https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 1, January 2023, pp. 101–118
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Measuring Founding Strategy

Jorge Guzman,[a] Aishen Li[b]

[a] Columbia University, New York, New York 10027; [b] Tsinghua University, Beijing 100190 China

Contact: jag2367@columbia.edu, https://orcid.org/0000-0002-0826-8306 (JG); las21@mails.tsinghua.edu.cn (AL)

Received: November 18, 2019
Revised: March 25, 2021; October 11, 2021; October 25, 2021
Accepted: October 31, 2021
Published Online in Articles in Advance: June 9, 2022

https://doi.org/10.1287/mnsc.2022.4369

Abstract. We introduce a novel approach to measure the founding strategic differentiation of startups and its relationship to follow-on performance. We use natural language processing and historical websites to estimate the similarity between the founding website of an individual startup, the historical website of public firms at the startup's founding year, and the founding website of other startups founded in the same year. We propose that distance in the value proposition stated in these websites represents differentiation in the market. Startup differentiation is estimated as the average text-based distance from the five closest incumbents (public firms). We implement this approach using a large sample of startups from Crunchbase. Our measure predicts a meaningful increase in early-stage financing and equity outcomes, unconditionally and controlling for cohort and industry fixed effects. The positive benefits of equity outcomes only evidence themselves after year 6 of age, suggesting more differentiated firms may take longer to prove themselves. Using out-of-sample tests, we also demonstrate that our measure is economically important, predicting 30% of the total variation in the receipt of early-stage financing and 20% of variation in equity outcomes. Public datasets of our differentiation score and scraped website data are provided, together with open-source code to replicate our approach in other settings.

History: Accepted by Joshua Gans, business strategy.

Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mnsc.2022.4369.

Keywords: entrepreneurship • strategy • strategic positioning • machine learning • text analysis

1. Introduction

Successful entrepreneurial strategies depend on both early-stage firm positioning and experimentation under uncertainty (Porter 1996, Gans et al. 2018, Gambardella et al. 2020, Koning et al. 2022). At the core of positioning is a startup's strategic differentiation: whether it can stake out a unique value proposition in the consumer market. However, the role and relative importance of this founding differentiation, and positioning more generally, has been a matter of debate. Although recent frameworks of entrepreneurial strategy posit an important role for early positioning (Siggelkow 2001, Gans et al. 2018), others, including leading practitioners, consider it hardly consequential (McGrath and MacMillan 2000, Reis 2011). The need for empirically assessing the strategic differentiation of startups and the relationship of this differentiation to performance outcomes appears evident. Yet, there is a key measurement challenge to do so: quantifying how unique startups are at their early stages requires observing all startups early in their lifecycle and developing a systematic way to understand their differences to the incumbent industry structure *at founding*.[1] How could one systematically evaluate founding differentiation, and positioning in general, for startups? What is the association of such measure with follow-on performance? More broadly, how can we measure a startup's founding strategy?

This paper introduces a novel approach to measure the founding differentiation of startups and assess its relationship to follow-on performance. To do so, we apply natural language processing methods to text written by startups and incumbents around the time of a startup's founding and create measures of the differentiation of each startup from the incumbent industry structure. Concretely, using the WaybackMachine (an online historical archive of Internet web pages), we download a copy of the website at or close to founding for a large sample of startups from Crunchbase and the website of all public companies during the startup's founding year. We then use the word-embeddings algorithm doc2vec (Le and Mikolov 2014) to create vectors of embeddings that represent the use of text in each website and estimate the cosine distance between these embeddings vectors. Using work in industrial organization as a heuristic, we aggregate distance measures to only consider the distance from the five incumbents closest to the focal firm (Bresnahan and Reiss 1991, Igami and Uetake 2020). We call this new measure the startup's

Differentiation Score. Conceptually, this measure represents the distance between the value proposition stated by a startup on its website at founding, and the value proposition stated by other companies in the market. After documenting our measure, we empirically show it predicts performance outcomes and statistically accounts for an important share of their variance. Together, our method and results allow us to measure founding strategy and its correlation to performance. These results are accompanied by public code and data to use our measures and expand on our approach in other settings.

Before delving into further details of our method and results, we illustrate the intuition of how firm statements can be used to measure differentiation by considering how a human analyst may assess the level of strategic differentiation amongst firms. An analyst could do so by studying company marketing statements, where companies tend to emphasize their differences and unique value propositions. Consider Southwest Airlines and Delta Airlines. Southwest's slogan is "Low fares. Nothing to hide. That's TransFarency!" This slogan emphasizes low cost and transparency, which will be particularly appealing to cost-sensitive customers tired of extra fees. Delta instead uses the slogan "World's Most Trusted Airline," which does not focus on low cost but instead on trust and global coverage. Trust and global coverage might not be as valuable for the cost-sensitive travelers to whom Southwest caters but will be for those travelers that seek to get anywhere reliably and on time and are willing to pay extra to do so.[2]

In this simple comparison, a strategy analyst can quickly and intuitively identify the differences between these two statements, even for companies in the same industry. These differences in statements do not simply reflect differentiation in the product features. The product in this case is ostensibly similar (a flight), and the statements instead capture the value proposition, be it variety (in route coverage, for Delta) or cost leadership (Southwest). Even for companies whose competitive advantage is not a unique product, but the ability to deliver a lower price, as in Southwest, these features are emphasized in the company's marketing to consumers. Now, if the strategy analyst is given a third company—such as Spirit Airlines, which has the slogan "Less Money, More Go."—they will also recognize a difference in the perceived distance between this new statement and the prior two. Spirit Airlines appears closer to Southwest, because both focus on the importance of low cost and will therefore impinge on the differentiation of Southwest more than of Delta.

Our approach expands this idea to all companies and a richer set of company statements. Given a consistent source of marketing materials—such as the website—of a large sample of companies in the United States, an analyst could expand on the previous approach to develop a more substantive characterization of the distance among firms and estimate the level of strategic differentiation of each one. In this paper, we build on this idea by using natural language processing as a scalable tool to assess distance in the historical founding websites of startups and public firms and measure strategic differentiation to public firms at founding.

After developing our measure of strategic differentiation, we begin validating it qualitatively by performing an in-depth look at the ranking of all startups in two Crunchbase categories: Consumer Electronics and Food and Beverage. The examples we highlight at the top of our measure, compared with others at the bottom, appear more meaningfully differentiated, and the public firms that are closest to them are (while relevant) only loosely related. We conclude our measure is consistent with the type of differences we would intuitively expect in a measure mapping the construct of strategic differentiation to the data.

Next, we empirically study how our measure predicts startup performance outcomes. Although these estimates are not causal, they represent useful comparative statics of the relationship between founding strategy and eventual firm success and are both validations of our measure and new facts on the descriptive association between founding strategy and performance.[3] We focus on three key results.

First, we consider how founding strategy predicts early-stage financing. We focus on regressions with founding year and industry fixed effects, where industries represent text-based industries replicating the approach of Hoberg and Phillips (2016). Compared with firms in the 10th percentile of our measure, firms in the 90th percentile are 10% more likely to raise early-stage financing and raise 117% more total early-stage financing.[4] In contrast—and consistent with the idea that we are capturing positioning vis-á-vis the consumer market and not the financing market—differentiation from other startups has much lower magnitudes and is negative when both types of differentiation measures are included in the same regression. Founding differentiation, particularly from the existing market structure, predicts the receipt of short-term financing.

Second, we use a similar specification to consider how founding strategy predicts long-term equity outcomes such as initial public offering (IPO) or acquisition. The relationship here is more nuanced. Differentiation is positive over the startup lifecycle, but there is a significant age dependency. Indeed, consistent with the literature documenting that more unique startups will struggle to achieve legitimacy initially but may ultimately perform better (Deephouse 1999, Marx et al. 2014), there is an initial negative relationship between our measure and the cumulative probability of equity outcomes that inverts after age 6, eventually predicting substantially higher outcomes for firms that have higher differentiation. Compared with firms in the 10th percentile, firms in the 90th percentile are 19% less likely to achieve an exit event in their year of birth (relative to the mean), but this changes to a positive cumulative difference of 8% by year 7 and 28% by year 10 of age. This inverted pattern holds for both IPOs and acquisitions, with similar relative magnitudes, and it is even more striking for high-value acquisitions (i.e., more than $100 million). Founding differentiation is a relevant predictor of equity performance.

Finally, third, we move beyond regressions to instead study the extent to which founding differentiation is economically relevant. Using out-of-sample receiver-operating-characteristic (ROC) scores to assess the variance explained, we show that a fully interacted model of four measures of founding differentiation[5] predicts 30% of the variation in the receipt of early-stage financing and 20% of the variation in the receipt of equity growth outcomes. Because our measures are naturally incomplete and imperfect, this estimate is best interpreted as a lower bound on the importance of founding strategy in this sample. Furthermore, if one takes seriously the argument in Teece et al. (1997) that high technology startups are the setting in which founding strategies should matter the least, this would also suggest that the extent to which differentiation predicts performance for firms in general is substantial. Founding strategy not only relates to performance but plays a meaningful role in statistically explaining variation in these outcomes.

Through these results, this paper contributes to two distinct areas of the strategy literature. The most important contribution is methodological. This paper provides a formal way to measure founding strategic differentiation, building from the tenets of strategy. The new measure we propose is different from prior attempts at formalizing strategy in that we focus specifically on how a firm is occupying a distinct value proposition from its competitors, as reflected in its marketing. Although prior work has instead sought to define the nature of what strategy is (Porter 1996, Van den Steen 2016), there are many applications where measurement itself is of fundamental value. We hope that our results provide useful guidance to researchers seeking to measure differences in the level of strategic positioning and startup founding differentiation. We suspect follow-on work will improve upon our approach. To support this effort, we have released through our Github repository all our code as open source and most of our data, including differentiation scores for startups, the raw website text used, and the trained word-embedding models that allow assessing the similarity of new firms to ours.[6]

Our second contribution is to the entrepreneurial strategy literature. Our results provide novel evidence on the importance of founding differentiation, and founding positioning, to performance outcomes. Although prior work has theorized that high technology companies such as those in Crunchbase may gain little from founding positioning, and instead achieve performance through either dynamic capabilities or experimentation (Teece et al. 1997, Eisenhardt and Martin 2000, McGrath and MacMillan 2000, Reis 2011), we show that this perspective may be too extreme. Rather, the essence of startup strategy may require recognizing both positioning and experimentation (Siggelkow 2001, Gans et al. 2018) and how they work together to develop a competitive advantage in a way that is sustainable but also quickly adaptable to rapidly changing contexts. Understanding better the interplay of static positioning and dynamic capabilities and experimentation is an important area for future work.

Perhaps the paper closest to ours is the influential work of Hoberg and Phillips (2016) (hereafter HP16). HP16 uses the text in the business description section of the annual reports of public firms to develop a new approach to understand industries and their dynamics. To review, HP16 uses the cosine similarity between word vectors weighted by the term frequency-inverse document frequency algorithm (tf-idf) to estimate a text-based distance between firm statements. Then, they implement clustering algorithms and propose each cluster represents a different text-based industry, ultimately showing that these industry definitions describe industry dynamics significantly better than Standard Industrial Classification (SIC) codes. Relative to this work, our paper offers several novel contributions.

First, conceptually, our paper's focus is on strategy rather than industry. This means that, although HP16 focuses on how companies agglomerate into groups of related firms, strategy focuses on what makes a company distinct from all potential competitors. Understanding those business-specific elements that drive firm performance beyond industry is at the core of strategy research (Rumelt 1991, McGahan and Porter 1997, Ruefli and Wiggins 2003). The precise construct we measure, strategic differentiation, is not measured in any of the papers by HP16, nor is our analysis of how these measures predict startup financing and equity outcomes or our estimates of the economic importance of founding strategy for technology-based startups. Empirically, our analyses also control for fixed effects of the text-based industries defined by HP16 (which we create by replicating their methodology within our data), and we cluster our standard errors by these text-based industries.

Second, methodologically, our machine learning model implements a more sophisticated natural language processing method than HP16. Our algorithm uses word embeddings, whereas HP16 used cosine similarity of relative frequency (tf-idf). The key difference between the two is that the word-embeddings approach also incorporates the context in which words are used when creating vectors to describe text documents. Concretely,[7] whereas tf-idf counts the presence of each word (or its stem) weighted by how infrequent the word is, word embeddings runs a neural network that tries to predict a word based on the words that surround it to develop a vector of weights representing each word (in our case, for each word, we use a window of the seven prior and seven follow-on words). This use of context makes word-embeddings algorithms give two synonyms similar weights if they are surrounded by similar words, and the same word gets very different weights if it is used in a different context. For example, the words "social media" would receive very distinct weights when they are new, because although both "social" and "media" are common words, they had not been used together in this context and are likely surrounded by other words that did not tend to be neighboring them before. However, the words "Facebook" and "Twitter" will get an embeddings vector similar to each other and to other social media words, because they are often used in the same context. Introducing the role of context turns out to be important in our analyses. In parallel to the idea in strategy that successful positioning requires not only using novel elements but also combining them in unique ways, we find that a measure equivalent to ours using tf-idf to measure distance has a weaker role in predicting equity outcomes. When we study the dynamic effects of differentiation on outcomes, the relationship of the tf-idf measure is negligible once we control for our measure, whereas ours remain robustly related to performance. Furthermore, when we study the variance explained, we show tf-idf only accounts for 3% of all variation in ROC estimates, whereas our measures, in contrast, account for 20%. We conclude that, in our setting, our measure is more meaningful in an economic sense.[8]

Finally, third, from a dat perspective, our paper is also the first to focus on startups and to do so using their founding websites. In contrast to 10K statements, which are created only by public firms and most often years after founding, websites allow observing a larger number of startups close to founding and therefore better understand the relationship of differentiation at founding to ex post performance. Understanding and unpacking further the differences in what is conceptually captured in different firm statements and whether they represent different elements of firm strategy and disclosure is a rich avenue for future work.

The rest of the paper proceeds as follows. Section 2 presents our formalized methodological approach. Section 3 reviews our data. Section 4 presents our results. Finally, Section 5 concludes.

2. Measuring Founding Strategy: A Text-Based Approach

Our approach is anchored around the idea that the relationship between text written by firms can be informative about the underlying market structure (Abrahamson and Hambrick 1997, Hoberg and Phillips 2016). This section overviews how we use firm websites to assess similarity in the value propositions of firms, how we translate this similarity to a measure of distance, and how we aggregate this distance into a measure of strategic differentiation, allowing us to score startup differentiation at founding.

2.1. Measuring Market Differentiation Through Firm Statements

Our approach to measuring strategic differentiation builds on four insights. First, although it is virtually impossible to observe the value a consumer sees in a product, it is possible to observe what the firm believes its value proposition to be. Firms constantly state their value proposition in their marketing statements to explain to consumers (or to some representative set of them) why their product or service should be purchased. Second, the similarity in these firm marketing statements is a good indicator of the substitutability between the value proposition of their offerings, thus allowing the assessment of how differentiated is a new startup from incumbent firms. Third, measuring distance between company statements is not merely a theoretical idea: There are standard text-analysis algorithms that allow us to quantify the relatedness of those statements to effectively create a measure of similarity in the stated value proposition of firms in the market. Distance is then simply the inverse of similarity. Fourth, observing at least some of these statements at or close to founding is possible through the use of archival websites. Because websites represent a de facto marketing channel for virtually all firms founded after a certain date, the distance between founding websites can be used to measure founding positioning.

Building on these insights, we define market relatedness as a measure representing the similarity between two firm statements. Given a startup and an incumbent statements s_i and s_j (one each) explaining their main value proposition, there exists some function h defined between zero and one that can measure a pairwise similarity between these two statements as

$$\sigma_{ij} = h(s_i, s_j), \sigma_{ij} \in [0, 1].$$

Companies with a value of similarity equal to one have completely equivalent statements, whereas companies with a similarity value of zero have no relationship to each other. Companies with partial similarity are in between.

2.2. Implementing Word and Paragraph Embeddings

To define the similarity function h, we focus on a specific natural language processing (NLP) algorithm called word embeddings (word2vec) (Mikolov et al. 2013). Compared with traditional bag-of-words approaches such as the term frequency-inverse document frequency algorithm (tf-idf) or topic modeling, the distinction of word-embeddings algorithms is incorporating the context in which words are used when characterizing them.[9] In essence, whereas in tf-idf a word is weighted only by how uncommon it is across documents, word2vec represents each word through a vector of N factors (embeddings), creating a factor-based description of the word. These factors are estimated through a neural network that predicts the probability a word is used based on other words that occur before or after it. Doing so means that, when the same word is used in a very different way (such as Casper, the friendly ghost, and Casper, the company), it results in a very different vector, whereas if a word is synonymous to another one (such as Nectar, also a mattress company, versus Casper, the mattress company), a similar vector would be estimated even though they are spelled completely different. In contrast, in this example, tf-idf would have delivered the opposite conclusion: scoring the two versions of Casper as the same and different from Nectar. In our implementation, we use the expanded version of word2vec, doc2vec (Le and Mikolov 2014), which allows us to take advantage of this technique to build document-level vectors. Finally, building from the insights of Mu et al. (2017), we include a postprocessing step and subtract the sample mean of each embedding to itself to make all of them mean zero. The expression $h(s_i, s_j)$ is estimated as the cosine similarity between the normalized version of the embeddings vector of any pair of firms i, j.[10]

2.3. From Similarity to Founding Strategy

The next step is to aggregate the pairwise similarity between all startups i and incumbents j into a firm-level measure of differentiation. We first define distance, δ_{ij}, by algebraically inverting σ_{ij}. Distance is a value between zero and one, where zero indicates that two companies are the same, and one means they are completely different.

$$\delta_{ij} = 1 - \sigma_{ij} \qquad (1)$$

Next, we aggregate distance across all incumbents to get an empirical measure of the differentiation score at founding. The mean or median are not good ways to aggregate measures of competitive overlap because most companies are unrelated to each other. Empirical studies in industrial organization highlight how the dynamics of competition are influenced by a small number of competitors and how, as this number increases, the ability of firms to charge margins quickly decreases, approximating a fully competitive economy (in strategy parlance, they lose their competitive advantage). We follow a simple heuristic and use the classic finding of Bresnahan and Reiss (1991) showing that markets become competitive after the first three to five competitors.[11] Although this heuristic is admittedly ad hoc and imperfect, it allows a tractable approach that is applicable across many firms.

The differentiation score is

$$\hat{S}_i = \frac{1}{5}\sum_{j \in J_i^5} \delta_{ij}\,,\ J_i^5 = \{5\ \ closest\ \ incumbents\}. \qquad (2)$$

We also estimate a startup's differentiation from the single closest incumbent, the five closest startups with the same founding year, and the single closest startup with the same founding year.

3. Data: Crunchbase, the Wayback Machine, and Industry Controls

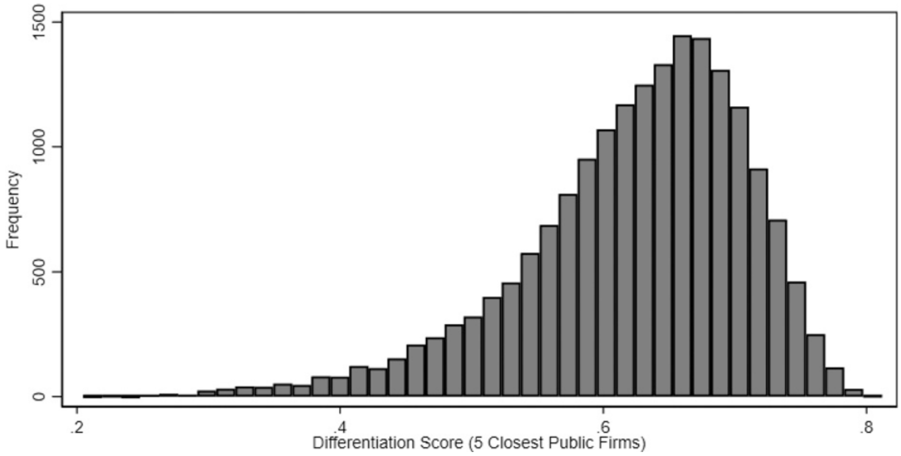
We implement this approach on a comprehensive list of startups from Crunchbase, which we complement with their historical websites at the time of founding, and the annual websites of publicly listed firms in the United States. We also include the industry of each startup, estimated by replicating Hoberg and Phillips (2016) within our data. We describe each data set below.

3.1. Crunchbase Startup Data

We obtained data on all companies available in Crunchbase founded between 2003 and 2019 that have a website. Crunchbase is a popular crowd-sourced data platform tracking a large number of technology-based startup companies. It is one of the main databases used in entrepreneurship and strategy research and performs particularly well in covering innovative firms that receive some form of institutional financing (Dalle et al. (2017) provide an overall assessment and examples of the use of Crunchbase in management and economics research). The data include both active and deceased companies, and we included all firms of any status in our analysis.[12]

For each company, we downloaded in April 2019 the company name, the founding date, the website address, the city and state of the main office, the date and amount of each financing round, whether the company achieved an equity event (IPO or an acquisition), the market valuation of the company at the exit event, the timing of the exit event, and the top-level Crunchbase category for this firm. Crunchbase categories are conceptually industry categorizations, but their focus is on characterizing firms across

Figure 1. Distribution of Strategic Differentiation Score (Five Closest Public Firms)



Notes. Reports the histogram of strategic differentiation score estimated as the mean distance in the founding website for the five closest public firms. Distance is one minus the similarity between websites, which is estimated using a word-embeddings algorithm of all public websites and startups in each cohort.

groups that better delineate startup industries than traditional SIC codes.

3.2. Website History Data with the Wayback Machine

We used the Wayback Machine, an online platform offered by the Internet Archive (archive.org), to download the initial website of each startup around the time of founding. The Wayback Machine provides access to a digital library containing more than 330 billion webpage snapshots occurring in history. These snapshots are taken at least a few times a year for all unique domain names on the Internet. We developed a web-scraping technology, available in our Github repository, to automatically query the Wayback Machine for the earliest version of the web page in the year after the year of founding in Crunchbase. We downloaded the homepage and the first-level links in the web page (up to 10 URLs to limit the size of the download). We excluded all pages that returned empty, that included too little text, that were not in English, that were not of MIME types “text/html” or “text/plain,” that reported an HTTP error such as a 403 or 303, or that appeared to be a boilerplate such as the default page of an Apache web server.

3.3. Incumbent Information

To consider the existing market structure at founding we focus on publicly listed firms. Specifically, using the IPO and de-listing dates in Preqin, we downloaded the first available website each year for all companies publicly listed in NASDAQ and the New York Stock Exchange using the same download algorithm used for the startups. This allows us to observe the market proposition of all public companies as stated at the time of startup founding and thus assess adequately the startup’s positioning in the market at this time.

3.4. Industry Controls

Finally, we develop industry categorization by implementing the method of HP16 to develop clusters of related industries based on the company’s own business descriptions within our data. To review, HP16 uses the term frequency inverse document frequency (tf-idf) algorithm in the business description of 10K annual reports to develop vectors of weighted words and then the cosine similarity between these vectors to estimate a scalar distance from one company to another. Then, they implement a k-means clustering algorithm and use the resulting cluster identifiers as the industry categorization. HP16 recommends using 300 clusters as the target number to mimic the distribution of SIC industries.[13] We implement this method with 300 industries using our website text rather than the 10K business descriptions to define industries within our data. The resulting variable *HP Industries* represents 300 indicators for the clusters created through this method. The median number of startups in an industry is 36 and the average is 50.

3.5. Estimates of Similarity and Founding Strategy

We call our preferred measure *Differentiation Score (5 Closest Public Firms)*, defined as the average distance from the focal startup’s website at founding to the five closest public firms. Figure 1 reports the full distribution of our measure. As is apparent, there are some outliers that have a significantly low differentiation score. To avoid having outliers drive our results, and instead focus on the core correlations of the data, we winsorize the distribution.[14] Table 2 reports summary statistics for four estimated differentiation scores after windsorizing. Our preferred measure has a mean value of 0.64 and a standard deviation of 0.064. The difference between the 10th and 90th percentile is 0.17. We also include measures for the distance from the single closest public firm, the average distance from the five closest Crunchbase startups founded in the same year, and the distance from the closest Crunchbase startup founded in the same year. Table A.1 in the online appendix reports the correlation between these four measures. Measures based on public firms and measures based on startups have very high correlations to each other, greater than 0.9, whereas the correlation between measures in these two groups is slightly lower, ranging from 0.59 to 0.74.

3.6. Summary Statistics

Table 1 presents summary statistics of our data. There are 12,406 startups in the data. The average founding website length for a startup is about 11,000 characters, but this measure is significantly skewed. Moving to outcomes, 68% of firms in Crunchbase receive early-stage financing, which we define as receiving any of angel financing, grant financing, crowdfunding, and seed and preseed rounds. The average early-stage financing received is $909,000. Twenty-nine percent get series A financing, with an average investment of $2.1 million. Eighteen percent of firms achieve an equity growth outcome, of which IPO represents only 1.7% and acquisition is 16%. There are 1.5% firms with a reported sales price of at least $100 million (about 10% of all acquisitions). We categorize these as high-value acquisitions.

4. Results

4.1. Qualitative Validations of the Estimated Differentiation Score

We begin by considering specific examples of how our measure describes companies within industries. To do so, we introduce in Online Appendix B the full list of all companies in our data for two Crunchbase categories, their differentiation score, and the startup's description as written in Crunchbase. We emphasize that both the Crunchbase description and categories are independent of the website text we use to develop our measures and not used anywhere in our approach estimating similarity or the follow-on regressions.

Table B1 in the online appendix is all companies in Consumer Electronics. We focus on a few examples in the extremes. The top-ranked companies appear innovative. For example, Zoi (now Runteq), ranked at the 98th percentile of our measure, was a company developing wearables to do personalized running coaching through sophisticated data analytics, and Healthy Stove, also at the 98th percentile, described as the world's first fully interactive oven. These appear, by and large, innovative consumer electronics products. Indeed, when the online magazine SportTechie covered Zoi, it highlighted particularly their distinctly innovative product ("Zoi is the first option to provide a virtual coach") and its promise ("We look forward to seeing future generations of this concept").[15] Table B3 in the online appendix delves deeper into our measure by reporting the public companies that score closest to Zoi in our algorithm. These include the health and device retailer Brookstone, the sports retailer Dicks Sporting Goods, and the medical device company Biozoom. By and large, they appear adequately related to Zoi as they operate in the broad areas of devices and health. However, they are also meaningfully different, and hence, Zoi scores high in our differentiation score.

At the bottom of the Consumer Electronics category, we find a contrasting set of products that are more typical of existing offerings and hence less differentiated from incumbent firms. These include, CorasWorks, a company developing websites using Microsoft Sharepoint technology, and Henge Docks, a

Table 1. Summary Statistics Crunchbase Firms

	Mean	Standard deviation	Minimum	Maximum
Website text length	11,227.623	16,377	100	99,668
Early-stage financing (thousands $)	909.157	3,344	0	191,000
Gets early-stage financing	0.682	0.466	0	1
Series A financing (thousands $)	2,113.331	5,849	0	100,000
Gets series A	0.290	0.454	0	1
IPO	0.017	0.128	0	1
Acquisition	0.160	0.367	0	1
High value acquisition (100m or more)	0.015	0.12	0	1
Growth	0.177	0.381	0	1
Observations	12,406			

Notes. Data set is all companies in Crunchbase founded since 2003 that raised financing and for whom we where able to download a founding website. Founding website is downloaded from the WaybackMachine as the earliest website available the year after founding. *Early Stage Financing* is defined as all financing that is seed financing, angel financing, or grants. *Website Text Length* is the number of total characters in the downloaded founding website text.

Table 2. Summary Statistics of Strategic Differentiation Score

	Mean	Standard deviation	p10	p50	p90	Minimum	Maximum
Differentiation score (five closest public firms)	0.635	0.0636	0.54	0.64	0.71	0.46	0.74
Differentiation score (five closest cohort startups)	0.654	0.0635	0.57	0.66	0.73	0.32	0.85
Differentiation score (closest public firm)	0.600	0.0722	0.5	0.61	0.69	0.11	0.73
Differentiation score (closest cohort startups)	0.610	0.0771	0.51	0.62	0.7	0.052	0.83
Observations	12,406						

Notes. Strategic differentiation score represents the conceptual distance in the market between a firm and some of its closest competitors. It is estimated in three steps. First, a measure of similarity is estimated between the founding website of all startups in a cohort and the website of all public firms during the startup year of founding. To do so, we use a word embeddings algorithm that accounts for both the incidence of words and their context. Next, distance is defined as one minus this similarity. Finally, differentiation is the average distance to the closest competitors. We report four measures. The distance to the five closest incumbent firms (public firms). Distance to the single closest public firm, distance to the five closest startups from the same cohort, and distance to the single closest startup.

company offering a docking station for the Apple Macbook. Apple itself does not offer a docking station for the Macbook, making Henge Docks's product a valuable one, but it is also in clear ways less distinct than Zoi. Indeed, when AppleInsider reviewed Henge Docks's product, it did not emphasize its distinctiveness or innovativeness, calling it simply "an expensive and elegant way to work at your desk."[16] In Table B4 in the online appendix, we see that closest public companies matched to Henge Docks are also companies that sell other docking stations and computer products including the electronics retailer BestBuy and the computer manufacturers AMX and Gateway. Generally, these appear to be more close competitors to Henge Docks than Zoi's matches. Hence, there is a lower differentiation estimate for Henge Docks to the ongoing market structure.[17]

We repeat this exercise in Table B2 in the online appendix but focus on a different type of category: Food and Beverage. The top companies once again include well-differentiated new ideas such as Coda Signature, a creator of cannabis-infused chocolates and high-end truffles, and Ripe.io, with the slogan "the blockchain of food." Consistent with the idea that Coda Signature's value proposition is innovative, it was awarded the Excellence in Innovation Award in 2019 by the National Cannabis Industry Association.[18] Comparing it to public companies, Coda Signature is markedly distinct. Although it matches to other public companies that may sell high-quality chocolates and health-oriented food, such as the supermarket Publix, the coffee chain Caribou Coffee, and the beverage company Pulse Beverage, the differences between these products and Coda Signature's are substantial. Although cannabis is at the core of Coda Signature's value proposition, none of the public companies operate with any kind of mind-altering substance. As in Consumer Electronics, we once again observe at the bottom of the list firms that are more typical in this industry, including the Olomomo Nut Company, a traditional producer of nut products, and Saucey Sauce, a company creating new sauces and marinades.

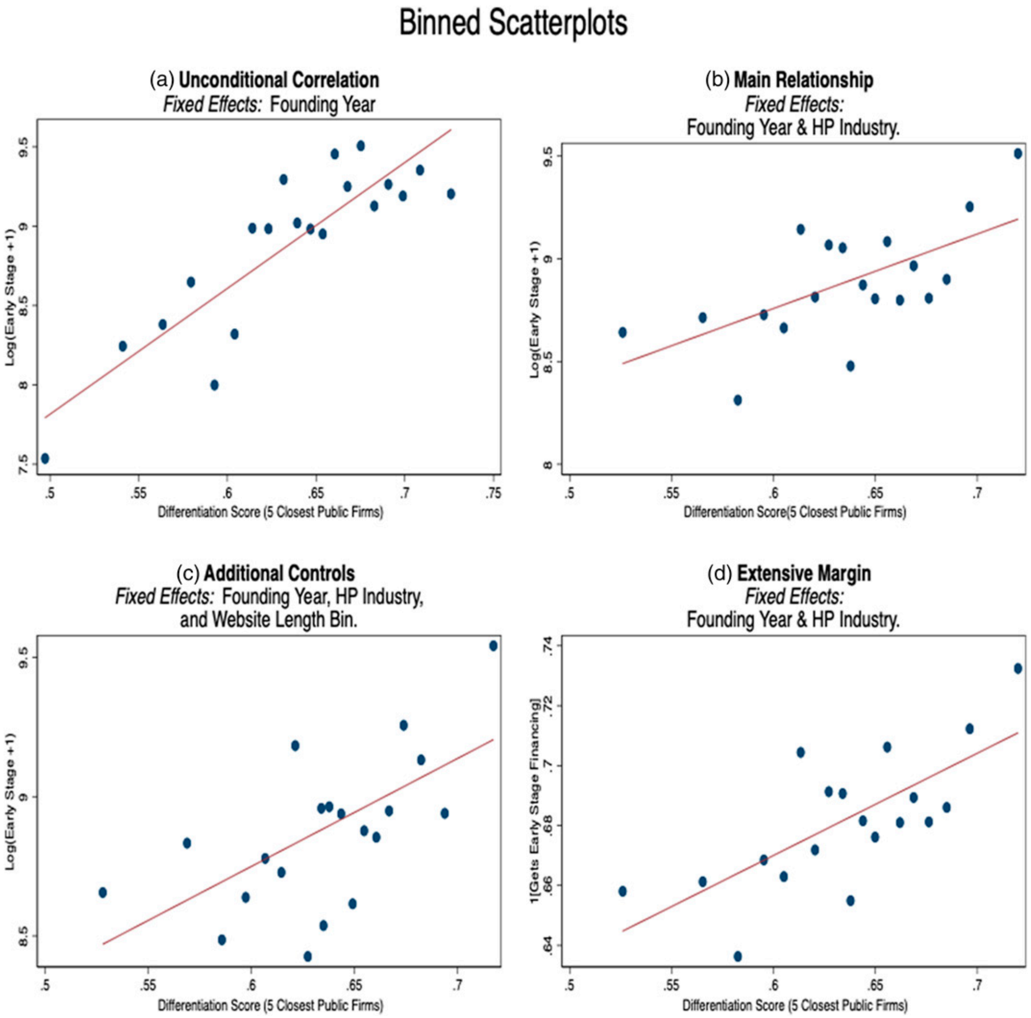
4.2. Founding Differentiation and Early-Stage Financing Outcomes

We next assess the association between our measures and startup financing. Figure 2 presents binned scatterplots correlating our main measure, the differentiation from the five closest public firms, to early-stage financing. Panel (a) includes only year of founding fixed effects. The relationship is positive and very precise. Within startup cohorts, startups with a higher differentiation score raise a higher amount of early-stage financing. Panel (b) replicates what will become our preferred specification, introducing both founding year and Hoberg and Phillips (HP) industry fixed effects. The pattern is slightly less pronounced but still meaningful. This relationship holds even within groups of related competitors. Panel (c) considers the possibility that there is something about the way the websites are built that correlates to both our measure and outcomes by controlling for the length of website text. To do so, we split our variable of website text length into 20 bins and include them as additional fixed effects. Reassuringly, there is little change in the relationship. Finally, Panel (d) reports the extensive margin, whether a startup gets financing at all, with a similarly positive result. Figures A1 and A2 in the online appendix consider other differentiation scores, such as those from one closest public firm or the closest startups, and include outcomes that incorporate series A financing events as early stage. The results are very similar. These graphs suggest significant positive relationships between our measure, estimated close to founding, and whether and how much early-stage financing startups get.

4.2.1. Regression Estimates.

We report the relationship of our measure to financing more precisely in Tables 3 and 4. In Table 3, we study the extensive margin of financing by reporting an ordinary least squares (OLS) regression with *Gets Early Stage Financing* as the dependent variable and the differentiation score as the independent variable. Standard errors are double clustered by HP industry and state to account for industry or location correlation in the error term.

Figure 2. (Color online) Differentiation Score and Early-Stage Financing



Notes. Early-stage financing is all financing events recorded in Crunchbase as "Seed," "Angel," "Crowdfunding," and "PreSeed." HP industries are the industries defined using the methodology of Hoberg and Phillips (2016) in our data. Figure A1 in the online appendix replicates these scatterplots with Series A financing events instead.

Column (1) shows a positive unconditional coefficient of 1.614. Column (2) shows that there are large cohort effects: the coefficient drops to 0.664 after including founding year fixed effects. Column (3) is our preferred specification, which includes founding year and HP industry fixed effects. The coefficient is 0.402.[19] To put this result into perspective, this implies that relative to firms at the 10th percentile of our

Table 3. Does Founding Differentiation Predict the Receipt of Early-Stage Financing?

	(1)	(2)	(3)	(4)
Differentiation score (five closest public firms)	1.614***	0.664***	0.402***	0.372***
	(0.143)	(0.141)	(0.0624)	(0.0529)
Founding year fixed effects	No	Yes	Yes	Yes
HP industry fixed effects	No	No	Yes	Yes
Year × state fixed effects	No	No	No	Yes
City fixed effects	No	No	No	Yes
Observations	12,406	12,406	12,406	12,406
R^2	0.049	0.139	0.198	0.352

Notes. OLS linear probability model. Dependent variable is equal to one if a startup gets early stage financing (seed or angel financing) and zero otherwise. HP industry fixed effects are fixed effects for 300 industries created by replicating the text-based industry approach of Hoberg and Phillips (2016) within our website data. Standard errors double clustered by HP industry and state.

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

measure, firms at the 90th percentile are 6.8 percentage points more likely to raise early-stage financing (about 10% of the mean). Finally, Column (4) is an additional robustness test that includes city fixed effects and state by year fixed effects to account for the possibility of geographic time-varying unobservables driving our effect. Our coefficient is very similar.

Next, in Table 4, we study the total amount of financing received, by using *Log(Early Stage+1)* as the dependent variable. The differences are more dramatic. Columns (1) and (2) show that there are similarly large cohort effects in our data. The coefficient for our preferred specification is column (3), using founding year and HP industry fixed effects. The estimate is 4.545. This implies that, on average, firms at the 90th percentile raise 117% more early-stage financing than those at the 10th percentile[20] and that moving by one standard deviation in our measure predicts 33% higher early-stage financing. Column (4) shows that this is robust to controlling for geography by including city and state by year fixed effects. We perform additional validations on columns (5)–(9). Columns (5)–(7) disaggregate early-stage financing into seed financing, grant financing, and angel financing. All coefficients are positive. Seed and grant financing are significant, whereas angel financing is slightly below the usual significance levels (p = 0.13). Column (8) includes series A financing events together with early-stage financing. The coefficient is positive with p = 0.11. Finally, column (9) reports a regression using series A financing only for firms that did not raise early-stage financing. The coefficient, although positive, is noisy and far from significant. We conclude from the evidence in Tables 3 and 4 that the relationship of founding differentiation to early-stage financing is positive, meaningful, and robust.[21]

4.2.2. Other Differentiation Scores.

Next, Table 5 considers the relationship of other differentiation scores to early-stage financing. Column (1) repeats the preferred estimate of Table 4 for comparability, using the differentiation score estimated from the five closest public firms. Column (2) instead uses the differentiation from the single closest firm. The coefficient is smaller at 3.203. Columns (3) and (4) focus on a different type of differentiation: differentiation from other startups in the same cohort. Although differentiation from public firms is intended to capture the market positioning in the extant U.S. economy, differentiation from other startups may better reflect the uniqueness and positioning in the venture financing market or access to other startup resources. Interestingly, these coefficients are only a third in magnitude from our main effect. The differentiation from the five closest startups has a coefficient of 1.282, whereas the differentiation from the closest startup has a coefficient of 1.478. Columns (5)–(7) introduce multiple measures at

Table 4. Does Founding Differentiation Predict the Amount of Early-Stage Financing?

	Dependent variable: *Early Stage*				Dependent variable: *Seed*	Dependent variable: *Grant*	Dependent variable: *Angel*	Dependent variable: *Series A + Early Stage*	Dependent variable: *Series A Subsample Firms without Early Stage*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Differentiation score (five closest public firms)	20.59*** (1.843)	7.934*** (1.744)	4.545*** (0.792)	4.185*** (0.715)	3.025*** (0.845)	0.809** (0.361)	0.823 (0.540)	1.931 (1.200)	1.663 (2.324)
Founding year fixed effects	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
HP industry fixed effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year by state fixed effects	No	No	No	Yes	No	No	No	No	No
City fixed effects	No	No	No	Yes	No	No	No	No	No
Observations	12,406	12,406	12,406	12,406	12,406	12,406	12,406	12,406	3,946
R^2	0.044	0.132	0.184	0.331	0.179	0.058	0.041	0.044	0.123

Notes. OLS model. Dependent variable is the log of total fundraised in early-stage financing plus one. HP industry fixed effects are fixed effects for 300 industries created by replicating the text-based industry approach of Hoberg and Phillips (2016) within our website data. Standard errors double clustered by HP industry and state. Column (7) is the Series A fundraising only for companies that did not raise early-stage financing.

$^*p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 5. Other Measures of Founding Differentiation and Early-Stage Financing

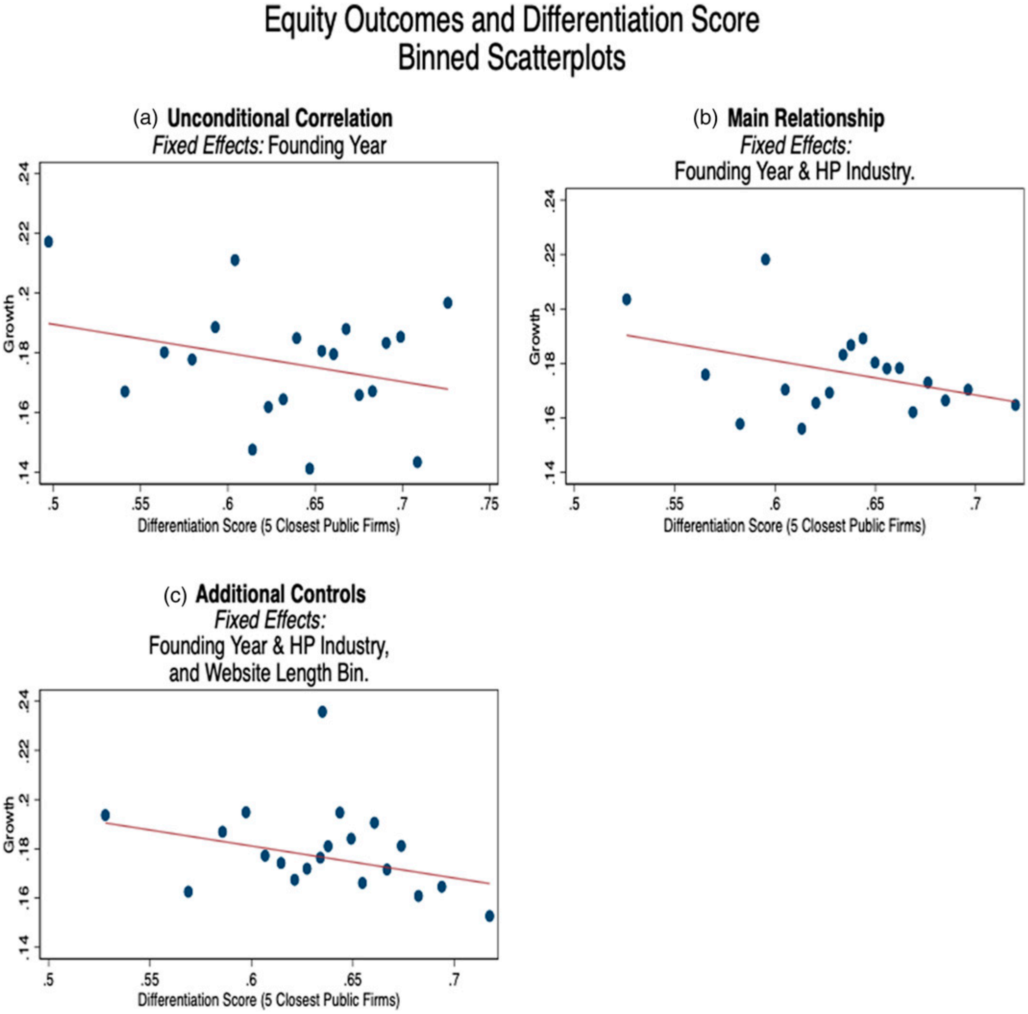
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Differentiation score (five closest public firms)	4.545***				6.661***	7.699***	
	(0.792)				(1.784)	(1.396)	
Differentiation score (closest public firm)		3.203***			−1.960		3.259***
		(0.712)			(1.627)		(0.800)
Differentiation score (five closest cohort startups)			1.282*			−4.223***	
			(0.607)			(1.169)	
Differentiation score (closest cohort startups)				1.478*			−0.0925
				(0.592)			(0.649)
Founding year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HP industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,406	12,406	12,406	12,406	12,406	12,406	12,406
R^2	0.184	0.184	0.183	0.183	0.184	0.185	0.184

Notes. OLS model. Dependent variable is the log of total fundraised in early-stage financing plus one. HP industry fixed effects are fixed effects for 300 industries created by replicating the text-based industry approach of Hoberg and Phillips (2016) within our website data. Standard errors double clustered by HP industry and state.

$^*p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

the same time to consider the correlation of one measure of differentiation conditional on others. In column (5), we note that when both differentiation from the five closest and single closest incumbent are included, only the former remains positive and significant. It seems including five incumbents rather than one gives more precision to our measure. More strikingly, when differentiation from both incumbent and startup firms is included

Figure 3. (Color online) Differentiation Score and Equity Growth Outcomes



Notes. Growth is IPO or acquisition. HP industries are the industries defined using the methodology of Hoberg and Phillips (2016) in our data.

Table 6. Does Founding Differentiation Predict Equity Performance?

	Dependent variable: *IPO or Acquisition*			Dependent variable: *IPO*	Dependent variable: *Acquisition*
	(1)	(2)	(3)	(4)	(5)
Differentiation score (five closest public firms)	−0.0961 (0.0811)	−0.112** (0.0463)	−0.127** (0.0456)	−0.0512** (0.0231)	−0.0607 (0.0416)
Founding year fixed effects	Yes	Yes	No	Yes	Yes
HP industry fixed effects	No	Yes	Yes	Yes	Yes
Year × state fixed effects	No	No	Yes	No	No
City fixed effects	No	No	Yes	No	No
Observations	12,406	12,406	12,406	12,406	12,406
R^2	0.099	0.142	0.262	0.111	0.126

Notes. OLS model. Dependent variable is a binary variable equal to one if a firm is IPO or acquired and zero otherwise. HP industry fixed effects are fixed effects for 300 industries created by replicating the text-based industry approach of Hoberg and Phillips (2016) within our website data. Standard errors double clustered by HP industry and state.

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

simultaneously in columns (6) and (7), the incumbent measure remains positive and significant, whereas the startup measure turns either negative or not significant. Together, we interpret these results as emphasizing that the positive relationship of our measure to early-stage outcomes is driven by underlying differences from the existing market structure, as proxied by public firms, rather than differences from startups or its implications on competition in the financing market itself. Strategic differentiation in the consumer market predicts performance.

4.3. Founding Differentiation and Equity Outcomes

We proceed to study how our differentiation score predicts long-term firm equity outcomes such as IPO or acquisition. Figure 3 reports binned scatterplots of our differentiation score with *Equity Growth*, a binary measure equal to one if the firm achieves IPO or acquisition, as the dependent variable, and *Differentiation Score (5 Closest Public Firms)* as the independent variable. Perhaps unintuitively, all relationships appear noisy and weak and have a negative slope in the fitted line.

4.3.1. Main Relationships. We study these relationships in more detail through regressions in Table 6. The coefficients show a different pattern than early-stage financing. Column (1) reports a noisy relationship of founding differentiation and equity outcomes within cohorts. This turns large and significant in column (2), using founding year and HP industry fixed effects. The coefficient is −0.112 and remains after including geography controls in column (3). On average, firms that score at the 90th percentile of our measure are 1.9 percentage points less likely to achieve a growth outcome than those at the 10th percentile

Table 7. Founding Differentiation and Equity Performance

	(1) Subsample: Drop firms founded 2012 or later	(2) Dependent variable: *IPO or Acquisition During First 5 Years*	(3) Dependent variable: *IPO or Acquisition After First 5 Years*	(4) Dependent variable: Log(Acquisition Price)	(5) Dependent variable: *High Value Acquisition*
Differentiation score (five closest public firms)	0.0189 (0.0693)	−0.125** (0.0375)	0.0639 (0.0431)	5.981** (2.238)	0.00767 (0.0148)
Founding year fixed effects	Yes	Yes	Yes	Yes	Yes
HP industry fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	5,559	12,406	12,406	374	12,406
R^2	0.128	0.054	0.134	0.500	0.040

Notes. OLS model. Dependent variable is a binary variable equal to one if a firm is IPO or acquired and zero otherwise. HP industry fixed effects are fixed effects for 300 industries created by replicating the text-based industry approach of Hoberg and Phillips (2016) within our website data. Standard errors double clustered by HP industry and state.

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 8. Dynamic Effects of Founding Differentiation on Equity Performance Across Age

	(1)	(2)	(3)	(4)
	Dependent variable: *IPO or Acquisition (Cumulative)*	Dependent variable: *IPO (Cumulative)*	Dependent variable: *Acquisition (Cumulative)*	Dependent variable: *High Value Acquisition (Cumulative)*
Age = 0 × differentiation score (five closest public firms)	−0.194**	−0.0195	−0.174**	−0.00517
	(0.0285)	(0.0165)	(0.0328)	(0.00776)
Age = 1 × differentiation score (five closest public firms)	−0.182**	−0.0189	−0.163**	−0.00494
	(0.0282)	(0.0165)	(0.0314)	(0.00769)
Age = 2 × differentiation score (five closest public firms)	−0.149**	−0.0169	−0.133**	−0.00363
	(0.0281)	(0.0166)	(0.0287)	(0.00751)
Age = 3 × differentiation score (five closest public firms)	−0.103**	−0.0148	−0.0885**	−0.000371
	(0.0307)	(0.0164)	(0.0273)	(0.00691)
Age = 4 × differentiation score (five closest public firms)	−0.0523	−0.0117	−0.0406	0.00297
	(0.0339)	(0.0164)	(0.0275)	(0.00645)
Age = 5 × differentiation score (five closest public firms)	0.00282	−0.00916	0.0120	0.00822
	(0.0383)	(0.0160)	(0.0296)	(0.00607)
Age = 6 × differentiation score (five closest public firms)	0.0566	−0.00163	0.0583*	0.0159**
	(0.0416)	(0.0170)	(0.0306)	(0.00585)
Age = 7 × differentiation score (five closest public firms)	0.122**	0.00863	0.114**	0.0211**
	(0.0470)	(0.0187)	(0.0337)	(0.00722)
Age = 8 × differentiation score (five closest public firms)	0.196**	0.0236	0.173**	0.0284**
	(0.0455)	(0.0207)	(0.0307)	(0.00847)
Age = 9 × differentiation score (five closest public firms)	0.269**	0.0337*	0.236**	0.0407**
	(0.0444)	(0.0197)	(0.0316)	(0.00919)
Age = 10 × differentiation score (five closest public firms)	0.346**	0.0434*	0.302**	0.0495**
	(0.0409)	(0.0227)	(0.0277)	(0.00854)
Age = 11 × differentiation score (five closest public firms)	0.412**	0.0542**	0.358**	0.0542**
	(0.0396)	(0.0217)	(0.0254)	(0.00935)
Founding year fixed effects	Yes	Yes	Yes	Yes
HP industry fixed effects	Yes	Yes	Yes	Yes
Observations	79,711	79,711	79,711	79,711
R^2	0.137	0.060	0.127	0.032

Notes. OLS model. Dependent variable is a binary variable equal to one if a firm has achieved IPO or acquired by age *t* and zero otherwise. Standard errors double clustered by HP industry and state.

$*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

(10% of the mean). Columns (4) and (5) separate IPOs and acquisitions. Column (4) shows a negative and meaningful coefficient for IPO, whereas column (5) reports a noisy but negative coefficient for acquisitions. Within our whole sample, the relationship from strategic differentiation to equity outcomes appears negative.

In Table 7, we perform additional analyses to provide more detail on this relationship across age and exit value. Column (1) drops all firms that are young in the data, and for whom, consequently, we cannot observe their whole lifecycle. To do so, we remove all firms born in 2012 or later. The coefficient is now positive but noisy and close to zero. Columns (2) and (3) separate the exits based on timing. Column (2) is the exits early in the startup lifecycle, within the first six years, whereas column (3) is those after six years.[22] Once again, although the coefficient for the early exits is negative, the one for the late exits is not and has a positive, but noisy, estimate. These timing differences foreshadow an additional analysis is needed: the relationship of differentiation to outcomes may change over time, and hence we would need to consider the dynamics of firms across their whole lifecycle.

Columns (4) and (5) consider instead the valuation of the acquisition outcomes to study heterogeneity in how "successful" an acquisition is. Although Crunchbase only reports acquisition values for a fraction of acquisitions, we assume that the misreporting is not correlated with differences in founding differentiation. Column (4) shows a large positive effect for differentiation on the acquisition price, conditional on being acquired. Column (5) shows that when we consider only high-value acquisitions, instead of all acquisitions, the coefficient turns to zero instead of negative. Overall, there appears to be evidence that the relationship between differentiation and equity outcomes is also more positive when we consider "larger" acquisitions.

4.3.2. Dynamics.

Thus far, we have limited ourselves to cross-sectional regressions, but this is unsatisfactory for equity outcomes given the regression results in

Tables 6 and 7 showing the role of founding differentiation on outcomes may vary across firm age. Not simply an empirical regularity, the existing literature also suggests the dynamics of acquisitions and IPOs may vary depending on how differentiated a firm is at founding. More unique startups will struggle to achieve legitimacy initially but may ultimately perform better (Deephouse 1999, Marx et al. 2014). If this is the case, our measure of differentiation could have a negative association to equity outcomes ocurring in the early years of the firms, which changes to a positive cumulative effect over time. Furthermore, in the analysis presented, this type of dynamics would create a negative bias, because many firms are only observed for a few years (i.e., the younger cohorts), and we do not get to observe the later years in their lifecycle.

We study this possibility in Table 8. To do so, we estimate our preferred specification in a panel format and report the individual coefficients for founding differentiation score against the cumulative probability that a firm has observed an equity exit by each year of age. The pattern we see is dramatic and consistent with the dynamics of highly differentiated companies. Firms with a higher differentiation score are initially less likely to exit, particularly during their first year (year 0). For our main growth outcome, a firm scoring at the 90th percentile of our measure is 19% less likely (relative to the mean) to achieve equity growth in year 0 compared with one at the 10th percentile. Yet, this pattern reverses as the firm ages. The coefficient turns positive by age 6 and continues increasing thereafter. By age 7, firms with a higher founding differentiation score are more likely to have had an equity exit. The coefficient, with a value of 0.086, implies that moving from the 10th to the 90th percentile is associated with a 1.4-percentage-point higher likelihood of exit, or 8% of the mean. By age 10, the difference is a 28% increase over the mean. Importantly, because the outcome measure is cumulative, the positive effect reflects the total success up to that year, including the negative impact of the early years. Differentiation predicts higher outcomes over the firm lifecycle, but it takes time for them to occur.

Columns (2)–(4) disaggregate our equity growth variable into different types of exits. Column (2) considers only IPOs. We observe the same pattern, although the coefficient is significant until age 9. By age 10, moving from the 10th to the 90th percentile increases IPO probability by 33% relative to the mean. Column (3) is acquisitions. This pattern is similar, but the coefficient turns statistically significant from age 6. By age 10, a startup at the 90th percentile is 27% more likely to be acquired than a startup at the 10th percentile. Finally, column (4) focuses only on high-value acquisitions (i.e., more than $100 million). Interestingly, the coefficient, in this case, is not negative in the early years, but it does become positive later on. By age 10, it represents an increase of 43% relative to the mean. These results provide further robustness validating the dynamic effects evidenced in column (1). The evidence is consistent with our differentiation score ultimately capturing the nature of more innovative or unique ideas, which perform better over the long term, but they take a longer time to achieve this performance, possibly because of the cost of educating and better understanding the market to either acquire legitimacy (Deephouse 1999) or to prove technological feasibility (Marx et al. 2014).

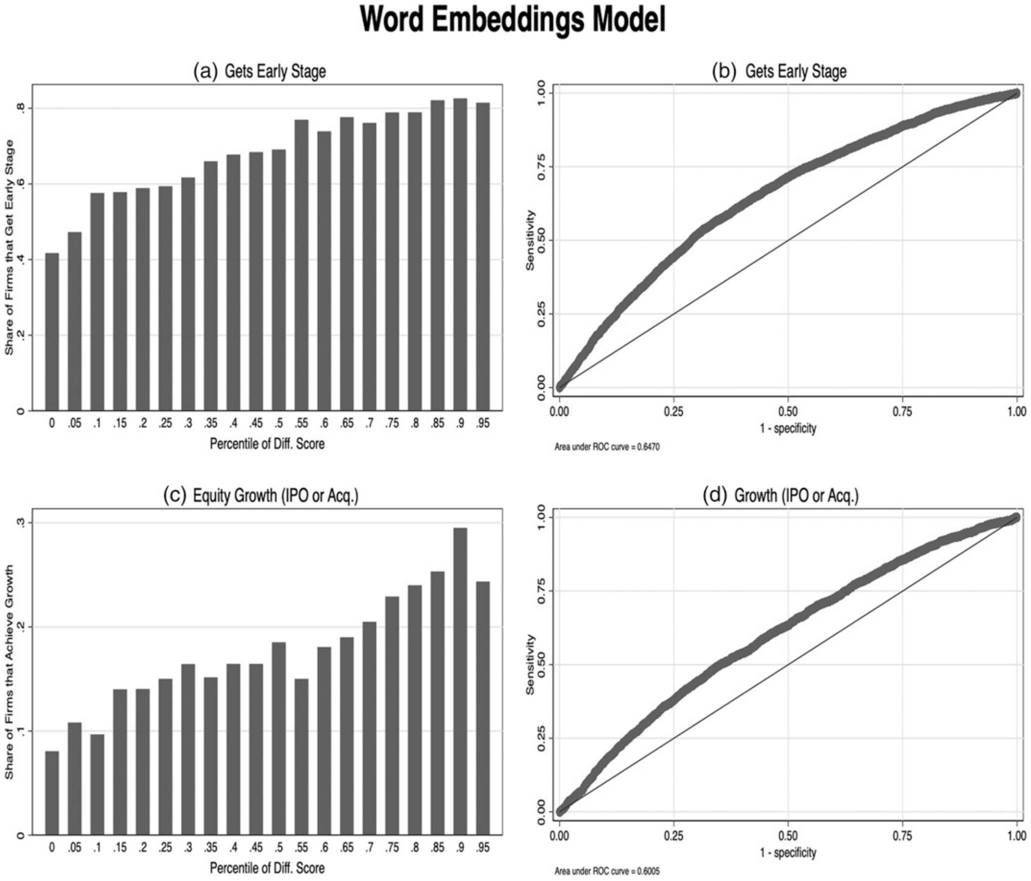
Finally, in Table A6 in the online appendix, we study how these dynamic outcomes relate instead to a differentiation measure developed using tf-idf similarity rather than our word-embeddings approach. The coefficients show our approach has a stronger association to outcomes. Even though both our measure and a tf-idf score show the dynamic effects of Table 8 when introduced in independent regressions, only our measure remains robust when both are used together in the same regression. The coefficients of the tf-idf measure, in contrast, become small in magnitude and mostly not significant.

4.4. How Much Does Founding Strategy Matter?

Finally, we study the extent to which our measures, and consequently founding strategy, are economically relevant. To do so, we perform an out-of-sample analysis to consider how much variation in outcomes can be predicted by our measures. Specifically, using a 10-fold approach, we regress a logit model with a fully interacted version of our four founding differentiation scores (five closest public firms, closest public firm, five closest cohort startups, and closest cohort startup) on the binary version of both of our outcomes: *Gets Early Stage Financing* and *Equity Growth*. We then store the out-of-sample predictions from these models[23] and study how well these out-of-sample predictions relate to realized outcomes. The results are reported in Figure 4.

Panels (a) and (b) consider early-stage financing. Panel (a) reports the share of firms that receive early-stage financing across the distribution of out-of-sample predicted probability. We observe a positive and increasing slope, with firms at the top end of the distribution being about twice as likely to get early-stage financing as firms at the bottom. Panel (b) is our preferred measure. It reports the out-of-sample ROC score (area under the curve) of this model. This is an established approach to assess the predictive fit of binary models. The ROC score conceptually answers the following question: if two firms, one with early-stage financing and one without, are fed to the model, what is the likelihood that the one with early-stage financing is scored higher by the model? A random model would have an ROC of 0.5, and a fully

Figure 4. Out-of-Sample Predictability of Performance from Founding Text



Notes. This figure reports out-of-sample tests of how well do our measures predict performance. To do so, we run a fully interacted model of our four differentiation measures on two binary outcomes, *Gets Early Stage Financing* and *Equity Growth* using a 10-fold approach where we split the data into 10 groups and use the regression of 9 groups to predict the remaining one out of sample. (a) and (c) Distribution of outcomes across the predicted probability of performance. (b) and (d) ROC (area under the curve) score that better measures the fit of the data.

informative one would be 1. The graph shows an ROC value of 0.65, implying the model can account for about 30% of the variation in outcomes.

Panels (c) and (d) consider the same two statistics for the equity growth outcome. Once again, we observe a meaningful ability of our index to predict performance. Firms in the top ventiles of the out-of-sample predicted index are about three times more likely to have an equity growth outcome than firms in the bottom ventiles. The ROC score is slightly lower at 0.60. Our model can account for 20% of variation in outcomes.

Figure A3 in the online appendix repeats the model using a simpler measure of similarly—the cosine of the term frequency-inverse document frequency (tf-idf) estimates. Interestingly, these estimates are lower. The ROC score for early-stage financing is 0.62, whereas the ROC for the equity growth outcome is only 0.516, implying the tf-idf measures only account for about 3.2% of total variation. Finally, Figure A4 in the online appendix reports a model that uses both HP industries and differentiation scores together. The ROC scores increase, but only moderately, to 0.68 for *Gets Venture Capital* and 0.63 for *Equity Growth*.

Together, these estimates provide a novel assessment of the importance of founding positioning on overall performance. We show that founding positioning accounts for 30% and 20% of variation in financing and equity outcomes, respectively, out-of-sample. Drawing a parallel to the theoretical concept that it is not only using unique resources but being able to use them together in novel ways that creates strong strategic positioning, the tf-idf model, which does not consider the context in which words are used, scores much lower than our doc2vec model, which does. Furthermore, because our measures are inherently noisy, our estimate on the economic importance of founding positioning is possibly a lower bound within our sample.

5. Conclusion

Building on existing work using text-based machine learning, we developed a novel approach to measure the strategic differentiation of startups and validated its role in predicting startup performance. Our approach focuses on the idea that companies state what their main value proposition is in marketing materials and that websites are core marketing channels for most firms. A historical version of these websites, paired with natural language processing methods, can allow measuring distance in the website text from one company to another at specific points in time, and this distance can be aggregated into a measure of market differentiation from a company's closest competitors. We implemented our approach on data from Crunchbase to measure startup founding differentiation from its closest incumbent firms. We show that founding differentiation predicts financing and long-term performance and accounts for a meaningful portion of the variation in outcomes. Founding strategy matters.

Our paper is designed to enable further research in this area. To do so, we have included several data and code appendices that allow incorporating our data set and approach into other contexts. Our paper is accompanied by the release of four distinct data sets preserved in the Harvard Dataverse: (i) a data set containing all website text downloaded for the specific companies in our sample, including individual snapshots of each public company by year; (ii) a data set containing the doc2vec models estimated through this data for each year in the sample and the Hoberg & Phillips tf-idf model using all websites; (iii) a data set using these models to estimate a matrix of text-based similarity between startups and other companies in the data each year; and (iv) a data set including the estimated strategic differentiation score for each company. Our purpose in releasing this is to allow any researcher to take advantage of the data we have provided and expand the knowledge around measuring and understanding startup strategy.

This data release is also accompanied by two distinct pieces of code, including (i) an open-source release of the code we use to scrape the Wayback Machine in Github and build our doc2vec models and (ii) a release of the regressions and code we have run to estimate all the tables in our paper. Because of sharing restrictions with Crunchbase, we are not able to release the main analysis data that includes Crunchbase private information, which was given to us through an academic license.

Building from these products, we hope researchers can continue expanding the analysis of text data, website information, and the way it predicts performance. Some of these avenues can include improving the NLP approach further by tuning specific parameters of the doc2vec model or introducing other, more advanced NLP models such as BERT; adding ancillary data beyond financing and equity outcomes, to understand other facets of firms and their evolution, such as patents, or workforce information; and expanding our approach beyond our sample to other contexts or countries. Whether, when, and how can strategy be measured are rich avenues for future work.

Moving to entrepreneurial strategy, we hope our work can bring back emphasis on the relative importance of founding positioning for startups and their success. Although strategy researchers had once pitted execution, dynamic capabilities, and founding positioning against each other as sources of competitive advantage, we are encouraged by how recent research instead recognizes the relative merits of each in startup performance (Gans et al. 2018, Koning et al. 2022). We provided an original estimate of how and how much does one facet of this equation matter, but we hope future work can continue applying analytical methods (such as machine learning) to these ideas to better elucidate what elements should go into an startup's ideal strategy.

Finally, considering our overall approach more broadly, we recognize that the role of data science and the information technology revolution in shaping the way management research and practice is done is only starting to take shape. We have focused here specifically on using data science to measure a traditional construct in a new way, but other applications can include causal inference using observational data and developing a new understanding of the key elements of strategic advantage that emerges from data-driven results rather than using data to map existing theory. Ultimately, a data-driven approach to strategy may be quite distinct from the canonical, framework-based approach. We look forward to continued work in this space.

Acknowledgments

The authors thank Mabel Abraham, Eric Abrahamson, Natarajan Balasubramanian, Joel Brockner, Raj Choudhury, Bo Cowgill, Enrico Forti, Kathy Harrigan, Aija Leiponen, Ulrich Kaiser, Daniel Keum, Bruce Kogut, Kristina McElheran, Stephan Meier, Patryk Perkowski, Christian Peukert, Robert Seamans, Ananya Sen, Scott Stern, Hannes Ullrich, and Zhiyuan Yu for helpful comments, as well as participants in the Columbia Data Science Day, the New York University Workshop on Machine Learning and Strategy, Academy of Management Workshops on Strategy and Machine Learning, Strategy Science, the Virtual Digital Economy Seminar, the Strategic Management Society Conference on Machine Learning, and seminars at Columbia Business School, Carnegie Mellon Tepper, and University of Texas Rio Grande Valley.

Endnotes

[1] Prior measures of differentiation have focused on using the presence of startups across industry-specific product categories, such as the number of categories in which video game startups release games (Cennamo and Santalo 2013).

[2] In fact, Southwest has a relatively low rate of on-time arrivals.

[3] Specifically, although our analyses show that firms that report a higher level of differentiation early on are (statistically) more likely to succeed afterward, they do not imply making a more distinct website directly would change the likelihood of success. Rather, this success is possibly driven by other variables, such as founder human capital and intellectual property, that drive both better strategy formulation and eventual outcomes. Studying the causal relationship of changes in strategic differentiation to firm performance is left for future work.

[4] We define early-stage financing as the sum of seed and preseed capital, grants, crowdfunding, and angel financing.

[5] We use the differentiation from the five closest public firms, the single closest public firm, the five closest startups of the same cohort, and the single closest startup of the same cohort.

[6] These are available at https://bit.ly/MeasuringFoundingStrategy.

[7] In our implementation, we use the doc2vec expansion of word2-vec developed by Le and Mikolov (2014), which allows developing embedding vectors for whole documents. However, we limit this explanation to the simpler word2vec model for ease of exposition.

[8] As we report in Table A3 in the online appendix, the correlation between these two measures is 0.19, highlighting that, although related, they are also meaningfully different.

[9] Bag-of-words approaches also allow incorporating semantics by increasing the number of words in each n-gram; however, this tends to quickly lead to sparsity.

[10] Our implementation uses the the gensim.models.doc2vec Python library with the following parameters: a vector size of 700, a word window of seven (both before and after the focal word), ignore all words that occur less than three times, using a distributed memory algorithm (PV-DM), no corpus file, and 10 total iterations over the corpus (epochs). All code is available at https://bit.ly/MeasuringFoundingStrategy.

[11] In more recent work, Igami and Uetake (2020) also study the impact of competition on incentives to innovate and find that incentives to innovative drop quickly, stabilizing after five competitors.

[12] As a matter of policy, Crunchbase keeps all companies ever recorded, except for special circumstances (see Quora (2013) for further explanation).

[13] This algorithm is the fixed-industry classification algorithm in Hoberg and Phillips (2016, p. 1435). Hoberg and Phillips also implement a network-based measure that our approach does not allow us to implement.

[14] Specifically, we exclude the top and bottom 5%. All results reported in this paper are robust (and often even higher and more statistically significant coefficients) when including these outliers, but the validations looking at the distribution of firms in the Consumer Electronics category suggested some firms at the bottom tail matched because of invalid text and error messages rather than actual firm text.

[15] See https://www.sporttechie.com/wearable-technology-and-the-way-we-run/.

[16] See https://appleinsider.com/articles/20/07/21/review-brydge-vertical-dock-is-an-expensive-and-elegant-way-to-work-at-your-desk.

[17] The relative rank of these companies when measuring differentiation using tf-idf rather than our method is different. Rank between measures is positively correlated at 0.23. Table A3 in the online appendix reports the correlation between each measure using our word-embeddings approach and tf-idf. Although all coefficients are positive, they are also far from one.

[18] See https://codasignature.com/press-release/coda-signature-wins-2019-ncia-excellence-in-innovation-award/.

[19] This drop of 40% when considering the explanatory power of industry on performance is lower than the classic estimate in Schmalensee (1985) on the role of industry to firm profitability but higher than the follow-on estimates in Rumelt (1991) and McGahan and Porter (1997).

[20] The 10th–90th percentile range is 0.17; $e^{(.17*4.545)} - 1 = 1.17$.

[21] Tables A4 and A5 in the online appendix report the key results while also including differentiation measures based on tf-idf; our results are robust to controlling for this measure.

[22] Fifty-eight percent of the startup exits in our sample occur within the first six years.

[23] In essence, we split the data into 10 random subsamples, and for each subsample, we use the predicted value from a regression using all other nine subsamples but excluding the focal one.

References

Abrahamson E, Hambrick DC (1997) Attentional homogeneity in industries: The effect of discretion. *J. Organ. Behav.* 18(S1):513–532.

Bresnahan TF, Reiss PC (1991) Entry and competition in concentrated markets. *J. Political Econom.* 99(5):977–1009.

Cennamo C, Santalo J (2013) Platform competition: Strategic trade-offs in platform markets. *Strategic Management J.* 34(11):1331–1350.

Dalle J-M, den Besten M, Menon C (2017) Using crunchbase for economic and managerial research, Working paper, OECD Science, Technology and Industry.

Deephouse DL (1999) To be different, or to be the same? It's a question (and theory) of strategic balance. *Strategic Management J.* 20(2):147–166.

Eisenhardt KM, Martin JA (2000) Dynamic capabilities: What are they? *Strategic Management J.* 21(10-11):1105–1121.

Gambardella A, Camuffo A, Cordova A, Spina C (2020) A scientific approach to entrepreneurial decision making: Evidence form a randomized control trial. *Management Sci.* 66(2): 564–586.

Gans J, Scott E, Stern S (2018) Strategy for start-ups. https://hbr.org/2018/05/strategy-for-start-ups.

Hoberg G, Phillips G (2016) Text-based network industries and endogenous product differentiation. *J. Political Econom.* 124(5): 1423–1465.

Igami M, Uetake K (2020) Mergers, innovation, and entry-exit dynamics: Consolidation of the hard disk drive industry, 1996-2016. *Rev. Econom. Stud.* 87(6), 2672–2702.

Koning R, Hasan S, Chatterji A (2022) Experimentation and startup performance: Evidence from A/B testing. *Management Sci.* Forthcoming.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. *Proc. Internat. Conf. on Machine Learn.* (PMLR), 1188–1196.

Marx M, Gans JS, Hsu DH (2014) Dynamic commercialization strategies for disruptive technologies: Evidence from the speech recognition industry. *Management Sci.* 60(12):3103–3123.

McGahan AM, Porter ME (1997) How much does industry matter, really? *Strategic Management J.* 18(S1):15–30.

McGrath RG, MacMillan IC (2000) *The Entrepreneurial Mindset: Strategies for Continuously Creating Opportunity in an Age of Uncertainty*, vol. 284 (Harvard Business Press, Cambridge, MA).

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. Preprint, submitted January 16, https://arxiv.org/abs/1301.3781.

Mu J, Bhat S, Viswanath P (2017) All-but-the-top: Simple and effective postprocessing for word representations. Preprint, submitted February 5, https://arxiv.org/abs/1702.01417.

Porter ME (1996) What is strategy? *Harvard Bus. Rev.* 6(74):61–78.

Quora (2013) How do i delete a company profile on crunchbase? Accessed April 1, 2020, https://www.quora.com/How-do-I-delete-a-company-Profile-on-Crunchbase.

Reis E (2011) *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (Crown Business, New York).

Ruefli TW, Wiggins RR (2003) Industry, corporate, and segment effects and business performance: A non-parametric approach. *Strategic Management J.* 24(9):861–879.

Rumelt RP (1991) How much does industry matter? *Strategic Management J.* 12(3):167–185.

Schmalensee R (1985) Do markets differ much? *Amer. Econom. Rev.* 75(3):341–351.

Siggelkow N (2001) Change in the presence of fit: The rise, the fall, and the renaissance of Liz Claiborne. *Acad. Management J.* 44(4):838–857.

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. *Strategic Management J.* 18(7):509–533.

Van den Steen E (2016) A formal theory of strategy. *Management Sci.* 63(8):2616–2636.