Innovation in Mature Firms: A Text-Based Analysis

Gustaf Bellstam, Sanjai Bhagat and J. Anthony Cookson*

August 31, 2017

Abstract

We develop a new measure of innovation using a textual analysis of analyst reports. Our text-based measure gives a useful description of innovation by mature firms with and without patenting and R&D. For non-patenting firms, the measure identifies 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 non-patenting firms.

*Previous versions of the paper were circulated under its former title, "A Text-Based Analysis of Corporate Innovation." All authors are from University of Colorado's Leeds School of Business. Cookson is the corresponding author. University of Colorado at Boulder, Leeds School of Business, Campus Box 419, Boulder, CO 80309, USA, (tony.cookson@colorado.edu). Bellstam can be contacted at gustaf.bellstam@colorado.edu, and Bhagat can be contacted at sanjai.bhagat@colorado.edu. This draft has benefited from helpful comments from Jamie Brown, Casey Dougal (discussant), Umit Gurun, Jerry Hoberg (discussant), Ryan Israelsen, Ross Levine, William Mann, Song Ma (discussant), Katie Moon, Jillian Papadak, Dimitris Papanikolaou (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 ITAM Finance Conference, the 2017 Midwest Finance Association Conference, the 2017 NBER Corporate Finance Spring Meeting, the 2017 University of Kentucky Finance Conference, the 2017 Western Finance Association Conference, the 2016 Northern Finance Association Conference, University of Exeter, University of New South Wales, University of Sydney, and the University of Colorado finance brownbag. Thank you to Alminas Zaldokas for generously sharing data on product announcements. All remaining errors are the authors' responsibility.

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; Bhagat and Welch, 1995; Brown, Fazzari, and Petersen, 2009; Cohen, Diether, and Malloy, 2013). Nevertheless, our empirical understanding of innovation is incomplete because existing innovation proxies – typically, 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 proxies for innovation 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 – i.e., firms in the S&P500. 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

[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, Gurun, and Kominers, 2014). In this vein, Saidi and Zaldokas (2016) provide evidence that patenting and trade secrets are substitutes depending on disclosure requirements for patenting, which indicates a significant amount of innovation is not patented.

introduced to the finance literature (Israelsen, 2014; Goldsmith-Pinkham, Hirtle, and Lucca, 2016; Hoberg and Lewis, 2017; Lowry, Michaely, and Volkova, 2016). Specifically, we employ the Latent Dirichlet Allocation (LDA) method of Blei, Ng, and Jordan (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 does not require a pre-specified word list related to innovation, and it 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, which is the centerpiece of our innovation measure.

Our main measure is derived from a fitted LDA model that allows for 15 distinct topics to a corpus of 665,714 analyst reports of 703 firms that were in the S&P500 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. 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, 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.

For studying innovation in mature firms, 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, which provides a reliable basis for comparing mature firms' innovation to one another. Even within our sample of 703 firms from the S&P500, 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 non-patenting firms, we present tangible examples of content from analyst reports for non-patenting firms that score high on

our measure. One such example, which highlights the value of our approach is Walmart. Walmart did not use patent protection in the early 1990s, but it has always been innovative with respect to how it organizes its cross-geography logistics (e.g., placement of warehouses and shipping logistics between locations). Taking an excerpt from a May 1993 analyst report (more detail 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 of text-based innovation, higher 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 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 non-patenting firms, providing further evidence that our measure extends in a useful manner beyond the set of firms that use patenting and R&D.

Even among the set of patenting firms, the text-based innovation measure provides useful additional information on innovation. We find that our text-based measure strongly correlates with the Kogan

et al. (2017) patent valuation measure within the set of firms that patent. In this way, our text-based approach distinguishes true innovation captured by valuable patents from patenting outcomes that are not as valuable.

Because the text-based innovation measure applies across many contexts, the measure captures whether a company has an innovative system or platform. Indeed, the text of the topic does not reflect language surrounding specific products, but the systematic use of technology to enhance revenue and decrease costs. This idea of innovative systems has been conceptually identified as important (see Egan, 2013), but traditional measurements have not captured this idea quantitatively. Empirically, we find that innovative firms are more acquisitive, especially of smaller firms, which is consistent with the incentives of a firm with an innovative system to acquire smaller firms as components to their revenue-generating system.

Beyond studying innovation in mature firms, our approach of using text to study innovation has a number of notable advantages, both in describing the nature of innovation, but also in ascribing value to those innovations. First, our text-based measure allows inclusion and measurement of non-patented 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, Mairesse, and Mohnen (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; Cohen, Gurun, and Kominers, 2014; Mann, 2016). Because our measure does not rely on patenting data, we enable measurement of innovation in 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, 2016). Though 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 both innovation in mature firms and non-patenting 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 IPO prospectuses (e.g., Hanley and Hoberg, 2010; Popadak, 2013; Hoberg, Phillips, and Prabhala, 2014; Hoberg and Maksimovic, 2015; Agarwal, Gupta, and Israelsen, 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, Garcia, and Norli, 2007; Garcia, 2013; Dougal et al., 2012; Israelsen, 2014; Cohen, Malloy, and Nguyen, 2016). 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 (Jegadeesh and Wu, 2017; Ganglmair and Wardlaw, 2017; Goldsmith-Pinkham, Hirtle, and Lucca, 2016; Hoberg and Lewis, 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.[2]

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, Mikhail, and Au (2005), Huang, Zang, and Zheng (2014), and Huang et al. (2015) who provide evidence, in a different context, that investors pay attention to the textual elements of analyst

[2]Even related work on innovation using text analysis has not constructed a similar measure of innovation. Specifically, Fresard, Hoberg, and Phillips (2017) studies 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.

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 measure of innovation to firm performance and value. Section 5 presents an application of our measure to M&A activity. The final section concludes with a summary of future research directions.

2 Data

We begin with a sample of mature firms that were a member of the S&P500 at some point between 1990 and 2012. This initial sample contains 797 firms. To obtain the set of analyst reports these firms, we download analyst reports from Investext via Thomson One for the years 1990 to 2012, which provides an initial sample of 807,309 analyst reports for 750 unique S&P500 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", "an") from the reports using a standard stopword list.[3] 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 and Bodnaruk, Loughran, and McDonald 2015) as an integral part of our textual

[3] We thank Bill McDonald for making these lists available on his website: http://www3.nd.edu/~mcdonald/Word_Lists.html.

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 measure we call the 'innovation' measure (as described in the Section 3.2) and aggregate it to the firm-year level before matching with accounting data from Compustat and patent data up to 2010 from Noah Stoffman's website (Kogan et al., 2017). The final sample has 6,200 observations from 703 unique firms for the period 1990-2010.

For our later analysis of mergers and acquisitions activity, acquisition data are from SDC Platinum. We count the number of completed acquisition during each fiscal year for each of the firms in our sample. In other words, we save records where the acquirer in SDC matches one of our sample firms. Compared with a sample of all Compustat firms, our sample firms are larger, older, have slightly higher R&D intensity, and higher returns on assets. They are similar in terms of asset tangibility and leverage. These are reasonable characteristics because we start with the S&P500 sample, which is comprised of larger firms with these characteristics.

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 (LDA) method of Blei, Ng, and Jordan (2003). To provide a foundation for the empirical work that follows, we describe some of the basic properties of the measure in our sample of S&P500 firms. The measure has desirable time series and cross-sectional properties for a measure of innovation.

As we described in the introduction, 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 the Appendix, Tables A.9 and A.10, the word-list measure exhibits slightly weaker valuation implications, and is not as robustly related to valuable patents as the more accurate

LDA-based measure. This is to be expected because the LDA methodology is better equipped at getting the context of innovative language correct.

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, our view is that the text of analyst reports is 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. Obviously, analysts cannot describe innovative activities that they cannot observe. Thus, we expect that the analyst will describe the publicly-available information relevant to innovation, which neglects insider information such as trade secrets. Still, 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. In addition to containing innovation-relevant information, 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.

One concern from using the analyst text is that analysts may exhibit biases in their evaluations of the firm. Though analysts may exhibit biases in how they evaluate the firms they cover, analysts have been long known to provide value-relevant information about firms (Womack, 1996). Further, our use of the analyst text is predicated on the idea that firms' innovative activities (i.e., the resources the firm uses to increase productivity and generate revenue) are something that analysts are supposed to describe qualitatively. By analyzing the textual content of analyst reports rather than their quantitative aspects, we expect that our innovation measure should be more immune to the usual

sources of analyst bias than alternative measures that take quantitative assessments directly from the analyst.

There is a growing literature that shows that the qualitative nature of analyst text contains useful information. In one of the earliest contributions in this vein, Asquith, Mikhail, and Au (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. Huang, Zang, and Zheng (2014) is the first large sample study of text in analyst reports. Using a sample that overlaps our sample, they find results consistent with Twedt and Rees. Specifically, they find a stock market reaction associated with the tone of reports of between 1.5% and 3.5% (2-day CAR) for reports in the top quintile relative to those in the bottom quintile. They also show that the tone of more qualitative topics (those with few uses of "$" or "%") are more important, a strong indication that the qualitative and descriptive portions of the analyst text are a valuable source of new information.

Based on existing studies of analyst text, it is clear that the qualitative aspects of analyst text contain value-relevant information about the firm. This fact suggests that the analyst reports will provide insight into innovation, which is a critical resource that helps firms generate value. 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 (LDA) model to a corpus of analyst reports following Blei, Ng, and Jordan (2003). This procedure assumes that documents are generated from a distribution 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, 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 a LDA model with $K = 15$ topics using the 665,714 analyst reports as the underlying corpus of documents.[4] 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 compute each topic's statistical distance from the word frequencies used in a popular textbook on innovation, and select the topic with the word distribution that has the smallest statistical distance from the innovation textbook's word distribution.[5] Specifically, we compute the Kullback-Liebler (KL) divergence of each topic's word distribution from the source text on innovation, similar to Lowry, Michaely, and Volkova (2016). 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

[4] 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), whereas models fit with a greater pre-specified number of topics exhibit redundancy (i.e., multiple topics about the same essential idea). Although the number of topics is the only degree of freedom we have in fitting a 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, Hirtle, and Lucca (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.

[5] The textbook we use for this validation exercise is Managing Innovation by Tidd, Bessant, and Pavitt (2005), which was the first hit on a Google search for an innovation textbook in pdf format. The readable pdf format was useful to produce a distribution of words used to describe innovation.

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 second panel of Figure 2 presents a placebo exercise in which the source text is a standard corporate finance textbook (Welch's "Corporate Finance: An Introduction"). Unlike the comparison to the innovation textbook, the innovation topic exhibits a similar KL divergence to other topics.

The measure also appears to intuitively measure the factors that describe innovative companies. For example, Figure 3 presents the topic distribution across words in the form of a word cloud (Appendix Figure A.3 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*, 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 helpful 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 stronger 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 analyst reports with sentiment below the 75th percentile, we set the topic loading at the analyst report level to be zero in the sentiment-adjusted topic measure. We aggregate this sentiment-adjusted topic measure to the firm-year level to construct our text-based measure of innovation, $innov\_text_{it}$. 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. Further, as robustness exercises, we have constructed the measure without the sentiment screen, and we have also controlled explicitly for average sentiment. In each case, the main results are similar.

3.3 Comparison to 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 mature firms in the S&P500, 219 firms have zero patenting throughout the full sample period (1990-2010). Although these firms do not patent, many are highly innovative. Panel (a) of Figure 4 presents side-by-side boxplots of our text-based innovation measure for patenting firms versus non-patenting firms. Although patenting firms have higher text-based innovation on average, the distribution of text-based innovation exhibits substantial overlap between non-patenting firms and patenting firms. Specific examples of highly innovative non-patenting firms are also consistent with this view.[6]

In columns 2 through 4 of Table 1, we present summary comparisons of text-based innovation for mature firms with and without patents. On average, patenting firms have higher text-based innovation than non-patenting firms by 0.27 standard deviations (0.20 sd at the firm level), a difference that is statistically significant at the one percent 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 patenting activity within the set of patenting firms.[7]

[6]The top three innovation firm-years among non-patenting 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 non-patenting firm. Second, in 2000, BroadVision was non-patenting 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.

[7]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 recently proposed rule-of-thumb adjustments to critical values (Harvey, Liu, and Zhu, 2016), and the statistical significance survives other more formal, multiple-comparisons adjustments (e.g., the Bonferroni correction). As we describe in the appendix, this topic explains nearly two times the variation of any other set of topic loadings among the 15 fitted LDA topics.

3.4 Comparison to 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 mature firms in the S&P500, 329 firms have zero R&D expenditures throughout the full sample period (1990-2010). Similar to the non-patenting firms, many non-R&D firms are highly innovative. Panel (b) of Figure 4 presents side-by-side boxplots of our text-based innovation measure for firms with and without R&D, which shows there is substantial overlap in the distribution of text-based innovation for firms with and without R&D.

In columns 5 through 7 of Table 1,[8] we present summary comparisons of text-based innovation for firms with and without R&D. 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 one percent 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 to the extent that the text-based measure is positively correlated with R&D intensity along these dimensions, but 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 captures the R&D boom and bust of the late 1990s and early 2000s, which is studied in Brown, Fazzari, and Petersen (2009). 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. There is a strong relationship between these two series, which have a correlation of 0.51. This correlation suggests that our measure of innovation captures the macro-level trends in innovative activity well. 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-

[8]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.

level R&D expenditures, 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 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 Appendix Table A.4). Even within narrowly-defined industries (4-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 measure of innovation to evaluate the impact of innovation on various measures of performance (e.g., return on assets, Tobin's Q), and examine what the analyst text about innovative firms reveals about the value of innovation. We further examine the relation between our text-based measure and future values of patenting, and perform several robustness checks on our measure.

4.1 Innovation and Performance

True innovation should reflect – as in the language of Drucker (1985) – the fact that a "resource" has been added to the firm. In this spirit, we evaluate whether our text-based measure relates positively to performance, and its impact on performance slowly declines as the innovation resource depreciates over time.

4.1.1 Operating Performance

We now turn to evaluating the performance implications of innovation using our text-based measure. In particular, we examine whether greater measured innovation today leads to greater operating performance (measured by return on assets) a year from now using the specification:

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

where the dependent variable ROA_{it+1} is return on assets (EBITDA/Assets) for firm i in year $t+1$. As above, specifications that include patenting outcomes also control for an indicator for whether the firm is a patenting firm. All specifications include year fixed effects (γ_t) and industry or firm fixed effects (ξ_s), and the coefficient of interest is β_1, which indicates how greater innovation according to our measure leads to changes in operating performance a year ahead. If innovation is valuable, our prediction is that $\beta_1 > 0$. Our specifications also control for standard control variables that are known to influence operating performance, and relate to innovation.

Columns 1 and 2 of Table 2 present the results of estimating equation (1). With industry and year dummies, there is a strong correlation between our text-based measure and the return on assets. A one standard deviation increase in the text measure is associated with a 0.9 percentage point increase in the return on assets in the following year. We find that this estimated effect is robust to including firm fixed effects, and thus, the within-firm variation in our text-based measure of innovation appears

to be valuable in terms of generating abnormal operating performance. Moreover, we see that the text-based measure is more robustly associated with increases in operating performance than patent counts and R&D intensity. Patent counts are not significantly and positively correlated with operating earnings in any specification. Although R&D intensity is positively correlated with future operating performance and the magnitudes are similar to our measure, the statistical significance is lower and the result is not robust across specifications. Moreover, as our estimates using our text-based measure control for alternative measures of innovation, the findings imply that the innovations captured by our measure are valued beyond what existing measures of innovation would predict.

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. Panel (b) of Table 2 shows the effects of innovation split by whether or not the firm uses patents. For patenting firms and non-patenting firms, we find similar point estimates for the coefficient on innovation, indicating that innovation is valued similarly for both types of firms. Moreover, we cannot reject that innovation affects operating performance differently for patenting and non-patenting firms, suggesting that our measure is informative beyond the set of patenting firms.

In Figure 8 (a), we present a plot that summarizes the effect of innovation on operating performance for one through four years into the future. Consistent with how innovation should affect operating performance as a resource that earns the firm revenue, the effects are positive and significant for up to four years after a shock to innovation according to our measure, and these effects decay over time. By contrast, when we evaluate the effects of other measures of innovation over time, patents is unrelated to future operating performance, and the effect of R&D intensity decays much more rapidly over time (see Appendix Table A.8 for details). As we expect that innovation generates persistent operating performance gains, this comparison suggests that our measure better captures a true effect of innovation (at least in the innovation-as-a-resource sense of Drucker, 1985)

4.1.2 Growth Opportunities

Beyond the effects on operating performance, we expect innovation to have longer-term implications for the firm's growth opportunities. The intuition is that 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, if text-based innovation is valuable in the same revealed preference sense, we should expect a significant effect on Tobin's Q because the market value will reflect this innovation premium. To evaluate this hypothesis, we examine the following specification:

$$Q_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}_{it}'\Gamma + \varepsilon_{it} \quad (2)$$

where Q_{it+1} is Tobin's Q (i.e., the ratio of market value to book value of the firm) as a measure of growth opportunities. As before, we include year and industry fixed effects, some specifications include firm fixed effects, and specifications that include patenting outcomes also control for an indicator for whether the firm is a patenting firm. Our coefficient of interest is β_1, which indicates how greater innovation according to our measure leads to changes in growth opportunities a year ahead.

Columns 3 and 4 of Table 2 present the results from estimating equation (2). We find a significant increase in market valuation relative to book valuation for firms that have greater text-based innovation. This is natural because the value of innovations are often difficult to account for in the book value of the firm. As in the operating performance specifications, it is useful to compare the predictability of our text-based measure with R&D intensity and patent counts. A standard deviation change in the text-based measure and patent counts lead to similar changes in future growth opportunities. A one standard deviation change in R&D intensity appears to have somewhat smaller effects on future growth opportunities than the text-based measure, and the effect is not as robust across specifications. Panel (b) of Table 2 shows the results split by whether the firm uses patents. We see

that an effect of innovation on Tobin's Q that is statistically indistinguishable between patenting and non-patenting firms. As with the results for operating performance, this finding highlights a notable advantage with our text-based innovation measure: it can be used for firms that do not use patents.

In Figure 8 (b), we present a plot that summarizes the effect of innovation on Q over time. Consistent with the idea that the market value of a firm reflects an innovation premium captured by our measure , the effects are positive and significant and these effects depreciate more slowly than the operating performance effects over time. For patents and R&D intensity, the effects over time are also persistent, but increase for some horizons (see Appendix Table A.8 for details). The nonlinearity of these effects is consistent with these alternative measures capturing innovation at a different time horizon (perhaps due to the delay between patent application and grant, or delay between R&D expenditure and innovative success).

4.1.3 Growth in Sales

Beyond the effects on operating performance, we expect innovation to have implications for the firm's sales insofar as the innovation reflects product differentiation or new product introductions. In this case, we should expect to see sales growth to increase following an increase in innovation.

$$Salesgrowth_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \quad (3)$$

where $Salesgrowth_{it+1}$ is the percentage growth in sales. As before, we include year and industry fixed effects, some specifications include firm fixed effects, and specifications that include patenting outcomes also control for an indicator for whether the firm is a patenting firm. Our coefficient of interest is β_1, which indicates how greater innovation according to our measure leads to growth in sales in the year ahead.

Columns 5 and 6 of Table 2 present the results from estimating equation (3). We find a statistically significant increase in sales for firms that have greater text-based innovation. As in the operating

performance specifications, it is useful to compare the predictability of our text-based measure with R&D intensity and patent counts. Patent counts appear to be negatively associated with sales growth while there is no apparent relationship between sales growth and R&D intensity. Table 2 (b) shows the results split by whether the firm uses patents. Sales growth seems somewhat more associated with non-patenting firm innovation, though the results are not statistically different between non-patenting and patenting firms.

In Figure 8 (c), we present a plot that summarizes the effect of innovation on sales growth over time. Gains in sales growth are transitory, only occurring in the year following the increase in innovation (see Appendix Table A.8 for details). Interpreting innovation as a resource that generates revenue, this transitory finding is natural. As operating performance increases persistently but sales growth experiences a one-time increase, the pattern of results indicates that our text-based measure reflects an increase in the innovation resource, rather than the growth of innovation over time.

4.1.4 Performance Results Using Rolling Window Version of the Measure

One concern with the innovation measure is the possibility of look-ahead bias in the performance regressions. Because we construct 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 5-year rolling window version of text-based measure, which completely 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 a LDA model fit only using analyst reports from the previous five years (1990-1994).

Table 3 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 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.

4.2 Forecasting Patent Values, Patent Counts, Citations, and Impact

In this subsection, we turn to examining the connection between the text-based innovation measure and patenting outcomes. Specifically, we examine the connection to standard patenting outcomes (patent counts, citations, and impact), as well as the value of patents within the universe of firms that use patents (described in Kogan et al., 2017).

4.2.1 Patent Value Measures

To estimate the relation between text-based innovation and patent value, we employ the following specification using data on the set of patenting firms:

$$Log(1+PatentValue_{it+1}) \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{4}$$

where $PatentValue_{it+1}$ is either the absolute dollar value of the market reaction of all patents granted to firm i during year $t+1$ (panel (a)), or that dollar value divided by the number of patents granted (panel (b)). The patent value is calculated as the cumulative abnormal return over the patent grant date multiplied by the market value (in millions) of the firm. We then sum the patent values for all granted patents for the firm over the fiscal year and evaluate how our text-based innovation measure predicts future patent values. Following similar specifications from the performance regressions, our specifications include controls for R&D, patenting, leverage, firm size, age, growth opportunities, firm or industry fixed effects, and year fixed effects.

Panel (a) of Table 4 presents results from estimating equation 4 using the absolute dollar value measure of patent value. We find a robust relationship where text-based innovation is associated with meaningful increases in future patent values. This relationship holds after controlling for patent citations, a measure that is often used as a proxy for patent value, and beyond being robust to granular industry fixed effects and firm fixed effects, it is also robust to controlling for other time-varying firm characteristics. In panel (b), we report results using the Value per Patent measure, which show a similarly robust relationship between text-based innovation and future patenting values.

4.2.2 Text-Based Innovation and Patenting

To evaluate how text-based innovation relates to patenting outcomes, we estimate the following specification for patenting outcomes one to three years into the future:

$$Log(1+\sum_{s=1}^{3} PatentingOutcome_{t+s}) \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{5}$$

where $\sum_{s=1}^{3} PatentingOutcome_{t+s}$ describes either the number of patent applications over the next three years, the number of patent citations over the next three years, or the number of citations per patent over the next three years. As with previous specifications, the $innov\_text_{it}$ variable is our text-based measure of innovation aggregated to the yearly level for firm i. All specifications use year fixed effects (γ_t) and industry fixed effects (ξ_s).

In Table 5, we present the results from estimating equation (5). The text-based innovation measure is positively related to patent counts, citations, and citations per patent over the next three years. All of these estimates are statistically significant at better than the five percent level, and are robust to broad industry classifications (2-digit SIC).

The findings in this section indicate that our measure contributes valuable information, even within the set of firms that use patents to protect their innovations. Within the set of patenting firms, our

text-based measure is strongly correlated with the most valuable patents, and it is a leading indicator of whether firms will patent in the coming years. Moreover, the text-based innovation measure can be computed using analyst reports in real time while 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 more traditional modes of innovative activity that take time to observe.

4.3 The Nature of Text-Based Innovation

In this section, we estimate the relationship between our text-based measure of innovation and two measures of product outputs: concentration/differentiation from similar competitors, and the number of product announcements. We find that our measure of innovation does not appear to reflect product-level innovations, but rather captures the idea of a firm having an innovative system or sets of processes. This notion of systems innovation is consistent with the nature of value maximization for the mature firms that comprise our sample. In addition, we provide examples where these innovations are patented (and correspond to valuable patents), but also examples where these innovations are not patented, and thus, cannot be spanned by existing innovation proxies.

4.3.1 Text-Based Innovation and Product Measures

First, we study the relationship between text-based innovation and an industry concentration measure constructed from product descriptions by Hoberg and Phillips (2016). The Hoberg and Phillips (2016) concentration measure captures the degree of differentiation within an industry, which would be greater if the firm's innovative activities were focused on distancing the firm from its nearest competitors. Specifically, we estimate the following specification:

$$Log(HHISimilarity_{i,t+1}) = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \quad (6)$$

where $HHISimilarity_{i,t+1}$ is taken from Hoberg and Phillips (2016) we look at how text-based innovation relates to how firms differentiate themselves from other firms in the product description in their 10-K filings. Specifically, we use their Hirfindahl-Hirschmann formulation based on industry classifications made from the product descriptions with the same coarseness as 3-digit SIC industries. The specifications also include the standard controls and 4-digit SIC fixed effects that we employed in the R&D and patent valuation specifications.

Results from estimating equation (6) are presented in columns 1 and 2 of Table 6. Inconsistent with text-based innovation reflecting greater differentiation of the final product, we find no statistically significant relationship between our text-based measure of innovation and the Hoberg-Phillips HHI measure. We are cautious about over-interpreting a failure to reject, but note that the point estimate is small in magnitude, and opposite in sign from an innovation-as-differentiation interpretation of our measure. In contrast, our measure appears to capture innovative systems, both from the context of notable examples like Walmart, and its relationship to valuable patents that correspond to innovative systems, see Figure 9.

In addition, we separately examine the relation between text-based innovation and a novel product announcements measure from Mukherjee, Singh, and Zaldokas (2016) using the specification:

$$Log(1+ProductIntroductions_{i,t+s}) \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \qquad (7)$$

where $ProductIntroductions_{i,t+1}$ is the count of firm i's product introductions (based on a textual analysis of firm press releases in Mukherjee, Singh, and Zaldokas, 2016) that are associated with a significant abnormal return on the announcement. As with the product differentiation tests above, we include the full suite of control variables and 4-digit industry fixed effects in this specification.

Columns 3 through 6 of Table 6 present results from estimating equation (7). Columns 3 and 4 show the contemporaneous relationship ($s = 0$) between text-based innovation and product announcements while columns 5 and 6 show how text-based innovation predicts future product an-

nouncements ($s = 1$). Similar to the product differentiation tests in columns 1 and 2, we find no statistically significant relationship between our text-based measure and product announcements.

As above, our null findings product introductions suggest that we capture a different notion of innovation than a more rapid introduction of new products, or greater differentiation of existing products. Our interpretation of these findings is that text-based innovation more accurately captures innovative systems. After all, text-based innovation is strongly related to patent values, future patenting, and performance outcomes in a manner that is theoretically consistent with innovation. Thus, it is useful to dig into the nature of text-based innovation – specifically, the nature of valuable patents and the nature of highly-innovative firms outside of the set of patenting firms.

4.3.2 Contextual Examples of Systems Innovation in Mature Firms

Within the set of patenting firms, it is useful to examine the content of valuable innovations. 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 either processes, components that can go in to one or several products, or components useful in the production process. Given that our measure appears to pick up on valuable patents with these characteristics that reflect innovative systems, these examples offer some insight into why we do not find a connection with product introductions or product differentiation.

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 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 mature firms like Walmart that have particularly innovative systems.

4.4 Topic Model Robustness

In this subsection, we present robustness to our main text-based innovation measure, which is based on a Latent Dirchlet Allocation model fit assuming 15 underlying topics. 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 spurious explanations unrelated to model fit (i.e., analyst sentiment, use of revenue/growth words)

4.4.1 Fitted Model Robustness

Table 7 presents robustness to the LDA model fit. First, in panel (a) we summarize the results of the 50-topic LDA robustness exercise. In our main specifications, we use relatively few topics to ensure that we capture the generality of the notion of innovation. If the 50-topic LDA has too many topics, the concern is that multiple topics could capture innovation in a similar way. To address this concern, we fit a topic model with 50 topics and identify the topic that is most similar to our main measure (Topic 6 from the 15-topic LDA). Two topics from the 50-topic model are highly correlated with our original topic, and the content of these topics is qualitatively similar (see Figure A.2). Table 7 (a) presents results with one of these two topics as the measure of innovation (using the other one makes no qualitative difference). We obtain results that are similar to Table 2(a) which suggests that the results in the paper are not driven by the choice of the number of topics.

Second, in panel (b), 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). To address this potential issue, we control for each of the other topic loadings aggregated to the firm-year level. As the results in Panel (b) of Table 7 indicate, the main results are qualitatively similar after controlling for other topic loadings, though in some cases, they become stronger.

4.4.2 Robustness to Alternative Explanations

Table 8 presents robustness to 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. Panel (a) of Table 8 presents the results controlling for analyst sentiment, which are similar to the main results.

In addition, 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. To address this issue, we construct word counts of analyst usage of the words "revenue" and "growth" to be used as controls in the specification. Panel (c) of Table 8 presents these results, which show that controlling for the relative word usage of "revenue" or "growth" does not explain the topic's relationship to firm performance. In panel (c) of Table 8, we conduct a similar exercise using words with the root "tech" in them. These word count controls indicate that the topic is not merely selecting the relative incidence of particular words, but consistent with the motivation to use LDA, our methodology seems to be picking up these words when they are used together contextually.

5 Innovation and Acquisition Activity

This section provides a useful application of the text-based innovation measure. Specifically, for the mature firms we study, we examine the relation between text-based innovation and subsequent acquisition activity. The results in this section are consistent with text-based innovation reflecting

an innovative system that generates productive merger opportunities, and are difficult to reconcile with agency-based rationales for merging.

5.1 Text-based Innovation and Acquisition Activity

In theory, innovation could relate to acquisition activity either positively or negatively. On one hand, innovation – either through development or adoption of novel technologies – and acquisitions can be thought of as alternative routes to obtain the technologies that enable the firm to be competitive. According to this view, innovation and acquisitions would be substitutes, and thus have a negative relation with one another (e.g., see Caskurlu, 2015). On the other hand, high-innovation firms could have greater possibilities for synergies with other firms with complementary resources, which would tend to encourage acquisitions that complement their firm's existing resources.

We evaluate how innovation is associated acquisitions using the following specification:

$$Log(1+\sum_{s=1}^{3} Acquisitions_{t+s}) = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{8}$$

where the dependent variable $Log(1+\sum_{s=1}^{3} Acquisitions_{t+s})$ is the log of one plus the number of acquisitions over $t+s$ for $s \in \{1,2,3\}$, and $innov\_text_{it}$ is our text-based innovation measure. As in the performance regressions, we include year and industry fixed effects and specifications that include patenting outcomes also control for an indicator for whether the firm is a patenting firm. The coefficient of interest is β_1, which is how greater innovation as measured by analyst text is associated with acquisition activity in the coming three years. If innovation generates synergies that lead to greater (fewer) acquisition opportunities, we expect $\beta_1 > 0$ ($\beta_1 < 0$). Whether the synergy view or the substitution view dominates is an empirical question.

Columns 1 and 2 of Table 9 Panel (a) presents the results from estimating equation (8). Across specifications, we find evidence that greater measured innovation today is associated with greater

acquisition activity over the next three years.[9] This estimated effect is robust to controlling for profitability and the market-to-book ratio, which proxy for free cash flow and relatively overvalued equity. Thus, the relation between innovation and acquisitions is unlikely to be driven by agency-based explanations for abnormal M&A activity. For the mature firms we study, this finding suggests that innovation is complementary to acquisition activity. This finding corroborates the underlying intuition described by Fresard, Hoberg, and Phillips (2017) in the context of vertical acquisitions, and extends this finding to mature firm innovation.

5.2 Text-based Innovation and Small Acquisitions

A possible reason for the positive relation between innovation and acquisitions is that firms with innovative systems tend to have greater opportunities to engage in productive acquisitions (because greater innovation complements factors in other firms as well). To provide evidence on this point, we separately consider the relation between text-based innovation and large acquisitions versus small acquisitions. A small acquisition is more likely to be a component to the firm's revenue generating system than a large acquisition from the standpoint of an acquiring firm. Following prior work that distinguishes among large and small acquisitions (Yim, 2013), we classify an acquisition as a small acquisition if the deal value according to SDC is less than 5 percent of the value of the acquiring firm, and large otherwise.

In columns 3 through 6 of Table 9 (a), we present the results for two splits of acquisition activity: large acquisitions (columns 3 and 4) and small acquisitions (columns 5 and 6). Across specifications, we find that greater text-based innovation is associated with significantly more small acquisitions in the future, but there is a much weaker relation between text-based innovation and large acquisitions.[10] Moreover, the positive relation between text-based innovation and small acquisitions

[9] In the appendix (Table A.6), we present estimates from an analogous specification, but using a linear probability model for whether the firm engages in any acquisitions in the coming 3 years (rather than the number of acquisitions during that time). We find that higher innovation is associated with the extensive margin (i.e., the main result is not just an increase in acquisitions among the set of firms that would conduct M&A anyway).

[10] In the appendix (Table A.7), we also relate merger announcement returns (CARs) to our text-based measure of innovation. For acquirers with high text-based innovation, we find that small acquisitions are viewed significantly more favorably than large acquisitions. Together with our finding that innovation is associated with more small acquisitions (but not large acquisitions), the CAR analysis suggests that text-based innovation generates synergies that are well recognized

is consistent with our overall interpretation that the text-based innovation measure captures innovative systems. This subsample finding is inconsistent with other prominent rationales to merge (e.g., empire building), which would tend to lead to larger acquisitions.[11]

5.3 Purging the Innovation Measure of Acquisition-Specific Language

One concern with the observed relation between our innovation measure and acquisition activity is that the words for "acquisition" and "merger" are commonly used by analysts. Thus, it is a natural concern that the relation to the text-based innovation is mechanically related to text-based innovation via the use of these acquisition words. To address this concern, we construct an alternative text-based measure of innovation purged of words that begin with "merg" and "acqui." To accomplish this task, we set these acquisition word frequencies to zero, and re-estimate the topic-by-document distribution without these acquisition words.

In Panel (b) of Table 9, we present the results on acquisition activity using this acquisitions-purged text-based measure of innovation. As is apparent from the results, the broad conclusions not only remain, but in some cases the magnitudes and statistical significance on the relation between text-based innovation and subsequent acquisition activity strengthens. In this way, these findings enhance our confidence that the underlying relation between our text-based innovation measure and acquisition activity reflects incentives faced by highly innovative firms rather than an artifact of the underlying textual descriptions.

6 Conclusions

In this paper, we have developed 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

by investors, and that corporate actions inconsistent with this synergy view (i.e., attempting to acquire a large firm, potentially with integration risk) are viewed negatively by investors.

[11] Beyond accounting for empire building via controls for ROA and market-to-book, we find that the acquisitions are focused on smaller firms where the potential for integration is greater, which suggests that empire building motives (e.g., see Harford, Humphery-Jenner, and Powell, 2012) are unlikely to drive the observed relation between text-based innovation and acquisitions.

innovation in mature firms without patents and with zero R&D expenditure. Such firms are common, even among our sample of 703 firms from the S&P500, 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 non-patenting 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 Fresard, Hoberg, and Phillips (2017) 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, unrelated to products. This mode of innovation likely exhibits a different relation to corporate outcomes that have been linked to product innovations. 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, Singh, and Zal-

dokas, 2016 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.

References

Agarwal, S., S. Gupta, and R. D. Israelsen. 2016. Public and private information: Firm disclosure, sec letters, and the JOBS Act. *Working Paper SSRN 2891089* .

Asquith, P., M. B. Mikhail, and A. S. Au. 2005. Information content of equity analyst reports. *Journal of Financial Economics* 75:245 – 282.

Atanassov, J. 2013. Do hostile takeovers stifle innovation? Evidence from antitakeover legislation and corporate patenting. *Journal of Finance* 68:1097–131.

Baier, S. L., G. P. Dwyer, and R. Tamura. 2006. How important are capital and total factor productivity for economic growth? *Economic Inquiry* 44:23–49.

Bernstein, S. 2015. Does going public affect innovation? *Journal of Finance* 70:1365–403.

Bhagat, S., M. Dong, D. Hirshleifer, and R. Noah. 2005. Do tender offers create value? new methods and evidence. *Journal of Financial Economics* 76:3–60.

Bhagat, S., and I. Welch. 1995. Corporate research & development investments international comparisons. *Journal of Accounting and Economics* 19:443–70.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bodnaruk, A., T. Loughran, and B. McDonald. 2015. Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis* 50:623–46.

Boudoukh, J., R. Feldman, S. Kogan, and M. Richardson. 2013. Which news moves stock prices? A textual analysis. *NBER Working Paper No. 18725* .

Bradley, D., I. Kim, and X. Tian. 2015. Do unions affect innovation? *Management Science, Forthcoming* .

Bradshaw, M. T. 2011. Analysts' forecasts: What do we know after decades of work? *Working Paper SSRN 1880339* .

Brown, J. R., S. M. Fazzari, and B. C. Petersen. 2009. Financing innovation and growth: Cash flow, external equity and the 1990s R&D boom. *Journal of Finance* 64:151–85.

Caskurlu, T. 2015. Effects of patent rights on industry structure and R&D. *Working Paper* .

Cohen, L., K. Diether, and C. Malloy. 2013. Misvaluing innovation. *Review of Financial Studies* 26:635–66.

Cohen, L., and A. Frazzini. 2008. Economic links and predictable returns. *Journal of Finance* 63:1977–2011.

Cohen, L., U. Gurun, and S. D. Kominers. 2014. Patent trolls: Evidence from targeted firms. *NBER Working Paper No. 20322* .

Cohen, L., C. Malloy, and Q. H. Nguyen. 2016. Lazy prices. *Working Paper SSRN 1658471* .

Cooper, M. J., A. M. Knott, and W. Yang. 2015. Measuring innovation. *Working Paper SSRN 2631655* .

Dougal, C., J. Engelberg, D. Garcia, and C. A. Parsons. 2012. Journalists and the stock market. *Review of Financial Studies* 25:639–79.

Drucker, P. 1985. *Innovation and entrepreneurship: Practice and principles*. Boston, MA: Butterworth Heinemann.

Edmans, A., D. Garcia, and Ø. Norli. 2007. Sports sentiment and stock returns. *Journal of Finance* 62:1967–98.

Egan, E. J. 2013. How start-up firms innovate: Technology strategy, commercialization strategy, and their relationship. *Working Paper SSRN 2364096* .

Fresard, L., G. Hoberg, and G. M. Phillips. 2017. Innovation activities and the incentives for vertical acquisitions and integration. *Working Paper SSRN 2242425* .

Fried, D., and D. Givoly. 1982. Financial analysts' forecasts of earnings: A better surrogate for market expectations. *Journal of Accounting and Economics* 4:85–107.

Galasso, A., and M. Schankerman. 2015. Patents rights and innovation by small and large firms. *Working Paper SSRN 2694725* .

Ganglmair, B., and M. Wardlaw. 2017. Complexity, standardization, and the design of loan agreements. *Working Paper* .

Garcia, D. 2013. Sentiment during recessions. *Journal of Finance* 68:1267–300.

Goldsmith-Pinkham, P., B. Hirtle, and D. Lucca. 2016. Parsing the content of bank supervision. *Working Paper FRBNY Staff Report No. 770* .

Griliches, Z. 1980. *New developments in productivity measurement*, chap. Returns to Research and Development Expenditures in the Private Sector, 419–62. University of Chicago Press.

Griliches, Z., ed. 1984. *R&D, patents, and productivity*. University of Chicago Press.

Griliches, Z. 1998. *The search for R&D spillovers, R&D and productivity: The econometric evidence*. University of Chicago Press.

Hall, B., C. Helmers, M. Rogers, and V. Sena. 2014. The choice between formal and informal intellectual property: A review. *Journal of Economic Literature* 52:375–423.

Hall, B. H. 1990. The impact of corporate restructuring on industrial research and development. *Brookings Papers on Economic Activity: Microeconomics* .

Hall, B. H., C. Helmers, M. Rogers, and V. Sena. 2013. The importance (or not) of patents to UK firms. *Oxford Economic Papers* 603 – 629.

Hall, B. H., A. Jaffe, and M. Trajtenberg. 2005. Market value and patent citations. *Rand Journal of Economics* 36:16–38.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg. 2001. The NBER patent citation data file: Lessons, insights and methodological tools. *NBER Working Paper No. 8408* .

Hall, B. H., J. Mairesse, and P. Mohnen. 2010. Measuring the returns to R&D. *Handbook of the Economics of Innovation* 2:1033–82.

Hanley, K. W., and G. Hoberg. 2010. The information content of IPO prospectuses. *Review of Financial Studies* 23:2821–64.

Harford, J., M. Humphery-Jenner, and R. Powell. 2012. The sources of value destruction in acquisitions by entrenched managers. *Journal of Financial Economics* 106:247–61.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

He, J., and X. Tian. 2013. The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics* 109:856 – 878.

Hellwig, M., and A. Irmen. 2001. Endogenous technical change in a competitive economy. *Journal of Economic Theory* 101:1 – 39.

Hirshleifer, D., A. Low, and S. H. Teoh. 2012. Are overconfident CEOs better innovators? *The Journal of Finance* 67:1457–98.

Hoberg, G., and C. Lewis. 2017. Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance, Forthcoming* .

Hoberg, G., and V. Maksimovic. 2015. Redefining financial constraints: A text-based analysis. *Review of Financial Studies* 28:1312–52.

Hoberg, G., and G. Phillips. 2010. Real and financial industry booms and busts. *Journal of Finance* 65:45–86.

Hoberg, G., G. Phillips, and N. Prabhala. 2014. Product market threats, payouts, and financial flexibility. *Journal of Finance* 69:293–324.

Hoberg, G., and G. M. Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124:1423–65.

Huang, A., R. Lehavy, A. Zang, and R. Zheng. 2015. Analyst information discovery and interpretation roles: A topic modeling approach. *Working Paper SSRN 2409482* .

Huang, A., A. Zang, and R. Zheng. 2014. Evidence on the information content of text in analyst reports. *The Accounting Review* 89:2151–80.

Israelsen, R. D. 2014. Tell it like it is: Disclosed risks and factor portfolios. *Working Paper SSRN 2504522* .

Jegadeesh, N., and D. Wu. 2017. Deciphering Fedspeak: The information content of FOMC meetings. *Working Paper SSRN 2939937* .

Jiang, J., and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proc. of the Int'l. Conf. on Research in Computational Linguistics* 19–33.

Jones, C. I. 2002. Sources of US economic growth in a world of ideas. *American Economic Review* 92:220–39.

Jurafsky, D., and J. H. Martin. 2009. *Speech and language processing, 2nd edition*. Pearson Education Inc.

Knott, A. M. 2008. R&D/returns causality: Absorptive capacity or organizational IQ. *Management Science* 54:2054–67.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman. 2017. Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics, Forthcoming* .

Kuznets, S., and J. T. Murphy. 1966. *Modern economic growth: Rate, structure, and spread*, vol. 2. Yale University Press New Haven.

Lee, C. M., P. Ma, and C. C. Wang. 2015. Search based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116:410–31.

Loh, R. K., and G. M. Mian. 2006. Do accurate earnings forecasts facilitate superior investment recommendations? *Journal of Financial Economics* 80:455–83.

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

Lowry, M., R. Michaely, and E. Volkova. 2016. Information revelation through regulatory process: Interactions between the SEC and companies ahead of the IPO. *Working Paper SSRN 2802599* .

Mann, W. 2016. Creditor rights and innovation: Evidence from patent collateral. *Working Paper SSRN 2356015* .

Manso, G. 2011. Motivating innovation. *Journal of Finance* 66:1823–60.

Miller, G. A. 1995. WordNet: A lexical database for english. *Communications of the ACM* 38:39–41.

Moser, P. 2012. Innovation without patents: Evidence from world's fairs. *Journal of Law and Economics* 55:43 – 74.

Mukherjee, A., M. Singh, and A. Zaldokas. 2016. Do Corporate Taxes Hinder Innovation? *Journal of Financial Economics, Forthcoming* .

Nicholas, T. 2008. Does innovation cause stock market runups? Evidence from the great crash. *American Economic Review* 98:1370–96.

Nordhaus, W. D. 1969. An economic theory of technological change. *American Economic Review* 59:18–28.

Popadak, J. A. 2013. A corporate culture channel: How increased shareholder governance reduces firm value. *Working Paper SSRN 2345384* .

Saidi, F., and A. Zaldokas. 2016. Patents as substitutes for relationships. *Working Paper SSRN 2735987* .

Scherer, F. M. 1965. Firm size, market structure, opportunity, and the output of patented inventions. *American Economic Review* 55:1097–125.

Schumpeter, J. A. 1934. *The theory of economic development: An inquiry into profits, credit, interest, and the business cycle*. Transaction Publishers.

———. 1939. *Business cycles: A theoretical, historical, and statistical analysis of the capitalist process*. McGraw-Hill New York.

Solow, R. M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70:65–94.

Swem, N. 2014. Information in financial markets: Who gets it first? *Working Paper SSRN 2437733* .

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirchlet Process. *Journal of the American Statistical Association* 101:1566–81.

Tian, X. 2012. The role of venture capital syndication in value creation for entrepreneurial firms. *Review of Finance* 16:245–83.

Tian, X., and T. Y. Wang. 2014. Tolerance for failure and corporate innovation. *Review of Financial Studies* 27:211–55.

Tidd, J., J. Bessant, and K. Pavitt. 2005. *Managing innovation: Integrating technological, market and organizational change*. John Wiley & Sons.

Trajtenberg, M. 1990. A penny for your quotes: Patent citations and the value of innovations. *Rand Journal of Economics* 21:172–87.

Tucker, C. 2014. Patent trolls and technology diffusion: The case of medical imaging. *Working Paper SSRN 1976593* .

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

Womack, K. L. 1996. Do brokerage analysts' recommendations have investment value? *Journal of Finance* 51:137–67.

Yim, S. 2013. The acquisitiveness of youth: CEO age and acquisition behavior. *Jounral of Financial Economics* 108:250–73.

7 Tables and Figures

7.1 Figures

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

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

(a) Low Patent Industries

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) High Patent Industries

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.

Figure 2: Selecting the Innovation Topic – Kullback-Liebler Divergence from an Innovation Textbook

Note: The first 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, Bessant and Pavitt), 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 second panel is a placebo exercise that uses a standard corporate finance textbook (Welch's "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.

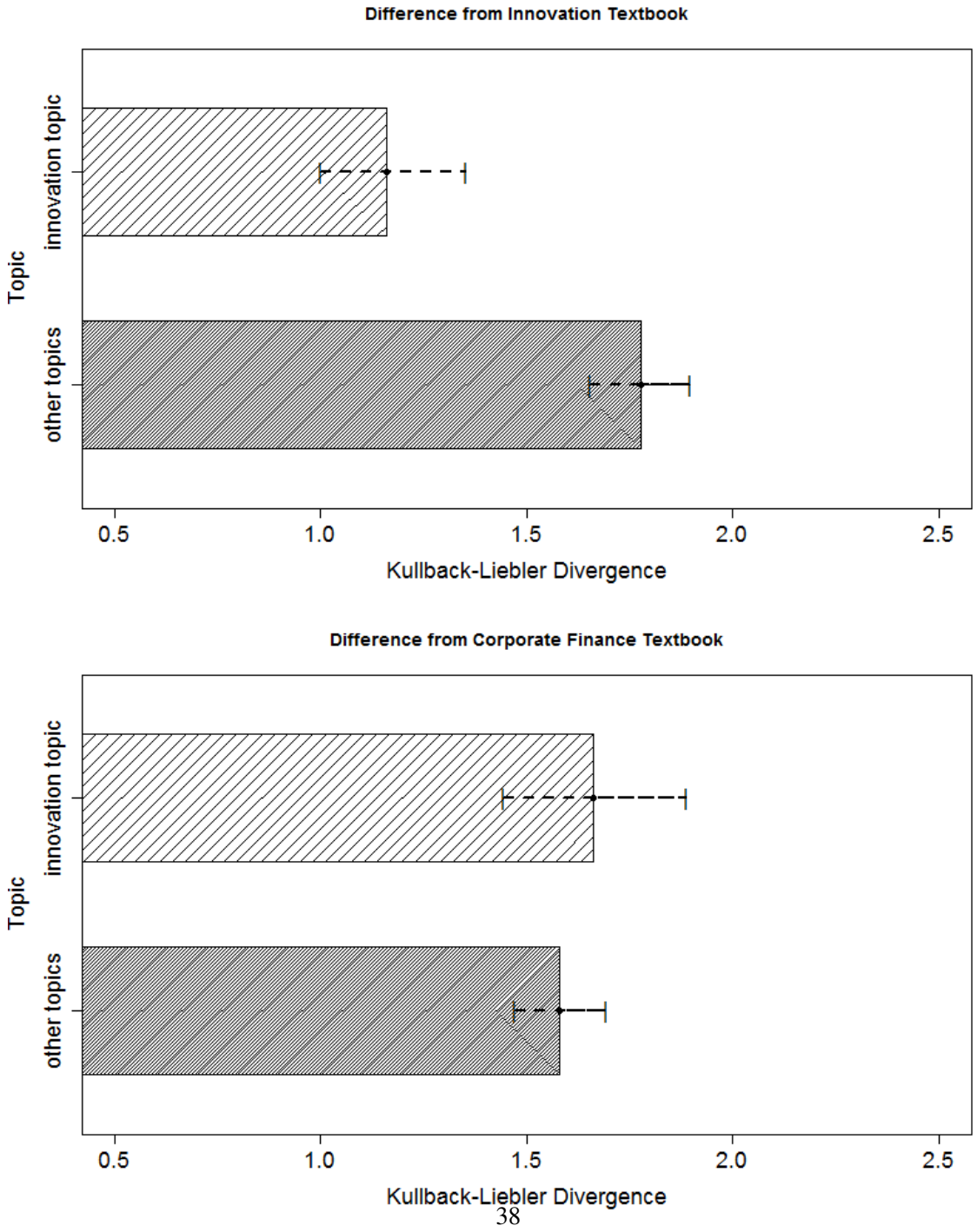


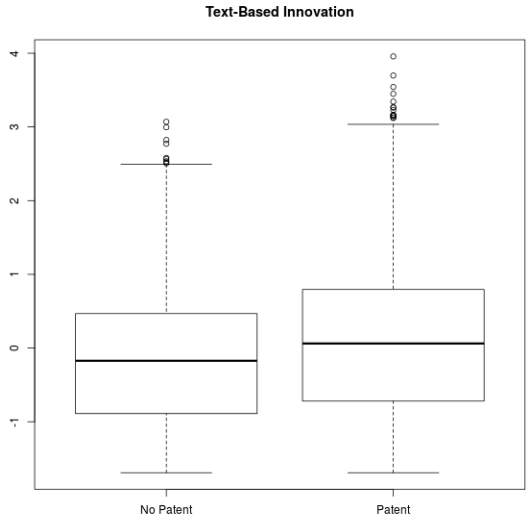
Figure 3: Text-Based Innovation Measure: Word Cloud

Note: 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&P500 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, Bessant, and Pavitt, 2005).

Figure 4: Distribution of Text-Based Innovation

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

(a) With and Without Patents



(b) With and Without R&D

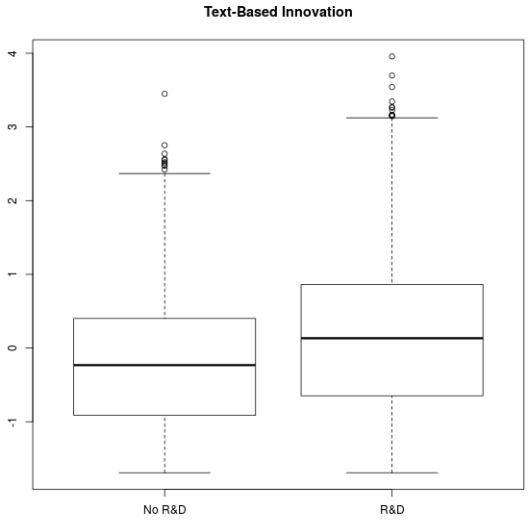
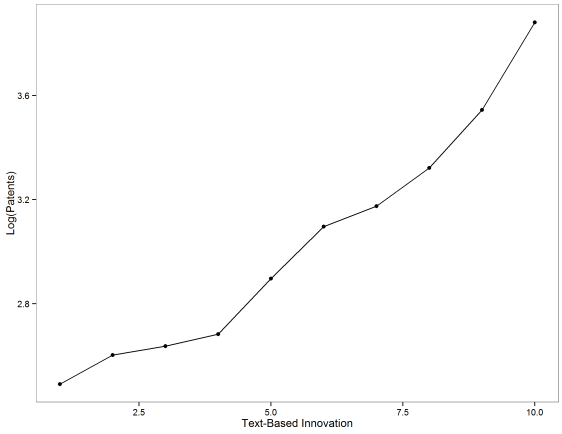


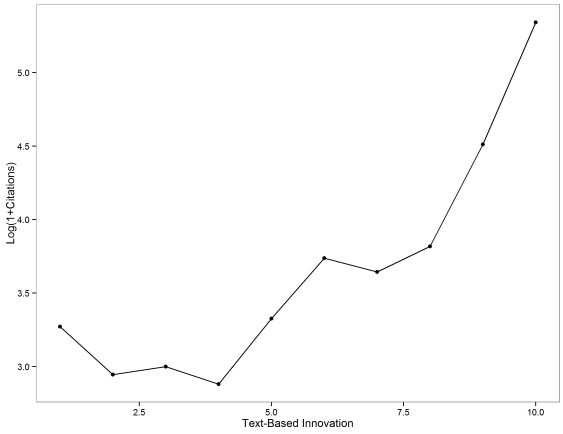
Figure 5: Relating Patent Counts and Patent Citations to the Text-Based Innovation Measure (Decile Bins)

Note: 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 ($\log(1+Citations)$), and Panel (c) presents the relation between text-based innovation and citations per patent ($\log\left(1+\frac{Citations}{Patent}\right)$).

(a) Patent Counts



(b) Patent Citations



(c) Citations per Patent

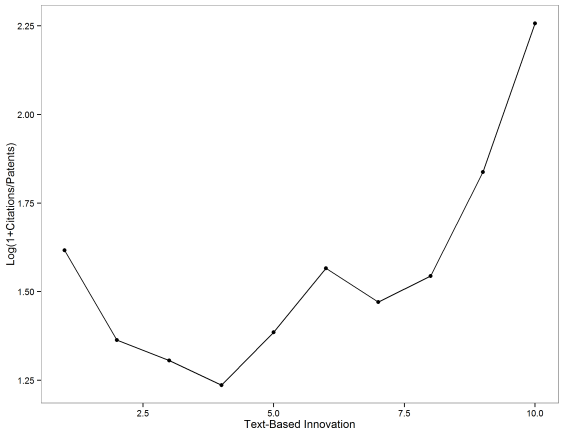


Figure 6: Time Series of Text-Based Innovation Measure and R&D (1990-2010)

Note: 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, which is statistically different from zero.

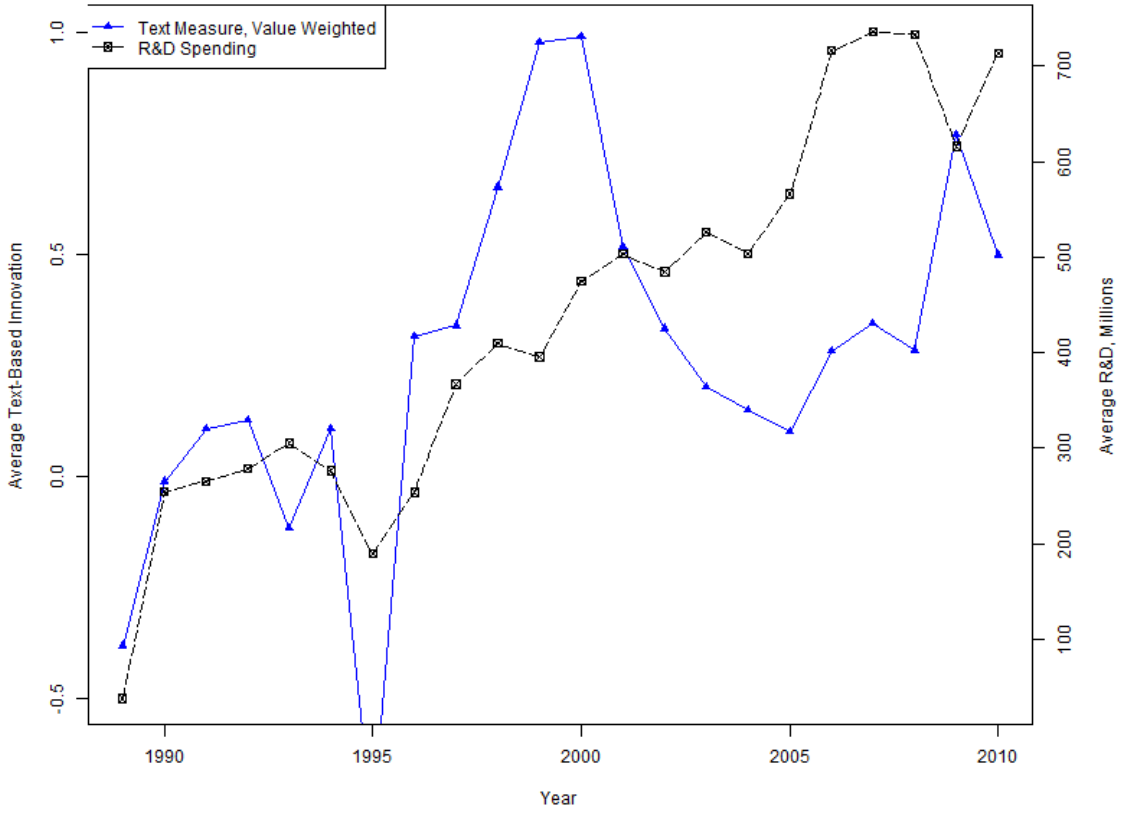


Figure 7: Cross-Industry Plot of R&D (1990-2004), Relationship to Text-Based Measure

Note: This figure provides a plot of R&D expenditures (demeaned by the average R&D/Assets) by industry covered in the sample of S&P500 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, and statistically different from 0.

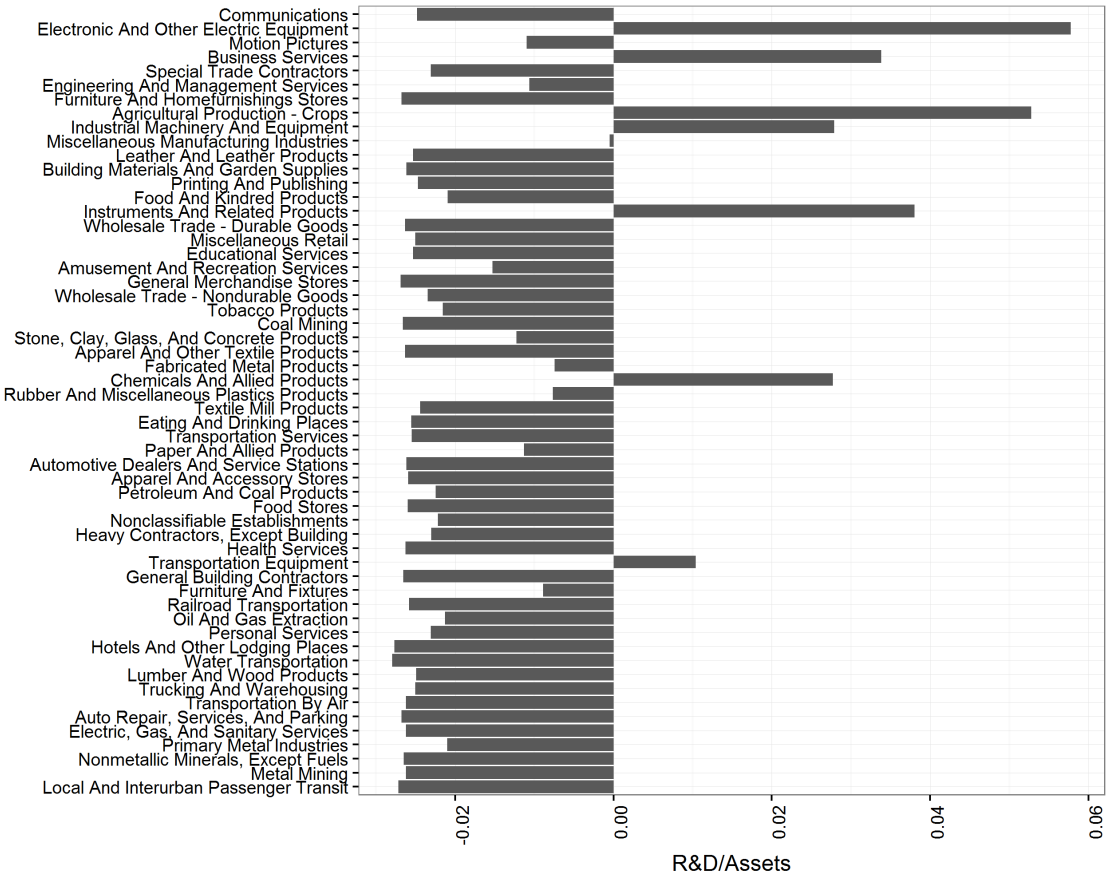
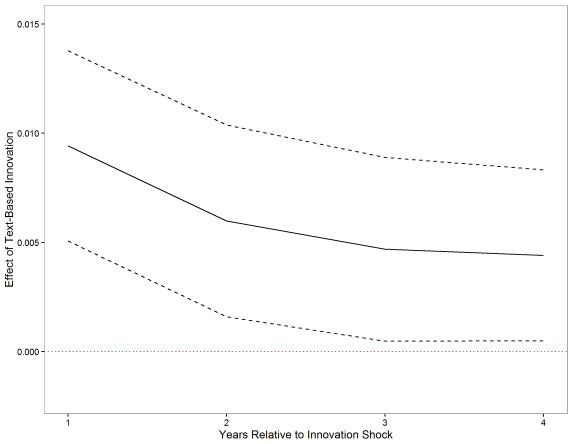


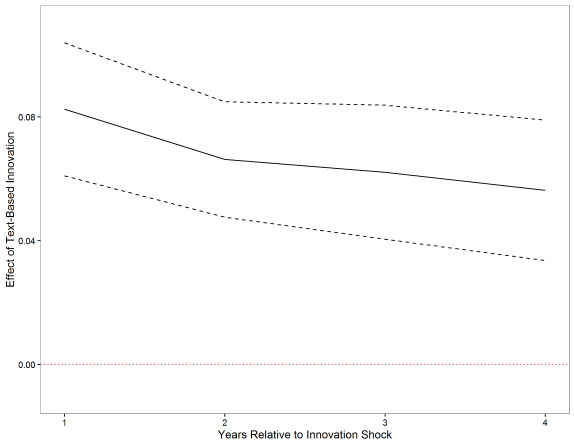
Figure 8: Long Run Effects of Innovation on Performance – Forecasting ROA and Tobin's Q up to Four Years Out

Note: These plots present the response in ROA, Q, and sales growth 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 appendix Table A.8. Dotted lines represent 95% confidence bands around the estimated effects.

(a) ROA



(b) Q



(c) Salesgrowth

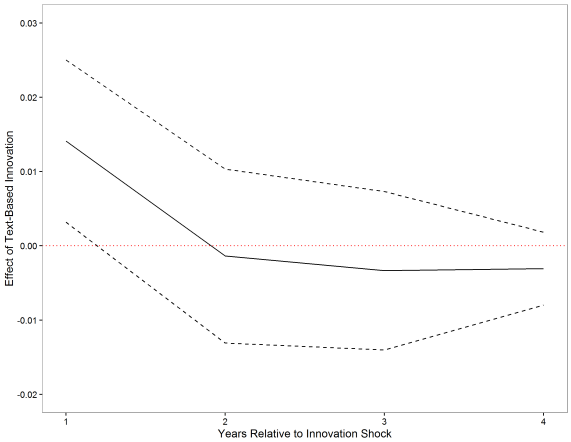


Figure 9: Valuable Patents (95th percentile)

Note: 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.

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.

7.2 Tables

Table 1: Summary Statistics

Note: The text-based innovation measure is presented in Z-score units because the scale of the measure (described in Appendix A.3) is not easily interpretable. Patents is the count of granted patents which were applied for during the year. Return on assets is EBITDA over total assets. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Age is the number of years since the firm entered Compustat (with the earliest date 1975). Panel (a) shows means of variables from the full sample and Panel (b) shows means of variables on the firm-level (i.e. after first taking the mean by firm). Columns 4 and 7 show differences between positive and zero R&D and positive and zero patents, respectively. Errors are calculated with firm and year clusters in panel (a). * denotes significance at the 10% level, ** at the 5% level, and *** at the 1% level.

(a) Summary Statistics

Variable	All	Patents>0	Patents=0	(5)-(6)	R&D>0	R&D=0	(2)-(3)
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***
Salesgrowth	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	6200	3586	2614		3268	2932	

(b) Firm-Level Summary Statistics

Variable	All	Patents>0	Patents=0	(5)-(6)	R&D>0	R&D=0	(2)-(3)
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***
Salesgrowth	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	

Table 2: Performance of Firms and Text-Based Innovation (1990-2010)

Note: This table presents OLS 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), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Full results are reported in the appendix (Table A.3). Variable definitions are presented in Table A.2. Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Firm Performance

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.082^{***}	0.051^{***}	0.014^{**}	0.010^{**}
	(0.002)	(0.002)	(0.011)	(0.008)	(0.006)	(0.005)
$Log(Patents)_t$	0.003	-0.003	0.028^{***}	-0.002	-0.012^{***}	-0.013^{**}
	(0.002)	(0.002)	(0.010)	(0.014)	(0.003)	(0.006)
Patenting Firm	0.010^{*}		0.041		-0.001	
	(0.005)		(0.032)		(0.009)	
R&D/Assets $(Z)_t$	0.006	0.010^{**}	0.075^{***}	0.028	-0.001	-0.007
	(0.005)	(0.004)	(0.021)	(0.025)	(0.006)	(0.008)
Other controls	X	X	X	X	X	X
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	6,066	6,066	5,933	5,933	6,070	6,070
Adjusted R^2	0.438	0.674	0.580	0.770	0.100	0.160
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Firm Performance - Patenting Firm Split

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$						
$\times$ Patenting Firm	0.009^{***}	0.005^{***}	0.082^{***}	0.052^{***}	0.013^{**}	0.008
	(0.002)	(0.002)	(0.012)	(0.009)	(0.005)	(0.005)
$\times$ Non-Patenting Firm	0.011^{***}	0.006^{*}	0.083^{***}	0.049^{***}	0.016^{**}	0.020^{**}
	(0.004)	(0.003)	(0.016)	(0.014)	(0.008)	(0.009)
$Log(Patents)_t$	0.003	-0.003	0.028^{***}	-0.002	-0.011^{***}	-0.013^{**}
	(0.002)	(0.002)	(0.010)	(0.014)	(0.003)	(0.006)
Patenting Firm	0.010^{*}		0.041		-0.002	
	(0.005)		(0.032)		(0.009)	
R&D/Assets $(Z)_t$	0.006	0.010^{**}	0.075^{***}	0.028	-0.001	-0.007
	(0.005)	(0.004)	(0.021)	(0.025)	(0.006)	(0.008)
Other controls	X	X	X	X	X	X
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	6,066	6,066	5,933	5,933	6,070	6,070
Adjusted R^2	0.438	0.674	0.579	0.770	0.100	0.160
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3: Performance of Firms and Text-Based Innovation – Rolling Window Version (1994-2010)

Note: This table presents OLS regressions that link the rolling window version of 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. The rolling window version of the text-measure is based on an LDA model of the 5 prior years of reports. Other innovation measures – log(patents), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Full results are reported in the appendix (Table A.3). Variable definitions are presented in Table A.2. Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Firm Performance

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.071^{***}	0.044^{***}	0.007	0.006
	(0.002)	(0.002)	(0.010)	(0.007)	(0.006)	(0.005)
$Log(Patents)_t$	0.003	-0.003	0.023^{**}	-0.024^{*}	-0.011^{***}	-0.009
	(0.002)	(0.002)	(0.012)	(0.014)	(0.004)	(0.007)
R&D/Assets $(Z)_t$	0.006	0.011^{**}	0.084^{***}	0.031	0.001	-0.008
	(0.006)	(0.005)	(0.024)	(0.026)	(0.006)	(0.010)
Patenting Firm	0.012^{*}		0.050		0.004	
	(0.006)		(0.034)		(0.011)	
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.428	0.680	0.580	0.796	0.102	0.165
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Firm Performance - Patenting Firm Split

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$						
$\times$ Patenting Firm	0.009^{***}	0.005^{**}	0.071^{***}	0.046^{***}	0.005	0.004
	(0.002)	(0.002)	(0.012)	(0.008)	(0.005)	(0.005)
$\times$ Non-Patenting Firm	0.012^{***}	0.008^{***}	0.072^{***}	0.039^{***}	0.011	0.015
	(0.005)	(0.003)	(0.014)	(0.011)	(0.014)	(0.016)
$Log(Patents)_t$	0.003	-0.003	0.023^{**}	-0.024^{*}	-0.011^{***}	-0.009
	(0.002)	(0.002)	(0.012)	(0.014)	(0.004)	(0.007)
R&D/Assets $(Z)_t$	0.006	0.011^{**}	0.084^{***}	0.031	0.001	-0.008
	(0.006)	(0.005)	(0.024)	(0.026)	(0.006)	(0.010)
Patenting Firm	0.011^{*}		0.050		0.002	
	(0.006)		(0.036)		(0.012)	
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.428	0.680	0.580	0.796	0.102	0.165
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 4: Patent Value and Text-Based Innovation (1990-2010)

Note: This table presents the output from OLS regressions that link our text-based innovation measure to existing proxies for patenting value. In panel (a), the dependent variable is the market value (i.e., the stock market jump on the day of the granted patent in $millions) aggregated over all patents granted during the year (taken from Kogan et al. (2017)). In panel (a), we scale this variable by patent count. Other controls are R&D intensity, leverage, the log of total assets, the log of age, and the log of Q. Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Patent Value

	Dependent variable:					
	Log(1 + Patent Value)$_t$					
	(1)	(2)	(3)	(4)	(5)	(6)
Text Innovation$_t$	0.271***	0.268***	0.055**	0.162***	0.160***	0.065***
	(0.046)	(0.044)	(0.023)	(0.036)	(0.035)	(0.021)
Log(1 + Patents)$_t$	1.034***	0.995***	0.672***	0.854***	0.818***	0.753***
	(0.035)	(0.032)	(0.032)	(0.049)	(0.044)	(0.043)
Log(1 + Citations)$_t$ (Z)		0.315***	0.367***		0.192***	0.186***
		(0.054)	(0.050)		(0.054)	(0.051)
Other Controls			X			X
4-digit SIC Dummies	X	X	X			
Firm FE				X	X	X
Year FE	X	X	X	X	X	X
Observations	3,587	3,587	3,587	3,587	3,587	3,587
Adjusted R^2	0.805	0.816	0.888	0.912	0.915	0.934
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Value Per Patent

	Dependent variable:					
	Log(1 + Value per Patent)$_t$					
	(1)	(2)	(3)	(4)	(5)	(6)
Text Innovation$_t$	0.238***	0.237***	0.077***	0.179***	0.179***	0.090***
	(0.037)	(0.037)	(0.026)	(0.037)	(0.037)	(0.027)
Log(1 + Citations)$_t$ (Z)		0.149***	0.124***		0.060**	0.050*
		(0.039)	(0.036)		(0.027)	(0.026)
Other Controls			X			X
4-digit SIC Dummies	X	X	X			
Firm FE				X	X	X
Year FE	X	X	X	X	X	X
Observations	2,999	2,999	2,999	2,999	2,999	2,999
Adjusted R^2	0.529	0.540	0.712	0.778	0.779	0.839
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 5: Patents and Text-Based Innovation (1990-2010)

Note: This table presents OLS regressions linking future patenting outcomes to current text-based innovation, accounting for standard controls. The dependent variables in this table are future patent counts, patent citations, and impact (i.e., citations per patent). As in other tables, the text-based measure is standardized to have a mean of 0 and a standard deviation of 1. Variable definitions are presented in Table A.2. Standard errors that are double clustered on firm and year are reported in parentheses.

	Dependent variable:					
	$\text{Log}(1+\sum_{s=1}^{3}\text{Patents}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{Citations}_{t+s})$		$\text{Log}(1+\frac{\sum_{s=1}^{3}\text{Citations}_{t+s}}{\sum_{s=1}^{3}\text{Patents}_{t+s}})$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation $(Z)_t$	0.182^{***}	0.046^{**}	0.315^{***}	0.148^{**}	0.149^{***}	0.082^{***}
	(0.055)	(0.020)	(0.075)	(0.061)	(0.025)	(0.022)
$\text{Log}(1+\text{Patents})_t$		1.018^{***}		0.961^{***}		-0.065^{***}
		(0.060)		(0.107)		(0.025)
$\text{R\&D/Assets}_t$		0.688		−0.310		0.017
		(0.485)		(1.994)		(0.650)
Log(Assets)_t	0.854^{***}	0.097^{***}	0.435^{***}	-0.246^{***}	-0.174^{***}	-0.074^{**}
	(0.080)	(0.021)	(0.115)	(0.088)	(0.040)	(0.033)
$\text{Return on Assets}_t$		−0.100		−0.532		0.759^{**}
		(0.320)		(0.963)		(0.351)
$\text{Asset Tangibility}_t$		0.076		-1.156^{**}		-0.965^{***}
		(0.194)		(0.561)		(0.205)
Leverage_t		-0.380^{***}		−0.558		−0.122
		(0.105)		(0.370)		(0.128)
Log(Age)_t		-0.112^{*}		-0.577^{***}		-0.353^{***}
		(0.059)		(0.193)		(0.088)
Log(Q)_t		0.138^{**}		0.146		0.089
		(0.056)		(0.188)		(0.058)
2-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	4,782	4,782	4,782	4,782	3,209	3,209
Adjusted R^2	0.580	0.869	0.443	0.591	0.590	0.622

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: Product Differentiation and Product Announcements (1990-2010)

Note: The dependent variable in columns 1 and 2 is the industry concentration measure from Hoberg and Phillips (2016), specifically the Hirfindahl-Hirschmann formulation based on industry classifications made from the product descriptions with the same coarseness as 3-digit SIC industries. Columns 3 through 6 use the count of product announcements when the stock market return was above the 75th percentile from Mukherjee, Singh, and Zaldokas (2016). As in other tables, the text-based measure is standardized to have a mean of 0 and a standard deviation of 1. Variable definitions are presented in Table A.2. Standard errors that are double clustered on firm and year are in parentheses.

	Dependent variable:					
	Log(Total Similarity)$_{t+1}$		Log(1 + Products)$_t$		Log(1 + Products)$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	−0.013	−0.013	0.005	0.005	0.028	0.028
	(0.020)	(0.020)	(0.022)	(0.022)	(0.029)	(0.029)
R&D/Assets (Z)$_t$	−0.026	−0.026	0.088**	0.088**	0.104**	0.104**
	(0.026)	(0.026)	(0.043)	(0.043)	(0.046)	(0.046)
Log(1 + Patents)$_t$	0.003	0.003	0.041***	0.041***	0.030*	0.030*
	(0.013)	(0.013)	(0.016)	(0.016)	(0.016)	(0.016)
Leverage$_t$	0.104	0.104	−0.094	−0.094	−0.106	−0.106
	(0.103)	(0.103)	(0.139)	(0.139)	(0.155)	(0.155)
Log(Total Assets)$_t$	0.013	0.013	0.289***	0.289***	0.278***	0.278***
	(0.023)	(0.023)	(0.044)	(0.044)	(0.046)	(0.046)
Log(Age)$_t$	0.088**	0.088**	−0.082	−0.082	0.015	0.015
	(0.042)	(0.042)	(0.082)	(0.082)	(0.100)	(0.100)
Asset Tangibility$_t$	0.035	0.035	−0.080	−0.080	−0.229	−0.229
	(0.124)	(0.124)	(0.282)	(0.282)	(0.255)	(0.255)
Log(Q)$_t$	0.005	0.005	0.151***	0.151***	0.126**	0.126**
	(0.049)	(0.049)	(0.054)	(0.054)	(0.058)	(0.058)
Return on Assets$_t$	0.193	0.193	0.487	0.487	0.386	0.386
	(0.182)	(0.182)	(0.364)	(0.364)	(0.466)	(0.466)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	4,488	4,488	2,030	2,030	1,897	1,897
Adjusted R^2	0.582	0.582	0.524	0.524	0.521	0.521
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 7: Robustness of LDA Model Fit

Note: The specifications and variable definitions for ROA, Q, and Salesgrowth are analogous to those in Table 2. Panel (a) reports the measure from a 50-topic LDA, panel (b) reports a 5-year rolling window version of the measure, and panel (c) reports the main measure (K=15) controlling for all other topic loadings. All specifications account for the full set of other controls, industry fixed effects (4-digit SIC), and year fixed effects. Standard errors that are double clustered on firm and year are in parentheses.

(a) Firm Performance, K=50 (1990-2010)

	Dependent variable:					
	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		Sales Growth$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.006***		0.057***		0.011*	
	(0.002)		(0.009)		(0.007)	
× Patenting Firm		0.005**		0.054***		0.010
		(0.002)		(0.010)		(0.007)
× Non-Patenting Firm		0.011***		0.071***		0.017*
		(0.002)		(0.015)		(0.009)
Controls, Industry FE, Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.432	0.433	0.569	0.569	0.099	0.098
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Controlling for other topics, K=15 (1990-2010)

	Dependent variable:					
	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		Sales Growth$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.012***		0.087***		0.020***	
	(0.002)		(0.010)		(0.004)	
× Patenting Firm		0.012***		0.088***		0.019***
		(0.002)		(0.010)		(0.005)
× Non-Patenting Firm		0.013***		0.086***		0.021***
		(0.003)		(0.016)		(0.007)
Controls, Industry FE, Year FE	X	X	X	X	X	X
Other Topics	X	X	X	X	X	X
Observations	6,066	6,066	5,933	5,933	6,070	6,070
Adjusted R^2	0.441	0.441	0.582	0.581	0.105	0.105
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 8: Accounting for Alternative Explanations (1990-2010)

Note: The specifications and variable definitions for ROA, Q, and Salesgrowth are the same as in Table 2. Panel (a) controls for analyst sentiment, panel (b) controls for the frequency of "revenue" and "growth" words, and panel (c) controls for the frequency of words with "tech" in their root. All specifications account for the standard set of other controls, industry fixed effects (4-digit SIC), and year fixed effects. Standard errors that are double clustered on firm and year are in parentheses.

(a) Controlling for average sentiment

	Dependent variable:					
	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		Sales Growth$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.009***		0.085***		0.009	
	(0.002)		(0.012)		(0.007)	
× Patenting Firm		0.009***		0.090***		0.008
		(0.003)		(0.013)		(0.007)
× Non-Patenting Firm		0.008*		0.066***		0.013
		(0.005)		(0.017)		(0.014)
Average Sentiment (Z)$_t$	0.010***	0.010***	0.047***	0.047***	0.016***	0.016***
	(0.002)	(0.002)	(0.009)	(0.009)	(0.004)	(0.004)
Controls, Industry FE, and Year FE	X	X	X	X	X	X
Observations	4,218	4,218	4,121	4,121	4,222	4,222
Adjusted R^2	0.444	0.444	0.605	0.605	0.098	0.098
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Controlling for words related to revenue and growth

	Dependent variable:					
	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		Sales Growth$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.007***		0.067***		0.011*	
	(0.002)		(0.010)		(0.006)	
× Patenting Firm		0.007***		0.070***		0.011*
		(0.002)		(0.011)		(0.006)
× Non-Patenting Firm		0.007*		0.060***		0.012
		(0.004)		(0.015)		(0.009)
Revenue Words (Z)$_t$	−0.006**	−0.006**	−0.031**	−0.031**	−0.005	−0.005
	(0.002)	(0.002)	(0.012)	(0.012)	(0.006)	(0.006)
Growth Words (Z)$_t$	0.011***	0.011***	0.075***	0.076***	0.015***	0.015***
	(0.002)	(0.002)	(0.013)	(0.013)	(0.004)	(0.004)
Controls, Industry FE, and Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.446	0.445	0.590	0.590	0.102	0.102
Note:						*p<0.1; **p<0.05; ***p<0.01

(c) Controlling for "technology" words

	Dependent variable:					
	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		Sales Growth$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.009***		0.084***		0.015***	
	(0.002)		(0.011)		(0.005)	
× Patenting Firm		0.009***		0.084***		0.015***
		(0.002)		(0.012)		(0.005)
× Non-Patenting Firm		0.010***		0.081***		0.016*
		(0.004)		(0.016)		(0.008)
Technology Words (Z)$_t$	0.0004	0.0005	0.017**	0.017**	−0.003	−0.003
	(0.002)	(0.002)	(0.009)	(0.009)	(0.006)	(0.006)
Controls, Industry FE, Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.436	0.435	0.577	0.577	0.099	0.099
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 9: Predicting Acquisition Activity Using the Text-Based Innovation Measure (1990-2010)

Note: The dependent variable in panel (a) is number of acquisitions completed in the next three years; this is the count of acquisition records from the SDC database which fall in the next three fiscal years. Panel (b) uses an alternative text-measure that is calculated without words that start with "merg" and "acqui". As in other tables, the text-based measure is standardized to have a mean of 0 and a standard deviation of 1. Other controls include log (patents), ROA, R&D intensity, log(assets), asset tangibility, leverage, log(age), log(Q), and a dummy for patenting firm. Full results are presented in the appendix (Table A.5). Variable definitions are presented in Table A.2. Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Acquisition Count – Main Innovation Measure

	Dependent variable:					
	$\text{Log}(1+\sum_{s=1}^{3}\text{\# Acquis}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{\# Big Acquis}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{\# Small Acquis}_{t+s})$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.088^{***}	0.030^{*}	0.005	0.012^{**}	0.087^{***}	0.024^{*}
	(0.019)	(0.015)	(0.005)	(0.005)	(0.019)	(0.014)
Other Controls		X		X		X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,200	6,200	6,200	6,200	6,200	6,200
Adjusted R^2	0.297	0.384	0.113	0.127	0.303	0.401
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Acquisition Count – Corpus Purged of Merger and Acquisition Words

	Dependent variable:					
	$\text{Log}(1+\sum_{s=1}^{3}\text{\# Acquis}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{\# Big Acquis}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{\# Small Acquis}_{t+s})$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.096^{***}	0.039^{**}	0.002	0.008^{*}	0.098^{***}	0.037^{**}
	(0.019)	(0.016)	(0.005)	(0.004)	(0.020)	(0.016)
Other Controls		X		X		X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,200	6,200	6,200	6,200	6,200	6,200
Adjusted R^2	0.298	0.384	0.113	0.127	0.304	0.402
Note:						*p<0.1; **p<0.05; ***p<0.01

Appendix to:

Innovation in Mature Firms: A Text-Based Analysis

A Appendix Tables and Figures

A.1 Additional Detail on LDA

Table A.1: Fit of Patenting Outcomes to Loadings for Every Topic in the 15-Topic LDA

Note: This table presents the t-statistics and adjusted R-squared on the linear relationship between a firm's patent applications and the loadings of each of the 15 topics from the fitted LDA model. Topic 6 is the innovation topic that we use for our text-based measure of innovation. This topic explains nearly two times the variation in patenting that any other topic can explain, and the word distribution is closest to the word frequencies in an innovation textbook (Tidd, Bessant, and Pavitt, 2005). Errors are double clustered on firm and year.

Topic	T-Stat	Adj R^2
6	12.372	0.047
15	8.697	0.024
12	6.915	0.015
2	6.773	0.014
11	4.718	0.007
7	2.908	0.002
10	-0.534	-0.0002
1	-3.764	0.004
5	-5.246	0.009
8	-5.722	0.010
4	-7.361	0.017
3	-7.394	0.017
13	-7.646	0.018
9	-7.888	0.020
14	-8.678	0.024

Figure A.1: Word Clouds of Two Other Fitted Topics

Note: These word clouds describe the frequency distribution of words used in the topic that is most strongly negatively correlated with patenting ("Underperforming Benchmark Topic," Topic 14), and the topic that bears the second strongest correlation with patenting ("Operating Performance Topic," Topic 15). As with the innovation topic, these topics are computed from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P500 firms. We set the number of topics in the fitted LDA model to be 15.

(a) Underperforming Benchmark Topic ($t = -8.678$)

(b) Operating Performance Topic ($t = 8.697$)

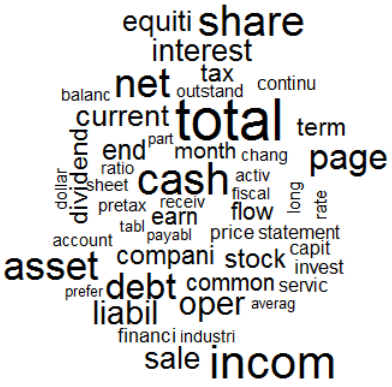
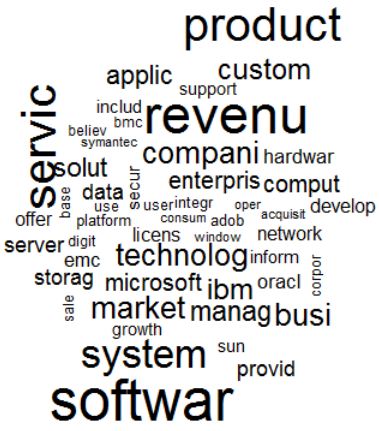


Figure A.2: Word Clouds of Two Innovation Topics from the 50-Topic LDA

Note: These word clouds describes the frequency distribution of words used in the two topics from the 50-topic LDA that are most strongly related to the innovation topic form the 15-topic LDA. As with the innovation topic, these topics are computed from the output of an Latent Dirchlet Allocation (LDA) model fit to the same corpus of analyst reports for S&P500 firms, but this time using 50 topics instead of 15.

(a) First Innovation Topic



(b) Second Innovation Topic

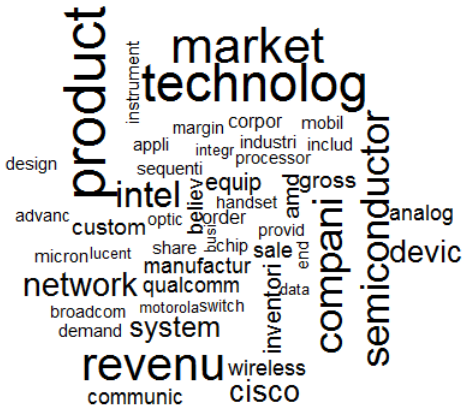


Figure A.3: Text-Based Innovation Measure: Word List

Note: This word list describes the frequency distribution of words used in the 'innovation' topic, the top 15 most common words from the topic are listed. The topic itself is from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P500 firms. We set the number of topics in the fitted LDA model to be 15, then selected the topic (out of these 15) for which the topic word distribution had the smallest Kullback-Liebler divergence with a benchmark innovation textbook (Tidd, Bessant, and Pavitt, 2005).

Word	Proportion
revenu	0.025
market	0.013
compani	0.012
servic	0.012
growth	0.011
technolog	0.009
product	0.009
network	0.009
system	0.008
softwar	0.007
data	0.007
busi	0.006
custom	0.006
wireless	0.006
total	0.006

A.2 Additional Tables and Full Results

Table A.3: Full Results on Performance of Innovative Firms (1990-2010)

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. Sales growth is defined as the percentage growth in sales between year t and year t+1 (in decimal form). The text-based innovation measure is converted to a Z-score for ease of interpretability. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Firm Performance

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.082^{***}	0.051^{***}	0.014^{**}	0.010^{**}
	(0.002)	(0.002)	(0.011)	(0.008)	(0.006)	(0.005)
$Log(Patents)_t$	0.003	-0.003	0.028^{***}	-0.002	-0.012^{***}	-0.013^{**}
	(0.002)	(0.002)	(0.010)	(0.014)	(0.003)	(0.006)
Patenting Firm	0.010^{*}		0.041		-0.001	
	(0.005)		(0.032)		(0.009)	
R&D/Assets $(Z)_t$	0.006	0.010^{**}	0.075^{***}	0.028	-0.001	-0.007
	(0.005)	(0.004)	(0.021)	(0.025)	(0.006)	(0.008)
$Log(Assets)_t$	-0.002	-0.027^{***}	-0.033^{**}	-0.207^{***}	-0.0002	-0.073^{***}
	(0.003)	(0.005)	(0.016)	(0.023)	(0.004)	(0.018)
Asset $Tangibility_t$	0.106^{***}	0.054^{**}	0.187^{**}	-0.034	-0.059	-0.305^{***}
	(0.017)	(0.022)	(0.091)	(0.108)	(0.036)	(0.106)
$Leverage_t$	-0.006	-0.008	-0.123	-0.129^{*}	-0.087^{***}	-0.059
	(0.021)	(0.020)	(0.083)	(0.069)	(0.029)	(0.040)
$Log(Age)_t$	0.004	-0.001	-0.089^{**}	-0.155	-0.028^{**}	-0.019
	(0.008)	(0.020)	(0.038)	(0.126)	(0.012)	(0.058)
$Cash/Assets_t$	0.101^{***}	0.038	0.979^{***}	0.387^{***}	0.049	0.011
	(0.031)	(0.030)	(0.124)	(0.093)	(0.050)	(0.055)
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	6,066	6,066	5,933	5,933	6,070	6,070
Adjusted R^2	0.438	0.674	0.580	0.770	0.100	0.160

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.2: Variable Definitions

Note: This table includes variable definitions and descriptions for outcome and control variables used throughout the paper. The data source is Compustat unless otherwise noted. As the main text includes a full discussion of the text-based innovation measure, the reader should refer to those sections for a description.

Variable	**Name**	**Description**
ROA	Return on assets	*EBITDA scaled by Total Assets*
Q	Tobin's Q	*Market value of equity plus total assets minus common equity and deferred taxes divided by total assets*
$Salesgrowth_t$	Sales growth	*The percentage change in sales in between year t and $t-1$ (decimal form)*
Tangibility	Asset tangibility	*Property plant and equipment divided by total assets*
Leverage	Leverage	*Total liabilities divided by assets, replacing book equity with market equity as of the last day of the fiscal year*
Age	Age	*The number of years since the first entered Compustat (earliest date 1975)*
Cash/Assets	Cash to assets ratio	*The ratio of cash to assets taken from Compustat for year t*
Patents	Patent count	*The number of patent applications in year t that correspond to an eventually granted patent*
Citations	Citation count	*The number of citations to patents applied for in year t*
Patenting Firm	Patenting Firm	*An indicator (=1) for whether a firm ever has a non-zero value of Patents.*
Patent Value	Patent Value	*The abnormal stock increase (in $millions) on the day of the granted patent (from Kogan et al. (2017))*
Products	Product Announcements	*The count of product amouncements in which the stock return exceeded the 75th percentile in Mukherjee, Singh, and Zaldokas (2016).*

Table A.3: Full Results on Performance of Innovative Firms (1990-2010)

(b) Firm Performance - Patenting Firm Split

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$						
× Patenting Firm	0.009^{***}	0.005^{***}	0.082^{***}	0.052^{***}	0.013^{**}	0.008
	(0.002)	(0.002)	(0.012)	(0.009)	(0.005)	(0.005)
× Non-Patenting Firm	0.011^{***}	0.006^{*}	0.083^{***}	0.049^{***}	0.016^{**}	0.020^{**}
	(0.004)	(0.003)	(0.016)	(0.014)	(0.008)	(0.009)
$Log(Patents)_t$	0.003	-0.003	0.028^{***}	-0.002	-0.011^{***}	-0.013^{**}
	(0.002)	(0.002)	(0.010)	(0.014)	(0.003)	(0.006)
Patenting Firm	0.010^{*}		0.041		-0.002	
	(0.005)		(0.032)		(0.009)	
$R\&D/Assets\ (Z)_t$	0.006	0.010^{**}	0.075^{***}	0.028	-0.001	-0.007
	(0.005)	(0.004)	(0.021)	(0.025)	(0.006)	(0.008)
$Log(Assets)_t$	-0.002	-0.027^{***}	-0.033^{**}	-0.207^{***}	-0.0001	-0.073^{***}
	(0.003)	(0.005)	(0.016)	(0.023)	(0.004)	(0.018)
$Asset\ Tangibility_t$	0.106^{***}	0.054^{**}	0.187^{**}	-0.034	-0.059	-0.305^{***}
	(0.017)	(0.022)	(0.091)	(0.108)	(0.036)	(0.106)
$Leverage_t$	-0.006	-0.008	-0.123	-0.129^{*}	-0.087^{***}	-0.060
	(0.021)	(0.020)	(0.083)	(0.069)	(0.029)	(0.040)
$Log(Age)_t$	0.004	-0.002	-0.089^{**}	-0.154	-0.028^{**}	-0.022
	(0.008)	(0.021)	(0.038)	(0.126)	(0.012)	(0.058)
$Cash/Assets_t$	0.101^{***}	0.038	0.979^{***}	0.387^{***}	0.050	0.011
	(0.031)	(0.030)	(0.124)	(0.093)	(0.050)	(0.055)
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	6,066	6,066	5,933	5,933	6,070	6,070
Adjusted R^2	0.438	0.674	0.579	0.770	0.100	0.160

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.4: Text-Based Innovation and R&D Expenses (1990-2010)

Note: The dependent variable is the ratio of R&D expenses to total assets. The text-based innovation measure is converted to a Z-score for ease of interpretability. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are double clustered on firm and year are reported in parentheses.

	Dependent variable:			
	R&D/Assets$_t$		R&D/Assets$_{t+1}$	
	(1)	(2)	(3)	(4)
Text-Based Innovation (Z)$_t$	0.010^{***}	0.002^{**}	0.010^{***}	0.001^{*}
	(0.002)	(0.001)	(0.002)	(0.001)
Log(Patents)$_t$		0.003^{***}		0.001
		(0.001)		(0.0004)
Patenting Firm		0.004^{**}		0.001^{*}
		(0.002)		(0.001)
Log(Assets)$_t$		-0.005^{***}		-0.001^{**}
		(0.001)		(0.001)
Return on Assets$_t$		-0.020		-0.003
		(0.020)		(0.008)
Asset Tangibility$_t$		0.001		-0.005
		(0.008)		(0.004)
Leverage$_t$		0.005		-0.004
		(0.005)		(0.004)
Log(Age)$_t$		-0.002		-0.002
		(0.003)		(0.001)
R&D/Assets$_t$				0.680^{***}
				(0.083)
Log(Q)$_t$		0.010^{***}		0.003^{**}
		(0.003)		(0.001)
4-digit SIC Dummies	X	X	X	X
Year FE	X	X	X	X
Observations	6,201	6,201	6,075	6,075
Adjusted R^2	0.450	0.713	0.433	0.817
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A.5: Full Results on Predicting Acquisition Activity (1990-2010)

Note: The dependent variable in panel (a) is number of acquisitions completed in the next three years; this is the count of acquisition records from the SDC database which fall in the next three fiscal years. Panel (b) uses an alternative text-measure that is calculated without words that start with "merg" and "acqui".The text-based innovation measure is converted to a Z-score for ease of interpretability. Return on assets is EBITDA scaled by total assets. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Acquisition Count

	Dependent variable:					
	$\text{Log}(1+\sum_{s=1}^{3}\text{\# Acquisitions}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{\# Big Acquisitions}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\text{\# Small Acquisitions}_{t+s})$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation (Z)$_t$	0.088^{***}	0.030^{*}	0.005	0.012^{**}	0.087^{***}	0.024^{*}
	(0.019)	(0.015)	(0.005)	(0.005)	(0.019)	(0.014)
Log(Patents)$_t$		0.062^{***}		0.002		0.062^{***}
		(0.018)		(0.005)		(0.017)
ROA$_t$		0.642^{**}		-0.014		0.647^{**}
		(0.268)		(0.067)		(0.264)
R&D/Assets$_t$		-0.862		-0.548^{***}		-0.560
		(0.729)		(0.158)		(0.754)
Log(Assets)$_t$		0.181^{***}		-0.026^{***}		0.198^{***}
		(0.026)		(0.006)		(0.026)
Asset Tangibility$_t$		-0.362^{***}		-0.072^{**}		-0.299^{**}
		(0.123)		(0.030)		(0.124)
Leverage$_t$		-0.538^{***}		-0.066^{**}		-0.497^{***}
		(0.082)		(0.030)		(0.078)
Log(Age)$_t$		0.103^{**}		-0.009		0.113^{**}
		(0.051)		(0.011)		(0.049)
Log(Q)$_t$		0.200^{***}		-0.038^{***}		0.229^{***}
		(0.044)		(0.011)		(0.044)
Patenting Firm		-0.112^{***}		-0.001		-0.105^{***}
		(0.039)		(0.013)		(0.038)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,200	6,200	6,200	6,200	6,200	6,200
Adjusted R^2	0.297	0.384	0.113	0.127	0.303	0.401
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.5: Full Results on Predicting Acquisition Activity (1990-2010)

(b) Acquisition Count - Purged Corpus

	Dependent variable:					
	$\text{Log}(1+\sum_{s=1}^{3}\#\text{ Acquisitions}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\#\text{ Big Acquisitions}_{t+s})$		$\text{Log}(1+\sum_{s=1}^{3}\#\text{ Small Acquisitions}_{t+s})$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.096^{***}	0.039^{**}	0.002	0.008^{*}	0.098^{***}	0.037^{**}
	(0.019)	(0.016)	(0.005)	(0.004)	(0.020)	(0.016)
Log(Patents)_t		0.062^{***}		0.002		0.063^{***}
		(0.018)		(0.005)		(0.017)
ROA_t		0.633^{**}		−0.016		0.640^{**}
		(0.267)		(0.067)		(0.263)
$\text{R\&D/Assets}_t$		−0.894		-0.546^{***}		−0.596
		(0.733)		(0.157)		(0.758)
Log(Assets)_t		0.180^{***}		-0.025^{***}		0.196^{***}
		(0.026)		(0.006)		(0.026)
$\text{Asset Tangibility}_t$		-0.363^{***}		-0.074^{**}		-0.299^{**}
		(0.123)		(0.030)		(0.124)
Leverage_t		-0.532^{***}		-0.067^{**}		-0.490^{***}
		(0.082)		(0.030)		(0.079)
Log(Age)_t		0.104^{**}		−0.008		0.114^{**}
		(0.051)		(0.011)		(0.049)
Log(Q)_t		0.198^{***}		-0.036^{***}		0.225^{***}
		(0.044)		(0.011)		(0.044)
Patenting Firm		-0.113^{***}		−0.001		-0.106^{***}
		(0.039)		(0.013)		(0.039)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,200	6,200	6,200	6,200	6,200	6,200
Adjusted R^2	0.298	0.384	0.113	0.127	0.304	0.402

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.6: Predicting Acquisition Activity - LPM (1990-2010)

Note: The dependent variable is an indicator variable that is set to 1 if there is an acquisition in the next year. Panel (a) uses the main text-based innovation measure. Panel (b) uses an alternative text-measure that is calculated without words that start with "merg" and "acqui". The text-based innovation measure is converted to a Z-score for ease of interpretability. Return on assets is EBITDA scaled by total assets. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are double clustered on firm and year are reported in parentheses.

(a) Linear Probability Model – Main Innovation Measure

	Dependent variable:					
	I(Acquisition)$_{t+1}$		I(Big Acquisition)$_{t+1}$		I(Small Acquisition)$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation (Z)$_t$	0.036***	0.014*	0.001	0.005	0.035***	0.010
	(0.008)	(0.007)	(0.003)	(0.004)	(0.008)	(0.007)
Log(Patents)$_t$		0.015**		0.001		0.016**
		(0.007)		(0.003)		(0.007)
ROA$_t$		0.241*		0.035		0.216*
		(0.132)		(0.053)		(0.131)
R&D/Assets$_t$		−0.529*		−0.274***		−0.392
		(0.282)		(0.099)		(0.301)
Log(Assets)$_t$		0.068***		−0.015***		0.078***
		(0.012)		(0.004)		(0.012)
Asset Tangibility$_t$		−0.174***		−0.035**		−0.145**
		(0.061)		(0.018)		(0.060)
Leverage$_t$		−0.251***		−0.024		−0.241***
		(0.043)		(0.022)		(0.036)
Log(Age)$_t$		0.065***		−0.005		0.070***
		(0.023)		(0.005)		(0.022)
Log(Q)$_t$		0.076***		−0.021**		0.090***
		(0.020)		(0.009)		(0.021)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,074	6,074	6,074	6,074	6,074	6,074
Adjusted R^2	0.165	0.202	0.032	0.038	0.170	0.216
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.6: Predicting Acquisition Activity - LPM (1990-2010)

(b) Linear Probability Model- Corpus Purged of Merger and Acquisition Words

	Dependent variable:					
	I(Acquisition)$_{t+1}$		I(Big Acquisition)$_{t+1}$		I(Small Acquisition)$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation (Z)$_t$	0.039^{***}	0.016^{**}	0.0004	0.004	0.039^{***}	0.014^{*}
	(0.008)	(0.007)	(0.002)	(0.003)	(0.008)	(0.007)
Log(Patents)$_t$		0.015^{**}		0.001		0.017^{**}
		(0.007)		(0.003)		(0.007)
ROA$_t$		0.237^{*}		0.034		0.213
		(0.132)		(0.054)		(0.131)
R&D/Assets$_t$		-0.539^{*}		-0.274^{***}		-0.404
		(0.282)		(0.099)		(0.302)
Log(Assets)$_t$		0.068^{***}		-0.015^{***}		0.077^{***}
		(0.012)		(0.004)		(0.012)
Asset Tangibility$_t$		-0.175^{***}		-0.036^{**}		-0.145^{**}
		(0.061)		(0.018)		(0.060)
Leverage$_t$		-0.249^{***}		-0.024		-0.239^{***}
		(0.043)		(0.022)		(0.037)
Log(Age)$_t$		0.065^{***}		-0.005		0.070^{***}
		(0.023)		(0.005)		(0.022)
Log(Q)$_t$		0.076^{***}		-0.020^{**}		0.089^{***}
		(0.020)		(0.009)		(0.021)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,074	6,074	6,074	6,074	6,074	6,074
Adjusted R^2	0.165	0.202	0.032	0.038	0.171	0.216

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.7: Relation of Text-Based Innovation to Merger Announcement CARs (-1 to +1 day)

Note: This table presents OLS regressions relating text-based innovation to subsequent M&A cumulative abnormal returns in a 3-day window (-1,1) around the merger announcement date. Small Acquisitions are acquisitions in which the deal value is less than 5 percent of the acquirer pre-merger value. As in other specifications, the text-based measure is standardized to have a mean of 0 and a standard deviation of 1. Standard errors that are double clustered on firm and year are reported in parentheses.

	Dependent variable:			
	CAR			
	(1)	(2)	(3)	(4)
Text-Innovation $(Z)_t$	-0.001	-0.010^{*}	-0.010^{**}	-0.012^{*}
	(0.001)	(0.006)	(0.005)	(0.006)
Text-Innovation $(Z)_t \times$ SmallAcq		0.009^{*}	0.010^{**}	0.012^{**}
		(0.005)	(0.005)	(0.006)
SmallAcq		-0.005	-0.005	-0.005
		(0.005)	(0.005)	(0.005)
Log(Patents)_t			-0.001	-0.001
			(0.001)	(0.001)
$\text{Return on Assets}_t$			0.014	0.038
			(0.018)	(0.029)
$\text{R\&D/Assets}_t$			0.035	0.038
			(0.031)	(0.076)
Log(Assets)_t			0.002	0.003
			(0.002)	(0.004)
$\text{Asset Tangibility}_t$			-0.008	-0.034
			(0.010)	(0.024)
Leverage_t			0.001	0.009
			(0.009)	(0.013)
Log(Age)_t			-0.004	-0.022
			(0.004)	(0.020)
Log(Q)_t			-0.005	-0.011^{*}
			(0.004)	(0.006)
Cash/Assets_t			-0.003	-0.015
			(0.010)	(0.013)
4-digit SIC Dummies	X	X	X	
Firm FE				X
Year FE	X	X	X	X
Observations	3,793	3,793	3,793	3,793
Adjusted R^2	0.065	0.068	0.068	0.154
Note:				*p<0.1; **p<0.05; ***p<0.01

A.3 Alternative Scaling of the Topic Loadings in Building the Text-Based Innovation Measure

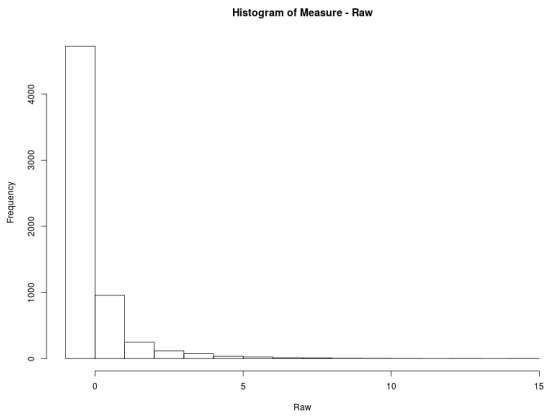
This appendix presents some additional detail on the scaling of the innovation topic loadings underlying the text-based innovation measure. Our primary measure transforms the topic loadings by taking the fourth root before applying the Loughran and McDonald (2011) sentiment filter. We use the fourth root transformation to mitigate skew in the text-based innovation measure.

To highlight the effect of the fourth root transformation, Figure A.4 presents histograms of the topic loadings across analyst reports for the raw topic loadings, an inverse hyperbolic sine transformation (IHS, approximately log), and the fourth root transformation. Both the raw topic loadings and the IHS transformation yield a measure that is highly skewed, whereas the fourth root transformation mitigates the skew, and accordingly should produce a measure with better properties. On this basis, we construct the text-based innovation measure using the less-skewed fourth root transformation.

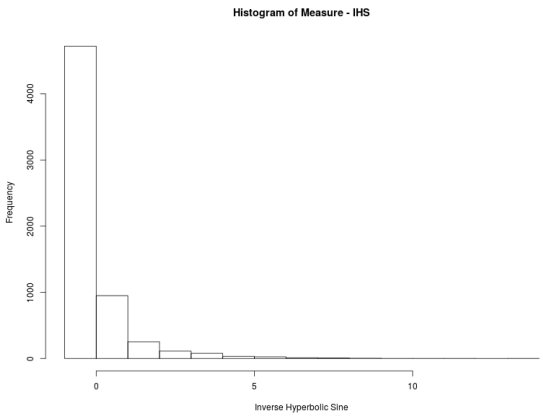
Figure A.4: Histograms of Innovation Topic Loadings and Transformations

Note: This figure presents sample histograms of the topic loadings used as a basis for the text-based innovation measure. In panel (a), the measure is the raw mean of the innovation topic loading for positive analyst reports about the firm over the fiscal year. Panel (b) uses the inverse hyperbolic sine transformation of the raw measure. Panel (c) uses the fourth root of the raw measure, as does the bulk of the paper.

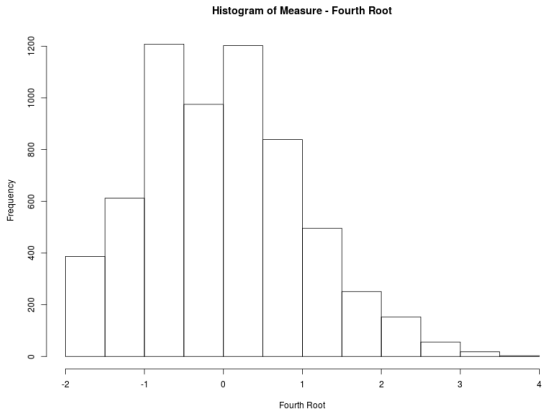
(a) Raw Topic Loadings



(b) Inverse Hyperbolic Sine of Topic Loadings



(c) Fourth Root of Topic Loadings



As a robustness check, we also recreate the measure based on the underlying skewed distributions (using both raw loadings, and IHS transformed loadings) and use these alternative measures in our main specifications. Table A.8 presents the main results using these alternative measures. The sign and significance of the main results are broadly consistent with our main measure, but the estimates tend to be more precise and stable using our primary measure, which confirms the rationale for using the less skewed measure in the first place.

	Dependent variable:					
	ROA_{t+1}	$Log(Q)_{t+1}$	SG_{t+1}	PV_{t+1}	V/P_{t+1}	FCite/FPat
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation IHS $(Z)_t$	0.006^{**}	0.072^{***}	0.009	0.054^{*}	0.063^{**}	0.033
	(0.003)	(0.013)	(0.007)	(0.029)	(0.025)	(0.023)
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,064	5,931	6,068	3,249	2,998	3,208
Adjusted R^2	0.432	0.574	0.098	0.805	0.712	0.667
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Main Results Using Inverse Hyperbolic Sine of Topic Loadings

Table A.8: Results Using Alternative Scaling of the Topic Loading to Construct the Text-Based Innovation Measure

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. SG is sales growth and is defined as the percentage growth in sales between year t and year t+1 (in decimal form). PV is the log of patent value, which it is defined as the stock market jump on the day of the granted patent (in millions) aggregated over all patents granted during the year (see Kogan et al. (2017) for details). V/P is value per patent which is computed as the log of patent value divided by number of patents plus one. FCite/FPat is future citations over future patents, it is computed as the log of the ratio between the number of cites and the number of patents granted in the next three years. The text-based innovation measure is the mean of the innovation topic loading for positive analyst reports about the firm over the fiscal year. In panel (b), the inverse hyperbolic sine of the measure is used. Errors are double clustered on firm and year.

(a) Main Results Using Raw Topic Loadings

	Dependent variable:					
	ROA_{t+1}	$Log(Q)_{t+1}$	SG_{t+1}	PV_{t+1}	V/P_{t+1}	FCite/FPat
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation (raw) $(Z)_t$	0.006^{**}	0.072^{***}	0.009	0.053^{*}	0.062^{**}	0.032
	(0.003)	(0.013)	(0.007)	(0.029)	(0.024)	(0.023)
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,064	5,931	6,068	3,249	2,998	3,208
Adjusted R^2	0.432	0.574	0.098	0.805	0.712	0.667
Note:						*p<0.1; **p<0.05; ***p<0.01

A.4 Long-Term Dynamics

Table A.8: Long-Term Tobin's Q, ROA, and Salesgrowth Using the Text-Based Innovation Measure

Note: Return on assets is EBITDA scaled by total assets. The text-based innovation measure is converted to a Z-score for ease of interpretability. All firms that have at least one patent during the sample period (1990-2004) are included in the regression. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Errors are double clustered on firm and year.

	Dependent variable:								
	ROA_{t+2}	ROA_{t+3}	ROA_{t+4}	$Log(Q)_{t+2}$	$Log(Q)_{t+3}$	$Log(Q)_{t+4}$	$Salesgrowth_{t+2}$	$Salesgrowth_{t+3}$	$Salesgrowth_{t+4}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation $(Z)_t$	0.006^{***}	0.005^{**}	0.004^{**}	0.066^{***}	0.062^{***}	0.056^{***}	-0.001	-0.003	-0.003
	(0.002)	(0.002)	(0.002)	(0.010)	(0.011)	(0.012)	(0.006)	(0.005)	(0.003)
Log(Patents) $(Z)_t$	0.009^{*}	0.012^{**}	0.011^{**}	0.055^{*}	0.063^{**}	0.065^{*}	0.013	0.023^{**}	-0.005
	(0.006)	(0.006)	(0.006)	(0.032)	(0.032)	(0.034)	(0.012)	(0.010)	(0.017)
Patenting Firm	0.006	0.002	0.0002	0.083^{***}	0.078^{***}	0.069^{***}	-0.003	-0.008	-0.011^{*}
	(0.005)	(0.004)	(0.004)	(0.023)	(0.022)	(0.023)	(0.007)	(0.006)	(0.006)
R&D/Assets $(Z)_t$	-0.001	-0.0002	0.001	-0.033^{**}	-0.029^{*}	-0.020	-0.017^{***}	-0.019^{***}	-0.014^{***}
	(0.003)	(0.003)	(0.003)	(0.017)	(0.017)	(0.018)	(0.007)	(0.007)	(0.004)
ROA_t	0.102^{***}	0.094^{***}	0.084^{***}	0.206^{**}	0.182^{**}	0.175^{*}	0.026	0.041	0.032
	(0.016)	(0.016)	(0.017)	(0.090)	(0.087)	(0.090)	(0.042)	(0.042)	(0.041)
$Log(Assets)_t$	-0.002	-0.005	-0.011	-0.046	-0.008	0.020	-0.118^{***}	-0.067^{***}	-0.062^{***}
	(0.019)	(0.017)	(0.017)	(0.074)	(0.075)	(0.076)	(0.024)	(0.018)	(0.020)
Asset $Tangibility_t$	0.001	-0.005	-0.008	-0.085^{**}	-0.098^{***}	-0.100^{***}	-0.021^{**}	-0.028^{***}	-0.010
	(0.008)	(0.007)	(0.007)	(0.034)	(0.036)	(0.036)	(0.010)	(0.009)	(0.011)
$Leverage_t$	0.100^{***}	0.093^{***}	0.078^{***}	0.856^{***}	0.806^{***}	0.725^{***}	0.042	0.082^{***}	0.104^{***}
	(0.028)	(0.028)	(0.029)	(0.126)	(0.125)	(0.128)	(0.039)	(0.031)	(0.040)
$Log(Age)_t$	0.004^{*}	0.004^{*}	0.003^{*}	0.031^{***}	0.026^{***}	0.026^{***}	-0.007^{*}	-0.004	-0.003
	(0.002)	(0.002)	(0.002)	(0.010)	(0.010)	(0.010)	(0.004)	(0.003)	(0.003)
4-digit SIC Dummies	X	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X	X
Observations	5,946	5,787	5,359	5,704	5,476	5,003	5,946	5,786	5,358
Adjusted R^2	0.452	0.445	0.454	0.570	0.572	0.573	0.090	0.089	0.098

Note: *p<0.1; **p<0.05; ***p<0.01

A.5 Word-List Measure versus Latent Dirchlet Allocation

An alternative technique for constructing a measure of innovation from text would be to create a word-list of words related to the idea of innovation. Using a word list of "innovation words," we could measure innovation in one of several ways, for example by counting the number of "innovative words" in each document scaled by the length of the document. As we will see, such an approach — though intuitive — suffers from a number of important limitations . Within the word-list paradigm of textual analysis, there are techniques to overcome these limitations, but these techniques lead to an increase in complexity, and an unsatisfactory level of researcher subjectivity. Our LDA-based method addresses these limitations in a different way, which allows us to avoid any influence of subjectivity on the part of the researcher. In this section, we build the simple word-list measure from the text of analyst reports, and by comparison, highlight some of the strengths of the LDA approach versus an augmented word-list approach..

The first challenge facing word-list approaches is to identify an appropriate list of words for the innovative word-list. Rather than hand classify words that are innovative versus not, we create an objective list by using Princeton University's WordNet database. WordNet is a lexical database available from Princeton University in which nouns, verbs, adjectives, and adverbs are grouped into so called synsets. Each synset contains a set of words with the same distinct meaning (a word is a member of multiple synsets if it has several distinct meanings). A synset represents a unique 'concept'. The database is built as a hierarchy where specific concepts are grouped under more general concepts. For example, rabbit would be grouped under mammals, which are grouped under animals, etc up to the root node 'entity' for all nouns. This type of relation is called hyponomy (or is-a relation, since a rabbit 'is a' mammal), and is the most commonly encoded relation in the WordNet database.[12] We filter out adjectives and adverbs for simplicity of the word-list construction.

To construct a list of "innovation words," we compute the relatedness between 'innovation' or 'innovate' and all other words in the WordNet database (the two are computed separately),[13] and

[12] Verbs are also grouped into hierarchies, such as hierarchies where the meaning gets more specific (in some sense) further down the tree. Verbs with opposite meaning are linked. In addition to hyponomy, the meronomy relation between nouns is classified, i.e. a part-whole relation

[13] The synset for 'innovation' is defined as 'a creation (a new device or process) resulting from study and experimenta-

restrict attention to the top 1% words of most related words. Specifically, we use the Jiang and Conrath (1997) distance to calculate how related two synsets are with each other. To obtain the Jiang and Conrath (1997) distance between two synsets, we compute the sum of all vertices between two synsets in the hierarchy, scaled by their information content. This is calculated as using the least common subsumer, the least general concept that encompasses both synsets. The formula is $JC_D = IC(a) + IC(b) - 2IC(lcs)$, where a and b denote the two synsets. The inverse of the distance is used as the relatedness measure.

Many words have multiple synsets, which indicates that these words have multiple meanings depending on context (e.g., "case" can mean "a small container," "to examine or check out," or "an instance or occurance"). Such words lead to noise in classifying whether words are truly corresponding to their innovative meaning, a problem that we do not have with the LDA-method, which groups words automatically depending on the context that is inferred from the structure of the document. In constructing the word-list measure, we partially address the multiple-meaning problem by using the highest relatedness score to capture the word most closely associated with innovation, but even this solution introduces noise to the extent that analysts are not always using words to mean their most innovative meaning.

We take the resulting word list and measure its similarity with each of our analyst reports by counting how many innovation words each document contains and scaling it by the document length.[14] For consistency with our main LDA-based measure, we aggregate the word-list measure across analyst reports written about the same firm in the same fiscal year for positive reports only (sentiment above the 75th percentile). Tables A.9 and A.10 respectivley present the results performance regressions and patenting regressions that are setup analogously to the tests in the paper. Following the analysis in the main text, we estimate following specifications:,

$$Performance_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{9}$$

tion'. The synset for 'innovate' is defined as 'bring something new to an environment'.

[14] A popular alternative is to use cosine similarity as in Hoberg and Phillips (2016).

and

$$Patenting_{it+1} \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{10}$$

where $Performance_{ti+1}$ is one of operating performance, log of Q, or salesgrowth; and $Patenting_{ti+1}$ is one of $Log(1+PatentValue_{ti+1})$, $Log(1+ValuePerPatent_{ti+1})$, or the log of the ratio of citations to patents over the next three years.

Results in Table A.9 show that this word-list based measure predicts future performance in a way that is quite similar to our LDA-based measure, both in terms of significance and magnitudes, which is consistent with how we think of innovation. Nevertheless, the word-list measure fails to correlate in a meaningful manner with more direct measures of innovation. For example, Table A.10 shows that the simplistic word-list measure fails to capture the value of patented innovation, and thus fails our tests that are designed to check whether valuable patented innovation is predicted by the measure of innovation.

It is plausible that the noise introduced by words with multiple meanings leads to enough noise that the word-list measure does not significantly predict the relevant patenting measures. Indeed, the coefficient estimates are of the same sign, just smaller in magnitude and less precisely estimated, by comparison to our LDA-based measure. In this case, refinements of the word-list measure could enhance precision on this dimension. In this spirit, one potential refinement of the word-list measure is called word-sense disambiguation, which is an algorithm aimed at finding the correct meaning of a word in a text. Using a limited sample of analyst reports and firms, we have used a simple Lesk algorithm in this spirit, and though it appears to work well, there is no compelling reason to use an augmented word-list algorithm in this vein over LDA because the augmented word-list algorithm is just as complex, it takes slightly longer to estimate, and it involves more researcher-directed choices that could ultimately influence the results. By contrast, LDA — though complex to estimate — requires much less researcher-input (only the number of topics is selected by the researcher), leading to a stronger, more objective text-based measure of innovation.

Table A.9: Patent Value, Word-List Measure (1990-2010)

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. Sales growth is defined as the percentage growth in sales between year t and year t+1 (in decimal form). In these tables, we compute the text-based innovation measure analogously as the mean of the innovation word-list loading for positive analyst reports about the firm over the fiscal year. To be consistent with the main measure, we take the fourth root of this measure and convert it to a Z-score. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Errors are double clustered on firm and year.

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
WordList-Innovation $(Z)_t$	0.011^{***}	0.006^{***}	0.073^{***}	0.040^{***}	0.018^{***}	0.015^{***}
	(0.002)	(0.001)	(0.010)	(0.007)	(0.003)	(0.003)
Patenting Firm	0.009		0.039		-0.003	
	(0.005)		(0.031)		(0.009)	
$Log(Patents)_t$	0.002	-0.003	0.022^{**}	-0.006	-0.012^{***}	-0.013^{**}
	(0.002)	(0.002)	(0.011)	(0.014)	(0.003)	(0.006)
$R\&D/Assets\ (Z)_t$	0.006	0.010^{**}	0.079^{***}	0.030	-0.001	-0.006
	(0.005)	(0.004)	(0.022)	(0.025)	(0.005)	(0.008)
$Log(Assets)_t$	-0.0004	-0.026^{***}	-0.022	-0.202^{***}	0.003	-0.069^{***}
	(0.003)	(0.005)	(0.016)	(0.022)	(0.005)	(0.018)
$Asset\ Tangibility_t$	0.102^{***}	0.055^{**}	0.158^{*}	-0.033	-0.061^{*}	-0.305^{***}
	(0.017)	(0.022)	(0.092)	(0.110)	(0.036)	(0.105)
$Leverage_t$	-0.007	-0.007	-0.132	-0.138^{*}	-0.083^{***}	-0.052
	(0.021)	(0.020)	(0.083)	(0.072)	(0.030)	(0.040)
$Log(Age)_t$	0.002	-0.005	-0.083^{**}	-0.166	-0.024^{**}	-0.024
	(0.007)	(0.018)	(0.032)	(0.115)	(0.010)	(0.051)
$Cash/Assets_t$	0.105^{***}	0.040	0.998^{***}	0.397^{***}	0.061	0.017
	(0.030)	(0.029)	(0.121)	(0.094)	(0.049)	(0.054)
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.441	0.676	0.577	0.770	0.102	0.161
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.10: Patent Value, Word-List Measure (1990-2010)

Note: The dependent variable is a patent value measure. The first four columns aggregate the value of all patents granted during the year, scaled by patent count in columns 3 and 4. Columns 5 and 6 use the citation weighted patents over the next three years as the measure of patent value. We use patent value data from Kogan et al. (2017) calculated as the abnormal stock market jump (in millions of dollars) on the day of a granted patent. We aggregate these patent values over the fiscal year. In these tables, we compute the text-based innovation measure analogously as the mean of the innovation word-list loading for positive analyst reports about the firm over the fiscal year. To be consistent with the main measure, we take the fourth root of this measure and convert it to a Z-score. Patents is the count of granted patents which were applied for during the year. Other controls are R&D intensity, leverage, the log of total assets, the log of age, and the log of Q. Errors are double clustered on firm and year.

	Dependent variable:					
	Log(1 + Patent Value)$_t$		Log(1 + Value per Patent)$_t$		Log(1 + $\frac{\sum_{s=1}^{3} \text{Citations}_{t+s}}{\sum_{s=1}^{3} \text{Patents}_{t+s}}$)	
	(1)	(2)	(3)	(4)	(5)	(6)
Text Innovation$_t$	0.017	0.033*	0.035	0.045**	0.0005	0.010
	(0.017)	(0.017)	(0.022)	(0.017)	(0.015)	(0.014)
Log(1 + Patents)$_t$	0.670***	0.746***				
	(0.032)	(0.042)				
Log(1 + Citations)$_t$ (Z)	0.368***	0.186***	0.124***	0.049*		
	(0.050)	(0.051)	(0.036)	(0.026)		
R&D/Assets$_t$	0.859	0.555	−1.211*	−0.037	−1.411**	−0.277
	(0.549)	(0.526)	(0.674)	(0.682)	(0.593)	(0.716)
Levarage$_t$	−0.791***	−0.687***	−0.845***	−0.767***	−0.053	−0.044
	(0.197)	(0.154)	(0.203)	(0.180)	(0.173)	(0.181)
Log(Assets)$_t$	0.823***	0.740***	0.420***	0.452***	−0.145***	−0.202***
	(0.049)	(0.070)	(0.041)	(0.073)	(0.039)	(0.061)
Log(Age)$_t$	−0.068	−0.180	−0.217**	−0.660*	−0.351***	−1.032***
	(0.112)	(0.326)	(0.108)	(0.383)	(0.084)	(0.291)
Log(Q)$_t$	1.011***	0.909***	0.966***	0.931***	0.156***	−0.0005
	(0.069)	(0.089)	(0.064)	(0.072)	(0.054)	(0.068)
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	3,587	3,587	2,999	2,999	3,209	3,209
Adjusted R^2	0.888	0.934	0.710	0.837	0.666	0.799
Note:						*p<0.1; **p<0.05; ***p<0.01