivalent, though the impact of each individual predictor is slightly attenuated.[18] We then extend this specification in (3-5) to include observables associated with early-stage milestones related to intellectual property. While the coefficients on the business registration observables remain similar (though once again slightly reduced in magnitude), each of the intellectual property observables is highly predictive. Given the high correlation between Delaware and Patent, we separately allow for the estimation of firms with a patent and no Delaware jurisdiction, firms with a Delaware jurisdiction and no patent, and firms with both.[19] In particular, receiving a patent is associated with a 23 times increase in the likelihood of growth for non-Delaware firms, and the combination of Delaware registration and patenting is associated with an 84 times increase in the likelihood of growth (simply registering in Delaware without a patent is associated with only a 15 times increase in the growth probability). Finally, firms successfully applying for a trademark in their first year after business registration are associated with a four-times increase in the probability of growth.[20]

These two models offer a tradeoff. On the one hand, the "richer" specification of model (3-5) involves an inherent lag in observability, since we are only able to observe early-stage milestones in the period after business registration (in the case of the patent applications, there is an additional 18-month lag due to the disclosure policies of the USPTO). While including a more informative set of regressors, model (3-5) is not as timely as model (3-4). Indeed, specifications that rely exclusively on information encoded within the business registration record can be calculated on a near real-time basis, and so provide the timeliest index for policymakers and other analysts. We will calculate indices based on both specifications; while our main historical analyses will be based off the results from model (3-5), model (3-4) can be used to provide our best estimate of changes

[18] Appendix Table A1 presents the complete set of coefficient estimates for the US CMP Clusters and US CMP High-Tech Cluster dummy variables. Briefly, as indicated in (A1-2), firms whose names indicate inclusion in a local industry (such as "restaurant", "realtor", etc) are 58% less likely to grow, firms associated with traded industries are not significant, and firms specifically associated with resource intensive traded industries are 12% less likely to grow. Names associated with specific high-technology sectors are also associated with growth: firms related to biotechnology are 3 times more likely to grow, firm associated with e-commerce are 44% more likely to grow, firms associated with IT 2.5 times, firms associated with medical devices 54%, and firms associated with semiconductors 2.3 times more likely to grow.

[19] An alternative way of presenting this would be to include only an interaction for both. The Delaware and Patent coefficients would stay the same, but the joint effect would require estimating $Delaware \times Patent$ interaction rather than providing the effect directly.

[20] It is worth noting that the coefficients in these two regressions are very similar to what we found in previous research in California (Guzman and Stern, 2015) and Massachusetts (Guzman and Stern, 2017). Figure A2 reports the coefficients associated with each state-level fixed effect; overall, our results are not sensitive to the inclusion or exclusion of these fixed effects in our regression analysis, predictive analytic estimates, or mapping of entrepreneurial quality.

in the last few years. We use the term *nowcasting* in reference to the estimates related to (3-4) and refer to (3-5) as the "full information" model (Scott and Varian, 2015).

### *D. Prediction Quality and Robustness*

In Figure 2, we evaluate the predictive quality of our estimates by undertaking a tenfold cross-validation test (Witten and Frank, 2005),[21] and report the out-of-sample share of realized growth outcomes at different portions of the entrepreneurial quality distribution. The results are striking. The share of growth firms in the top 5% of our estimated growth probability distribution ranges from 51% to 54%, with an average of 53%. The share of growth firms in the top 1% ranges from 34% to 38%, with 36% on average. Growth, however, is still a relatively rare event even among the elite: the average firm within the top 1% of estimated entrepreneurial quality has only a 2.4% chance of realizing a growth outcome.

In Appendix Table A2, we repeat our full information model with a series of robustness tests to verify that the magnitudes in our model are not driven by variation across years or states. In (A2-1) we report a variation of our model after also including year fixed-effects, (A2-2) includes state-specific time trends, and (A2-3) includes both year fixed-effects and state-specific time trends. While there is some variation in the magnitude of the coefficients, these changes are relatively small, suggesting that the estimates are not driven by idiosyncratic variation across years or states.[22]

## IV. The State of American Entrepreneurship

We now leverage these prediction models to calculate the centerpiece of our analysis: evaluating trends in entrepreneurial quality (EQI), entrepreneurial potential (RECPI), and regional economic performance (REAI) across the 32 states in our sample from 1988 through 2014. We estimate two RECPI indexes, a full information index based on (3-5) using information in

[21] Specifically, we divide our sample into 10 random subsamples, using the first subsample as a testing sample and using the other 9 to train the model. For the retained test sample, we compare realized performance with entrepreneurial quality estimates from the model resulting from the 9 training samples. We then repeat this process 9 additional times, using each subsample as the test sample exactly once. This approach allows us to estimate average out of sample performance, as well as the distribution of out of sample test statistics for our model specification.

[22] Table B1 assesses the robustness of the index across states by estimating the out-of-sample shares of firms in the top 5% and top 10% of quality by state, and the correlation between entrepreneurial quality estimates performed individually for each state and the national estimate. While there is variation in each of these statistics across states, all of them indicate a relatively strong correlation between quality at the state level and our national measure.

intellectual property and business registration records which we simply call RECPI, and a nowcasting index that uses only business registration records (3-4), which we call Nowcasted RECPI. U.S. RECPI, reported in Figure 3, is RECPI adjusted by the aggregate GDP of the 32 states in the sample.[23] Finally, we also include a confidence interval estimated through a Monte Carlo process repeating our procedure for 30 bootstrapped random samples (i.e. with replacement) of the same size as our original sample. Before analyzing trends in the indexes, we note that both U.S. RECPI and Nowcasted U.S. RECPI move very close to each other and that the confidence interval of U.S. RECPI is narrow.

Both indexes indicate a rise of entrepreneurial potential in the 1990s through the year 2000, with a rapid drop between 2000 and 2002. However, the level observed during the 2000s through 2008 is consistently higher than the level observed during the first half of the 1990s. After a decline during the Great Recession (2008 and 2009), we observe a sharp upward spring starting in 2010.[24] Interestingly, Nowcasted U.S. RECPI is observed at its third highest level in 2014. Relative to quantity-based measures of entrepreneurship such as the BDS, these estimates seem to reflect broad patterns in the environment for growth entrepreneurship, such as capturing the dot-com boom and bust of the late 1990s and early 2000s, and capturing the rise of high-growth start-ups over the early years of this decade.

Our index of entrepreneurial potential does show gaps relative to realized entrepreneurial performance. Though the statistics of GDP Growth in Figure 1B as well as the number of growth firms in Figure 1A peak in the years 1995 and 1996 (respectively), U.S. RECPI instead peaks in the year 2000. This offers insight into the possible sensitivity of entrepreneurial potential to credit market cycles. While the 1996 cohort may have had lower initial potential, those firms were able to take advantage of the robust financing environment during the early years of their growth; in contrast, the peak U.S. RECPI start-up cohorts of 1999 and 2000 may have been limited in their ability to reach their potential due to the "financial guillotine" that followed the crash of the dot-com bubble (Nanda and Rhodes-Kropf, 2013, 2014).

U.S. RECPI offers a new perspective on the "state" of entrepreneurship (at least for these fifteen states). Specifically, our Nowcasting index suggests that there has been a steep rise in

[23] It is also possible to adjust by population instead of GDP. RECPI / population shows a starker positive increase than RECPI / GDP, as GDP per capita has also increased through the time period represented.

[24] These broad patterns closely accord with the patterns we found for Massachusetts in Guzman and Stern (2017).

entrepreneurial potential over the last several years, and 2014 is the first year to begin to reach the peaks of the dot-com boom. Indeed, it is useful to recall that our measure is *relative to GDP*: on an absolute scale, U.S. RECPI 2014 is at the highest level ever registered. Finally, we emphasize that, though there are small deviations, both the nowcasted and full information indexes have a very high concordance.

### *A. Geographic Variation in Entrepreneurial Quality*

Figure 4 illustrates the geographic variation in entrepreneurial quality for the 32 states in our sample. We present RECPI by ZIP Code for all ZIP Codes with at least ten new firms (to avoid overcrowding the image), where the size of each point is equal to the quantity of entrepreneurship, and the color of the point indicates the EQI for that zip code (with darker coloring indicating a higher EQI).

This map offers insight into the distribution of entrepreneurial quantity and quality across the United States. First, the most intense areas for entrepreneurial potential are in well-known entrepreneurial ecosystems such as Silicon Valley, Boston, and Austin. Second, several large cities, including Los Angeles, Houston, Dallas and even Detroit host not simply a high level of new registrants but a high average level of entrepreneurial quality among their start-ups. Third, a number of other well known locations such as Seattle, northern Virginia (in the Washington DC area), and register a high average EQI. At the same time, there are large areas of the United States that host a high level of entrepreneurship but where estimates of start-up quality are relatively low. Florida, in particular, seems to have a very high average quantity with low average quality. Many of the Mountain States (e.g., Wyoming, Idaho, and Utah), as well as Northern New England (Vermont and Maine) also seem to have a relatively low average estimated quality even within key cities such as Salt Lake City.

Overall, this evidence supports three interrelated conclusions. First, relative to a perspective emphasizing a worrisome secular decline in "shots on goal" (Hathaway and Litan, 2014b), our approach and evidence suggest that there has been a more variable pattern of entrepreneurship from 1988 to 2014, and that the last five years have been associated with an accumulation of entrepreneurial potential similar to that which marked the late 1990s. Second, this variation in potential has a clear relationship with later entrepreneurship performance of such cohorts as measured by the number of realized growth firms as well as market value created by firms in those

cohorts. Finally, given the more gently sloped shape of the entrepreneurial boom of recent years, it may be the case that this accumulation of entrepreneurial potential is more sustainable than earlier periods.

## V. Trends in the Effect of the US Entrepreneurial Ecosystem (REAI)

Entrepreneurship performance depends on not simply founding new enterprises, but the scaling of those enterprises in a way that is economically meaningful. This insight motivates our second set of findings where we examine "ecosystem" performance across the United States, as measured by the Regional Ecosystem Acceleration Index (REAI). REAI captures the relative ability of a given start-up cohort to realize its potential, relative to the expectation for growth events as measured by RECPI (i.e., REAI = Number of Growth Events / RECPI). A value of 1 in the index indicates no ecosystem effect. A value above 1 indicates a positive ecosystem effect, and a value under 1 indicates a negative effect. In contrast to RECPI, this index reflects the impact of the economic and entrepreneurial environment in which a start-up cohort participates (i.e., the "ecosystem" in which it participates). This ecosystem will include the location in which the firm is founded (e.g., Silicon Valley versus Miami) as well as the environment for funding and growth at the time of founding. In Figure 5, we examine the changing environment for entrepreneurship in the United States (i.e., change in the US Ecosystem, as reflected in the 32 states for which we have data), we plot REAI over time from 1988-2008, and we develop a projected measure of REAI for years 2009-2012.[25]

Three distinct periods stand out. The early portion of our sample saw a significant increase in REAI from a slight negative level to a peak of 1.58 for the 1995 cohort. This is consistent with our evidence from Figure 1, in which the 1995-1996 start-up cohort was indeed the most "successful." This peak was followed by a steady decline through 2000, in which, conditional on the estimated quality of a given start-up, the probability of growth was declining as a result of the environment (i.e., time) in which that start-up was trying to grow. From 2000-2007, there is a period of slight decline, with REAI moving from 0.95 down to 0.63. These differences are economically meaningful: a start-up at a given quality level is estimated to be 3 times more likely

[25] Because our approach requires that we observe the *realized* growth firms we can only measure our index with a 6 year lag, thus, up to 2008. For years 2009 to 2012, we estimate our model with a varying lag of $n = 2014 - year$ and calculate RECPI using such lag.

to experience a growth event in the six years after founding if it was founded in 1995 rather than in 2007. Finally, though still a preliminary estimate, we observe a resurgence in REAI for cohorts from 2007 to 2012, highlighting a potential improvement in the entrepreneurial ecosystem in recent years in parallel with the boom in the availability of entrepreneurial finance. While this rise is economically important, its ultimate impact once all growth outcomes are realized remain to be seen.

## VI. Equity versus Employment Growth Outcomes

While equity growth is a measure of success for founders and investors, realizing significant employment growth is an alternative measure of entrepreneurial success more closely tied to broader economic performance (e.g. Krishnan, Nandy, and Puri, 2015; Davis and Haltiwanger, 1992). While a full analysis of the relationship between business registration observables and comprehensive employment outcomes is beyond the scope of this paper (as such an analysis could more naturally be conducted in the context of an integrated longitudinal database such as the LBD), we undertake a preliminary robustness check to evaluate how the use of an employment-based success metric influences our analysis and findings. To do so, we take advantage of a dataset of employment levels for more than 10 million firms available from Infogroup USA between 1997 and 2014.[26] We construct two new outcome variables, *Employment Growth 500* and *Employment Growth 1000*, each equal to 1 for all firms recorded as having greater than 500 or 1000 or more employees, respectively, within 6 years, and 0 otherwise. Though this measure does not capture the employment levels of all firms (and all employment data are themselves categorical estimates rather than the fine-grained measures available through administrative data), this rough cut allows us to identify the vast majority of firms that experience the (rare and usually highly observable) event of becoming a large employer in a relatively short period of time. As context, the 500 employee threshold is used to differentiate small and medium-sized enterprises from large firms by the Small Business Administration, and so it is useful to consider this transition within six years from one category to the other.

[26] Infogroup is a private sector business database similar to Dunn and Bradstreet. An overview of the dataset and our variable construction, as well as references to prior work using these data, is provided in Appendix D. We utilize the annual snapshot data maintained by MIT Libraries from 1997 to 2014. We match firms by name and states and then examine, for each firm name/state combination, whether that firm achieves a given employment outcome (500 or 1000) within six years of its business registration. As well, to avoid duplicates, we focus only on headquarter locations (as indicated by Infogroup), deleting all non-headquarter establishments. Infogroup reports employment for the entire company in the 'headquarter' entry.

We use these data to conduct three interrelated exercises. First, in Table 5, we compare our baseline entrepreneurial quality model using *Growth* versus the *Employment Growth* measures as the dependent variable. The estimates are surprisingly similar not just in sign but also in relative magnitude, with a higher concordance between *Growth* and *Employment Growth 500*. For example, firms with a trademark are 6.2 times more likely to get 1000 employees (4 times for equity growth), firms with a patent 46.8 times (20.8 times for equity growth), firms registered in Delaware 13.3 times (14.0 for equity growth), and firms with both a patent and Delaware registration 131.4 times (80.56 for equity growth). This similarity between coefficients suggests that our baseline model not only captures financial outcomes but also captures significant variation across firms in their potential to achieve a rare and outsized employment growth outcome.

As a second exercise, we use the model with the lower level of concordance (*Employment Growth 500*) as an alternative baseline for our predictive approach to form a quality estimate for each firm in our sample and compare our initial entrepreneurial quality estimates with this alternative. The correlation between a predictive analytic based on *Growth* versus *Employment Growth 500* is 0.84. Finally, we examine how the incidence of *Employment Growth 500* is predicted by our estimates of entrepreneurial quality using our baseline equity growth regression and report the share of firms that achieve employment growth in the top 5% and 10% of quality. The results are striking: more than 48% of all measured employment growth outcomes occur within the top 10% of our entrepreneurial quality distribution, with around 40% in the top 5%.

While we emphasize that this analysis is incomplete insofar as our measures of employment growth may be incomplete, it nonetheless suggests that there is a meaningful relationship between equity and employment growth, and that both of these highly skewed outcome variables have a predictable relationship with measures of underlying entrepreneurial quality.

## VII. The Impact of the Business Cycle on Entrepreneurial Quantity and Quality

We now proceed to evaluate how the business cycle influences US entrepreneurial quality and quantity. To do so, we implement an SVAR regression that models the interdependent relationship between RECPI and the business cycle, and allows us to estimate the impact of GDP growth on entrepreneurship. The impact of business cycles on entrepreneurship has been long-debated in economics, most notably in the contrast between the debt deflation theory of Fisher (1933) and the "cleansing effect" of recessions emphasized by Schumpeter (1939).

Bernanke and Gertler (1989) offer the first formal account of the debt deflation hypothesis, demonstrating how random macroeconomic shocks influence the balance sheets of would-be entrepreneurs and consequently change their ability to undertake the ambitious projects that represent new high-quality entrepreneurship (see also Carlstrom and Fuerst (1997) and Rampini (2004)). Relatedly, business cycles also change investor expectations concerning the future availability of capital, which in turn can lead to a reduction in the riskiness of financed projects (Caballero, Farhi, and Hammour, 2006), a dynamic with particular implications for the availability of early-stage venture capital (Nanda and Rhodes-Kropf, 2016). On the other hand, a smaller theoretical literature has focused on the potential for more growth-oriented entrepreneurship to be founded in recessions due to a "cleansing effect", under the possibility that the bankruptcy or financial stress of marginal incumbent firms during a downturn might enhance entry opportunities for new productive firms (Schumpeter 1939; Caballero and Hammour, 1994).

Existing empirical evidence has yet to precisely support one hypothesis or the other. For example, while Nanda and Rhodes-Kropf (2013) show that increases in the supply of venture capital lead VCs to invest in more innovative firms, and Moreira (2015) highlights that there are procyclical differences in the initial size of firms across the business cycle, which persist, Koellinger and Thurik (2012) do not observe any relationship between GDP growth and the subsequent quantity of entrepreneurship (using surveys of business ownership) in a panel of 22 OECD countries.[27] More generally, none of these papers provide a direct test of the key underlying hypotheses in Bernanke and Gertler (1989), or other theories. Do positive changes in the GDP growth rate predict no change in the overall quantity of entrepreneurship and a positive change in the quality-adjusted quantity of entrepreneurship? Conversely, do recessions result in a downward shift in the distribution of entrepreneurial quality-adjusted quality?

To evaluate such relationship, we propose a structural vector autoregression model (SVAR) that models the interdependent nature of entrepreneurship and economic growth. Economic growth influences entrepreneurship contemporaneously and with a lag, while entrepreneurship only responds to growth with a lag, reflecting the time it takes to undertake an investment, which can be modeled as follows:

[27] Koellinger and Thurik (2012) do find, however, a relationship in the opposite causal direction, that entrepreneurship predicts (Granger-causes) economic growth.

$$(5)\ Ln(\Delta GDP_t) = a_1 Ln\left(\frac{RECPI_{t-1}}{GDP_{t-1}}\right) + \cdots + a_n Ln\left(\frac{RECPI_{t-n}}{GDP_{t-n}}\right) + b_1 Ln(\Delta GDP_{t-1}) + \ldots + b_n Ln(\Delta GDP_{t-n}) + u_t$$

$$(6)\ Ln(\frac{RECPI_t}{GDP_t}) = c_1 Ln(\frac{RECPI_{t-1}}{GDP_{t-1}}) + \cdots + c_n Ln(\frac{RECPI_{t-n}}{GDP_{t-n}}) + d_0 Ln(\Delta GDP_t) + d_1 Ln(\Delta GDP_{t-1}) + \cdots + d_n Ln(\Delta GDP_{t-n}) + v_t$$

where $Ln(\Delta GDP_t)$ represents GDP growth (in log) at time $t$ and $Ln\left(\frac{RECPI_t}{GDP_t}\right)$ represents the quality-adjusted flow of entrepreneurship at *t*, and $u_t$ and $v_t$ are the idiosyncratic disturbances in the growth rate and the entrepreneurship rate, respectively. We fit a recursive SVAR to estimate $d_n$ as the percent increase in quality-adjusted entrepreneurship from a percentage increase in the annual GDP growth rate.

Using the (admittedly small) sample of 27 annual observations for the United States from 1988-2014, Appendix Table A4 reports coefficient estimates from this approach as well as equivalent regressions using only the quantity of firms instead of RECPI. Figure 6 presents the impulse response functions. We begin in (A4-1) and (A4-2) with a reduced-form single-lag VAR model. While (A4-1) reports a positive relationship between GDP growth and RECPI, (A4-2) indicates no relationship between changes in GDP growth and start-up quantity. We then turn to a three-lag SVAR model that allows not only to capture dynamics but also allows for contemporaneous impact between GDP growth and entrepreneurship.[28] The parameter estimates are similar. Throughout we observe a positive relationship between GDP growth and subsequent RECPI and no relationship between GDP growth and the raw quantity of entrepreneurship. We see this more clearly in Figures 6A and 6B, which present the impulse response function for regression. The figures indicate that a doubling of the GDP growth rate leads to a 2% increase in $\frac{RECPI}{GDP}$ in current year *t*, and a 4% increase in years *t+1* and *t+2*, which then tapers off. In contrast, as illustrated in Figure 6B, there is no net relationship between GDP growth and $Ln\left(\frac{Obs}{GDP}\right)$. We further this analysis in (A4-5) and (A4-6) by considering an alternative measure of business cycles, the presence of an economic recession as determined by the NBER Business Cycle Dating

[28] The three-lag structure minimizes both the Aikake Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (SBIC).

Committee. As shown in Figures 6C and 6D, the onset of a recession decreases $\frac{RECPI}{GDP}$ by 5% in *t* and *t+1* (with a subsequent tapering off), while having no net impact on $Ln(\frac{Obs}{GDP})$.

We emphasize that these results should be viewed with caution. We are basing our inferences on only a relatively short time-series, and it is of course possible that the relationship between economic performance and entrepreneurship changes over time and place. With that important caveat, these results are consistent with the hypothesis that, while economic shocks have an ambiguous and noisy impact on the overall start-up rate, there is a meaningful relationship between economic shocks and the propensity to start ventures with high growth potential at founding.

## VIII. Conclusion

This paper develops a quality-based approach with business registration records for 32 states to create and evaluate novel indices of the quantity and quality-adjusted quantity of entrepreneurship. Not simply a matter of data, the predictive analytics approach allows us to focus on a more rigorous examination of variation over time and across places in the potential from a given start-up cohort (RECPI), the ability of an entrepreneurial ecosystem to realize that potential over time (REAI) and the relationship between entrepreneurship and economic fluctuation.

This analysis offers several new findings about the state of American entrepreneurship. First, in contrast to the secular decline observed in aggregate quantity-oriented measures of business dynamism (Decker et al, 2014), the expected number of growth outcomes in the United States has followed a cyclical pattern that appears sensitive to the capital market environment and overall market conditions. U.S. RECPI reflects broad and well-known changes in the environment for startups, such as the dotcom boom and bust of the late 1990s and early 2000s. As well, a quality-adjusted predictive analytics approach captures striking regional variation in the growth potential of start-ups across the United States, including the presence of strong ecosystems such as Silicon Valley or Boston and relatively quantity-oriented entrepreneurship regions such as Miami.

By accounting for quality, our estimates offer a different perspective on the role of start-ups in the US economy over the past thirty years. While the expected number of high-growth startups peaked in 2000 and then fell dramatically with the dot-com bust, starting in 2010 there has been a sharp upward swing in the expected number of successful startups formed and the accumulation of entrepreneurial potential for growth (even after controlling for the change in the overall size of

the economy). Indeed, in contrast to the secular decline in start-up activity observable in the BDS, our estimates of U.S. RECPI indicate a net *upward* trend across the full time-series of our sample. For example, the rate of expected successful startups fell to its lowest point in 1991, and reached its second highest level in 2014 (the final year of our sample). This finding suggests that the challenges to growth arising from entrepreneurship may be less directly related to the lack of formation of high-growth potential startups, and instead more related to other dynamics or ecosystem concerns. In particular, while there is high cyclicality in RECPI / GDP, REAI—the likelihood of startups to reach their potential—declined in the late 1990s and did not recover through 2008. Relative to the mid 1990s, the 2000s was a period in which a lower level of entrepreneurial potential was realized. For example, conditional on the same estimated potential, a 1995 startup was 3 times more likely to achieve a growth event in 6 years than a startup founded in 2007.

Accounting for entrepreneurial quality through a predictive analytics approach is not simply a question of more nuanced measurement of the same phenomena. Instead, a shift towards entrepreneurial quality allows one to more directly connect entrepreneurship and overall economic performance. Using our measures in a structural VAR model, we find economic shocks are associated with a procyclical impact on the quality-adjusted quantity of entrepreneurship, while there is no relationship with quantity alone. These results provide novel empirical evidence on the way entrepreneurship is shaped by economic conditions, and allow us to begin to adjudicate between competing theories of this relationship.

More generally, our analysis suggests that directly taking a quantitative approach to the measurement of entrepreneurial quality can yield new insight into the precursors and consequences of entrepreneurial ecosystems and the impact of entrepreneurship on economic and social progress. Several follow-on research directions are possible. First, our data reveal striking variation across regions and time in both the quality-adjusted quantity of entrepreneurship as well as the potential for growth condition on initial quality. Examining how regional and temporal determinants of entrepreneurial ecosystems impact both entrepreneurial quality, the growth process, and even the migration of firms between regions is a promising area for future research (Guzman, 2018). Second, while our current analysis examines the link between entrepreneurial quality at founding and subsequent growth (measured as either equity or employment growth), it is separately possible to examine how particular institutions that impact start-ups after founding (such as the receipt of

venture capital) impact that growth process. For example, in Catalini, Guzman and Stern (2019), we examine both the selection into and impact of venture capital on start-up firms by exploiting this predictive analytics approach. Further work connecting firm founding, capital investment, and growth is likely to allow for a more structured understanding of the role of external capital in start-up growth. Finally, a striking feature of our predictive analytics results is the unusual level of skewness in the entrepreneurial quality distribution (e.g., around 40% of all equity growth outcomes are contained within the top 1% of the estimated quality distribution).

Directly measuring the high level of skewness in the initial distribution of firms likely offers new insight into a number of areas, such as industrial organization, finance, and organizational economics. These benefits are likely to be enhanced by ongoing technological improvements in data storage and processing capabilities, which will likely improve the precision and applicability of these estimates. To this end, these estimates are an initial implementation pointing towards a more general approach using founding observables and ex-post performance to estimate founding quality. For example, it may be possible to develop integrated datasets including measures based on firm founding statements, online job postings, media mentions, and the degree and nature of online or social media presence (e.g., presence or absence of a web page, functionality of that webpage, etc.). Combined with more sophisticated predictive algorithms (e.g., machine learning approaches as developed in Guzman (2018)), it may be possible to capture different types of performance and the linkage between the initial conditions at founding and different types of economic and social impact.

## REFERENCES


**Acs, Zoltan J., and David B. Audretsch.** 1988. "Innovation in Large and Small Firms: An Empirical Analysis." *The American Economic Review* 78 (4): 678–90.

**Akcigit, Ufuk, and William Kerr.** 2018. "Growth Through Heterogeneous Innovations." *Journal of Political Economy* 126 (4): 1374-1443.

**Angrist, Joshua D., and Jörn-Steffen Pischke.** 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

**Aulet, William, and Fiona E. Murray.** 2013. "A Tale of Two Entrepreneurs: Understanding Differences in the Types of Entrepreneurship in the Economy." *SSRN Electronic Journal*. 10.2139/ssrn.2259740.

**Axtell, Robert L.** 2001. "Zipf distribution of US firm sizes." *Science* 293(5536): 1818-20.

**Barnes, Beau Grant and Nancy Harp and Derek Oler.** 2014. "Evaluating the SDC Mergers and Acquisitions Database." *The Financial Review* 49(4): 793.

**Bernanke, Ben, and Mark Gertler.** 1989. "Agency Costs, Net Worth, and Business Fluctuations." *The American Economic Review* 79 (1): 14-31.

**Belenzon, Sharon, Aaron Chatterji, and Brendan Daley.** 2017. "Eponymous Entrepreneurs" *The American Economic Review* 107 (6): 1638-55.

**Belenzon, Sharon, Aaron Chatterji, and Brendan Daley.** 2018. "Choosing Between Growth and Glory" National Bureau of Economic Research (NBER) Working Paper 24901.

**Caballero, Ricardo J., and Mohamad L. Hammour.** 1994. "The Cleansing Effect of Recessions," *The American Economic Review* 84 (5): 1350-1368.

**Caballero, Ricardo J., Emmanuel Farhi, and Mohamad L. Hammour.** 2006. "Speculative Growth. Hints from the U.S. Economy" *The American Economic Review* 96 (4): 1159-1192.

**Carlstrom, Charles T, and Timothy S. Fuerst.** 1997. "Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis." *The American Economic Review* 87 (5): 893-910.

**Catalini, Christian, Jorge Guzman, and Scott Stern. 2019**. "Passive Versus Active Growth: Evidence from Founder Choices and Venture Capital Investment." Working Paper.

**Chatterji, Aaron.** 2018. "The Main Street Fund: Investing in an Entrepreneurial Economy". *The Hamilton Project*, June.

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

**Davis, Steven, and John Haltiwanger.** 1992. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation." *The Quarterly Journal of Economics*. 107 (3): 819-862.

**Davis, Steven, John Haltiwanger, and Scott Shuh.** 1996. *Job Creation and Destruction*. Cambridge: MIT Press.

**Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda.** 2014. "The Role of Entrepreneurship in US Job Creation and Economic Dynamism." *Journal of Economic Perspectives* 28 (3): 3-24.

**Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda.** 2016. "Where Has All The Skewness Gone? The Decline in High-Growth (Young) Firms." *European Economic*

*Review* 86: 4-23.

**Delgado, Mercedes, Michael E. Porter, and Scott Stern.** 2016. "Defining Clusters of Related Industries." *Journal of Economic Geography*. 16 (1): 1-38.

**Dunne, Timothy, Mark J. Roberts, and Larry Samuelson.** 1988. "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries." *RAND Journal of Economics* 19(4): 495-515.

**Ericson, R., and A. Pakes.** 1995. "Markov-Perfect Industry Dynamics: A Framework for Empirical Work." *The Review of Economic Studies*, 62(1): 53-82.

**Evans, David S.** 1987. "The Relationship Between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries." *The Journal of Industrial Economics* 35 (4): 567–81.

**Fazio, Catherine, Jorge Guzman, and Scott Stern.** 2019. "The Impact of State-Level R&D Tax Credits on the Quantity and Quality of Entrepreneurship." Working Paper.

**Fisher, I.** 1933. "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1(4): 337-357.

**Gibrat, Robert.** 1931. *Les Inégalités Économiques.* Paris: Librairie du Recueil Sirey

**Gornall, Will, and Ilya A. Strebulaev.** 2015. "The Economic Impact of Venture Capital: Evidence from Public Companies." Stanford Graduate School of Business Working Paper 3362.

**Guzman, Jorge.** 2019. "Go West Young Firm: Agglomeration and Embeddedness in Startup Migrations to Silicon Valley." Columbia Business School Research Paper 18-49.

**Guzman, Jorge, and Aleksandra Kacperczyk.** 2019. "Gender Gap in Entrepreneurship." *Research Policy.* 46 (7): 1666-1680.

**Guzman, Jorge, and Scott Stern. 2015.** "Where is Silicon Valley?" *Science* 347(6222): 606-609.

**Guzman, Jorge, and Scott Stern.** 2017. "Nowcasting and Placecasting Entrepreneurial Quality and Performance." *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, edited by John Haltiwanger, Erik Hurst, Javier Miranda, and Antoinette Schoar. Chicago: University of Chicago Press. Chapter 2.

**Haltiwanger, John, Ron Jarmin, and Javier Miranda.** 2013. "Who Creates Jobs? Small versus Large versus Young." *The Review of Economics and Statistics* 95 (2): 347-361.

**Hathaway, Ian, and Robert E. Litan.** 2014a. "Declining Business Dynamism in the United States: A Look at States and Metros." *Economic Studies at Brookings Series*. May, 2014.

**Hathaway, Ian, and Robert E. Litan.** 2014b. "Declining Business Dynamism: It's for Real." *Economic Studies at Brookings Series*. May, 2014.

**Hathaway, Ian, and Robert E. Litan.** 2014c. "The Other Aging of America: The Increasing

Dominance of Older Firms." *Economic Studies at Brookings Series*. July, 2014

**Hopenhayn, H.** 1992. "Entry, Exit, and firm Dynamics in Long Run Equilibrium." *Econometrica* 60(5): 1127-1150.

**Hurst, Erik & Benjamin Wild Pugsley**. 2011. "What do Small Businesses Do?" *Brookings Papers on Economic Activity* 43(2): 73-142.

**Jovanovic, Boyan.** 1982. "Selection and the Evolution of Industry." *Econometrica* 50(3): 649-70.

**Kaplan, Steven, and Josh Lerner.** 2010. "It Ain't Broke: The Past, Present, and Future of Venture Capital." *Journal of Applied Corporate Finance* 22 (2): 36-47.

**Kerr, William, Ramana Nanda, and Matthew Rhodes-Kropf.** 2014. "Entrepreneurship as Experimentation." *Journal of Economic Perspectives* 28 (3): 25-48.

**Koellinger, Philipp and Roy Thurik.** 2012. "Entrepreneurship and the Business Cycle." *The Review of Economics and Statistics* 94(4): 1143-1156.

**Klapper, Leora, Raphael Amit and Mauro F. Guillén,** 2010. "Entrepreneurship and Firm Formation across Countries." *International Differences in Entrepreneurship*, 129-158. Chicago: University of Chicago Press.

**Klepper, Steven**. 1996. "Entry, Exit, Growth, and Innovation over the Product Life Cycle." *The American Economic Review* 86(3): 562-583.

**Klette, Tor and Samuel Kortum.** 2004. "Innovating Firms and Aggregate Innovation." *Journal of Political Economy* 112(5): 986-1018.

**Kortum, Samuel, and Josh Lerner.** 2000. "Assessing the contribution of venture capital to innovation." *RAND Journal of Economics* 31 (4): 674-692.

**Krishnan, Karthik, Debarshi K. Nandy and Manju Puri.** 2015. "Does Financing Spur Small Business Productivity? Evidence from a Natural Experiment." *Review of Financial Studies* 28 (6): 1768-1809.

**Mansfield, Edwin.** 1962. "Entry, Gibrat's Law, Innovation, and the Growth of Firms." *The American Economic Review* 52(5): 1023–1051.

**McFadden, Daniel.** 1974. "Conditional logit analysis of qualitative choice behavior." *Frontiers in Econometrics.* 105-142. New York: Academic Press.

**Moreria, Sara.** 2016. "Firm Dynamics, Persistent Effects of Entry Conditions, and Business Cycles." SSRN Working Paper 3037178.

**Mills, Karen Gordon, and Brayden McCarthy**. 2016. "The State of Small Business Lending:

Innovation and Technology and the Implications for Regulation." Harvard Business School (HBS) Working Paper 17-042, November.

**Nanda, Ramana, and Matthew Rhodes-Kropf.** 2013. "Investment Cycles and Startup Innovation." *Journal of Financial Economics* 110(2): 403-418.

**Nanda, Ramana, and Matthew Rhodes-Kropf.** 2014. "Financing Risk and Innovation." Harvard Business School (HBS) Working Paper 11-013.

**Nanda, Ramana and Matthew Rhodes-Kropf.** 2016. "Financing Entrepreneurial Experimentation." *Innovation Policy and the Economy*. 16(1): 1-23. Chicago: University of Chicago Press.

**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.

**Pohlman, John T and Dennis W. Leitner.** 2003. "A Comparison of Ordinary Least Squares and Logistic Regression." *The Ohio Journal of Science* 103(5): 118-125.

**Rampini, Adriano and Amir Sufi and S. Viswanathan.** 2014. "Dynamic risk management." *Journal of Financial Economics* 111 (2): 271-296

**Sutton, John.** 1997. "Gibrat's Legacy." *Journal of Economic Literature* 35: 40-59.

**Samila, Sampsa, and Olav Sorenson.** 2011. "Venture Capital, Entrepreneurship, and Economic Growth." *The Review of Economics and Statistics*. 93(1): 338-349.

**Schoar, Antoinette.** 2010. "The Divide between Subsistence and Transformational Entrepreneurship." *Innovation Policy and the Economy*, edited by Josh Lerner and Scott Stern, 57 – 81. Chicago: University of Chicago Press.

**Schumpeter, J.A.** 1939. *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process*. New York: McGraw-Hill.

**Stern, Scott and Valentina Tartari**. 2018. "The Role of Universities in Local Entrepreneurial Ecosystems." Working Paper.

**Stone, Brad.** 2013. *The Everything Store: Jeff Bezos and the Age of Amazon.* New York: Little, Brown and Company.

**Witten, Ian H., and Eibe Frank.** 2005. *Data Mining: Practical machine learning tools and techniques.* Burlington: Morgan Kaufmann.

TABLE 1. SUMMARY STATISTICS (1988-2014)

| | Source | Mean | Std Dev |
|---|---|---|---|
| **Outcome Variable** | | | |
| Growth | SDC Platinum | 0.00071 | 0.02672 |
| **Corporate Form Observables** | | | |
| Corporation | Bus. Reg. Records | 0.48 | 0.50 |
| Delaware | Bus. Reg. Records | 0.024 | 0.153 |
| **Name-Based Observables** | | | |
| Short Name | Bus. Reg. Records | 0.46 | 0.50 |
| Eponymous | Bus. Reg. Records | 0.0707 | 0.2563 |
| **Intellectual Property Observables** | | | |
| Patent | USPTO | 0.0019 | 0.0439 |
| Trademark | USPTO | 0.0014 | 0.0374 |
| **Industry Measures (US CMP Clusters)** | | | |
| Local | Estimated from name | 0.19 | 0.39 |
| Traded (3) | Estimated from name | 0.537 | 0.499 |
| Traded Resource Int. | Estimated from name | 0.133 | 0.339 |
| **Industry Measures (US CMP High-Tech Clusters)** | | | |
| Biotech Sector | Estimated from name | 0.002 | 0.044 |
| Ecommerce Sector | Estimated from name | 0.050 | 0.218 |
| IT Sector | Estimated from name | 0.022 | 0.147 |
| Medical Dev. Sector | Estimated from name | 0.028 | 0.166 |
| Semiconductor Sector | Estimated from name | 0.000 | 0.020 |
| Observations | | 27,976,477 | |

(2) US CMP Cluster Dummies are estimated by using a sample of 10M firms and comparing the incidence of each word in the name within and outside a cluster, then selecting the words that have the highest relative incidence as informative of a cluster. Firms get a value of 1 if they have any of those words in their name. The procedure is explained in detail in our Data Appendix.
(3) Note that there are also firms that we cannot associate with local nor traded industries.

TABLE 2. LOGIT UNIVARIATE REGRESSIONS

**Firm Name Measures:**

| *Variable* | *Coefficient* | *Pseudo R2* |
|---|---|---|
| Short Name | 3.147*** | 0.018 |
| | (0.0612) | |
| Eponymous | 0.299*** | 0.003 |
| | (0.0168) | |

**Corporate Form Measures:**

| *Variable* | *Coefficient* | *Pseudo R2* |
|---|---|---|
| Corporation | 3.375*** | 0.016 |
| | (0.0769) | |
| Delaware | 23.72*** | 0.088 |
| | (0.427) | |

**IP Measures:**

| *Variable* | *Coefficient* | *Pseudo R2* |
|---|---|---|
| Patent | 88.50*** | 0.059 |
| | (2.225) | |
| Trademark | 45.30*** | 0.016 |
| | (1.882) | |

**Industry Measures (US CMP Clusters):**

| *Variable* | *Coefficient* | *Pseudo R2* |
|---|---|---|
| Local | 0.206*** | 0.011 |
| | (0.00848) | |
| Traded Resource Intensive | 0.952 | 0.000 |
| | (0.0243) | |
| Traded | 1.208*** | 0.001 |
| | (0.0212) | |
| Biotech Sector | 12.22*** | 0.004 |
| | (0.723) | |
| Ecommerce Sector | 1.823*** | 0.002 |
| | (0.542) | |
| IT Sector | 5.463*** | 0.012 |
| | (0.146) | |
| Medical Dev. Sector | 3.486*** | 0.006 |
| | (0.102) | |
| Semiconductor Sector | 12.61*** | 0.001 |
| | (1.517) | |

| | |
|---|---|
| Observations | 18,764,856 |

Logit univariate regressions of Growth (IPO or Acquisition within 6 years) with each of the observables we develop for our dataset. Incidence rate ratios reported; Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001

TABLE 3. GROWTH PREDICTIVE MODEL - LOGIT REGRESSION ON IPO OR ACQUISITION WITHIN 6 YEARS

| | | Preliminary Models | | Nowcasting (up to real-time) | Full (2 year lag) |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| **Corporate Governance Measures** | | | | | |
| Corporation | 4.070*** | | | 3.276*** | 3.061*** |
| | (0.0975) | | | (0.0788) | (0.0739) |
| Delaware | 23.28*** | | | 18.22*** | |
| | (0.451) | | | (0.363) | |
| **Name-Based Measures** | | | | | |
| Short Name | | 2.815*** | | 2.487*** | 2.263*** |
| | | (0.0546) | | (0.0491) | (0.0456) |
| Eponymous | | 0.218*** | | 0.296*** | 0.315*** |
| | | (0.0123) | | (0.0168) | (0.0179) |
| **Intellectual Property Measures** | | | | | |
| Patent | | | 49.57*** | | |
| | | | (1.572) | | |
| Trademark | | | 7.772*** | | 3.964*** |
| | | | (0.484) | | (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 Clusters** | | | | Yes | Yes |
| **US CMP High-Tech Clusters** | | | | Yes | Yes |
| N | 18,764,856 | 18,764,856 | 18,764,856 | 18,764,856 | 18,764,856 |
| R-squared | 0.135 | 0.057 | 0.094 | 0.163 | 0.187 |

We estimate a logit model with Growth as the dependent variable. Growth is a binary indicator equal to 1 if a firm achieves IPO or acquisition within 6 years and 0 otherwise. Growth is only defined for firms born in the cohorts of 1988 to 2008. This model forms the basis of our entrepreneurial quality estimates, which are the predicted values of the model. Incidence ratios reported; Robust standard errors in parenthesis. * p<0.05 ** p<0.01 *** p<0.001

TABLE 4. 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) |
| **Patent - Delaware Interaction** | | | |
| 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 Clusters** | Yes | Yes | Yes |
| **US CMP High-Tech Clusters** | Yes | Yes | Yes |
| 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. * $p < .05$ , ** $p < .01$, *** $p < .001$

FIGURE 1

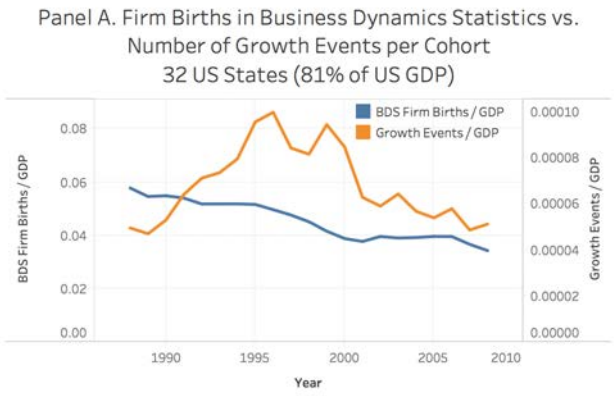
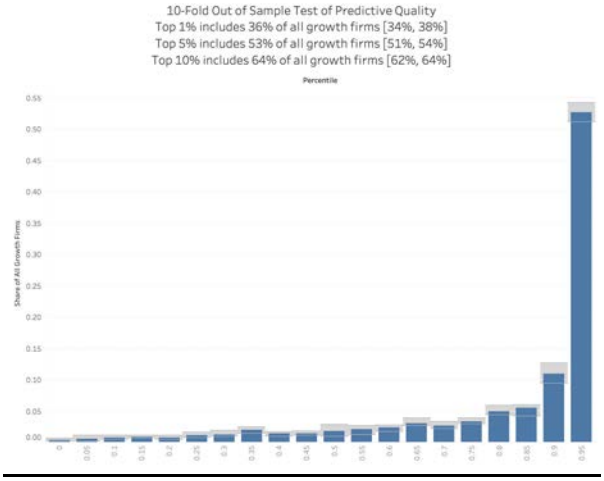
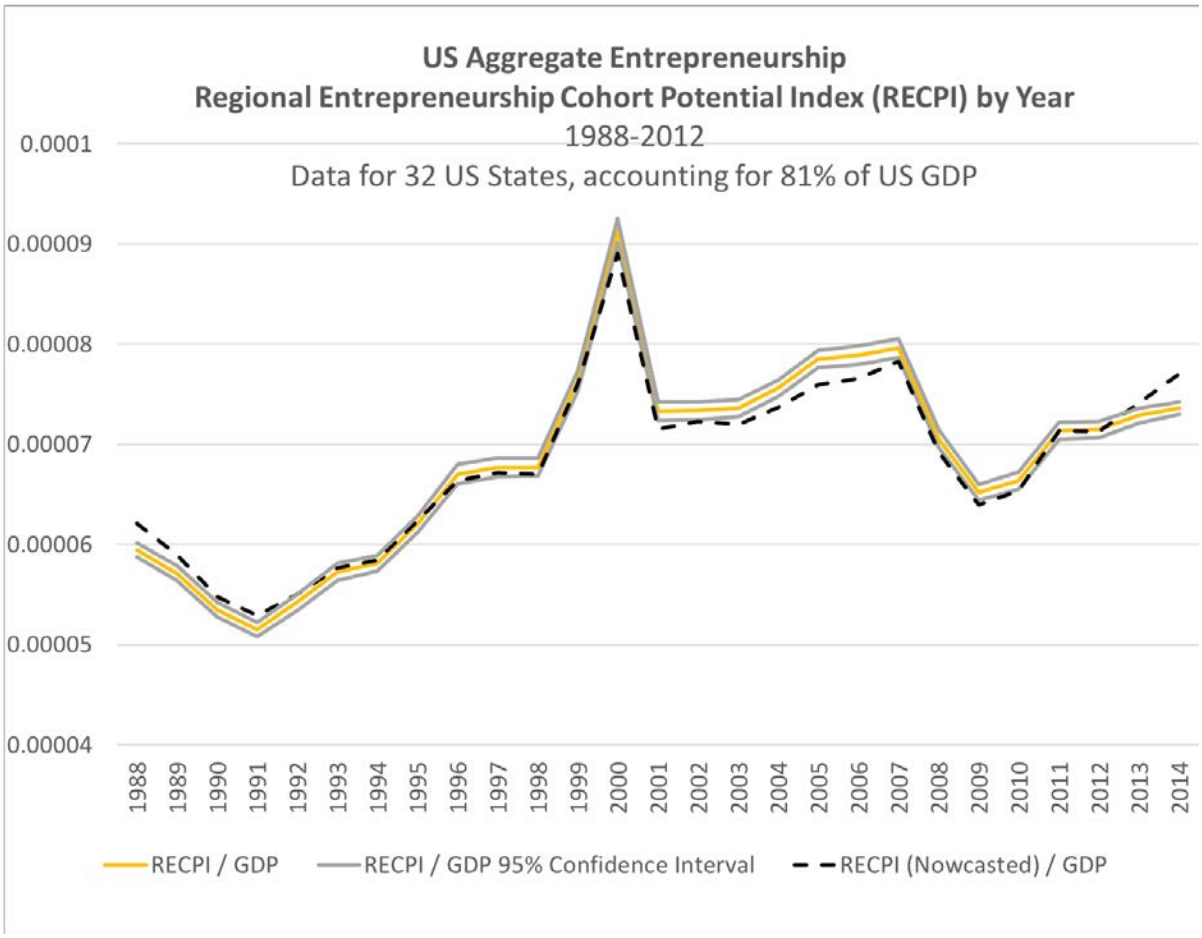



FIGURE 2



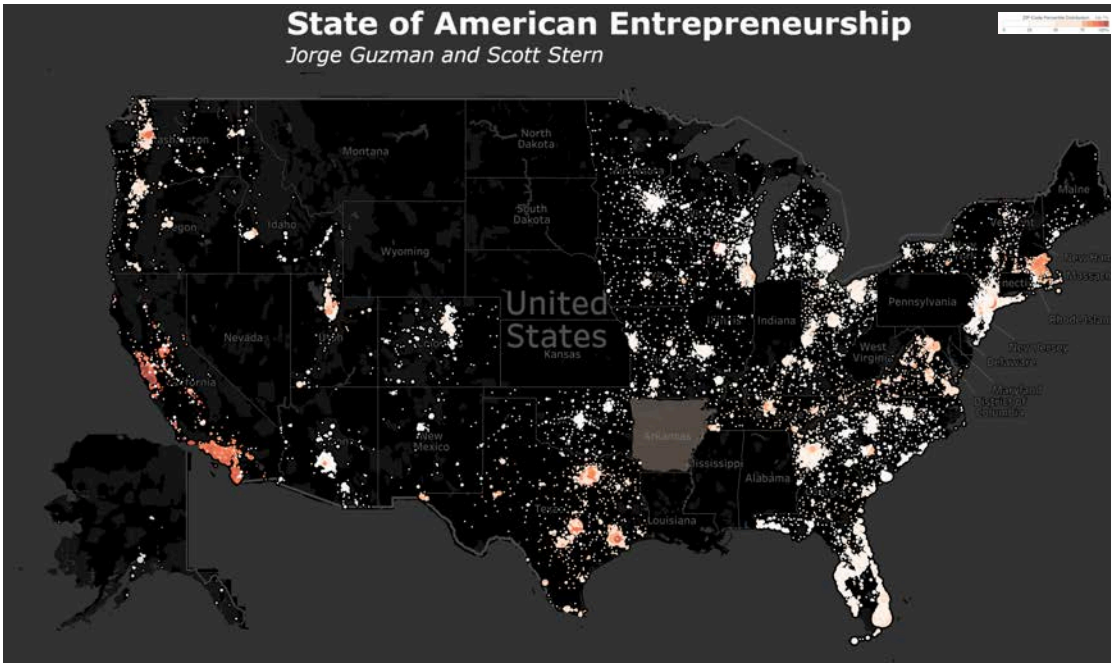

Notes: This Figure presents the results of an out of sample cross validation procedure performed on all firms born between 1988 and 2008 in our database. We use a 10-fold cross validation and plot the incidence of growth across each 5-percent bin.

FIGURE 3



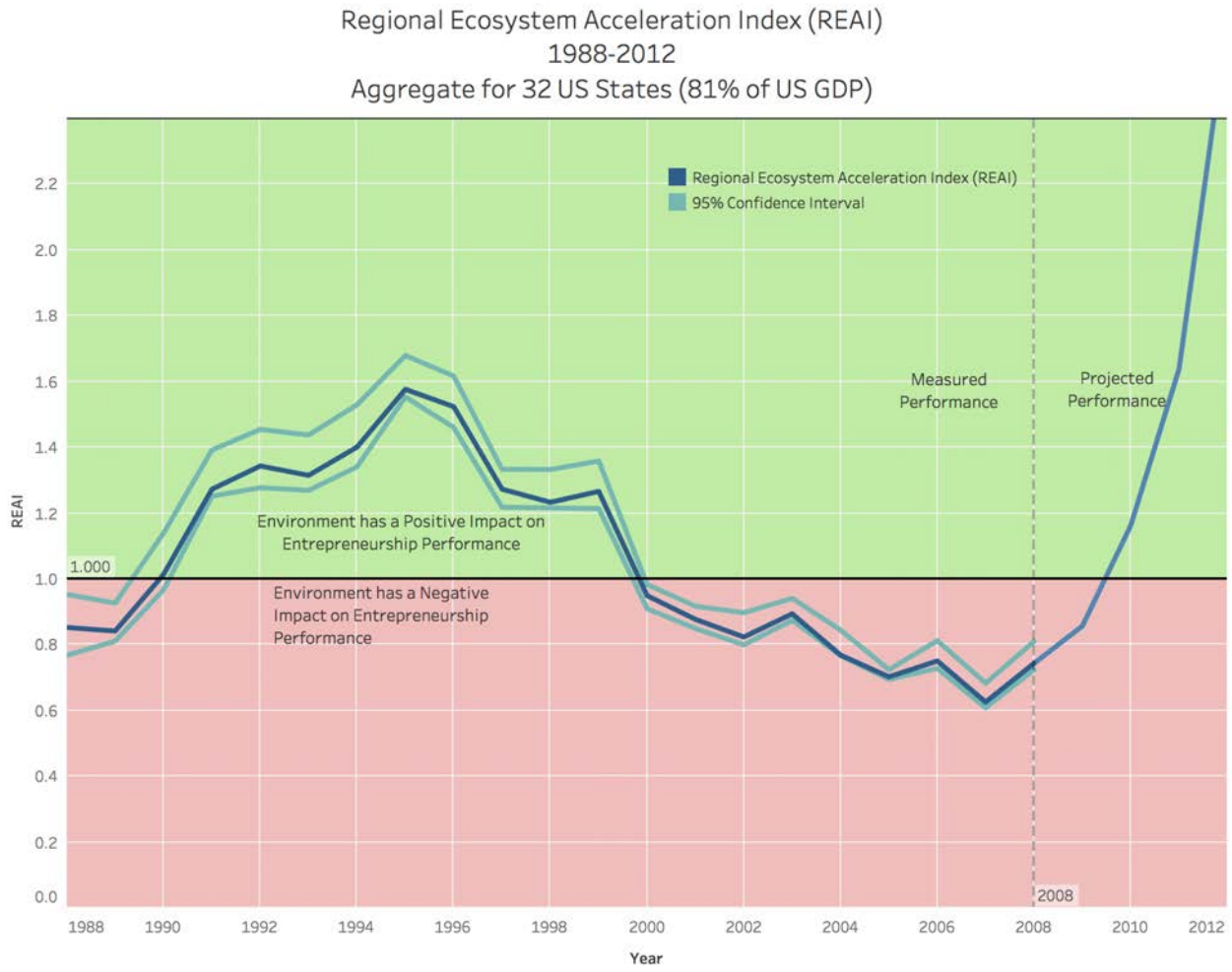

Notes: RECPI / GDP represents the total, quality adjusted, entrepreneurship production in a region after controlling for the size of the economy in that year.

FIGURE 4. THE STATE OF AMERICAN ENTREPRENEURSHIP




Notes: This map represents the quality and quantity of entrepreneurship in 2012 across all 32 US states for which we have data in our sample. Data is presented by ZIP Code. The size of the point represents the quantity of firms and the color of the point represents the average quality of entrepreneurship in that ZIP Code (white is the lowest average quality and dark red the highest). The entrepreneurial quality model includes state fixed-effects to account for the institutional differences in firm registration. For six states, Washington, Oklahoma, Arkansas, Ohio, Virginia, and New York, our data does not allow us to track the precise ZIP Code in which the firm is located for all firms.

FIGURE 5

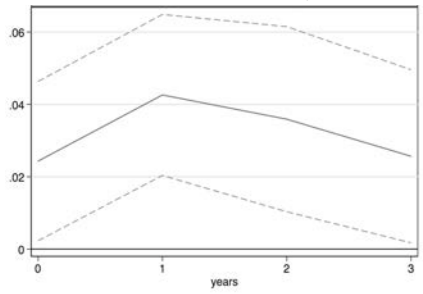



Notes: The Regional Entrepreneurship Acceleration Index (REAI) measures the realized performance of an entrepreneurial ecosystem relative to the expected potential of that ecosystem. It is defined as the number of growth events (IPO or acquisition within six years of founding) to occur from a cohort over the RECPI of that cohort. Confidence intervals are estimated through a Monte Carlo process drawing 30 random samples of size N. Projected performance is a preliminary estimate given the number of growth events that have occurred so far for that cohort.

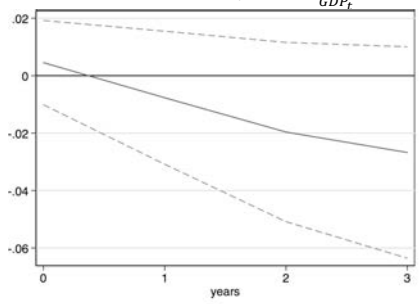
FIGURE 6. EFFECT OF ECONOMIC CONDITIONS ON ENTREPRENEURSHIP

IMPULSE RESPONSE FUNCTIONS

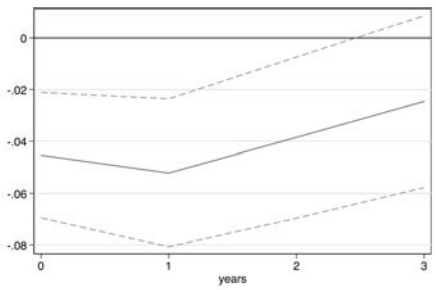
A. Effect of $Ln(\Delta GDP_t)$ on $Ln(\frac{RECPI_t}{GDP_t})$



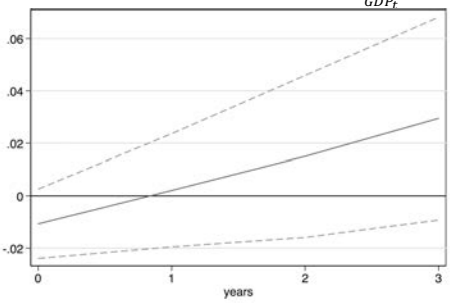

B. Effect of $Ln(\Delta GDP_t)$ on $Ln(\frac{Obs_t}{GDP_t})$




C. Effect of $1[US\_Recession_t]$ on $Ln\left(\frac{RECPI_t}{GDP_t}\right)$




D. Effect of $1[US\_Recession_t]$ on $Ln(\frac{Obs_t}{GDP_t})$




Notes: This figure reports the relationship of economic conditions to entrepreneurship production in the United States. ΔLn(GDP(t)) is the log change in US GDP between years t-1 and t. 1[US Recession(t)] is a dummy variable equal to 1 if the US was in a recession in that year. The years of recession are 1990, 2001, 2008, and 2009. All lag structures are chosen as those that maximize the Akaike's Information Criterion (AIC) and the Schwarz's Bayesian Information Criterion (SBIC), which agree in all cases. All models pass a VAR stability test, with all eigenvalues within the unit circle. A Granger causality test rejects the null of no relationships ($p < 0.01$) for the ΔLn(GDP) models, and is marginally unable to reject it for the US Recession models ($p = 0.13$).

PRINT APPENDIX

***Data Coverage US States - Ranked by GDP***

| *Rank in US GDP* | *State* | *GDP* | *Share of GDP* |
|---|---|---|---|
| 1 | *California* | $2,287,021 | 13.0% |
| 2 | *Texas* | $1,602,584 | 9.1% |
| 3 | *New York* | $1,350,286 | 7.7% |
| 4 | *Florida* | $833,511 | 4.7% |
| 5 | *Illinois* | $742,407 | 4.2% |
| 7 | *Ohio* | $584,696 | 3.3% |
| 8 | *New Jersey* | $560,667 | 3.2% |
| 9 | *North Carolina* | $491,572 | 2.8% |
| 10 | *Georgia* | $472,423 | 2.7% |
| 11 | *Virginia* | $464,606 | 2.6% |
| 12 | *Massachusetts* | $462,748 | 2.6% |
| 13 | *Michigan* | $449,218 | 2.6% |
| 14 | *Washington* | $425,017 | 2.4% |
| 17 | *Minnesota* | $326,125 | 1.9% |
| 18 | *Colorado* | $309,721 | 1.8% |
| 19 | *Tennessee* | $296,602 | 1.7% |
| 20 | *Wisconsin* | $293,126 | 1.7% |
| 21 | *Arizona* | $288,924 | 1.6% |
| 22 | *Missouri* | $285,135 | 1.6% |
| 25 | *Oregon* | $229,241 | 1.3% |
| 27 | *Oklahoma* | $192,176 | 1.1% |
| 28 | *South Carolina* | $190,176 | 1.1% |
| 29 | *Kentucky* | $189,667 | 1.1% |
| 30 | *Iowa* | $174,512 | 1.0% |
| 32 | *Utah* | $148,017 | 0.8% |
| 34 | *Arkansas* | $129,745 | 0.7% |
| 39 | *New Mexico* | $95,310 | 0.5% |
| 43 | *Idaho* | $66,548 | 0.4% |
| 46 | *Alaska* | $60,542 | 0.3% |
| 47 | *Maine* | $56,163 | 0.3% |
| 50 | *Rhode Island* | $45,962 | 0.3% |
| 52 | *Vermont* | $30,723 | 0.2% |
| | *Total GDP in Sample* | $13,941,781 | |
| | *Number of States* | 32 | |
| | *US GDP* | $17,565,783 | |
| | ***Share of GDP in Sample*** | **81%** | |

**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, MIT**
**Scott Stern, MIT and NBER**

**For online publication**

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) (1) | Full (2 year lag) (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

| Dependent Variable | (1) Equity Growth (IPO or Acquisition) | (2) Employment > 500 | (3) 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) | | | 0.011 | 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

## 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 $H = \{h_1, \dots, h_N\}, h_i \in \{0,1\}\ \forall h_i$.

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.

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

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

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

*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 $\theta$ 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. More details are available at http://thomsonreuters.com/sdc-platinum/

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., $\rho_i = \frac{\sum_{j=\{\text{local firms}\}} 1[w_i \subseteq \text{name}_j]}{\sum_{j=\{\text{non-local firms}\}} 1[w_i \subseteq \text{name}_j]}$). We then define a list of Top Local Words, defined as those words that are (a) within the top quartile of $\rho_i$ 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, and one from November 24, 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, 2015b 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 in November 24, 2014). We also geocode each ZIP Code to assess the distance of any geographic

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

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 2.4% and 3.3%. 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 4 includes this result. For the cohort of 10-year-old firms, we estimate 88% of the records to still contain the original ZIP Code, and for 26 year old firms we estimate this share at 83%. 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 76% of records still contain the original ZIP Code by 10 years and 72% by 26 years. In unreported tests, we find the share of firms that move in the top 1% is not meaningfully higher than the top 10%.

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

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 ($25^{th}$ percentile is 3.8), 50% of all firm moves are on less than 8 miles ($50^{th}$ percentile is 7.8), and 90% of all moves are 35 miles or less ($90^{th}$ percentile is 35.24). The top 10% has a similar median (6.8) though higher variance ($90^{th}$ percentile is 330 miles).

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 .017 with a p-value of .27 using robust standard errors and an $R^2$ of .0005. This effect is (basically) indistinguishable from zero.

We also regress log-age on difference in ZIP Code quality to get a coefficient of -.016 with a p-value of .40 and $R^2$ of .0002.

*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 7,485 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, 3,348 firms (44.73%) are re-registrations, which are 3,035 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, 591,423 firms, suggesting that at most 0.5% of firms can potentially have a bias. We identify 1,932 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 0.76% of all Delaware firms are re-registrations of firms changing corporate form, while the other 99.4% 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 (78%), and New Jersey (73%) and lowest in Florida (35%); California (55%) 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 New York (.973) and the lowest in North Carolina (.528), all other states are between .598 and .960. 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 41% to 64%. 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. 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.

## 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.

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.

## TABLE B1

Goodness of Fit Measures of Entrepreneurial Quality Model Across States

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.

| 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 | 66.7% | 66.7% | 63.8% |
| Arizona | 95 | 61.5% | 72.7% | 87.8% |
| California | 4166 | 55.3% | 62.5% | 94.7% |
| Colorado | 49 | 100.0% | 100.0% | 65.9% |
| Florida | 1148 | 34.8% | 43.0% | 91.5% |
| Georgia | 475 | 63.0% | 71.7% | 96.0% |
| Iowa | 60 | 66.7% | 66.7% | 79.6% |
| Idaho | 43 | 83.3% | 100.0% | 91.6% |
| Illinois | 332 | 68.3% | 78.0% | 88.6% |
| Kentucky | 111 | 57.1% | 72.7% | 82.3% |
| Massachusetts | 1069 | 77.7% | 84.5% | 94.9% |
| Maine | 22 | - | - | - |
| Michigan | 176 | 42.1% | 47.4% | 84.9% |
| Minnesota | 298 | 55.6% | 70.0% | 94.7% |
| Missouri | 134 | 53.8% | 60.0% | 84.5% |
| North Carolina | 185 | 42.1% | 47.4% | 52.8% |
| New Jersey | 431 | 73.3% | 74.5% | 94.6% |
| New Mexico | 29 | 100.0% | 100.0% | 84.1% |
| New York | 883 | 67.4% | 70.5% | 97.0% |
| Ohio | 344 | 45.7% | 58.6% | 93.6% |
| Oklahoma | 109 | 62.5% | 62.5% | 85.2% |
| Oregon | 218 | 70.0% | 80.0% | 94.5% |
| Rhode Island | 3 | - | - | - |
| South Carolina | 103 | 58.3% | 75.0% | 59.8% |
| Tennessee | 177 | 63.6% | 63.6% | 91.6% |
| Texas | 1785 | 45.4% | 57.4% | 97.3% |
| Utah | 220 | 59.1% | 68.8% | 94.9% |
| Virginia | 212 | 62.1% | 71.4% | 93.5% |
| Vermont | 17 | - | - | - |
| Washington | 326 | 58.8% | 70.6% | 91.5% |
| Wisconsin | 140 | 56.3% | 57.1% | 78.2% |
| Average | | 63.8% | 67.4% | 86.0% |

## 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 | 65% | 44% | 28% | 0.87 |
| 1989 | 225 | 68% | 51% | 31% | 0.94 |
| 1990 | 273 | 61% | 45% | 26% | 0.94 |
| 1991 | 320 | 57% | 47% | 22% | 0.89 |
| 1992 | 373 | 61% | 47% | 25% | 0.95 |
| 1993 | 404 | 62% | 44% | 26% | 0.94 |
| 1994 | 442 | 62% | 47% | 23% | 0.93 |
| 1995 | 557 | 57% | 41% | 20% | 0.94 |
| 1996 | 612 | 67% | 53% | 27% | 0.93 |
| 1997 | 528 | 67% | 53% | 32% | 0.93 |
| 1998 | 534 | 69% | 57% | 37% | 0.96 |
| 1999 | 648 | 72% | 59% | 41% | 0.89 |
| 2000 | 627 | 73% | 64% | 52% | 0.96 |
| 2001 | 451 | 66% | 51% | 34% | 0.96 |
| 2002 | 442 | 58% | 48% | 31% | 0.94 |
| 2003 | 493 | 61% | 47% | 29% | 0.93 |
| 2004 | 452 | 60% | 50% | 35% | 0.95 |
| 2005 | 428 | 62% | 51% | 34% | 0.94 |
| 2006 | 493 | 61% | 54% | 39% | 0.93 |
| 2007 | 409 | 65% | 58% | 42% | 0.95 |
| 2008 | 433 | 63% | 53% | 38% | 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 | 2.5% | 97.5% | 7.5% | 92.5% |
| 2-4 | 3.3% | 94.3% | 5.7% | 87.2% |
| 4-6 | 2.4% | 92.0% | 4.3% | 83.4% |
| 6-8 | 1.7% | 90.4% | 4.1% | 80.0% |
| 8 | 1.4% | 89.2% | 2.0% | 78.4% |
| 10 | 1.1% | 88.2% | 2.4% | 76.5% |
| 12 | 1.0% | 87.3% | 1.4% | 75.4% |
| 14 | 1.0% | 86.4% | 1.6% | 74.2% |
| 16 | 0.8% | 85.7% | 0.7% | 73.7% |
| 18 | 0.7% | 85.1% | 0.6% | 73.3% |
| 20 | 0.6% | 84.6% | 0.6% | 72.8% |
| 22 | 0.6% | 84.1% | 0.7% | 72.3% |
| 24 | 0.4% | 83.8% | 0.2% | 72.2% |
| 26 | 0.3% | 83.5% | 0.4% | 71.9% |

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 Original Regression | 2 Growth (Only Acq >= 100M) | 3 IPO in 6 Years | 4 Growth in 9 Years | 5 IPO (Ever) | 6 Growth (Only Acq >= 100M) | 7 IPO in 6 Years | 8 Growth in 9 Years | 9 IPO (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 |

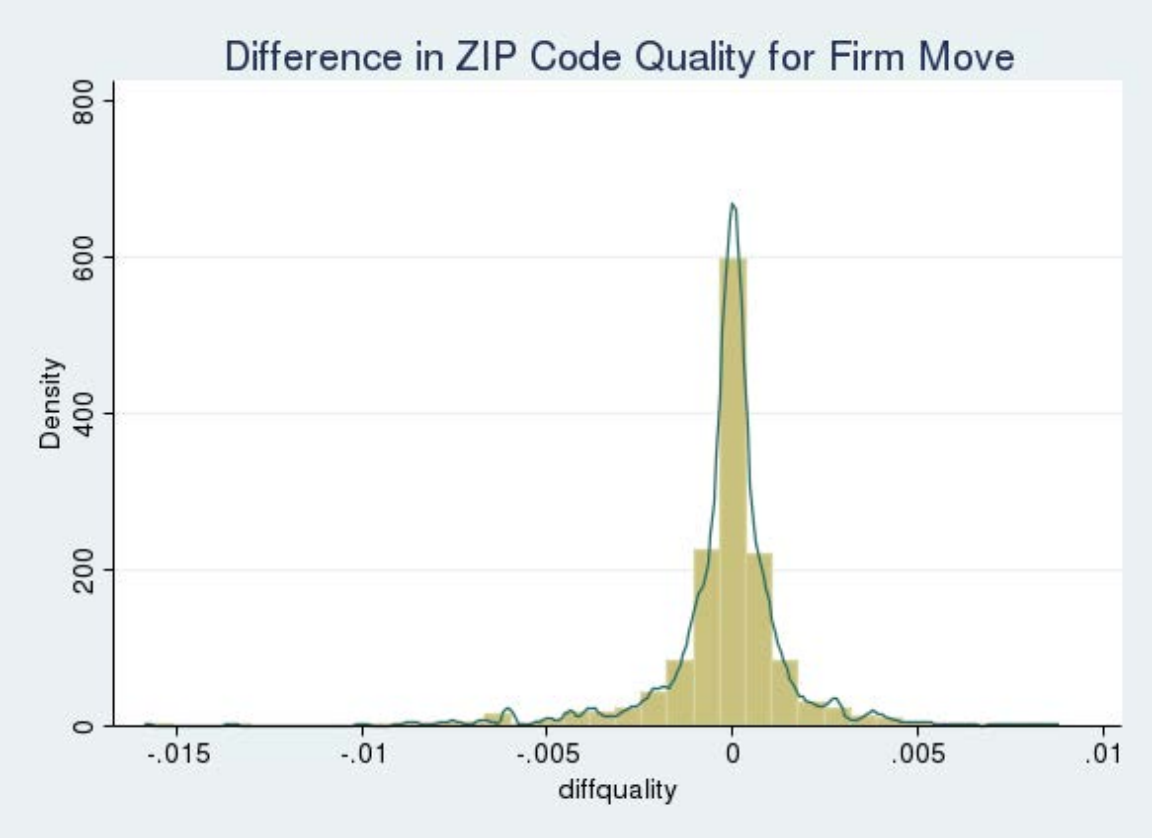
# TABLE B5

**Re-Registrations in Massachusetts**

| | |
|---|---|
| *General Statistics* | |
| Total Massachusetts Firms in Sample | 591,423 |
| Firms founded through a re-registration | 3,035 |
| Share of Firms Founded through re-registration | 0.51% |
| Re-incorporations with source and destination firm in sample | 1,932 |
| *Corporations* | |
| Firms that Gain Corporation = 1 | 640 |
| Total Corporations in Sample | 358,978 |
| Share | 0.18% |
| *Delaware Jurisdiction* | |
| Firms that Gain Delaware = 1 | 245 |
| Total Delaware Firms in Sample | 32,194 |
| Share | 0.76% |
| *Patents* | |
| Firms that Gain Patent = 1 | 43 |
| Total Patent Firms in Sample | 2,373 |
| Share | 1.81% |
| *Trademark* | |
| Firms that Gain Trademark = 1 | 30 |
| Total Trademark Firms in Sample | 1,365 |
| Share | 2.20% |
| *Short Name* | |
| Firms that Gain Short Name = 1 | 222 |
| Total Short Name Firms in Sample | 265102 |
| Share | 0.08% |

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
200
400
600
800
-.015
-.01
-.005
0
.005
.01
diffquality