Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Text-Based Analysis of Corporate Innovation

Gustaf Bellstam, Sanjai Bhagat, J. Anthony Cookson

To cite this article:
Gustaf Bellstam, Sanjai Bhagat, J. Anthony Cookson (2020) A Text-Based Analysis of Corporate Innovation. Management Science

Published online in Articles in Advance 16 Sep 2020

. https://doi.org/10.1287/mnsc.2020.3682

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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–28
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

A Text-Based Analysis of Corporate Innovation

Gustaf Bellstam,[a] Sanjai Bhagat,[b] J. Anthony Cookson[a,*]

[a] Facebook, Seattle, Washington 98109; [b] Leeds School of Business, University of Colorado, Boulder, Colorado 80309
*Corresponding author
Contact: bellstam@gmail.com (GB); sanjai.bhagat@colorado.edu (SB); tony.cookson@colorado.edu,
https://orcid.org/0000-0001-6920-9353 (JAC)

Received: December 15, 2017
Revised: April 10, 2019; January 6, 2020
Accepted: March 9, 2020
Published Online in Articles in Advance: September 16, 2020

https://doi.org/10.1287/mnsc.2020.3682

Abstract. We develop a new measure of innovation using the text of analyst reports of S&P 500 firms. Our text-based measure gives a useful description of innovation by firms with and without patenting and R&D (research and development). For nonpatenting firms, the measure identifies innovative firms that adopt novel technologies and innovative business practices (e.g., Walmart's cross-geography logistics). For patenting firms, the text-based measure strongly correlates with valuable patents, which likely capture true innovation. The text-based measure robustly forecasts greater firm performance and growth opportunities for up to four years, and these value implications hold just as strongly for innovative nonpatenting firms.

History: Accepted by Gustavo Manso, finance.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2020.3682.

Keywords: innovation • textual analysis • machine learning • natural language processing • latent Dirchlet allocation

1. Introduction

Innovation has long been thought to play a central role both for economic growth and short-term fluctuations (Schumpeter 1939, Kuznets and Murphy 1966, Nordhaus 1969). Owing to its fundamental importance, innovation has attracted significant academic attention (e.g., Hall 1990, Brown et al. 2009, and Cohen et al. 2013). Nevertheless, our empirical understanding of innovation is incomplete because existing innovation measures—typically, research and development (R&D) intensity or outcomes related to patenting—do not fully capture the nature and scope of innovative output.

Taking a classical view, innovation can reflect a wide array of firm activities beyond product introductions, including new production methods, new supply sources, exploitation of new markets, and new organizational forms (Schumpeter 1934). In contrast to this general view of innovation, most existing innovation measures are specific to particular industries and production processes that rely on R&D expenditures and patenting (e.g., high-tech or pharmaceutical). In this way, the widespread use of R&D and patenting proxies has led innovation research to focus on innovation related to new-product introductions and to neglect studying other forms of innovation.[1]

To help bridge this gap, we propose a new measure of corporate innovation derived from textual descriptions of firm activities by financial analysts. Our measure encapsulates a broad notion of innovative processes, products, and systems, which well describes innovation in mature firms—that is, firms in the S&P 500. Innovation in mature firms has been sparsely studied, despite these firms comprising the most valuable corporations in the economy. One reason for this lack of academic attention is because mature firm innovation involves much more than developing and introducing new products. By offering a measure of innovation beyond products, our analysis provides a useful first step toward understanding mature firm innovation.

We construct the text-based innovation measure using topic-modeling tools that have been recently introduced to the finance literature (Israelsen 2014, Goldsmith-Pinkham et al. 2016, Hoberg and Lewis 2017, Lowry et al. 2020). Specifically, we employ the latent Dirichlet allocation (LDA) method of Blei et al. (2003) on the text of a large corpus of analyst reports. The underlying assumption behind LDA is that each analyst report is generated by drawing content from a common set of topics, or clusters of words. According to this modeling intuition, analyst reports have different content because they reflect a different mix of these underlying topics. A fitted LDA model recovers the set of topics (common across analyst reports) that best describe the empirical distribution of word groupings across analyst reports. The LDA routine automatically accounts for the possibility that words have different meanings depending on context, an advantage over count-based word-list techniques. The fitted LDA also provides an intensity with which each analyst report discusses each topic. These report-level intensities form

the centerpiece of our innovation measure because they capture the extent to which an analyst uses the innovation topic to describe the firm.

Our main measure is derived from a fitted LDA model that allows for 15 distinct topics. We fit the LDA model to a corpus of 665,714 analyst reports of 703 firms that were in the S&P 500 during 1990–2012. From this fitted topic model, we compute the Kullback–Liebler divergence of each topic from the language used in a mainstream textbook on innovation, and we select the topic that has the lowest divergence from the textbook language. Beyond this selection criterion, the selected topic stands out as a reliable innovation proxy, both qualitatively and quantitatively. Qualitatively, the words in the innovation topic are also words that analysts should use to describe innovations (e.g., service, system, technology, product, and solution). Quantitatively, the topic correlates strongly with patenting and R&D intensity among patenting firms. Beyond basic correlations, all of our findings using the text-based measure are robust to controlling for patenting, implying that the correlation with patenting does not drive our findings.

An important advantage of our text-based innovation measure is that it can be computed for firms that do not patent and do not use R&D. Even within our sample of 703 firms from the S&P 500, a total of 329 firms have zero R&D, and 219 firms have zero patents for the entire sample period (1990–2010). To illustrate that the measure is useful for nonpatenting firms, we present tangible examples of content from analyst reports for nonpatenting firms that score high on our measure. One such example that highlights the value of our approach is Walmart. Walmart did not use patent protection in the early 1990s, but this was a peak period of innovation for Walmart, which transformed the low-cost retail sector during the 1990s via innovative processes (e.g., placement of warehouses and shipping logistics between locations). Taking an excerpt from a May 1993 analyst report (more details are in Figure 1), Walmart was described as "at the leading edge of retail store technology," very broadly in terms of tracking inventory, procurement, and theft prevention. Our topic analysis captures this language, and, as a result, we correctly classify Walmart as one of the most innovative companies in 1993, even though this was a time period when Walmart did not use patents at all.

In addition, the text-based innovation measure captures the innovative use of technology, which includes both innovative technology adoption and in-house technology development. Industry-level comparisons of our text-based measure and R&D intensity provide useful insight into these different modes of innovation. Industries that have high text-based innovation and high R&D intensity tend to be industries in which in-house technology development is more common (e.g., Electronic Equipment and Business Services). In contrast, industries with high text-based innovation, but low R&D intensity, are industries in which the most innovative companies are skilled at technology adoption (e.g., Communications and Motion Pictures). These industry-level examples show that our text-based innovation measure is most useful beyond standard expense-based measures in settings or industries where it is important to measure the firm's ability to adopt new technologies.

Turning to corporate-valuation implications, higher text-based innovation forecasts an increase in future operating performance and an increase in measured growth opportunities embedded in Tobin's Q, results that are robust to firm and two-digit Standard Industrial Classification (SIC) industry-year fixed effects. Consistent with the nature of innovations that generate persistent improved performance and opportunities for growth, we find that both operating performance and Tobin's Q are significantly greater for up to four years after an increase in text-based innovation. Importantly, the valuation implications of innovation are similar for both patenting and nonpatenting firms, providing further evidence that our measure extends in a useful manner beyond the set of firms that use patenting and R&D. These performance implications of text-based innovation are robust to controlling for words related to revenue, growth, and technology, accounting for analyst sentiment, and alternative specifications of the LDA model.

Even within the set of patenting firms, the text-based innovation measure provides useful additional information on innovation, distinct from patenting activity. Indeed, we find that firms with high text-based innovation do not generate more patents in the following three years after controlling for firm characteristics and industry fixed effects. However, firms with high text-based innovation have significantly better innovation quality: Their patents have greater impact (i.e., more citations per patent); they have more product announcements, using data from Mukherjee et al. (2017); and their patents have significantly greater patent valuation, using the Kogan et al. (2017) patent-valuation measure. In this way, our text-based approach distinguishes true innovation captured by valuable patents and products from low-value patenting activities.

As an illustration of how the text-based innovation measure can provide insight in addition to existing innovation measures, we replicate and extend a recent finding in the innovation literature: the finding in Custódio et al. (2019) that generalist CEOs (i.e., those with diverse industry experience) produce greater patenting and patent citations. After we verify that there is a similarly positive relation between general

manager experience and patenting in our sample of S&P 500 firms, we show that general manager ability bears a strong *negative* relation to text-based innovation. Although generalist CEOs tend to generate more technical innovation in the form of patent counts and citations, our findings suggest that generalist CEOs do not increase all types of corporate innovation. This negative relation between general manager ability and text-based innovation is important to consider, especially given our robust evidence that firms with high text-based innovation perform better.

We subject the innovation measure to careful scrutiny to ensure that it is reliable and valid. Notably, a central concern with our use of analyst text is that firms might disclose innovative activities strategically, and, thus, analyst assessments will tend to reflect firms' strategic disclosures rather than true innovation. We address this possibility with several empirical tests. First, to evaluate whether variation in innovation disclosures reflects strategic timing of disclosures rather than innovation, we consider the subsample of firms that have highly forward-looking analyst reports (adapting a measure of forward-looking intensity from Muslu et al. 2015). Restricting attention to forward-looking firms, we find very similar results to our main specifications, suggesting that strategic timing of disclosures does not substantively affect our innovation measure. Second, a related possibility is that analyst text is biased toward positive aspects of innovation because managers are strategically less likely to discuss failed innovation activities. We evaluate whether our measure can speak to negative realizations of innovation by constructing a "negative" text-based innovation measure, which aggregates the innovation topic loadings across analyst reports with negative sentiment. Consistent with these analyst assessments containing useful negative information about innovation, this negative innovation measure is a robust predictor of worse firm performance. Taken together with our main results, this finding indicates that the innovation topic contains useful information about both positive and negative aspects of innovation.[2]

Our approach of using text to study innovation has a number of notable advantages, both in describing the nature of innovation and in ascribing value to those innovations. First, our text-based measure allows inclusion and measurement of nonpatented innovation, which has been a significant limitation of recent work utilizing patenting measures to proxy for innovativeness. Second, our measure is not subject to the problems inherent in the use of Cobb–Douglas-type production function to measure the impact of innovation (see Knott 2008 and Hall et al. 2010 for discussions and criticism of this method). Third, our measure is not subject to concerns about strategic disclosure of patents. In fact, because we focus on the language of analysts who are unlikely to time their reports, we avoid sources of bias from managerial disclosures as well.

Our work contributes to an emerging line of research that draws a distinction between patenting measures and innovation (e.g., Kogan et al. 2017, Mann 2018, and Cohen et al. 2019). Because our measure does not rely on patenting data, we enable measurement of innovation in firms and industries that do not patent or use R&D. In this respect, our findings are related to recent research that shows innovation is not well measured by patents, particularly in the case of trade secrets (Saidi and Zaldokas 2020). Although the notion of innovative systems in mature firms studied in our paper is distinct from trade secrets, both kinds of innovation extend beyond the set of patenting firms. As nonpatenting firms' innovative activities are understudied, we expect significant interest in approaches like ours to extend the analysis of innovation to new subsamples and types of innovation.

Beyond offering a useful measure of innovation, our work is part of a growing literature within finance and accounting that makes use of text descriptions to study important aspects of corporate behavior. Recent text-based analyses in corporate finance have examined linkages between firms and industries, the value of corporate culture, product market fluidity, financial constraints, and the information content in initial public offering prospectuses (e.g., Hanley and Hoberg 2010, Popadak 2013, Hoberg et al. 2014, Hoberg and Maksimovic 2015, and Agarwal et al. 2016). At the same time, the asset-pricing literature has employed kindred text-analysis procedures to measure sentiment and other asset-pricing risks and anomalies (Edmans et al. 2007, Dougal et al. 2012, Garcia 2013, Israelsen 2014, Cohen et al. 2020). Within the broader literature on text analysis in finance, our work is most closely related to the growing set of papers that use latent Dirchlet allocation (Goldsmith-Pinkham et al. 2016, Ganglmair and Wardlaw 2017, Hoberg and Lewis 2017, Jegadeesh and Wu 2017). Although there has been significant interest among finance scholars in text analysis in general and LDA in particular, our analysis is the first to systematically use a text analysis to construct a measure of innovation.[3]

In another vein, our use of the text of analyst reports relates to the study of the behavior and impact of analysts more broadly. Much of this work has focused on quantitative aspects of analyst reports (Loh and Mian 2006), what information analysts actually produce (Swem 2014), or the influence of analyst coverage on the real decisions of investors or firms (e.g., see analyst coverage tests in Cohen and Frazzini 2008). Some of this work has shown how analyst coverage influences the innovativeness of firms (He and Tian 2013), but none of this work has examined

the information from the text of analyst reports as it relates to innovation. In this sense, our contribution is related to Asquith et al. (2005), Huang et al. (2014), and Huang et al. (2015), who provide evidence, in a different context, that investors pay attention to the textual elements of analyst reports, rather than just the quantitative analyst forecasts. Our analysis suggests a new reason for investors to pay attention to the text of analyst reports: valuable information on firm innovation.

The remainder of the paper is organized as follows. Section 2 describes our data sources and sampling scope. Section 3 details how we construct our measure and presents evidence on its time-series and cross-sectional properties. Section 4 presents the main results linking our text-based innovation measure to firm performance and other innovation outcomes. Section 5 presents an illustrative application of our measure—a conceptual replication and extension of Custódio et al. (2019). The final section concludes with a summary of future research directions.

2. Data

We begin with a sample of firms that were a member of the S&P 500 at some point between 1990 and 2012. This initial sample contains 797 firms. To obtain the set of analyst reports from these firms, we download analyst reports from Investext via Thomson One for the years 1990–2012, which provides an initial sample of 807,309 analyst reports for 750 unique S&P 500 firms searchable in Thomson One.

After downloading the reports, we remove common stopwords (e.g., words commonly used in text without contextual meaning, like "the," "that," and "an") from the reports using a standard stopword list.[4] Prior to any textual analysis, we use a standard algorithm to stem the words contained in the analyst reports (i.e., group words into the same root as in "technolog" captures "technology" and "technological," among other related terms). To focus on a homogenous set of analyst reports, we drop reports with under 100 words remaining after the cleaning or over 5,847 words (the 98th percentile). After processing the text and matching with Compustat identifies, we obtain a final sample of 665,714 reports, on which we base our textual analysis.

We combine the pure textual data from Thomson One with sentiment word lists (Loughran and McDonald 2011, Bodnaruk et al. 2015) as an integral part of our textual classification of innovation. These lists have been adjusted for financial language and have been shown to be more appropriate than other sentiment word lists when reading financial text.

After constructing the main text sample, we calculate the text-based innovation measure (following the procedure we describe in Section 3.2) at the firm-year level. To obtain our final sample, we merge this innovation measure with accounting data from Compustat and patenting data from Noah Stoffman's website (Kogan et al. 2017), which are available until the year 2010. The final sample has 6,200 observations from 703 unique firms for the period 1990–2010.

3. Text-Based Measure of Innovation

In this section, we describe how we construct the text-based measure of innovation using the latent Dirchlet allocation method of Blei et al. (2003).[5] Specifically, we describe the nature of information about innovation that is likely contained in analyst reports, as well as how we implement LDA on the corpus of analyst reports. After outlining the details of the measure's construction, we describe some of the basic properties of the measure in our sample of S&P 500 firms. Particularly in relating to contemporaneous patenting and R&D outcomes, the measure has desirable time-series and cross-sectional properties for a measure of innovation.

3.1. Informativeness of Analyst Text

Before parsing the information content of analyst reports into information about innovation and other topics, it is important to consider the incentives and information environment that lead the analysts to write about firms in the first place. Broadly, the text of an analyst report represents the analyst's best attempt at providing a qualitative description of the firm's value-relevant activities. As innovation is one of these activities, we expect that analysts' text descriptions about firms will contain information about innovation. Indeed, recent work has shown that the qualitative aspects of analyst text contain value-relevant information about the firm's activities (Asquith et al. 2005, Twedt and Rees 2012, Huang et al. 2014), which suggests that the content of analyst reports ought to provide useful insight into the nature of innovation.[6]

A potential concern regarding building the measure from analyst reports is that analysts cannot describe innovative activities that they cannot observe. Thus, our measure of innovation can only reflect publicly observable information about the firm that was either disclosed by the firm or inferred by the analyst. In this respect, an important potential limitation to using text from analyst reports is that firms might disclose innovative activities strategically and that analyst assessments will tend to reflect these strategic disclosures by firms. For example, some firms may disclose the innovation at the very final stages so that they are not scooped by others. Alternatively, some firms can disclose very early (even prematurely) to attract funds from investors or deter rivals to undertake the same innovation projects. We empirically address the possibility that analyst reports are biased via strategic disclosures by restricting

the sample to firms with forward-looking disclosures and conducting an analysis in which we use industry-rival disclosures to instrument for firm-level disclosures. In both of these tests, we find that our results are similar to our main results (see Tables A.13 and A.14 in the online appendix), which suggests that analysts can infer meaningful information beyond what firms would strategically disclose about innovation.[7]

Further, the text-based innovation measure will naturally capture innovation activities beyond patents, which includes some trade secrets (e.g., although Coca-Cola's secret formula is a trade secret, the value of this secret is well known to analysts). Beyond trade secrets, there are myriad ways for a firm to be innovative without filing for a patent or investing in R&D (e.g., see the Walmart example from the introduction and Figure 1). We expect that our analysis of the text of analyst reports reveals these innovative activities. Indeed, our measure identifies high-innovation firms and industries that do not patent or use R&D, which suggests that existing proxies overlook an important subset of innovative activities (see the discussion in Sections 3.3 and 3.4).[8]

Our use of the analyst text is predicated on the idea that firms' innovative activities are something that analysts are supposed to describe qualitatively. By extracting the qualitative aspects of analyst reports rather than their quantitative aspects, our innovation measure should be less subject to the usual sources of analyst bias than alternative measures that take quantitative assessments directly from the analyst. Despite this focus, a potential concern is that analyst text is biased toward successful innovation because managers are unlikely to disclose innovative activities that do not work out. In this case, analyst reports, which rely on public information, may not contain useful negative information about firms' innovative activities. To address this concern, we build a negative text-based measure from

Figure 1. High Text-Based Innovation: Excepts from Selected Reports

(a)

Firm	Date	Excerpt
WAL-MART	1993-05-14	Technology also will play an important part in Wal-Mart's growth from $55 billion in sales in 1992 to more than $200 billion in sales in the year 2000. In fact, Wal-Mart already is at the leading edge of retail store technology. The company generally uses technology to improve productivity and at the same time reduce costs. As an example, Wal-Mart is using radio frequency technology in its stores to track sales and inventory information more closely, providing better information faster, enabling the company to better control its inventories and purchases, and concurrently make more purchases closer to need. Wal-Mart also recently initiated a system to track refunds and check authorizations, which should reduce the shrinkage level. This system can help the retailer to identify an item stolen from one store that is submitted for refund at a nearby store, for example. We expect Wal-Mart to remain at the leading edge of technology for retailing and distribution systems, keeping it a step ahead of its competitors.
DILLARD	1993-03-01	We also continue to like very much Dillard's long-term earnings outlook, believing that the Company's singular strengths in such areas as automated control systems, store design and vendor relationships will help it to gain market share, over time.
KOHL'S	2006-11-09	We continue to believe KSS is in the relatively early stages of a broad-based and sustainable turnaround – that is being driven by real fundamental improvements in merchandise design, assortment, systems, marketing, inventory control, and store design.
DARDEN RESTAURANT	2002-12-01	Emerging restaurant concepts add opportunity for continued expansion and reinvestment of operating earnings.

(b)

Firm	Date	Excerpt
GOOGLE	2009-07-01	Google Apps is competitive in the managed application market, because the company offers an alternative model to the development and deployment of enterprise applications that exploits the cloud delivery concept to provide an aggressively priced and innovative subscription-based collaborative alternative to the conventional licensed software models. The company's Web 2.0 integration concepts, brand clout and marketplace momentum does not hurt the company either.
AMD	1998-11-13	For the first time, we believe that AMD could be poised for a differentiated product versus Intel. The K6-3 will have a 6-month lead over Intel's Katmai and will be mechanically similar to Slot 1 called Slot A. The K7, which will be introduced in 1999, will have a faster system bus based on the Alpha. AMD will target the small and medium business segment for the K7 and seek to improve the penetration of notebooks in 1999.
SYMBOL TECHNOLOGIES	2002-01-04	The integration of barcode scanning with wireless LANs and handheld computers is something that no other company can offer. However, to better understand the company's full suite of products, we will look at Symbols' products and position in the scanning, wireless LAN and handheld appliance businesses.

Notes. This figure shows excerpts from reports classified as highly indicative of innovation according to our text-based innovation measure. Panel (a) lists four example reports from industries with limited or no overall patenting. Panel (b) shows examples from firms in industries that rely heavily on patenting.

the subset of negative-sentiment analyst reports. Consistent with analysts inferring negative innovation outcomes from the firm's information environment, we find that negative information about innovation relates negatively to performance (Section 4.2.3). This finding shows that analysts can infer valuable negative information about innovation, despite the natural incentives for the firm to strategically disclose positive realizations of innovative outcomes.

In addition to containing value-relevant information about innovation, the language of analyst reports has relatively common textual structure (i.e., similar word usage, jargon, specificity, and topics covered) relative to media reports about the firm, or even disclosures by the firm itself. This feature of analyst reports is convenient from the standpoint of our topic-modeling approach described in the next subsection, which assumes that each report is built from a common set of latent topics. With this understanding of the qualitative content of analyst reports, we now turn to describing how we measure innovation using the analyst text.

3.2. Measuring Innovation with Latent Dirchlet Allocation

We fit a latent Dirichlet allocation model to a corpus of analyst reports following Blei et al. (2003). The LDA methodology assumes that each document is generated from a mixture of a fixed set of topics, where each topic is a distribution of words. LDA is a so-called "bag of words" method, which means that the order within documents is not important. To fit an LDA model, the researcher only needs to specify the total number of topics K, and the routine produces two outputs from the corpus of documents: (i) a distribution of word frequencies for each of the K topics, common across documents; and (ii) a distribution of topics across documents (i.e., the frequencies with which the topics are used in each document).

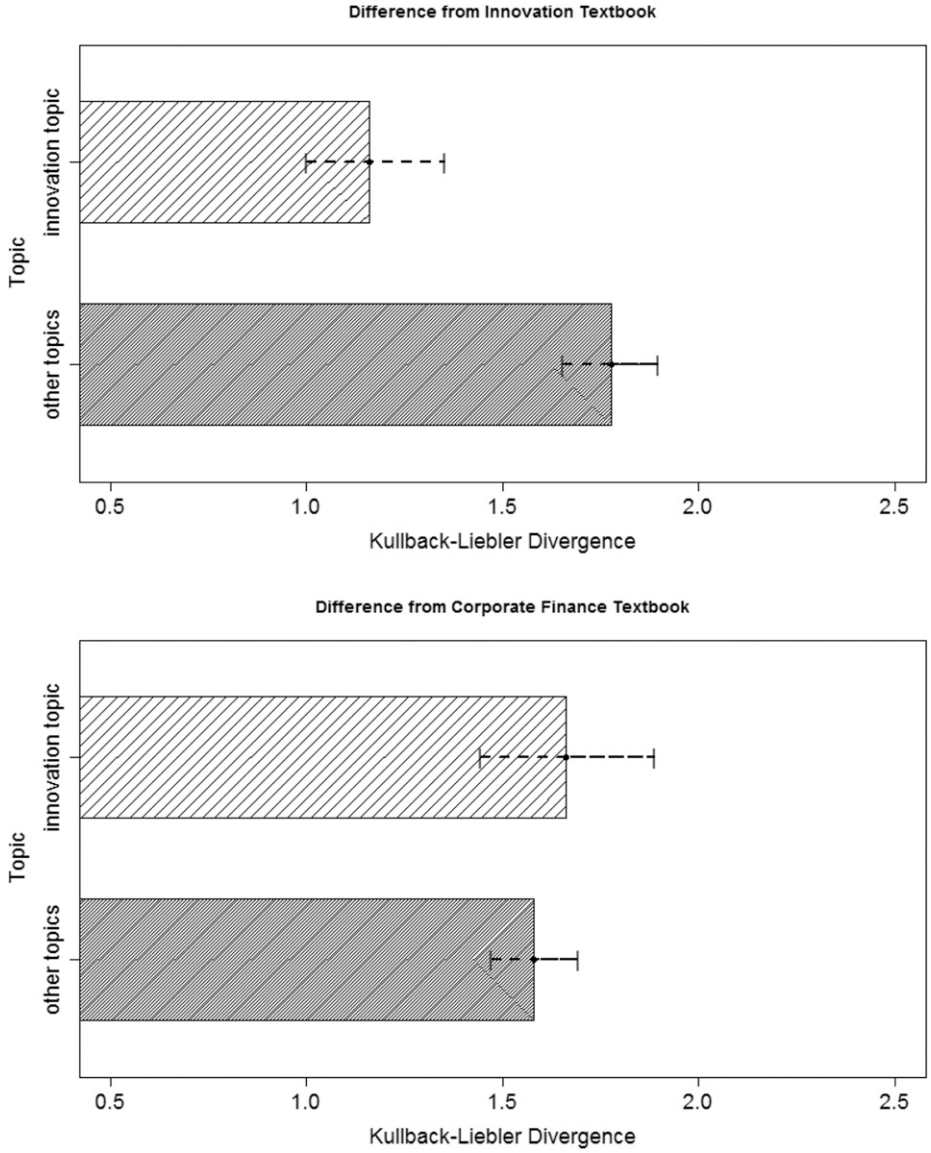
The content of each topic emerges endogenously as the set (and frequency) of words that tend to group together in the analyst reports. For each document, the topic distribution is a vector of loadings that describe how intensively the topic is being used in a particular document. Equivalently, the underlying method assigns a likelihood that the document is about that topic, such that if a document has a higher loading for a particular topic, it is more likely associated with the topic.

To construct our innovation measure, we estimate an LDA model with $K = 15$ topics using the 665,714 analyst reports as the underlying corpus of documents.[9] Fitting this LDA model gives the 15 topics—each a frequency distribution over words—that best fit the context of the analyst reports. To identify the topic that most accurately captures innovation, we select the topic with the word distribution that has the smallest statistical distance from the a popular innovation textbook's word distribution.[10] Specifically, we compute the Kullback–Liebler (KL) divergence of each topic's word distribution from the source text on innovation, similar to Lowry et al. (2020). In our context, the KL divergence is useful because it is a measure of the expected information loss from using the topic distribution to proxy for the distribution of words in the textbook. Thus, selecting the topic with the lowest KL divergence is equivalent to picking the most informative topic about the source text. Figure 2 presents a summary of these KL-divergence calculations, together with bootstrapped 95% confidence bands for the innovation topic and the average of the other topics. Using this method, the innovation topic is significantly more informative about textbook innovation than the typical topic written by analysts. To argue that this lower KL divergence is because of innovation rather than general finance language, the lower panel of Figure 2 presents a placebo exercise in which the source text is a standard corporate finance textbook (Welch's 2008 *Corporate Finance: An Introduction*). Unlike the comparison with the innovation textbook, the innovation topic exhibits a similar KL divergence to other topics.

Contextually, this innovation topic relates intuitively to the factors that describe innovative companies. For example, Figure 3 presents the topic distribution across words in the form of a word cloud (Table A.3 in the online appendix provides word frequencies for the 10 most common words in the topic). When writing about this topic, analysts most frequently use words such as *revenue*, *growth*, *services*, *network*, *market*, *company*, and *technology*. Beyond the contextual word usage, we show that firms that have high values of this measure have the hallmarks of innovative firms.

Before using the loadings as a measure of innovation, it is important to refine the measure to account for analysts who write about the innovative activities of the firm in a negative or neutral tone. Specifically, if an analyst is talking with neutral or ambivalent sentiment about the company, it is less likely that the strong loading on the "innovation" topic reflects more innovation by the company. We address this source of noise by focusing on the analyst reports that have relatively strong positive sentiment (i.e., those in the top quartile of sentiment, measured by $\frac{\#positive\_words - \#negative\_words}{\#total\_words}$ from the word list in Loughran and McDonald 2011). For our main measure, we disregard innovation language in analyst reports with sentiment below the 75th percentile by setting the topic loading at the analyst report level

Figure 2. Selecting the Innovation Topic—Kullback–Liebler Divergence from an Innovation Textbook



Notes. The upper panel in this figure presents the Kullback–Liebler (KL) divergence of our selected innovation topic and the source textbook on innovation (*Measuring Innovation* by Tidd et al. 2005) and compares it to the average KL divergence from the source textbook on innovation across all of the other topics in the 15-topic LDA fit. The lower panel is a placebo exercise that uses a standard corporate finance textbook (Welch's 2008 *Corporate Finance: An Introduction*) as the source text instead. The bars indicate the mean KL divergence, and the bands provide 95% confidence intervals computed from the 2.5% and 97.5% percentiles of a bootstrapped sampling distribution with 500 replications.

to be zero before aggregating these topic loadings to the firm-year level. This firm-year level measure is our main text-based measure of innovation, $innov\_text_{it}$.[11]

We perform several empirical tests to ensure that it is the content of the topic, rather than the screen on sentiment, that drives the properties of our measure. Indeed, the innovation topic loadings and the sentiment have a low correlation, equal to 0.08. Thus, reports that load on the innovation topic are unlikely to merely reflect positivity about earnings or revenue. Specific to this point, as robustness exercises, we have also controlled explicitly for average sentiment as well as words that relate to revenue or growth (see the discussion in Section 4.2.2). More powerfully, we rerun the analysis on the subset of firm-years that have below-median analyst sentiment (see Table A.12 in the online appendix), and we have constructed a version of the measure that is based on an LDA fitted to a corpus with revenue and growth words dropped from the text (see Table A.16 in the online appendix). In each case, the properties of our innovation measure are similar.

Figure 3. Text-Based Innovation Measure: Word Cloud

Notes. This word cloud describes the frequency distribution of words used in the "innovation" topic. The topic itself is from the output of an latent Dirchlet allocation (LDA) model fit to a corpus of analyst reports for S&P 500 firms. We set the number of topics in the fitted LDA model to be 15, then select the topic (out of these 15) for which the distribution of words in the topic is closest to an innovation textbook (Tidd et al. 2005).

3.3. Comparison with Patenting Outcomes

An important advantage of the text-based innovation measure is that it captures innovative activities of firms that do not patent. In our sample of 703 firms in the S&P 500, a total of 219 firms have zero patenting throughout the full sample period (1990–2010). Although these firms do not patent, many are highly innovative. Figure 4(a) presents side-by-side boxplots of our text-based innovation measure for patenting firms versus nonpatenting firms. Although patenting firms have higher text-based innovation on average, the distribution of text-based innovation exhibits substantial overlap between nonpatenting firms and patenting firms. Specific examples of highly innovative nonpatenting firms are also consistent with this view.[12]

In columns (2)–(4) of Tables 1 and 2, we present summary comparisons of text-based innovation for firms with and without patents. On average, patenting firms have higher text-based innovation than nonpatenting firms by 0.27 standard deviations (0.20 SD at the firm level), a difference that is statistically significant at the 1% level, indicating a significant positive correlation between our text-based measure and whether a firm engages in patenting. Within the set of patenting firms, our text-based measure and patenting outcomes are also positively correlated. To this end, Figure 5 presents a graphical depiction of how the text-based measure fits patenting outcomes by plotting the log of patenting measures against decile bins of the text-based innovation measure. Regardless of the measure of patenting employed (counts, citations, or citations per patent), the text measure correlates strongly with contemporaneous patenting intensity within the set of patenting firms.[13]

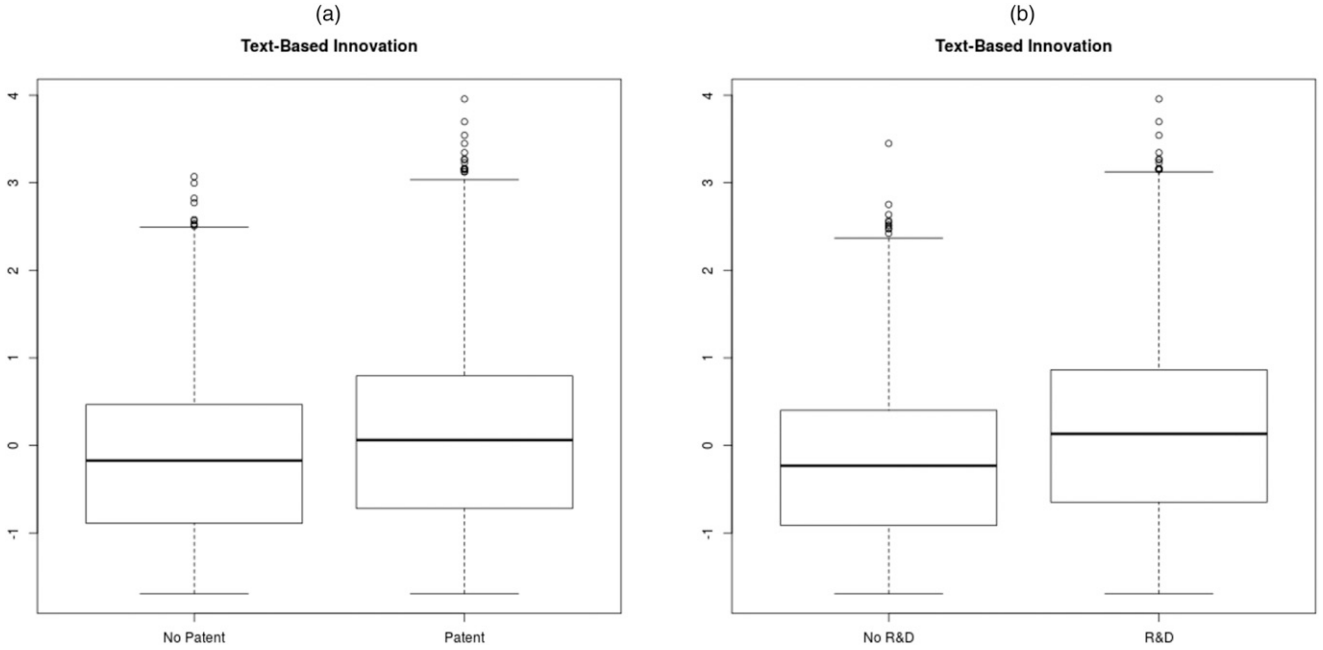
3.4. Comparison with Technology Development via R&D

The text-based innovation measure also captures innovative activities of firms that do not perform R&D. In our sample of 703 firms in the S&P 500, a total of 329 firms have zero R&D expenditures throughout the full sample period (1990–2010). Similar to the nonpatenting firms, many non-R&D firms are highly innovative. Figure 4(b) presents side-by-side boxplots of our text-based innovation measure for firms with and without R&D, which shows that there is substantial overlap in the distribution of text-based innovation for firms with and without R&D.

In columns (5)–(7) of Tables 1 and 2, we present summary comparisons of text-based innovation for firms with and without R&D.[14] Firms with positive R&D expenditure have higher text-based innovation by 0.39 standard deviations (0.43 at the firm level), a difference that is statistically significant at the 1% level, indicating a significant positive correlation between our text-based measure and R&D expenditure. The time-series and cross-industry correlations are also informative, both as a point of validation and also to highlight specific industries and time periods in which text-based innovation is high and R&D intensity is low. Our interpretation of this section's results is that the text-based measure of innovation measures the adoption of technology, even in industries that have low R&D intensity.

In the time series (1990–2010), the text-based innovation measure captures the macro-level trends in innovative activity well. Figure 6 presents the plot of the text-based measure of innovation over time (a value-weighted average across firms). For comparison, the time series of average R&D expenditures by year is also presented on the same plot. In the time series, there is a strong relationship between text-based innovation and aggregate R&D intensity, which have a correlation of 0.58. In the cross-section, the text-based innovation measure also matches cross-industry differences in R&D expenditures well. Figure 7 presents a bar plot of industry-level R&D expenditures (demeaned by the average R&D intensity), with the industries sorted from the highest value to the lowest value of innovation using our text-based measure. The figure shows a significant relationship between R&D and the innovation measure at the industry

Figure 4. Distribution of Text-Based Innovation



Notes. This figure shows the distribution of the text-based innovation measure. Panel (a) shows boxplots of the text-based innovation measure for patenting firms and for nonpatenting firms. Panel (b) shows boxplots of the text-based innovation measure for R&D firms and for non-R&D firms.

level, which is also indicated by the correlation of 0.47.

Examining the fit industry by industry yields additional qualitative insight into what the text-based measure of innovation adds to existing proxies. Notably, industries with high text-based innovation and high R&D intensity tend to be industries in which it is more natural to develop technologies in-house (e.g., Electronic Equipment and Business Services). In contrast, the ill-fitting industries with high text-based innovation are industries in which the most innovative companies are skilled at technology adoption (e.g., Communications and Motion Pictures). These patterns suggest that the text-based measure is useful to identify firms that utilize technology to support a revenue-generating system and that the measure is most useful beyond standard measures when it reflects the firm's ability to adopt technology productively.

We have also estimated the relation between R&D intensity and the text-based measure more systematically in a panel data context (results presented in Table A.19 in the online appendix). Even within narrowly defined industries (four-digit SIC), there is a strong statistically significant link between R&D intensity and text-based innovation. The link between text-based innovation and R&D intensity persists after controlling for other firm-specific factors, and text-based innovation reliably forecasts R&D intensity one year ahead, even holding constant this year's R&D intensity. These within-industry findings are consistent with the text-based innovation measure, capturing technology adoption decisions that are broader than the decision to develop technology via R&D expenditure.

4. Empirical Results

In this section, we use our text-based innovation measure to evaluate the impact of innovation on various firm-performance measures (i.e., return on assets, Tobin's Q, and sales growth). Beyond subjecting this relation to regression analysis with firm and industry-year fixed effects, we perform several robustness checks on this innovation–performance relation. Finally, in relating the measure to other innovation outcomes, we show that the text-based innovation measure primarily captures innovation quality, both within the set of patenting firms and more broadly.

4.1. Innovation and Performance

In this section, we empirically relate the text-based innovation measure to firm-performance measures with three goals in mind: (1) showing that the text-based innovation measure contains value-relevant information; (2) drawing a comparison of the text-based measure to patenting and R&D measures of innovation; and (3) understanding whether the measure's

Table 1. Summary Statistics: Firm-Year Summary Statistics

	All	Patents > 0	Patents = 0	(2) − (3)	R&D > 0	R&D = 0	(5) − (6)
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Innovation measures							
Text-Based Innovation	0.00	0.12	−0.16	0.27***	0.18	−0.21	0.39***
Patents	62.9	109	0.00	109***	118	1.71	116***
R&D/Assets	0.03	0.04	0.00	0.04***	0.05	0.00	0.05***
Performance Measures							
ROA	0.15	0.16	0.15	0.00	0.16	0.15	0.01
Log(*Q*)	0.55	0.61	0.47	0.15***	0.66	0.43	0.24***
Sales Growth	0.09	0.08	0.10	−0.02	0.08	0.09	−0.01
Characteristics							
Log(*Assets*)	8.78	8.86	8.67	0.19**	8.75	8.81	−0.05
Asset Tangibility	0.36	0.30	0.43	−0.13***	0.27	0.46	−0.19***
Leverage	0.58	0.57	0.61	−0.04**	0.56	0.61	−0.05***
Log(*Age*)	3.18	3.20	3.15	0.06**	3.16	3.20	−0.04
Observations	6,200	3,586	2,614		3,268	2,932	

Notes. This table presents sample means at the firm-year level for the main variables of interest using the full sample (column (1)) and using subsamples that are indicated in the column heading in columns (2), (3), (5) and (6). Columns (4) and (7) report the difference in means for the indicated subsamples, together with statistical significance from a two-sample *t*-test. Standard errors are clustered by firm.
**$p < 0.05$; ***$p < 0.01$.

properties are similar for patenting firms versus nonpatenting firms.

Specifically, we estimate the relation between greater innovation at date t and firm performance on date $t+1$ using the specification:

$$Y_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it}. \quad (1)$$

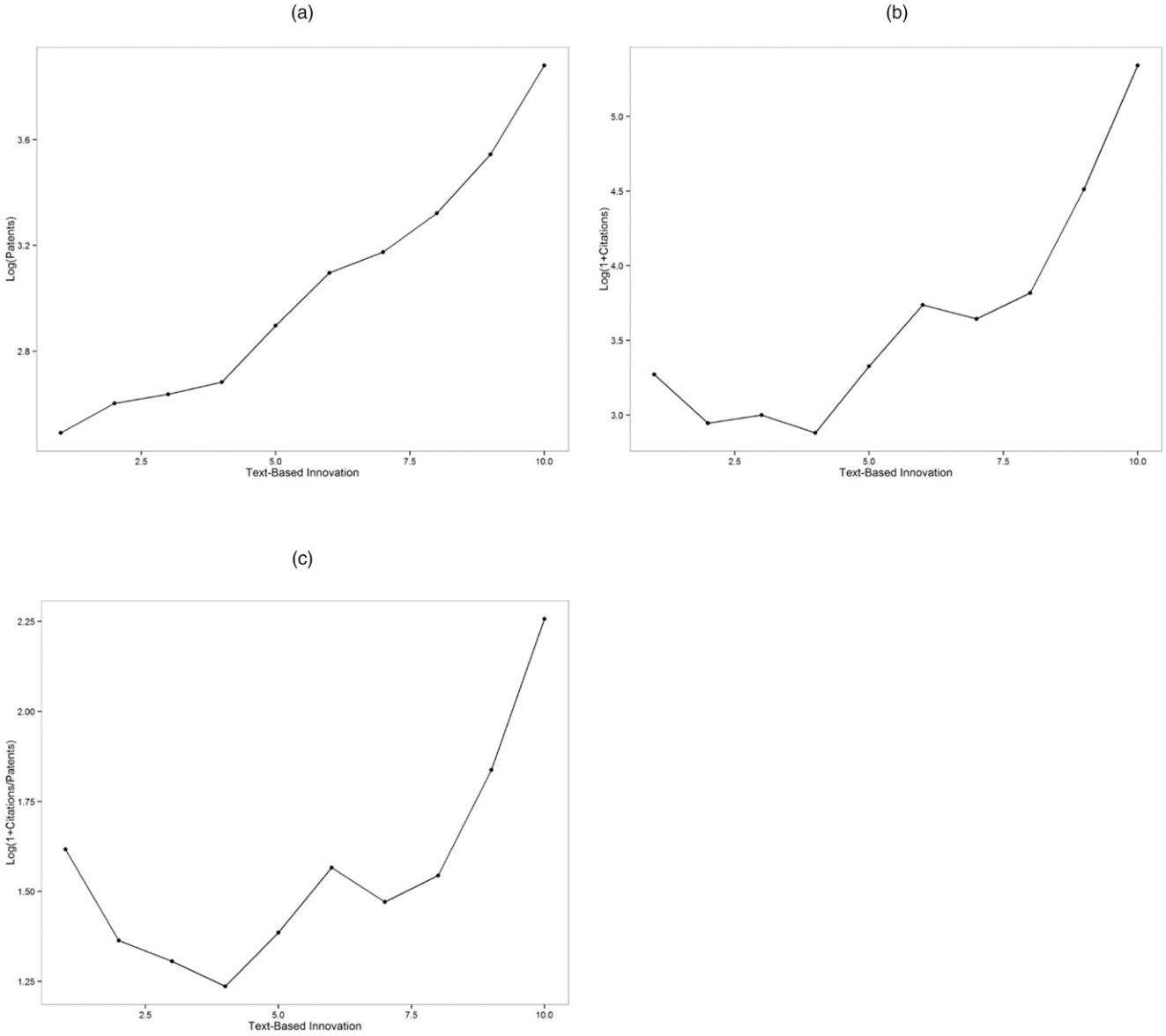
The dependent variable Y_{it+1} is a measure of firm performance for firm i in year $t+1$. We use this specification to evaluate three distinct dependent variables: *return on assets*, *Tobin's Q*, and *sales growth*. The coefficient of interest is β_1, which indicates how greater text-based innovation associates with changes in future operating performance. If innovation is valuable, our prediction is that $\beta_1 > 0$. Our base specification includes year and industry (four-digit SIC (SIC4)) fixed effects (γ_t and ξ_s), but we also include firm fixed effects (ξ_i) and industry-year fixed effects (γ_{st}) in some specifications to account for

Table 2. Summary Statistics: Firm-Level Summary Statistics

	All	Patents > 0	Patents = 0	(2) − (3)	R&D > 0	R&D = 0	(5) − (6)
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Innovation measures							
Text-Based Innovation	0.00	0.06	−0.14	0.20***	0.20	−0.23	0.43***
Patents	44.0	64.0	0.00	64.0***	81.5	1.49	80.0***
R&D/Assets	0.03	0.04	0.00	0.03***	0.05	0.00	0.05***
Performance measures							
ROA	0.15	0.15	0.15	−0.00	0.15	0.14	0.01
Log(*Q*)	0.55	0.58	0.49	0.08**	0.67	0.41	0.26***
Sales Growth	0.09	0.09	0.11	−0.02	0.09	0.10	−0.01
Characteristics							
Log(*Assets*)	8.57	8.67	8.36	0.31***	8.50	8.65	−0.14*
Asset Tangibility	0.35	0.33	0.39	−0.07***	0.26	0.45	−0.19***
Leverage	0.57	0.58	0.57	0.00	0.55	0.60	−0.05***
Log(*Age*)	3.05	3.09	2.96	0.13***	3.01	3.10	−0.08**
Firms	703	484	219		374	329	

Notes. This table presents sample means at the firm level for the main variables of interest using the full sample (column (1)) and using subsamples that are indicated in the column heading in columns (2), (3), (5) and (6). Columns (4) and (7) report the difference in means for the indicated subsamples, together with statistical significance from a two-sample *t*-test. Standard errors are clustered by firm.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Figure 5. Relating Patent Counts and Patent Citations to the Text-Based Innovation Measure (Decile Bins)



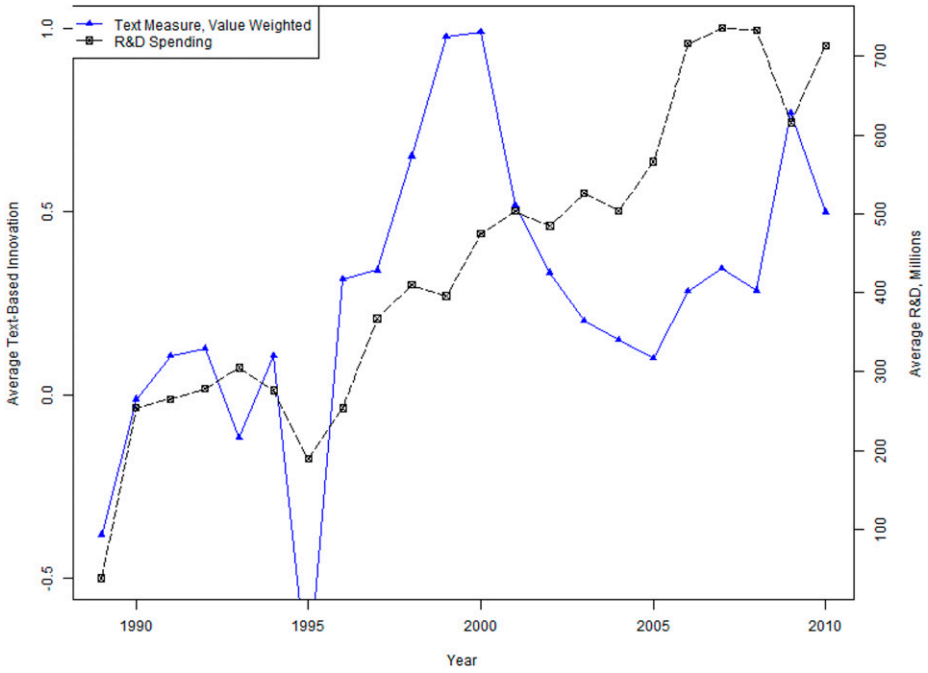
Notes. This figure plots the relation between the text-based innovation measure and commonly used patenting measures. In each panel, the text-based innovation measure is grouped into 10 deciles. Panel (a) presents the relation between text-based innovation and logged patent counts (log(1 + *Patents*)). Panel (b) presents the relation between text-based innovation and patent citations (log(1 + *Citations*)). Panel (c) presents the relation between text-based innovation and citations per patent ($\log(1 + \frac{Citations}{Patent})$).

firm-specific unobservables and industry-specific time-varying unobservables.[15] To account for correlated errors, the specifications cluster standard errors by firm.[16]

To facilitate comparisons of our estimated magnitudes to existing measures of innovation, our specifications for Equation (1) control for patenting outcomes (counts and citations), R&D intensity, and an indicator for nonpatenting firm. The specifications also include standard time-varying firm controls that relate to operating performance and innovation. For ease of interpretation, the text-based innovation measure is standardized to have a mean of zero and a standard deviation of one. In the discussion that follows, we use this specification to explicitly contrast the estimated magnitude of the relation between text-based innovation and performance with the estimated magnitudes from other innovation measures.

4.1.1. Innovation and Firm Performance. We first evaluate the relation between text-based innovation and future operating performance. Columns (1)–(3) in Table 3 present the results from estimating Equation (1), in which the dependent variable is *return on assets* (earnings before interest, taxes, depreciation,

Figure 6. (Color online) Time Series of Text-Based Innovation Measure and R&D (1990–2010)



Notes. This figure provides a time-series plot of the text-based innovation measure, which is aggregated to a yearly figure by computing the value-weighted average. The time-series plot average R&D expenditure for firms in the sample is also presented in this figure. The two series have a time-series correlation of 0.58.

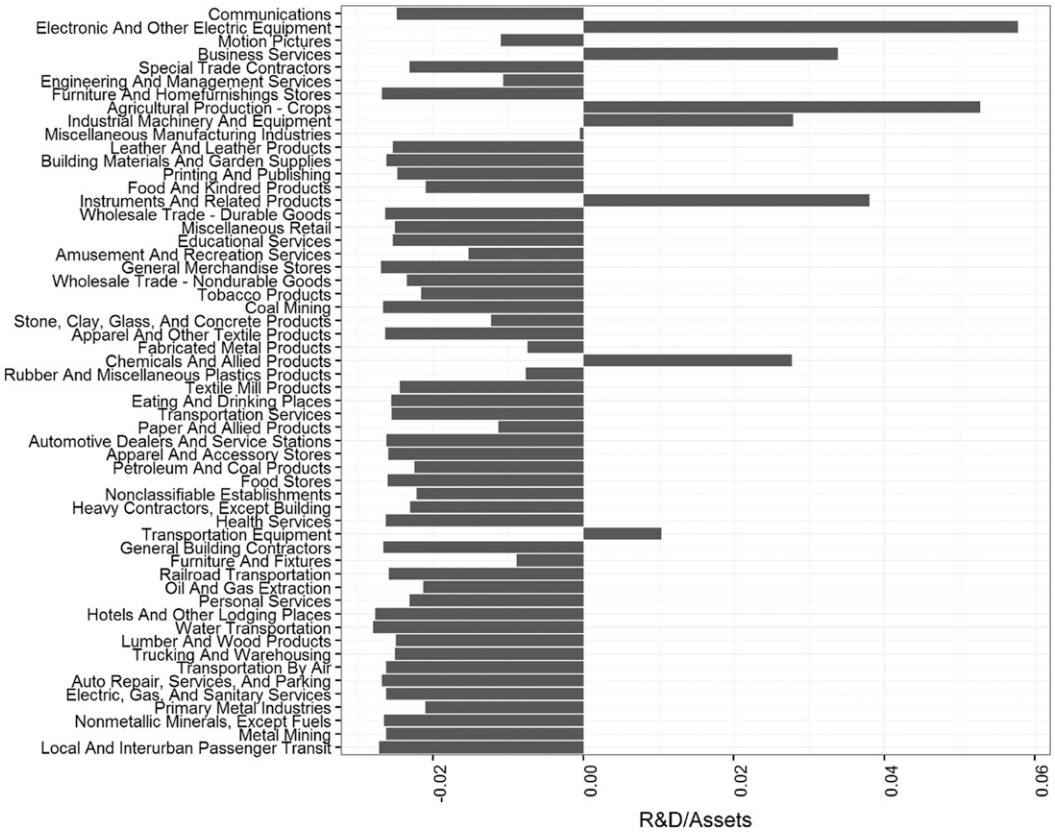
and amortization (EBITDA)/Assets). Across specifications, the estimates imply that a one-standard-deviation increase in the text measure is associated with between 0.5- and 0.9-percentage-points higher return on assets in the following year. This estimated effect is robust to including firm fixed effects (columns (2) and (3)) and industry-year fixed effects (column (3)), and the magnitude of the coefficient estimate is stable upon including these more granular fixed effects. Thus, within-firm variation in the text-based measure is a useful predictor for future operating performance, and this relation is not explained away by time-varying unobservables at the industry level.

Relating to existing measures of innovation, the text-based measure is more robustly associated with increases in operating performance than patent counts, patent citations, and R&D intensity. Patent counts and patent citations are not significantly correlated with future operating performance in any specification. The statistical significance for R&D intensity is marginal and not robust across specifications, though R&D intensity is positively correlated with future operating performance, and the estimated magnitudes are similar to our measure. Moreover, as our specifications control for alternative measures of innovation, these findings imply that text-based innovation is valuable beyond what existing innovation measures would predict.

We also expect corporate innovation to have implications for the firm's growth opportunities. Intuitively, investors recognize an innovative firm when they see it, and rationally estimate an increase in the firm's future cash flows, thus enhancing its market valuation. In line with this intuition, we should expect a significant relation between text-based innovation and future Tobin's Q because the market value will come to embed this innovation premium. To evaluate this hypothesis, columns (4)–(6) in Table 3 present the results from estimating Equation (1), in which the firm-performance measure is the natural log of Tobin's Q for firm i at date $t + 1$. Across specifications, we find that text-based innovation is strongly related to future Tobin's Q. Specifically, a standard-deviation increase in text-based innovation is associated with between 2.2% and 8.3% greater Tobin's Q. This estimated effect is robust to including firm fixed effects (columns (2) and (3)) and two-digit SIC (SIC2)-year fixed effects (column (3)), which accounts for time-varying unobserved characteristics at the industry level.

In contrast, patenting outcomes (counts and citations) exhibit a much weaker and less robust relation to future growth opportunities: Patent counts are unrelated to Tobin's Q, whereas patent citations exhibit a positive relation (roughly 2%) to future Tobin's Q that vanishes upon including industry-year fixed effects. In the specification with industry fixed effects, a standard-deviation change in R&D intensity exhibits a

Figure 7. Cross-Industry Plot of R&D (1990–2004), Relationship to Text-Based Measure



Notes. This figure provides a plot of R&D expenditures (demeaned by the average R&D/Assets) by industry covered in the sample of S&P 500 firms. To show the relation between text-based innovation and R&D expenditures across industries, the industries in the plot are ordered from the highest value of text-based innovation to the lowest value. The correlation between R&D expenditures and the text-based measure across industries is 0.40. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1.

similar relation to future growth opportunities as the text-based measure, but the relation is less robust upon including firm and industry-year fixed effects (columns (5) and (6) in Table 3).

Finally, we expect innovation to have implications for the firm's sales insofar as increases in innovation reflect the introduction of new or better-quality products. In this case, we should expect to see sales growth increase following an increase in innovation. To evaluate this hypothesis, columns (7)–(9) in Table 3 present the results from estimating Equation (1), in which the firm-performance measure is the sales growth for firm i at date $t+1$. Across specifications, we find that text-based innovation exhibits a positive and statistically significant link to future sales growth. A standard-deviation increase in text-based innovation is associated with between 1.0- and 2.0-percentage-points greater sales growth. This estimated effect is robust to including firm fixed effects (columns (2) and (3)) and SIC2-year fixed effects (column (3)), which accounts for time-varying unobserved characteristics at the industry level.

In contrast to the robust significance of the text-based measure, we find that patent counts, patent citations, and R&D intensity are inconsistently related to sales growth. Indeed, patent counts appear to be negatively associated with sales growth in the specification with firm fixed effects, whereas there is no apparent relationship between sales growth and R&D intensity.

4.1.2. Patenting Firms vs. Nonpatenting Firms. A notable advantage of our text-based measure is that it can be computed for firms without patents and, thus, can help evaluate innovation for a broader set of firms than patenting firms. In Table 4, we highlight this feature of text-based innovation by including the interaction between text-based innovation and an indicator for *Nonpatenting Firm* (= one if a firm has zero patents for the entire sample period). The coefficient on this interaction provides a test for significant differences in the innovation–performance relation between patenting firms and nonpatenting firms.[17]

Across specifications in Table 4, we find similar point estimates for the coefficient on innovation for

Table 3. Performance of Firms and Text-Based Innovation (1990–2010): Firm Performance

	Dependent variable								
	ROA_{t+1}			$Log(Q)_{t+1}$			*Sales Growth*$_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Innovation $(Z)_t$	0.009***	0.005***	0.005***	0.083***	0.049***	0.039***	0.015***	0.010**	0.013**
	(0.002)	(0.001)	(0.002)	(0.009)	(0.007)	(0.007)	(0.004)	(0.004)	(0.005)
Log(*Patents*)$_t$	0.002	−0.002	−0.0001	0.003	−0.027	−0.009	−0.007	−0.015**	−0.016*
	(0.003)	(0.003)	(0.003)	(0.015)	(0.016)	(0.016)	(0.005)	(0.007)	(0.008)
Log(*Citations*)$_t$	0.001	−0.0004	−0.001	0.016*	0.020**	0.007	−0.003	0.002	0.004
	(0.002)	(0.002)	(0.002)	(0.008)	(0.008)	(0.009)	(0.004)	(0.004)	(0.005)
R&D/Assets $(Z)_t$	0.006	0.010***	0.008**	0.074***	0.027	0.023	−0.001	−0.007	−0.010
	(0.005)	(0.004)	(0.004)	(0.020)	(0.022)	(0.023)	(0.006)	(0.009)	(0.010)
Nonpatenting Firm	−0.009*				−0.037				0.00000
	(0.006)				(0.032)				(0.012)
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.674	0.725	0.577	0.771	0.800	0.099	0.159	0.225

Notes. This table presents ordinary least squares regressions that link the text-based innovation measure to measures of performance: *ROA*, log(*Q*), and *sales growth*. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures—log(*patents*) and log(*citations*)—are included in the specification to provide a basis for comparison. Full results are reported in the online appendix (Table A.6). Variable definitions are presented in Table A.1 in the online appendix. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

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

patenting firms versus nonpatenting firms, indicating that innovation has similar performance consequences for both types of firms. Indeed, regardless of the measure of firm performance, we cannot reject the hypothesis that innovation exhibits the same relation to performance for patenting and nonpatenting firms. Not only is the interaction statistically insignificant, but the magnitude of the difference is small, particularly for the regressions of return on assets (roughly 0.1 percentage points). This quantitative similarity of the estimates suggests that the text-based innovation measure is informative beyond the set of patenting firms.[18]

4.1.3. Dynamics of the Innovation–Performance Relation.

In this section, we examine how the empirical relation between text-based innovation and firm performance

Table 4. Performance of Firms and Text-Based Innovation (1990–2010): Patenting Firm Split

	Dependent variable								
	ROA_{t+1}			$Log(Q)_{t+1}$			*Sales Growth*$_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation $(Z)_t$	0.009***	0.005***	0.005***	0.085***	0.052***	0.041***	0.015***	0.009*	0.012**
	(0.002)	(0.002)	(0.002)	(0.010)	(0.008)	(0.008)	(0.005)	(0.005)	(0.006)
× *Nonpatenting Firm*	0.001	0.001	−0.001	−0.007	−0.003	−0.006	0.003	0.011	0.011
	(0.004)	(0.003)	(0.003)	(0.017)	(0.016)	(0.015)	(0.009)	(0.011)	(0.012)
R&D/Assets $(Z)_t$	0.007	0.010***	0.008**	0.081***	0.029	0.023	−0.005	−0.007	−0.010
	(0.004)	(0.004)	(0.004)	(0.019)	(0.023)	(0.023)	(0.006)	(0.009)	(0.010)
Nonpatenting Firm	−0.011**			−0.063**			0.011		
	(0.006)			(0.031)			(0.012)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.674	0.725	0.577	0.771	0.800	0.099	0.159	0.224

Notes. This table presents ordinary last squares regressions that link the text-based innovation measure to measures of performance: *ROA*, log(*Q*), and *sales growth*. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Full results are reported in the online appendix (Table A.6). Variable definitions are presented in Table A.1 in the online appendix. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

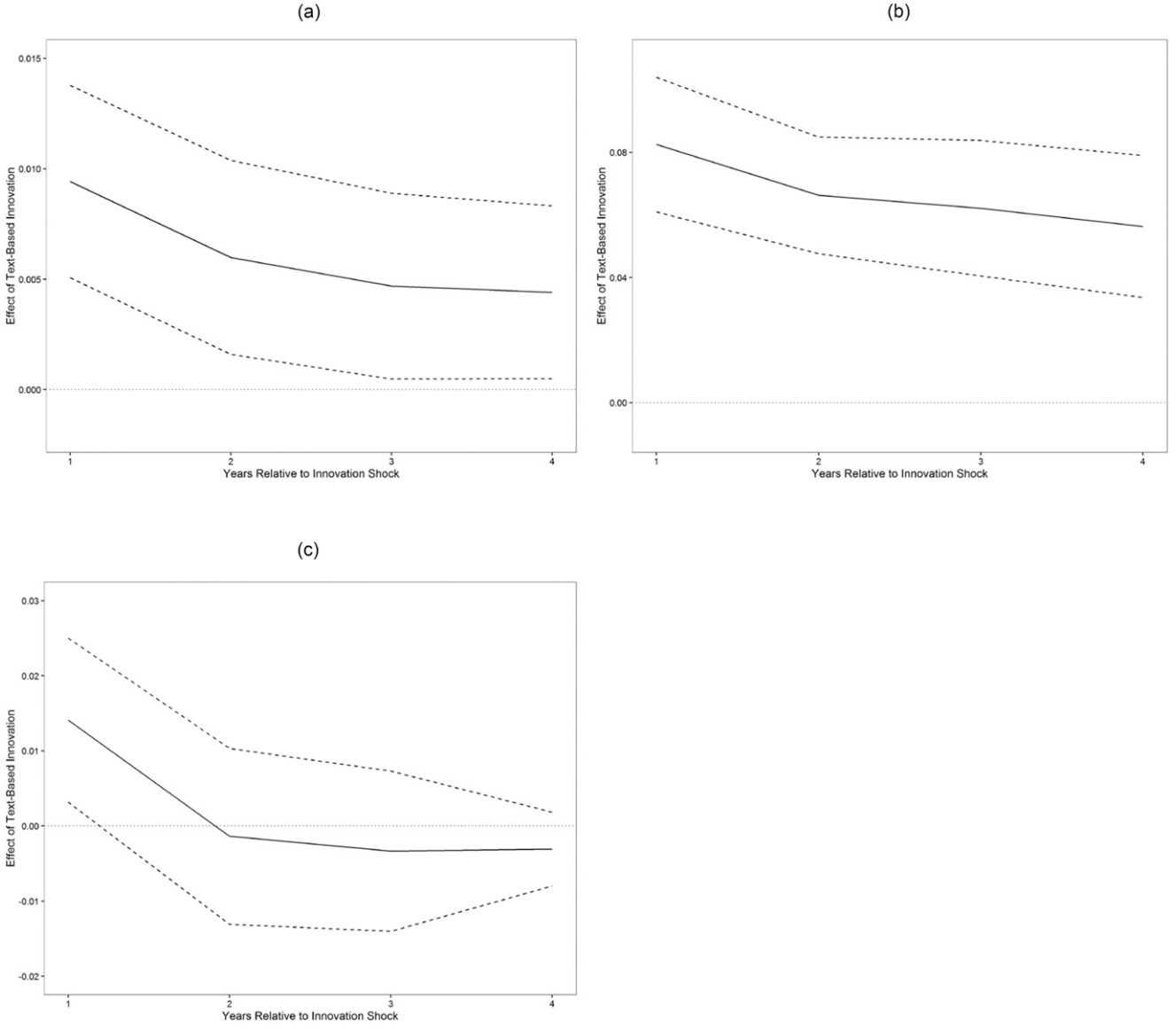
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

holds up at time horizons beyond one year. To this end, Figure 8 plots the estimated coefficient of text-based innovation in performance regressions that vary the timing of the performance measure from one through four years into the future. The underlying regression specification for these plots is Equation (1), with industry and year fixed effects, and the dotted lines represent 95% confidence bands.

The time pattern of the results in Figure 8 is consistent with the text-based measure providing a useful description of—to borrow language from Drucker (1985)—innovation as a "resource" that has been added to the firm. Specifically, Figure 8, (a) and (b) show that the estimated coefficients in operating performance and Q regressions are positive and significant, but for both measures of performance, the estimated coefficient declines smoothly over time. By contrast, Figure 8(c) shows that text-based innovation is related to a transitory increase in sales growth, which is only positive and significant in the year following the increase in text-based innovation. Collectively, these findings indicate that text-based innovation represents a one-time increase in sales that leads to greater (but depreciating) performance gains in the intermediate term.

Table A.10 in the online appendix presents full detail on the regression specifications for years $t+1$ through $t+4$ that underlie Figure 8. In contrast to our

Figure 8. Long-Run Effects of Innovation on Performance—Forecasting *ROA* and Tobin's Q up to Four Years Out



Notes. These plots present the response in *ROA* (panel (a)), Q (panel (b)), and *sales growth* (panel (c)) to a one-standard-deviation increase in the text-based measure of innovation. The x-axis represents the number of years ahead, and the y-axis is the beta estimate from Table A.10 in the online appendix. Dotted lines represent 95% confidence bands around the estimated effects.

findings for text-based innovation, other innovation measures exhibit nonrobust and inconsistent time patterns of results over the one- to four-year horizon. Specifically, patenting outcomes (counts and citations) exhibit a statistically insignificant relation to operating performance and sales growth, and only a small, marginally significant positive relation to Q for patent citations. R&D intensity is positively related to Q over the four-year horizon, but it bears an insignificant relation to operating performance and a slight negative relation to sales growth at time horizons $t+3$ and $t+4$. These findings for other innovation measures are difficult to reconcile with innovation as a resource for the firm. One rationale for these inconsistent time patterns for other innovation measures is that both R&D intensity and patenting outcomes measure innovation at an uncertain lag relative to firm performance (e.g., patents take time to commercialize or may reflect past innovations; R&D investments similarly take time to materialize into useful outcomes).

Related to this point of timing, it is helpful to examine robustness in the timing of text-based innovation's link to performance. Especially if firm performance and innovation are persistent within a firm, a potential concern is that performance drives text-based innovation rather than the other way around. For this reason, we empirically evaluate the reverse relation between text-based innovation and return on assets for up to four time lags.[19] Table 5 presents regression results from a regression of text-based innovation on four time lags of return on assets, as well as fixed effects and controls that were included in our main specifications. In specifications with industry and year fixed effects, we find a significant relation between one-year lagged performance and text-based innovation (but not longer time lags). As we enrich the specification, however, the positive relation between lagged performance and text-based innovation is not robust to firm or industry-year fixed effects. Contrasting these findings with our main results, our evidence appears to be most consistent with the notion that text-based innovation drives performance, rather than the other way around.

4.2. Robustness

In this section, we present robustness to our main text-based innovation measure. Broadly, we conduct two types of robustness exercises—robustness to the LDA model fit (i.e., choices of sample frame, number of topics, and meaning of topics) and robustness to explanations unrelated to model fit (i.e., analyst sentiment, analyst information sets, and use of revenue/growth words).

4.2.1. Fitted Model Robustness. Tables 6, 7, and 8 present estimates of the firm-performance specifications for three notable robustness exercises on the fit of the LDA model. Specifically, we report robustness on the specifications from Tables 3 and 4 that include

Table 5. Text-Based Innovation and Lagged Firm Performance (1990–2010)

	Dependent variable: *Text-Based Innovation*$_t$				
Explanatory variables	(1)	(2)	(3)	(4)	(5)
ROA_{t-1}	1.219***	1.120***	0.656**	0.521	0.598
	(0.322)	(0.376)	(0.310)	(0.322)	(0.366)
ROA_{t-2}	0.126	0.068	−0.030	0.011	0.039
	(0.316)	(0.397)	(0.327)	(0.332)	(0.409)
ROA_{t-3}	0.144	0.133	−0.008	0.013	−0.042
	(0.362)	(0.415)	(0.406)	(0.405)	(0.446)
ROA_{t-4}	−0.190	−0.276	−0.484	−0.499	−0.655**
	(0.300)	(0.329)	(0.314)	(0.308)	(0.322)
Lagged R&D and Patenting	X	X	X	X	X
Other controls		X		X	X
Industry (SIC4) FE	X	X			
Firm FE			X	X	X
Year FE	X	X	X	X	
SIC2-Year FE					X
Observations	3,621	3,621	3,621	3,621	3,621
Adjusted R^2	0.469	0.480	0.557	0.563	0.610

Notes. This table presents ordinary least squares regressions of the text-based innovation measure on lagged measures of performance (*ROA*), patenting, and R&D activity. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Variable definitions are presented in Table A.1 in the online appendix. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

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

firm and SIC2-year fixed effects (i.e., column (3) for *ROA*, column (6) for log(*Q*), and column (9) for *sales growth*).

First, we address the concern of look-ahead bias in the performance regressions. Because we construct the innovation topic from an LDA model fit on the entire sample period (1990–2010), a reader may be concerned that the innovation topic merely reflects factors that are eventually revealed to be valuable for firms, but that the information would not be viewed as innovation at the time of observation. To address this potential concern, we reproduce the performance results using a five-year rolling window version of text-based measure, which alleviates the look-ahead bias concern because the rolling-window measure is based solely on past data. For example, in the rolling-window version of the analysis, we construct the measure for a firm in 1995 using the topic loadings from an LDA model fit only using analyst reports from the previous five years (1990–1994).

Table 6 presents the performance results using the rolling-window measure in place of the main measure. Results on operating performance and Tobin's *Q* are nearly identical in magnitude and statistical significance using the rolling-window version, whereas the findings using sales growth are less robust (albeit with the same sign and similar magnitude to the main result). These findings suggest that the relation between text-based innovation measure and firm performance reflects the value of true innovative activity rather than look-ahead bias.

Second, in Table 7, we summarize the results of using a 10-topic LDA to construct the text-based innovation measure. Because LDA requires that the researcher specify the number of latent topics, it is important to show that the essential findings of our paper are not driven by this choice. When using the 10-topic LDA, we select a qualitatively similar topic, and we obtain results that are similar to Tables 3 and 4, which suggests that the results in the paper are not driven by the choice of the number of topics.

Third, in Table 8, we address the concern that the other topics in the 15-topic LDA are correlated with our measure and, thus, drive the result for a more mechanical reason (e.g., an "operating performance" topic emerges in the 15-topic LDA; see Figure A.1 in the online appendix). To address this potential issue, we control for each of the other topic loadings aggregated to the firm-year level. As the results in Table 8 indicate, the main results are qualitatively similar after controlling for other topic loadings, though in some cases, they become stronger.[20]

4.2.2. Robustness to Alternative Explanations. Table 9 presents a specification that accounts for three other alternative explanations. In particular, because construction of the measure relies on only the reports with high analyst sentiment, a reader may be concerned that the sentiment of the reports rather than their content is driving the relation of text-based innovation to the performance measures. Relatedly, given the words most prominently used in the innovation topic, a reader may have a separate concern that the LDA topic is merely a crude technique to approximate for whether analysts discuss the firm's revenue or growth prospects, unrelated to innovation. Finally, one might be concerned that mentions of technology drive the innovation topic (i.e., that a simpler approach would be to count words related to technology).

To address these related issues, we compute the average sentiment of analyst reports at the firm-year level (*Sentiment*), as well as construct word counts of analyst usage of the words "revenue," "growth," and

Table 6. Robustness of LDA Model Fit: Rolling Window Version (1994–2010)

	Dependent variable					
	Return on Assets$_{t+1}$		Log(*Q*)$_{t+1}$		*Sales Growth*$_{t+1}$	
Explanatory variables	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (*Z*)$_t$	0.006***	0.006***	0.040***	0.043***	0.010*	0.006
	(0.002)	(0.002)	(0.007)	(0.008)	(0.006)	(0.006)
× *Nonpatenting Firm*		–0.0003		–0.011		0.020
		(0.003)		(0.014)		(0.015)
Controls, firm FE, SIC2-year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.730	0.730	0.819	0.818	0.227	0.227

Notes. The specifications and variable definitions for *Return on Assets*, *Q*, and *Sales Growth* are analogous to those in Tables 3 and 4. This table reports a five-year rolling window version of the measure. All specifications account for the full set of other controls, firm fixed effects (FE), and two-digit SIC (SIC2) industry-year fixed effects. Standard errors that are clustered by firm are in parentheses.

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

Table 7. Robustness of LDA Model Fit: Firm Performance, $K = 10$ (1990–2010)

	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		*Sales Growth*$_{t+1}$	
Explanatory variables	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation $(Z)_t$	0.005***	0.005***	0.041***	0.042***	0.012**	0.011*
	(0.002)	(0.002)	(0.007)	(0.008)	(0.006)	(0.006)
× *Nonpatenting Firm*		−0.001		−0.005		0.010
		(0.003)		(0.015)		(0.012)
Controls, Firm FE, SIC2-Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.725	0.725	0.800	0.800	0.225	0.224

Notes. The specifications and variable definitions for *Return on Assets*, *Q*, and *Sales Growth* are analogous to those in Tables 3 and 4. This table reports the measure from a 10-topic LDA. All specifications account for the full set of other controls, firm fixed effects (FE), and two-digit SIC (SIC2) industry-year fixed effects. Standard errors that are clustered by firm are in parentheses.

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

"technology" to be used as controls in the firm-performance specifications (*Revenue Words*, *Growth Words*, and *Technology Words*). Table 9 presents the results of specifications that control for these measures. We find that controlling for these alternative explanations does not affect the nature of the results on operating performance or growth opportunities, but our findings on sales growth are not robust to controlling for these characteristics.

Related to sentiment and word usage, we perform two additional robustness exercises, which we report in the online appendix. First, to account for sentiment, we restrict attention to the firm-year observations for which the analyst sentiment is below the median. Table A.12 in the online appendix shows that the main findings are similar on this low-sentiment subsample. Second, rather than control for word counts, we construct an alternative (purged) text-based innovation measure by deleting "revenue" and "growth" words in the original corpus. This technique accounts for direct mentions of these terms without over-controlling for them in a regression specification. When we estimate the main specifications with this purged measure (see Table A.16 in the online appendix), all of the main findings, including the link between text-based innovation and sales growth, are robust. Taken together, these robustness exercises indicate that the topic does not merely reflect the relative incidence of particular words, but consistent with our motivation to use LDA, the measure captures the appropriate context in which these words appear together.

4.2.3. Negative Information About Innovation. As a final robustness exercise, we construct an additional measure of innovation—negative text-based innovation—which aggregates to the firm-year level the innovation topic loadings from analyst reports that have a negative sentiment. This measure is interesting in its own right, as it has the potential to describe firms

Table 8. Controlling for other topics, $K = 15$ (1990–2010)

	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		*Sales Growth*$_{t+1}$	
Explanatory variables	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation $(Z)_t$	0.006***	0.006***	0.039***	0.041***	0.016**	0.014**
	(0.002)	(0.002)	(0.008)	(0.009)	(0.006)	(0.007)
× *Nonpatenting Firm*		−0.001		−0.005		0.011
		(0.003)		(0.015)		(0.012)
Controls, Firm FE, SIC2-Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.725	0.725	0.800	0.800	0.227	0.226

Notes. The specifications and variable definitions for *Return on Assets*, *Q*, and *Sales Growth* are analogous to those in Tables 3 and 4. This table reports the main measure ($K = 15$) controlling for all other topic loadings. All specifications account for the full set of other controls, firm fixed effects (FE), and two-digit SIC (SIC2) industry-year fixed effects. Standard errors that are clustered by firm are in parentheses.

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

Table 9. Accounting for Alternative Explanations (1990–2010)

	Dependent variable					
	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		*Sales Growth*$_{t+1}$	
Explanatory variables	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.004*	0.005**	0.039***	0.045***	0.002	−0.003
	(0.002)	(0.003)	(0.010)	(0.010)	(0.008)	(0.008)
× *Nonpatenting Firm*		−0.005		−0.032		0.031*
		(0.004)		(0.022)		(0.018)
Sentiment (Z)$_t$	0.003**	0.004**	0.014**	0.014**	0.008	0.008
	(0.001)	(0.001)	(0.006)	(0.006)	(0.006)	(0.006)
Revenue Words (Z)$_t$	0.001	0.001	−0.005	−0.006	−0.013*	−0.013*
	(0.003)	(0.003)	(0.012)	(0.012)	(0.008)	(0.008)
Growth Words (Z)$_t$	0.007***	0.007***	0.037***	0.037***	0.012*	0.012*
	(0.002)	(0.002)	(0.009)	(0.009)	(0.007)	(0.007)
Technology Words (Z)$_t$	0.001	0.001	0.008	0.008	0.001	0.002
	(0.002)	(0.002)	(0.011)	(0.011)	(0.008)	(0.007)
Controls, Firm FE, SIC2-Year FE	X	X	X	X	X	X
Observations	4,218	4,218	4,121	4,121	4,222	4,222
Adjusted R^2	0.738	0.738	0.819	0.819	0.190	0.191

Notes. The specifications and variable definitions for *Return on Assets*, Q, and *Sales Growth* are the same as in Tables 3 and 4. In addition, these specifications include controls for analyst sentiment, word count frequencies of "revenue" and "growth" words (words with "gro" or "rev" as their root), and the word count frequency of "technology" words (words with "tech" as their root). All specifications account for the standard set of other controls, firm fixed effects (FE), and year fixed effects. Standard errors that are clustered by firm are in parentheses.

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

that innovate poorly, but this exercise also helps to provide a validation that the main measure's construction is reliable.

Tables 10 and 11 present the results from estimating the main specification in Equation (1), but replacing text-based innovation with the negative text-based innovation measure. Consistent with the motivating intuition, we find a robust and strongly significant inverse relation between negative text-based innovation and future firm performance.[21] Table 11 presents specifications that include an interaction between the innovation measure and a *Nonpatenting Firm* indicator to test for significant differences between patenting firms and nonpatenting firms. Although eight out of nine specifications exhibit a statistically insignificant interaction, the magnitude of the

Table 10. Performance of Firms and Negative Text-Based Innovation (1990–2010)

	Dependent variable								
	Return on Assets$_{t+1}$			Log(Q)$_{t+1}$			*Sales Growth*$_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Neg. Text-Based Innovation (Z)$_t$	−0.009***	−0.006***	−0.005***	−0.040***	−0.022***	−0.021***	−0.023***	−0.023***	−0.023***
	(0.002)	(0.001)	(0.002)	(0.008)	(0.006)	(0.007)	(0.004)	(0.004)	(0.004)
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.675	0.725	0.567	0.768	0.798	0.103	0.164	0.229

Notes. This table presents ordinary least squares regressions that link the "negative" text-based innovation measure to measures of performance: *Return on Assets*, log(Q), and *sales growth*. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures—log(*patents*), log(*citations*), an indicator for patenting firm, and *R&D intensity*—are included in the specification, but not reported for brevity of presentation. Other controls include log(*assets*), *asset tangibility*, *leverage*, log(*age*), and *cash/assets*. Variable definitions are presented in Table A.1 in the online appendix. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

***$p < 0.01$.

Table 11. Performance of Firms and Negative Text-Based Innovation (1990–2010): Patenting Firm Split

	Dependent variable								
	Return on Assets$_{t+1}$			Log(Q)$_{t+1}$			*Sales Growth*$_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Neg. Text-Based Innovation (Z)$_t$	−0.009***	−0.008***	−0.006***	−0.036***	−0.024***	−0.023***	−0.026***	−0.025***	−0.025***
	(0.002)	(0.002)	(0.002)	(0.008)	(0.006)	(0.006)	(0.005)	(0.006)	(0.007)
× *Nonpatenting Firm*	0.003	0.007**	0.005	−0.010	0.011	0.013	0.012	0.011	0.015
	(0.004)	(0.003)	(0.003)	(0.020)	(0.015)	(0.017)	(0.009)	(0.011)	(0.014)
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.435	0.676	0.726	0.563	0.766	0.798	0.101	0.163	0.229

Notes. This table presents ordinary least squares regressions that link the "negative" text-based innovation measure to measures of performance: *Return on Assets*, log(Q), and *sales growth*. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures—log(*patents*), log(*citations*), an indicator for patenting firm, and *R&D intensity*—are included in the specification, but not reported for brevity of presentation. Other controls include log(*assets*), *asset tangibility*, *leverage*, log(*age*), and *cash/assets*. Variable definitions are presented in Table A.1 in the online appendix. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

***$p < 0.01$.

interactions is often quite substantial (half of the main effect or more for the ROA specifications). On account of this, the overall relation between negative text-based innovation and firm performance is statistically insignificant among the set of nonpatenting firms. Still, within the set of patenting firms, the relation between negative text-based innovation and future firm performance (given by the main effect) is negative and robust.[22]

4.3. Text-Based Innovation vs. Other Aspects of Innovation

In this subsection, we examine the connection between the text-based innovation measure and other aspects of innovation. Specifically, we estimate the connection between text-based innovation and future patenting outcomes (patent counts and citations-per-patent), as well as contemporaneous values of the patent value measure introduced by in Kogan et al. (2017) and a measure of product introductions introduced by Mukherjee et al. (2017).

We empirically relate the text-based innovation measure to these other innovation measures with the goal of understanding how text-based innovation relates to existing measures. Specifically, we estimate the relation between greater innovation at date t and other innovation measures using the specifications:

$$future\_innovation\_measure_{i,t+1\to t+3} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it}, \quad (2)$$

$$current\_innovation\_measure_{it} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it}. \quad (3)$$

The dependent variable $future\_innovation\_measure_{i,t+1\to t+3}$ is a measure of innovation for firm i that is aggregated over the subsequent three years. Specifically, we employ two such future indicators of innovative output: logged patent counts summed over years $t+1, t+2$ and $t+3$ ($\log(1+Patents_{t+1\to t+3})$) and logged citations-per-patent over years $t+1, t+2$ and $t+3$ ($\log(1+\frac{Citations_{t+1\to t+3}}{Patents_{t+1\to t+3}})$). Relatedly, the dependent variable $current\_innovation\_measure_{it}$ is a measure of innovation for firm i that is observed on date t. We employ two current measures of innovation: logged market value of patents in year t using the Kogan et al. (2017) measure and the logged number of product introductions in year t taken from Mukherjee et al. (2017).

The coefficient of interest is β_1, which indicates how greater text-based innovation associates with other measures of innovation. To focus on within-industry variation, our base specification includes industry (SIC4) and year fixed effects (γ_t and ξ_s), but we also include firm fixed effects (ξ_i) and SIC2-year fixed effects in some specifications to understand whether the relation between text-based innovation and other innovation measures is driven by within-firm variation or across-firm variation. To account for correlated errors, the specifications cluster standard errors by firm.

Table 12 presents results from estimating Equation (2) for the patent counts and citations-per-patent over the subsequent three years. The relation between text-based innovation and future patent counts is inconsistent in sign across specifications and is only statistically significant in column (3), when it is negative. However, there is a robust positive relationship between text-based innovation and citations-per-patent.

Table 12. Text-Based Innovation vs. Other Aspects of Innovation (1990–2010): Text-Based Innovation, Patents, and Citation Impact

	Dependent variable							
	$\text{Log}(1 + Patents_{t+1\to t+3})$				$\text{Log}(1 + \frac{Citations_{t+1\to t+3}}{Patents_{t+1\to t+3}})$			
Explanatory variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Text Innovation$_t$	0.053	0.026	−0.070**	−0.016	0.098***	0.062***	0.039***	0.024*
	(0.051)	(0.020)	(0.034)	(0.024)	(0.017)	(0.013)	(0.014)	(0.013)
Log(1 + *Patents*)$_t$		0.634***		0.416***		−0.287***		−0.165***
		(0.048)		(0.063)		(0.031)		(0.037)
Log(1 + *Citations*)$_t$		0.279***		0.254***		0.279***		0.125***
		(0.032)		(0.038)		(0.028)		(0.022)
Other controls		X		X		X		X
Industry (SIC4) FE	X	X			X	X		
Firm FE			X	X			X	X
Year FE	X	X			X	X		
SIC2-Year FE			X	X			X	X
Observations	4,264	4,264	4,264	4,264	3,208	3,208	3,208	3,208
Adjusted R^2	0.604	0.883	0.857	0.908	0.802	0.848	0.910	0.918

Notes. This table presents output from ordinary least squares regressions that link our text-based innovation measure to patent counts, citation impact, and patenting value. To focus on the within-patenting properties of the innovation measure, the sample is restricted to patenting firms. In this table, the dependent variables we consider are logged patent counts over the following three years ($t + 1$ to $t + 3$), $\text{Log}(1 + \text{Patents}_{t+1\to t+3})$, and logged citation impact of patents over the following three years, $\text{Log}(1 + \frac{\text{Citations}_{t+1\to t+3}}{\text{Patents}_{t+1\to t+3}})$. Controls include other innovation measures—log(*patents*), log(*citations*), an indicator for patenting firm, *R&D intensity*, *date t values* of log(*assets*), *asset tangibility*, *leverage*, log(*age*), and *cash/assets*. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

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

Using within-industry variation (columns (5) and (6)), a standard-deviation increase in text-based innovation is associated with 6.2%–9.8% more citation impact, an effect that is statistically significant at the 1% level. The relation remains statistically significant when we also include firm fixed effects and SIC2-year fixed effects, though the magnitude of the estimate identified from this within-firm and within-industry-year variation is smaller.[23]

Table 13 presents results from estimating Equation (3) for the contemporaneous values of Kogan et al. (2017) patent value and Mukherjee et al. (2017) product introductions. In contrast to the extent of future patenting, we find a positive relation between text-based innovation and both patent value and product introductions, though this relation is more robust for patent value than for product introductions. In the patent-value specifications with firm characteristic controls (columns (2) and (4)), a standard-deviation increase in text-based innovation at date t is associated with an increase of 6.4%–9.0% of the value of the firm's patents during year t. Similarly, for the product-introductions specifications that control for firm characteristics (columns (6) and (8)), a standard-deviation increase in text-based innovation is associated with approximately 2.5%–4.4% more product introductions. When focusing on within-industry variation (i.e., the specifications with SIC4 fixed effects in columns (1), (2), (5), and (6)), the estimate is statistically significant. However, the relation between text-based innovation and product introductions is not robust to including industry-year fixed effects, though we obtain a positive point estimate across specifications.[24]

Taken together, the results in this section indicate that our measure contributes valuable information about the quality of innovation, even within the set of firms that use patents to protect their innovations. Although the text-based innovation measure is not robustly related to future patent counts, it is strongly correlated with the most valuable patents, and it is positively and significantly related to the citation impact of future innovations. Moreover, the text-based innovation measure can be computed by using analyst reports in real time, whereas patenting outcomes take longer (e.g., even counts of applications for eventually granted patents must wait for the patent to be granted or denied). Thus, our text-based measure is useful in providing a leading indicator for innovation, which contrasts with patenting outcomes that take time to observe. Finally, in showing a generally positive relation between text-based innovation and product introductions, we provide useful evidence that captures innovative activities that are not well spanned by patenting measures.

4.3.1. Contextual Examples of Systems Innovation in Mature Firms. As a complement to our regression evidence above, it is useful to examine the content of

Table 13. Text-Based Innovation vs. Other Aspects of Innovation (1990–2010): Text-Based Innovation, Patent Value, and Product Announcements

	Dependent variable							
	Log(1 + *Patent Value*)$_t$				Log(1 + *Products*)$_t$			
Explanatory variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Text Innovation$_t$	0.171***	0.064***	0.016	0.090***	0.100***	0.044*	0.015	0.025
	(0.061)	(0.023)	(0.041)	(0.025)	(0.032)	(0.026)	(0.033)	(0.031)
Log(1 + *Patents*)$_t$		0.417***		0.565***		0.033		−0.032
		(0.067)		(0.094)		(0.031)		(0.042)
Log(1 + *Citations*)$_t$		0.668***		0.637***		0.009		0.034
		(0.039)		(0.054)		(0.019)		(0.022)
Other controls		X		X		X		X
Industry (SIC4) FE	X	X			X	X		
Firm FE			X	X			X	X
Year FE	X	X			X	X		
SIC2-Year FE			X	X			X	X
Observations	3,586	3,586	3,586	3,586	1,715	1,715	1,715	1,715
Adjusted R^2	0.573	0.915	0.832	0.945	0.435	0.537	0.656	0.663

Notes. This table presents output from ordinary least squares regressions that link our text-based innovation measure to patent counts, citation impact, and patenting value. To focus on the within-patenting properties of the innovation measure, the sample is restricted to patenting firms. In this table, the dependent variable is the Kogan et al. (2017) measure of market value of patents (i.e., the stock market jump on the day of the granted patent in $millions) aggregated over all patents granted during the year in columns (1)–(4). The dependent variable in columns (5)–(8) is the log of the number of product announcements when the stock market return was above the 75th percentile from Mukherjee et al. (2017). Controls include other innovation measures—log(*patents*), log(*citations*), an indicator for patenting firm, *R&D intensity*, *date t values* of log(*assets*), *asset tangibility*, *leverage*, log(*age*), and *cash/assets*. Standard errors that are clustered by firm are reported in parentheses. FE, fixed effects.

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

valuable innovations, which relate strongly to our text-based innovation measure. Figure 9 presents a list of valuable patents in order of value starting at the 95th percentile of patent values. Most of these highly valuable patented innovations are not particular to a specific product, but, rather, reflect a valuable component or the patenting of a valuable process. In fact, only one patent in this list is directly related to a specific product—a vaccine. Other patents are processes, components that can go into one or several products, or components useful in the production process.

Taking a step outside of the universe of patenting firms, we turn our attention to the retail sector in 1993, which our measure indicates as highly innovative, but, nonetheless, is a low-patenting industry at the time. Figure 1 presents two excerpts from analyst reports of firms that are considered particularly innovative. These are firms that do not rely heavily on patents but are considered innovative by the analyst. Consistent with our interpretation that the innovation we measure reflects innovative systems, the reports describe the firms as innovative in ways that are separate from bringing new products to market. For example, the analyst report about Walmart describes how Walmart "uses technology to improve productivity and at the same time reduce costs." The report describes several dimensions along which Walmart is innovative and is an industry leader, in the way that they use technology in their supply chain management and theft prevention. Because these innovations were not discovered using R&D expenditures and were not patented, our measure is in a unique position to capture this type of innovation, which is a common for firms like Walmart that have particularly innovative systems.

5. Illustrative Application

This section provides an illustrative application of the text-based innovation measure, in which we replicate and extend a recent finding in the innovation literature: the finding by Custódio et al. (2019) that generalist CEOs (i.e., those with more diverse managing experience) generate more patenting innovation.

First, we conceptually replicate the main result from Custódio et al. (2019) in our sample of S&P 500 firms. Custódio et al. (2019) showed that generalist managers innovate more, as measured by both patent

Figure 9. Valuable Patents (95th Percentile)

Firm	Patent	Date	Title	Abstract
WASTE MANAGEMENT	4,927,317	1990-05-22	Apparatus for temporarily covering a large land area	A method for temporarily covering a large land area and an apparatus for suspending a flexible cover from a front loader bucket of an earth-moving vehicle.
COMPAQ	5,454,081	1995-09-26	Expansion bus type determination apparatus	A circuit that automatically detects whether an input/output expansion board is connected to an EISA system or an ISA system.
TEXACO	5,644,244	1997-07-01	Method for analyzing a petroleum stream	Methods are provided for determining a solids to liquids ratio in a flowing petroleum stream having an immiscible solids, oil and water flow.
3COM CORP	5,651,002	1997-07-22	Internetworking device with enhanced packet header translation and memory	An internetworking device providing enhanced packet header translation for translating the format of a header associated with a source network into a header format associated with a destination network of a different type than the source network.
ERICSSON	5,706,301	1998-01-06	Laser wavelength control system	A laser wavelength control system (20) stabilizes laser output wavelength. The control system includes a reflector/filter device (40) upon which laser radiation is incident for yielding both a filtered-transmitted signal (FS) and a reflected signal (RS).
HALLIBURTON	5,716,910	1998-02-10	Foamable drilling fluid and methods of use in well drilling operations	A foamable drilling fluid for use in well operations such as deep water offshore drilling where risers are not employed in returning the fluid to the surface mud pit.
ELECTRONIC DATA SYSTEMS	5,801,366	1998-09-01	Automated system and method for point-of-sale (POS) check processing	An automated check processing system includes an input device receiving checking account information and a check amount of a check provided for payment in a translation.
LILLY (ELI)	7,138,521	2006-11-21	Crystalline of N-[4-[2-(2-Amino-4,7-dihydro-4oxo-3H-pyrrolo[2,3-D]pyrimidin-5-YL)ethyl]benzoyl]-L-glutamic acid	The invention relates to the field of pharmaceutical and organic chemistry and provides an improved process for preparing the novel heptahydrate crystalline salt of multitargeted antifolate N-[4-[2-(2-amino-4,7-dihydro-4-oxo-3H-pyrrolo[2,3-d]-pyrimidin-5-yl)ethyl]benzoyl]-L-glutamic acid.
FEDEX	7,429,057	2008-09-30	Lifting systems and methods for use with a hitch mechanism	A lifting system for a hitch mechanism is provided.
BRISTOL-MYERS SQUIBB	7,825,097	2010-11-02	Nucleotide vector vaccine for immunization against hepatitis	Nucleotide vector comprising at least one gene or one complementary DNA coding for at least a portion of a virus, and a promoter providing for the expression of such gene in muscle cells.

Notes. This is a list of patents on the 95th percentile of patent values ($80 million). Observations with only one patent grant during the day are shown. EISA, extended industry standard architecture; ISA, industry standard architecture.

counts and citations. Specifically, their base specification is of the form:

$$patenting\_outcome_{it} = \gamma_{st} + \beta_1 General\ Ability\ Index_{it} + \mathbf{X}'\gamma + \varepsilon_{it}, \quad (4)$$

where $patenting\_outcome_{it}$ is either the count of patents filed in year t by firm i or the number of patent citations for patents filed in year t by firm i. The key explanatory variable of interest is $General\ Ability\ Index_{it}$, which measures the diversity of managing experience for the CEO of firm i at date t.[25] For this illustrative application, we conceptually replicate the specification in Custódio et al. (2019) that employs firm and industry-year fixed effects. A notable difference is that our sample contains S&P 500 firms, whereas Custódio et al. (2019) study a broader set of firms, those in the S&P 1500. After merging with their measure, we obtain 3,860 firm-year observations in comparison with 8,297 in Custódio et al. (2019). Given these differences, our objective is not to replicate exactly the coefficient estimates from their paper, but, rather, to present their results in our sample of S&P 500 firms. To maintain consistency with the analysis in the rest of our paper, all specifications include our standard set of control variables.

The first four columns of Table 14 present the conceptual replication results for patent counts and patent citations as dependent variable. In the specification in column (1), the estimate on the *General Ability Index* is statistically significant at the 10% level, and the magnitude is statistically indistinguishable from the published estimate from Custódio et al. (2019). When we include the standard set of controls in column (2), the coefficient estimate on *General Ability Index* is marginally statistically insignificant (p-value of 0.14). Although the magnitude of 0.040 is slightly smaller in magnitude than the comparable estimate from Custódio et al. (2019) of 0.073, their published estimate is within a 95% confidence interval constructed around our estimate $(-0.015, 0.095)$. Therefore, we cannot reject the null hypothesis that the published result is different from what we obtain, despite finding a statistically insignificant result. Similarly, we obtain statistically insignificant estimates ranging from 0.063 to 0.080 in the patent-citations regressions, which are quantitatively similar to the

Table 14. Replication and Extension of Custódio et al. (2019)

	Dependent variable					
	Log Patents$_t$		Log Citations$_t$		*Text-Based* Innovation$_t$	
	(1)	(2)	(3)	(4)	(5)	(6)
General Ability Index	0.055*	0.040	0.080	0.063	−0.075***	−0.072**
	(0.030)	(0.028)	(0.051)	(0.049)	(0.028)	(0.029)
Other controls		X		X		X
Firm FE	X	X	X	X	X	X
SIC2-Year FE	X	X	X	X	X	X
Observations	3,860	3,860	3,860	3,860	3,860	3,860
Adjusted R^2	0.935	0.938	0.908	0.909	0.637	0.638

Notes. This table presents ordinary least squares regressions of innovation measures on the general ability index from Custódio et al. (2013). For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Consistent with the main result in Custódio et al. (2019), table 3, column (3), we include SIC2-year and firm fixed effects (FE) in all specifications. Other controls include log(*assets*), *asset tangibility*, *leverage*, log(*age*), *cash/assets*, and an indicator for whether the firm is a patenting firm. Variable definitions are presented in Table A.1 in the online appendix. Standard errors that are clustered by firm are reported in parentheses.

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

comparable specifications in Custódio et al. (2019), who report a coefficient estimate of 0.062 in their comparable specification.

In contrast to the results on patenting outcomes, we estimate a *negative* and statistically significant relationship between general managerial ability and text-based innovation (columns (5) and (6)).[26] Beyond providing additional evidence that text-based innovation captures a distinct aspect of innovation from patenting, this finding suggests that additional patenting innovation from generalist managers (noted in Custódio et al. 2019) may reflect a tradeoff away from other types of innovation that are captured by our text-based measure. The generally lower level of text-based innovation in firms with generalist CEOs is important to consider, especially given the robust link between text-based innovation and firm performance that we have shown in our main tests.

6. Conclusions

In this paper, we develop a useful new measure of corporate innovation based on a textual analysis of analyst reports. Our text-based innovation measure provides a useful description of innovation in firms without patents and with zero R&D expenditure. Such firms are common: Even among our sample of 703 firms from the S&P 500, there are 219 firms with no patents and 329 firms that had zero R&D expenditure for our entire sample period (1990–2012). Moreover, there is a substantial overlap between the distribution of innovation for patenting firms and the distribution of innovation for nonpatenting firms (similarly for R&D versus zero-R&D), which indicates that important innovative activities are overlooked by using patenting and R&D as proxies for innovation. Indeed, this view is confirmed by notable examples of firms that do not patent or use R&D but are nonetheless identified as highly innovative by our measure (e.g., Walmart).

Beyond expanding the sample of innovative firms to study, our textual analysis provides a useful step toward understanding innovation in the spirit of Schumpeter (1934), who described five types of innovation: new products, new methods of production, new sources of supply, exploitation of new markets, and new ways to organize business. Patenting and R&D expenditure typically pertain to product innovation, and the literature's focus on these measures has left the other categories understudied. To take one example of how adopting this broader view (and measurement) of innovation is useful, recent research by Frésard et al. (2020) argues that firms with realized innovations are more likely to be acquired in a vertical merger because realized innovations are easier to commercialize than innovations in progress. The authors use patenting outcomes to proxy for realized innovation, and, thus, their focus is primarily on innovation and commercialization of products. As our analysis shows, the text-based innovation measure captures important innovative activity in business systems. This mode of innovation likely exhibits a different relation to corporate outcomes that have been linked to product innovations (indeed, we illustrate this feature of our measure in the context of a conceptual replication and extension of Custódio et al. 2019). In this vein, future research could use textual measures of innovation to examine the extent to which the lessons learned from studying product innovations translate into other types of corporate innovation.

Finally, although our analysis is applied to the text of analyst reports, our textual approach could be applied to other settings to identify complementary measures of innovation. Media articles, required firm

disclosures (10Ks), and press releases may also contain information about firms' innovative activities. Recent work has considered some of these textual databases as a source of information on corporate innovation (e.g., see the analysis of product innovation in Mukherjee et al. 2017 using press releases), but given the available wealth of textual sources of information about firms, much more progress is possible. Our text-based innovation measure suggests that examining these sources of textual information about firms is fertile ground for future research.

Acknowledgments

This draft has benefited from helpful comments from two anonymous referees, an anonymous associate editor, Gustavo Manso (the editor), Jamie Brown, Casey Dougal (discussant), Umit Gurun, Jerry Hoberg (discussant), Ryan Israelsen, Ross Levine, William Mann, Song Ma (discussant), Katie Moon, Jillian Popadak, Dimitris Papanikolaou (discussant), Tomas Thornquist (discussant), Ting Xu (discussant), Farzad Saidi, Ed Van Wesep, Brian Wolfe (discussant), and Jaime Zender; as well as the conference and seminar participants at Babson College, Boston University, the 2016 European Summer Symposium for Financial Markets (evening session), the 2016 Front Range Finance Seminar, Iowa State University, the 2016 Instituto Tecnológico Autónomo de México Finance Conference, the 2017 Midwest Finance Association Conference, the 2017 National Bureau of Economic Research Corporate Finance Spring Meeting, the 2017 University of Kentucky Finance Conference, the 2017 Western Finance Association Conference, the 2017 European Finance Association Conference, the 2016 Northern Finance Association Conference, the University of Exeter, the University of New South Wales, the University of Sydney, and the University of Colorado finance brownbag. The authors thank Alminas Zaldokas for generously sharing data on product announcements. All remaining errors are the authors' responsibility.

Endnotes

[1] As a measure of innovation, patents have a number of additional well-known weaknesses. For example, not all innovations are put under patent protection or can be put under patent protection (Moser 2012, Hall et al. 2014), and some patents are filed for defensive reasons (e.g., see work on "patent trolls" by Tucker 2014 and Cohen et al. 2019). In this vein, Saidi and Zaldokas (2020) provide evidence that patenting and trade secrets are substitutes depending on disclosure requirements for patenting, which indicates that a significant amount of innovation is not patented.

[2] In another test related to firm-specific incentives to strategically disclose innovation activities, we instrument for firm innovation disclosures using industry-rival disclosures, which are not subject to the same disclosure incentives. In these specifications, we find similar results to our main findings, which helps to alleviate concerns about strategic disclosures. See Table A.14 in the online appendix for these tests.

[3] Even related work on innovation using text analysis has not constructed a similar measure of innovation. Specifically, Frésard et al. (2020) study how innovation and vertical integration relate to one another while making use of text analysis, but the text-analysis component of their work is confined to vertical relatedness rather than innovation. Their innovative outcomes are the more standard R&D intensity and patenting outcomes from the literature.

[4] The authors thank Bill McDonald for making these lists available on his website: https://sraf.nd.edu.

[5] LDA has a number of advantages over naive word-list techniques (e.g., Loughran and McDonald 2011). For our purposes, the most important advantage is that LDA accurately reflects context of the word usage, whereas a naive word-list textual analysis does not. As we show in Tables A.21 and A.22 in the online appendix, the word-list measure delivers qualitatively similar conclusions, but exhibits slightly weaker valuation implications and is not as robustly related to valuable patents as the more accurate LDA-based measure. This is consistent with the LDA methodology more accurately accounting for the context of innovative language.

[6] Asquith et al. (2005) hand-classify a limited sample of analyst reports into various categories and show that some categories have investment value. More recently, authors have worked on parsing the text of analyst reports in a more systematic fashion. Using a sample of initiation reports, Twedt and Rees (2012) show that, controlling for recommendation changes and other factors, the tone of reports has an associated stock market reaction. Using a large sample of analyst reports, Huang et al. (2014) find a stock market reaction of between 1.5% and 3.5% (two-day cumulative abnormal return) for reports in the top quintile of analyst tone relative to those in the bottom quintile. They also show that the tone of more qualitative topics (those with few uses of "$" or "%") is more important, which suggests that qualitative aspects of the analyst text are a valuable source of new information.

[7] Related to the point of strategic disclosures, analysts may be less informed about the firm's innovation activities than the firm's managers. This increases the noise in our innovation text measure and biases the coefficients of our innovation text measure (in the regressions) toward zero. To the extent that we find our innovation text measure as statistically significant, it would be even more so if we could reduce this source of noise.

[8] Relative to the patenting measures of innovation, one notable limitation of the text-based measure is that it is observed at the firm-year level, which prevents within-firm, cross-sector analyses of innovation. From this standpoint, the text-based measure of innovation would not be useful to describe product-market positioning or evaluate the determinants of innovation-level valuation. These potential concerns apply to other widely used text-based measures of product competition (e.g., Hoberg et al. 2014 and Hoberg and Phillips 2016), and they are consistent with our interpretation of the text-based innovation measure as a measure of systems innovation.

[9] We experimented with other numbers of topics. Fitted LDA models with fewer topics tended to work similarly well (the model with K = 10 delivers all of the quantitative insights we report in the main text; see Table 4B), whereas models fit with a greater prespecified number of topics exhibit redundancy (i.e., multiple topics about the same essential idea; see Figure A.3 in the online appendix for word clouds of two similar innovation topics from a model with K = 50 topics). Although the number of topics is the only degree of freedom we have in fitting an LDA model, the extensive literature on LDA does not offer standardized guidance on how to select the appropriate number of topics because the appropriate number of topics depends on the application. Some applications of LDA have optimized an objective function to obtain an optimal number of topics in their context. For example, Goldsmith-Pinkham et al. (2016) maximize saliency of topics from one another, and other authors have estimated Hierarchical Dirchlet Process models (HDP-LDA), which obtains a likelihood-maximizing number of topics (Teh et al. 2006). Our objective is to select the number of topics to capture a general notion of innovation to apply across different contexts. Automated routines that seek to maximize a likelihood function will tend to overfit by selecting a larger number of topics that adapt to different contexts. Thus, automated routines will tend to lead to topics that are too granular to capture a broad notion of innovation.

[10] The textbook we use for this validation exercise is *Managing Innovation* by Tidd et al. (2005), a widely adopted innovation textbook that was available in PDF format via Google search. The readable PDF format was useful to produce a distribution of words used to describe innovation. One potential disadvantage of this benchmark textbook is that it was published during the sample period and could influence or be influenced by the innovative activities in our sample. To alleviate this concern, we consider an alternative to this benchmark textbook by processing a textbook published prior to our sample period, *Innovation and Entrepreneurship*, by Drucker (1985), as robustness. As we show in the online appendix (Figure A.2), our selection of the innovation topic is not sensitive to the choice of the benchmark.

[11] As a complement to this main measure, we also construct a negative text-based innovation measure that is constructed by using the subset of reports that have negative sentiment. In contrast to our main measure, negative text-based innovation loads negatively on firm performance and exhibits a weak relation to other innovation outcomes. This measure provides useful and distinct information about innovative failures, which we highlight in the context of a replication and extension of Custódio et al. (2019) in Section 5.

[12] The top three innovation firm-years among nonpatenting firms in our sample highlight the ability of our measure to identify overlooked high-innovation firms. First, in 1996, Shared Medical Systems Corporation produced information-processing systems for the healthcare industry at a time when Internet technology was emerging, but was a nonpatenting firm. Second, in 2000, BroadVision was a nonpatenting firm that was a software vendor for web applications that enhanced internal management systems of firms (HR, sales processing, online shopping, etc.). Finally, in 1994, Alltel Wireless was a wireless service provider that developed a large network of subscribers across much of the United States by adopting network technology manufactured by Lucent, Motorola, Nortel, Cisco, and Juniper Networks.

[13] The innovation topic and patenting outcomes have a strong correlation within the set of patenting firms. Specifically, we find that the innovation topic exhibits a stronger correlation to patenting than any of the other topics from the LDA. The statistical significance of the relation between our innovation topic and patenting is present, even after taking into account the multiple-comparisons problem of searching over 15 topics. Indeed, the test statistic in a linear regression is $t = 12.37$, which far exceeds rule-of-thumb adjustments to critical values (Harvey et al. 2016), and the statistical significance survives other more formal, multiple-comparisons adjustments (e.g., the Bonferroni correction). As we describe in the online appendix (Table A.2), this topic explains nearly two times the variation of any other set of topic loadings among the 15 fitted LDA topics.

[14] The table presents comparisons of other characteristics as well, which are consistent with intuition about R&D and patenting. For example, there is a strong correlation between patenting and R&D expenditures. Both patenting and R&D firms have lower asset tangibility and lower leverage. In addition, R&D firms tend to be younger than non-R&D firms, and firms with patents tend to be older.

[15] Throughout our empirical analysis, we use two-digit SIC industries when we construct industry-year fixed effects. In our sample, there are 223 distinct SIC4 industries across 21 years, which would imply nearly as many SIC4-year fixed effects (4,683) as observations (~6,000). We note that the recent study of Custódio et al. (2019) employs a similar strategy to account for industry-year variations using SIC2-year fixed effects.

[16] Our results are nearly identical when double clustering by firm and year (see Table A.11 in the online appendix), but we adopt the one-way clustering to avoid the concern that the year dimension of our data set has only 21 clusters, and, thus, the asymptotic approximation to the variance–covariance matrix may not be valid.

[17] For these interactive specifications, we drop the patenting controls because patent counts and citations are equal to zero for nonpatenting firms, meaning that the interaction with these terms included could mechanically capture a slope effect of the amount of patenting rather than the raw difference between patenting and nonpatenting firms. The specification that includes these controls delivers nearly identical results (see Table A.9 in the online appendix). We thank an anonymous referee for identifying this specification issue.

[18] In the online appendix (Table A.7), we present a complementary exercise in which we interact text-based innovation with a continuous measure of patenting intensity, estimated on the subset of patenting firms. This specification tests whether the performance implications are different for firms with high versus low patenting intensity. We find that text-based innovation exhibits a similar-magnitude relation to performance for firms with high versus low patenting intensity, similar to our main comparison between patenting and nonpatenting firms. Relatedly, in Table A.8 in the online appendix, we re-estimate the main noninteractive performance specification, but *without* controlling for the patenting measures. Regardless of the controls employed, our findings are virtually unchanged, which alleviates the concern that the significant relation between text-based innovation and performance is an artifact of how we account for patenting outcomes as control variables.

[19] These specifications also include four lags of patent counts and R&D intensity to address the separate empirical concern that patenting and R&D outcomes influence text-based innovation. To this point, we find that patenting does not predict text-based innovation at any lag (conditional on fixed effects and controls), suggesting that our findings are not driven by a lagged correlation with patenting outcomes. We also find that R&D exhibits a robust relation to text-based innovation at a two-year lag (but not significant at other lags), consistent with the notion that R&D investments take time to materialize. See Table A.20 in the online appendix for details on these full results.

[20] In addition, we perform several additional robustness exercises related to selecting the innovation topic, which are reported in the online appendix. Figure A.3 and Table A.5 in the online appendix present the results from an analysis of a 50-topic LDA. In addition, we conduct three robustness exercises that pertain to how we select the innovation topic. First, in Table A.4 in the online appendix, we present evidence that the topic most correlated with patenting (the "Patenting Topic") exhibits a stable and high rank in terms of its correlation with the words in the *Managing Innovation* textbook (Tidd et al. 2005) we use as a benchmark. Second, we parse the text of an alternative innovation textbook, *Innovation and Entrepreneurship* by Drucker (1985), which was published before our sample period. Figure A.2 in the online appendix shows that our choice of the innovation topic is the same if we employ this benchmark. Third, in the performance regressions, we present an interaction between our text-based innovation measure and a postpublication indicator for *Managing Innovation* (Tidd et al. 2005). In these specifications, we find similar results before versus after the innovation book's publication (Table A.15 in the online appendix). Collectively, these tests help to alleviate the concern that the innovation textbook is influenced by innovations specific to this time period or that the innovation book influences the language that analysts use.

[21] Although this conclusion is sensible, it is important to emphasize that this measure does not merely reflect low sentiment about the firm, but, rather, low sentiment by analysts who write intensively about innovation. Indeed, the main (positive) innovation measure produces a robust positive relation to future firm performance, even conditioning on low average sentiment of analysts (see Table A.12 in the online appendix).

[22] A rationale for this pattern of results is that patenting firms, by virtue of engaging in a public innovation process, are less opaque with respect to their innovative failures. For this set of firms, analysts

are better able to provide insight into the nature of innovative failures. For firms that do not patent, there is additional noise in the measure. With this feature of the measure in mind, we expect that some of the more promising applications of the negative text-based innovation measure are to the study of firms that have patents.

[23] Because we control for current logged patents and logged citations in the specification in column (8) with firm fixed effects, it is technically inappropriate to interpret this coefficient estimate because this is a version of controlling for a lagged dependent variable in a panel data context. Such a panel regression model with feedback violates the strict exogeneity condition that is required for fixed-effects estimation to be consistent (e.g., see Wooldridge 2003). On this basis, we place greater weight on the result in column (7). We choose to report this specification, and the specification column (4), which exhibits the same empirical problem, because we wish to maintain a consistent structure for the empirical tests in the paper.

[24] The nonsignificant result in columns (7) and (8) is due to the inclusion of industry-year fixed effects. In unreported results (not reported for brevity), we find a statistically significant relation between text-based innovation and product innovation in a regression with firm fixed effects. In addition, the online appendix presents two additional results related to how text-based innovation relates to other innovation measures. First, we estimate the relation between text-based innovation and future patent value and future product introductions (one year ahead) in Table A.17 in the online appendix. These results indicate a similar positive relation, albeit slightly weaker than the contemporaneous relation we describe here. Second, we estimate the relation between negative text-based innovation and these other aspects of innovation in Table A.18 in the online appendix. Consistent with the negative measure capturing the lack of innovation success, we find that negative text-based innovation is generally insignificantly related to innovation quality, particularly for specifications that include firm characteristics.

[25] This index, which was developed by Custódio et al. (2013), is equal to the first principal component of five characteristics of the CEOs experience profile: (1) the past number of positions, (2) the past number of firms, (3) the past number of industries in which he worked, (4) whether he held a CEO position at a different company, and (5) whether he worked for a conglomerate firm.

[26] Apart from being negative and statistically significant, the coefficient on the *General Ability Index* is significantly lower for text-based innovation than for patents, with a t-statistic of −3.62 using a bootstrapped standard error for the difference-in-coefficient test.

References

Agarwal S, Gupta S, Israelsen RD (2016) Public and private information: Firm disclosure, sec letters, and the JOBS Act. Working paper, National University of Singapore, Singapore.

Asquith P, Mikhail MB, Au AS (2005) Information content of equity analyst reports. *J. Financial Econom.* 75(2):245–282.

Bodnaruk A, Loughran T, McDonald B (2015) Using 10-K text to gauge financial constraints. *J. Financial Quant. Anal.* 50(4):623–646.

Boudoukh J, Feldman R, Kogan S, Richardson M (2013) Which news moves stock prices? A textual analysis. NBER Working Paper 18725, National Bureau of Economic Research, Cambridge, MA.

Brown JR, Fazzari SM, Petersen BC (2009) Financing innovation and growth: Cash flow, external equity and the 1990s R&D boom. *J. Finance* 64(1):151–185.

Cohen L, Frazzini A (2008) Economic links and predictable returns. *J. Finance* 63(4):1977–2011.

Cohen L, Diether K, Malloy C (2013) Misvaluing innovation. *Rev. Financial Stud.* 26(3):635–666.

Cohen L, Gurun U, Kominers SD (2019) Patent trolls: Evidence from targeted firms. *Management Sci.* 65(12):5449–5956.

Cohen L, Malloy C, Nguyen QH (2020) Lazy prices. *J. Finance* 75(3):1371–1415.

Custódio C, Ferreira MA, Matos P (2013) Generalists vs. specialists: Lifetime work experience and CEO pay. *J. Financial Econom.* 108(2):471–492.

Custódio C, Ferreira MA, Matos P (2019) Do general managerial skills spur innovation? *Management Sci.* 65(2):459–476.

Dougal C, Engelberg J, Garcia D, Parsons CA (2012) Journalists and the stock market. *Rev. Financial Stud.* 25(3):639–679.

Drucker P (1985) *Innovation and Entrepreneurship: Practice and Principles* (Butterworth Heinemann, Boston).

Edmans A, Garcia D, Norli Ø (2007) Sports sentiment and stock returns. *J. Finance* 62(4):1967–1998.

Frésard L, Hoberg G, Phillips GM (2020) Innovation activities and the incentives for vertical acquisitions and integration. *Rev. Financial Stud*. Forthcoming.

Galasso A, Schankerman M (2015) Patents rights and innovation by small and large firms. Working paper, University of Toronto, Toronto.

Ganglmair B, Wardlaw M (2017) Complexity, standardization, and the design of loan agreements. Working paper, ZEW–Leibniz Centre for European Economic Research, Mannheim, Germany.

Garcia D (2013) Sentiment during recessions. *J. Finance* 68(3):1267–1300.

Goldsmith-Pinkham P, Hirtle B, Lucca D (2016) Parsing the content of bank supervision. FRBNY Staff Report 770, Federal Reserve Bank of New York, New York.

Grennan JA (2013) A corporate culture channel: How increased shareholder governance reduces firm value. Working paper, Duke University, Durham, NC.

Hall B, Helmers C, Rogers M, Sena V (2014) The choice between formal and informal intellectual property: A review. *J. Econom. Lit.* 52(2):375–423.

Hall BH (1990) The impact of corporate restructuring on industrial research and development. *Brookings Papers on Economic Activity: Microeconomics* 1990:85–135.

Hall BH, Mairesse J, Mohnen P (2010) Measuring the returns to R&D. Hall BH, Rosenberg N, eds. *Handbook of the Economics of Innovation*, vol. 2 (Elsevier, Amsterdam), 1033–1082.

Hanley KW, Hoberg G (2010) The information content of IPO prospectuses. *Rev. Financial Stud.* 23(7):2821–2864.

Harvey CR, Liu Y, Zhu H (2016) . . .And the cross-section of expected returns. *Rev. Financial Stud.* 29(1):5–68.

He J, Tian X (2013) The dark side of analyst coverage: The case of innovation. *J. Financial Econom.* 109(3):856–878.

Hoberg G, Lewis C (2017) Do fraudulent firms produce abnormal disclosure? *J. Corporate Finance* 43:58–85.

Hoberg G, Maksimovic V (2015) Redefining financial constraints: A text-based analysis. *Rev. Financial Stud.* 28(5):1312–1352.

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

Hoberg G, Phillips G, Prabhala N (2014) Product market threats, payouts, and financial flexibility. *J. Finance* 69(1):293–324.

Huang A, Zang A, Zheng R (2014) Evidence on the information content of text in analyst reports. *Accounting Rev.* 89(6):2151–2180.

Huang A, Lehavy R, Zang A, Zheng R (2015) Analyst information discovery and interpretation roles: A topic modeling approach. Working paper, Hong Kong University of Science and Technology, Hong Kong.

Israelsen RD (2014) Tell it like it is: Disclosed risks and factor portfolios. Working paper, Michigan State University, East Lansing, MI.

Jegadeesh N, Wu D (2017) Deciphering Fedspeak: The information content of FOMC meetings. Working paper, Emory University, Atlanta.

Knott AM (2008) R&D/returns causality: Absorptive capacity or organizational IQ. *Manage. Sci.* 54(12):2054–2067.

Kogan L, Papanikolaou D, Seru A, Stoffman N (2017) Technological innovation, resource allocation, and growth. *Quart. J. Econom.* 132(2):665–712.

Kuznets S, Murphy JT (1966) *Modern Economic Growth: Rate, Structure, and Spread*, vol. 2 (Yale University Press, New Haven, CT).

Loh RK, Mian GM (2006) Do accurate earnings forecasts facilitate superior investment recommendations? *J. Financial Econom.* 80(2): 455–483.

Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66(1):35–65.

Lowry M, Michaely R, Volkova E (2020) Information revelation through regulatory process: Interactions between the SEC and companies ahead of the IPO. *Rev. Financial Stud.* Forthcoming.

Mann W (2018) Creditor rights and innovation: Evidence from patent collateral. *J. Financial Econom.* 130(1):25–47.

Moser P (2012) Innovation without patents: Evidence from world's fairs. *J. Law Econom.* 55(1):43–74.

Mukherjee A, Singh M, Zaldokas A (2017) Do corporate taxes hinder innovation? *J. Financial Econom.* 124(1):195–221.

Muslu V, Radhakrishnan S, Subramanyam K, Lim D (2015) Forward-looking MD&A disclosures and the information environment. *Management Sci.* 61(5):931–948.

Nordhaus WD (1969) An economic theory of technological change. *Amer. Econom. Rev.* 59(2):18–28.

Saidi F, Zaldokas A (2020) Patents as substitutes for relationships. *Management Sci.* Forthcoming.

Schumpeter JA (1934) *The Theory of Economic Development: An Inquiry into Profits, Credit, Interest, and the Business Cycle* (Transaction Publishers, Piscataway, NJ).

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

Swem N (2014) Information in financial markets: Who gets it first? FEDS Working Paper No. 2017-023, Federal Reserve Board, Washington, DC.

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirchlet process. *J. Amer. Statist. Assoc.* 101(476):1566–1581.

Tian X (2012) The role of venture capital syndication in value creation for entrepreneurial firms. *Rev. Finance* 16(1):245–283.

Tian X, Wang TY (2014) Tolerance for failure and corporate innovation. *Rev. Financial Stud.* 27(1):211–255.

Tidd J, Bessant J, Pavitt K (2005) *Managing Innovation: Integrating Technological, Market and Organizational Change* (John Wiley & Sons, New York).

Trajtenberg M (1990) A penny for your quotes: Patent citations and the value of innovations. *RAND J. Econom.* 21(1):172–187.

Tucker C (2014) Patent trolls and technology diffusion: The case of medical imaging. Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Twedt B, Rees L (2012) Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *J. Accounting Public Policy* 31(1):1–21.

Welch I (2008) *Corporate Finance: An Introduction* (Prentice-Hall, Upper Saddle River, NJ).

Wooldridge JM (2003) *Introductory Econometrics: A Modern Approach*, 2nd ed. (Southwestern Publishing/Thompson Learning, Mason, OH).