A Text-Based Analysis of Corporate Innovation

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

February 1, 2017

Abstract

We construct a new text-based measure of innovation using the content of analyst reports of S&P 500 firms. Our text-based measure captures innovation that is not measured by existing proxies, which is the case when innovation is not financed by R&D and is not patented. The text-based innovation measure is useful even within the set of patenting firms because it strongly correlates with valuable patents, which likely capture true innovation. Indeed, the text-based innovation measure is robustly related to greater firm performance and growth opportunities for up to four years, and these value implications hold just as strongly for non-patenting firms. Digging deeper, highly-innovative firms according to our text-based measure become innovative by producing innovative systems (e.g., Walmart's cross-geography logistics). Consistent with this interpretation, we find that highly-innovative firms are more acquisitive, using acquisitions of relatively smaller firms to augment their innovative systems. Taken together, these findings provide deeper insight into the value of innovation more broadly, not just innovation that can be patented.

[*] All authors are from University of Colorado's Leeds School of Business. Cookson is the corresponding author (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), Katie Moon, Ting Xu (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 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

Inventive activity has long been thought to play a central role both for economic growth and short-term fluctuations (Schumpeter, 1939; Kuznets and Murphy, 1966).[1] Owing to its fundamental importance, innovation has attracted significant academic attention (e.g., Hall, 1990; Bhagat and Welch, 1995; Nordhaus, 1969; Brown et al., 2009; Tian and Wang, 2011; Moser, 2012). Nevertheless, our empirical understanding of innovation is incomplete because existing proxies for innovation – typically, R&D intensity or outcomes related to patenting – do not fully capture the nature and scope of innovative output.

According to Drucker (1985), "innovation is the specific instrument of entrepreneurship. It is the act that endows resources with a new capacity to create wealth. Innovation, indeed, creates a resource." Taking this view, innovation is important for firm performance across many industries, and it can be accomplished in many different ways. In contrast to this general nature of innovation itself, 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, see Tian and Wang, 2011).

We fill this gap between the specificity of innovation proxies and the general nature of innovation by proposing a new measure of innovation that infers the amount of corporate innovation using the content of text descriptions of firm activities by the financial analysts. Our measure encapsulates what analysts *broadly* describe as innovative processes, components of products, and systems. 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 (Hall et al., 2014; Cooper et al., 2015), and some patents are filed for defensive reasons (e.g., see work on 'patent trolls' by Tucker, 2014, and Cohen, Gurun and Kominers, 2014). Because we use the language of external observers, our measure does not suffer from this set of biases. Indeed, we show using our measure that much of the value of corporate innovation is overlooked when attention is restricted

[1] Solow (1956) postulated an economy's output growth as a positive function of physical and human capital growth, showing that deviations of an economy's actual output growth from the implied growth would be attributed to changes in technology and institutional change. For countries with a stable political climate, deviations of actual output growth from the implied growth would be attributed to growth in total factor productivity (TFP). Solow (1956) found that more than 90 percent of output growth per worker in the U.S. could be attributed to TFP. More recent TFP estimates for the U.S. are lower, but still substantial, 34 percent; for example, see Jones (2002), and Baier, Dwyer and Tamura (2006).

only to patented innovations.[2]

We construct the text-based measure of innovation using topic modeling tools that have been recently introduced to the finance literature (Hoberg and Lewis, 2015; Ganglmair and Wardlaw, 2015; Goldsmith-Pinkham et al., 2016). Specifically, we employ the Latent Dirichlet Allocation (LDA) method of Blei et al. (2003) on the text of a large corpus of analyst reports. The underlying assumption behind LDA is that each analyst report is generated by drawing content from a common set of topics, or clusters of words. According to this modeling intuition, analyst reports have different content because they reflect a different mix of these underlying topics. A fitted LDA model recovers the set of topics (common across analyst reports) that best describe the empirical distribution of word groupings across analyst reports. The LDA routine 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 standard word-list techniques. The fitted LDA also provides an intensity with which each analyst report discusses each topic. After fitting the LDA model, we inspect content of the resulting topics, identifying the topic that best reflects the nature of innovation (details below). With this topic in hand, we measure the level of a firm's corporate innovation by the intensity with which analysts write about the innovation topic.

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 723 firms that were in the S&P 500 during 1990-2012. From this fitted topic model, one topic stands out as a measure of innovation, both qualitatively and quantitatively. Qualitatively, the words analysts use to describe the firm are also those that should describe innovations (e.g., product, service, system, technology, solution). Quantitatively, the topic correlates strongly with existing innovation measures for the set of patenting firms. Beyond this basic correlation, which helped us identify the innovation topic, the innovation topic also exhibits sensible time-series and cross-industry patterns that correspond to a well-performing measure of innovation. For our 1990-2012 sample, the measure tracks closely with the the aggregate time series of R&D intensity (e.g., the R&D boom and bust of the 1990s and early 2000s studied in Brown

[2] There are well-known selection biases that arise from focusing on patenting outcomes. For example, Hall et al. (2014) argue that patents are more important for firms that create "discrete" products, such as pharmaceuticals and chemicals, and that patents are least important for process innovations since process innovations often are easy to innovate around. Patents are more beneficial when they offer higher degrees of protection, all else equal. Higher protection is related to how the innovation can be delimited, the ease of innovating around the patent, and legal enforcement. In this vein, Saidi and Zaldokas (2016) provide evidence that patenting and trade secrets are substitutes depending on disclosure requirements for patenting, a finding that indicates that a significant amount of innovation is not patented.

et al., 2009), bearing a 0.51 time series correlation with aggregate R&D expenditures. Across industries, the measure also exhibits a strong correlation with R&D intensity (0.47 correlation). Even within industry, the text-based innovation measure correlates strongly with R&D expenditures, and provides a useful forecast for future R&D expenditures after accounting for conventional determinants of R&D intensity (i.e., tangibility, size, Tobin's Q, and age). Although patenting outcomes are also correlated with future R&D expenditures, the relation between R&D to patenting is less robust than the relation between R&D and the text-based innovation measure.These findings imply that the correlation with patenting measures of innovation does not drive our findings for our text-based measure of innovation. In other words, the insights gleaned from our measure could not be obtained from analogous patenting measures, likely because patents cover a limited set of innovations, and do not always reflect innovative expertise, but other factors like strategic patenting or patent collateral (e.g., see Cohen et al., 2014; Mann, 2016).

An important advantage of our text-based innovation measure is that it can be computed for firms that do not patent, which expands the scope of innovation that can be studied. To show the measure is useful for these firms, we present tangible examples of content from analyst reports for non-patenting firms that score high on our measure. For example, Walmart, which scores high on our measure, is an informative example of the value of our approach. Walmart did not apply for many patents during the 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 extensively.

Turning to corporate valuation implications of text-based innovation, higher levels of innovation forecast an increase in future operating performance, as well as an increase in measured growth opportunities embedded in Tobin's Q, results that are robust to firm fixed effects. These increases in performance likely stem from the innovations themselves, and from synergies with other aspects of the corporation. 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 traditionally innovative firms.

We provide additional evidence on the nature of our innovation measure. Within the set of patenting firms, we find that our text-based measure strongly correlates with the Kogan et al. (2012) patent valuation measure. That is, our text-based approach enables us to distinguish true innovation captured by valuable patents from patenting outcomes that do not reflect true innovation (e.g., those filed for strategic purposes). Extending beyond the set of patenting firms, we note that our measure does not rely on patenting outcomes, and is useful to capture innovations beyond patenting firms.

Appealing to the text of the topic itself, text-based innovation captures whether a company has successfully constructed an innovative system or platform rather than producing differentiated products themselves. This idea of innovative systems has been conceptually identified as being important (see Egan, 2013), but traditional measurements have not captured this idea quantitatively. Consistent with the innovative systems interpretation, we find that innovative firms are more acquisitive, especially of smaller firms. This increase in acquisitions tilted toward smaller firms makes economic sense because smaller firms likely have components that augment well the innovative firm's system.

The text-based innovation measure 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 et al. (2010) for discussions and criticism of this method). Third, our measure is not subject to concerns about strategic disclosure of patents using the stock market response to announcements of a patent award. In fact, because we focus on the language of analysts who are unlikely to time their reports, we avoid other sources of bias from firm disclosures as well.[3]

[3] Specifically, our measure is less susceptible to endogeneity concerns compared to measures constructed from managerial disclosures, such as annual reports, since analysts are outsiders vis-à-vis firm managers.

Our work contributes to an emerging line of research that draws a distinction between patenting measures and innovation (e.g., Kogan et al., 2012; Cohen et al., 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 studied in our paper is distinct from trade secrets, both kinds of innovation extend beyond the set of patenting firms. As non-patenting firms' innovative activities are understudied, we expect significant interest in approaches like ours to extend the analysis of innovation beyond patenting firms or industries.

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, and the information content in IPO prospectuses (Hanley and Hoberg, 2010; Hoberg and Phillips, 2010; Loughran and McDonald, 2011a; Popadak, 2013; Hoberg et al., 2014). At the same time, the asset pricing literature has employed kindred text-analysis procedures to measure sentiment and other asset pricing risks and anomalies (Edmans et al., 2007; Garcia, 2013; Dougal et al., 2012). 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 (Hoberg and Lewis, 2015; Jegadesh and Wu, 2015; Ganglmair and Wardlaw, 2015; Huang et al., 2015; Goldsmith-Pinkham et al., 2016). 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.[4]

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

[4] Even related work on innovation using text analysis has not constructed a similar measure of innovation. Specifically, Fresard et al. (2016) 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.

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) and Huang, Zang and Zheng (2012) who provide evidence, in a different context, that investors pay attention to the textual elements of analyst reports, rather than just the quantitative analyst forecasts. Our analysis suggests a new reason for investors to pay attention to the text of analyst reports: Valuable information on corporate 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 focuses on the relation between our text-based measure of innovation, and firm performance and value. The final section concludes with a summary.

2 Data

We start by selecting a sample of firms using the criterion that all firms must have been a member of the S&P 500 at some point during our sample period, from 1990 to 2012, this leaves us with an initial sample of 797 firms. To obtain the set of analyst reports these firms, we wrote a webcrawler to automatically search in Investext via Thomson One to download analyst reports 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 the stopword lists provided by Bill McDonald.[5] Prior to any textual analysis, we use a standard algorithm to stem (i.e., group words into the same root as in “technolog” captures “technology” and “technological,” among other related terms) the words contained in the analyst reports. 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

[5] http://www3.nd.edu/~mcdonald/Word_Lists.html

McDonald 2011b and Bodnaruk, Loughran and McDonald 2013) as an integral part of our textual classification of innovation. These lists have been adjusted for financial language and have been shown to be more appropriate than other sentiment word lists when reading financial text.

After constructing the main text sample, we calculate the measure we call the ’product revenue’ measure (as described in the ’Methods’ section) 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., 2012). We are left with 6,201 observations from 703 unique firms for the period 1990-2010.

For our later analysis of M&A 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 the text of analyst reports, and how to construct the text-based measure of innovation using the Latent Dirchlet Allocation (LDA) method of Blei et al. (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&P 500 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 standard word-list techniques (e.g., Loughran and McDonald, 2011b). 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.8 and A.9, the wordlist 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 provide an account of 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 ought to have a 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.

A natural 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.

Our approach of using analyst text to provide useful insight into firm activities is part of a growing literature to use analyst text in innovative ways. 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 (2012) 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 strongly 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 et al. (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).

Intuitively, 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 to be associated with the topic.

To construct our innovation measure, we set the total number of topics to be $K = 15$ using the 665,714 analyst reports as the underlying corpus of documents.[6] After conducting the LDA with 15 topics, we compute the average loading by firm-year, and correlate the logged value of this average loading with the patent count at the firm-year level. We select the topic with the highest correlation with patenting to be the basis for our test-based measure of innovation. Figure 2 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*. Aside from having a strong correlation with patents, this measure appears to also intuitively measure the factors that describe innovative companies. We hereafter refer to the topic as the ’product revenue’ topic. As we will show, firms that have high values of this measure have the hallmarks of innovative firms.[7]

Although we constructed the topic in a mechanical manner (by selecting the strongest correlation topic with patenting), the patent revenue topic has a strongly significant relationship to patenting, even after taking into account the obvious multiple comparisons problem of searching over 15 topics to select the best fitting one. Indeed, the test statistic in a linear regression is $t = 12.37$, which dramatically exceeds rule-of-thumb adjustments recently discussed in the finance literature (Harvey et al., 2015), and it survives other more formal, multiple-comparisons adjustments (e.g., the Bonferroni correction).[8] 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. Moreover, the word clouds of the next strongest fitting topics do not as closely resemble innovation of products, systems, or services. Further, Figure 3 presents a graphical depiction of how the text-based measure fits patenting outcomes by plotting logged 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 matches well the patenting patterns.

[6] We experimented with other numbers of initial topic loadings. Fewer topics tended to work similarly well, whereas a greater pre-specified number of topics led to redundancy (i.e., multiple topics about the same essential idea).

[7] Importantly, it is not using one or two of the most proment words in isolation that drives the results. All of our results on the relation between our text-based measure of innovation and valuation or revenue hold, even after accounting for the frequency of words that contain “revenue” or “growth” as their root (see Table 6, Panel (c)). These findings point to one of the distinctive features of LDA fitted topics – the content of the word cloud is how the words work together rather than merely reflecting a linear combination of the word frequencies used in the topic.

[8] For example, the null-hypothesis that none of the topics describe innovation can be rejected at a p-value of α if one or more of the topics are significant at a level $\frac{\alpha}{k}$ where k is the number of topics (this is called the Bonferroni correction).

Before using the loadings as a measure of innovation, it is important to refine the measure to account for analysts who write about the innovative activities of the firm in a negative or neutral tone. Specifically, if an analyst is talking with neutral or ambivalent sentiment about the company, it is less likely that the strong loading on product revenue 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, 2011b). 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. As a robustness exercise, we have constructed the measure without the sentiment screen, and we have also controlled explicitly for average sentiment. In each case, the results are quantitatively similar.

3.3 Comparison to Technology Development via R&D

Before describing the main findings in Section 4, we evaluate the text-based measure of innovation by comparing it with R&D intensity in the time series, in the cross-section of industries, and within industry. R&D intensity has been widely used as a measure of development of new technologies, and thus, as a measure of innovation (Griliches, 1980; Brown et al., 2009). In relation to the word usage captured by the innovation measure, our interpretation of this section's results is that the text-based measure of innovation measures the adoption of technology, even in industries that have low R&D intensity.

In the time series (1990-2010), the text-based innovation measure appears to capture the R&D boom and bust of the late 1990s and early 2000s, which is studied in Brown et al. (2009). Figure 4 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 (5) presents a bar plot of industry-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 indicates 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 technlogy 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.

Beyond these broad patterns, 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 in-house.

4 Empirical Results

In this section, we use our text-based measure of innovation to predict and forecast firm innovative effort as measued by R&D expenditures, evaluate the impact of innovation on various measures of performance (e.g., future 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 relationship between our text-based measure and future values of patenting, and examine several checks on the robustness of our measure.

4.1 Innovation and Performance

If our measure captures true innovation, it should reflect – as in the language of Drucker (1985) – the fact that a "resource" has been added to the firm. In this case, innovation should strongly relate to performance, and its impact on performance should slowly decline 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:

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

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 the more traditional innovation measures of 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 6 (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.7 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 rationale for this is captured by intuition behind the methodology in the recent Forbes list of innovative companies.[9] The intuition is that investors rec-

[9] http://www.forbes.com/sites/innovatorsdna/2015/08/19/how-we-rank-the-worlds-most-innovative-companies-2015/#71c616864524

ognize an innovative firm when they see it, and rationally place buying pressuring on the firm's stock, embedding a stock market premium into the 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:

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

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 indistinguishible between patenting and non-patenting firms. As with the analogous results for operating performance, this finding points to a notable advantage with our text-based innovation measure: it can be used for firms that do not use patents.

In Panel (b) of Figure 6(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.7 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. If innovations are improvements to firm products, or new valuable products, we should expect to see sales growth to increase following an increase in innovation.

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

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 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 apperant 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 panel (c) of Figure 6, we present a plot that summarizes the effect of innovation on sales growth over time. Gains in salesgrowth are transitory, only occuring in the year following the increase in innovation (see Appendix Table A.7 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.

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

At this point, we have established a connection between our text-based innovation measure and both operating performance, market value of the firm, and the propensity to invest in R&D. Next, we turn to the connection between our 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., 2012). We evaluate this connection by looking at four sets of outcome variables. We start with patent value data from Kogan et al. (2012).

4.2.1 Patent Value Measures

Turning to the relationship between text-based innovation and patent value, we estimate the following specification within the set of patenting firms:

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

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. 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 3 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 the estimates using the Value per Patent measure, and a similarly robust relationship between text-based innovation and future patenting values.

4.2.2 Text-Based Innovation and Patenting

Next, we turn to evaluating the relationship of text-based innovation to more traditional patent based measures. Specifically, 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 4, we present the results from estimating equation (5). The text-based innovation measure is strongly 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 are a strong indication 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. 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 5. 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. Moreover, 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 7. We have more to say about this interpretation of our innovation measure when we describe the acquisition results.

In another examination of whether our measure relates to product or demand-side innovation, we look at a novel product announcements measure constructed by Mukherjee, Singh and Zaldokas (2016) by estimating 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 from Mukherjee et al., 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)-(6) of Table 5 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 announcements ($s = 1$). 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 (which correlate strongly with text-based innovation) and the nature of highly-innovative firms outside of the set of patenting firms.

4.3.2 Contextual Examples of Systems Innovations

Within the set of patenting firms, it is useful to examine the content of valuable innovations. Figure 7 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, it is not surprising we do not find a connection with product introductions or product diffreentiation.

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 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)

Table 6 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 6 (a) presents results with one of these two topics as the measure of innovation (using the other one makes no qualitative difference). We see results that are very 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 present a version of the results that estimate the 15-topic LDA on a five-year rolling window basis. For example, to construct the measure for a firm in 1995, we use the topic loadings from the LDA model is fit on analyst reports from 1990-1994. For each five-year rolling window, we select the topic that most strongly correlates with patenting applications, and construct the measure in a similar manner to the main analysis. This exercise allows us to alleviate any concerns that the analyst reports were spuriously correlated with eventually successful technologies. As the results in panel (b) indicate, the findings are broadly similar for the five-year rolling average measure. For the main analysis, we prefer the single LDA fit measure because it relies on one fitted LDA, which enables us to more easily assess the qualitative content of the innovation topic.

Third, in panel (c), 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 6 indicate, the main results are qualitatively similar after controlling for other topic loadings, though in some cases, they become stronger.

Table 7 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 7 presents the results controlling for analyst sentiment, which are very 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 (b) of Table 7 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 7, we coduct 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

In this section, we examine the relationship between mergers and acquisitions activity and our text-based innovation measure. The results in this section serve two purposes. First, they provide an application of our new innovation measure that relates to other studies of innovation. Second, the findings on the pattern of acquisitions across big and small firms support our interpretation that the text-based measure is a useful proxy for innovative systems.

In theory, innovation could relate to acquisition activity either positively or negatively. On one hand, less innovative firms could substitute away from innovation toward acquisitions to obtain the technologies that enable the firm to be competitive. Under this innovate-versus-acquire decision, innovation and acquisitions would be substitutes, and thus have a negative relationship to one another (e.g., see Caskurlu, 2015). This would tend to be the case for product-related innovations. On the other hand, more innovation by a firm could open up greater possibilities for synergies with other firms with complementary innovations, as would be the case for firms with innovative systems. Under this innovate-to-synergy view, innovation and acquisitions would tend to be positively correlated because more innovation today would lead to greater productive acquisitions in the future (e.g., see Egan, 2013).

We evaluate whether innovation is associated with more or fewer acquisitions using the following specifications:

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

$$I(Acquisitions_{t+1}) = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{9}$$

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. $I(Acquisitions_{t+1})$ is an indicator variable equal to 1 when the firm acquired another firm in year $t+1$. As before, 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 acquisition opportunities, we expect $\beta_1 > 0$, but if innovation and acquisitions are substitute methods to obtain competitive technological processes, $\beta_1 < 0$. Whether the synergy view or the substitution view dominates is an empirical question we evaluate with these specifications.

Columns (1) and (2) of Table 8 Panel (a) presents the results from estimating equation (8). Across specifications, we find robust evidence that greater measured innovation today is associated with greater acquisition activity over the course of the next three years. A one standard deviation increase in text-based innovation leads to an increase in acquisition activity according to the specifications that include industry fixed effects. Moreover, 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 effect of innovation that we isolate is unlikely to be related to agency-based explanations for abnormal M&A activity. Results are similar in Table 8 Panel (b) where we estimate the linear probability model described by (9).

In Column (3) through Column (6) of Table 8, we present the results for two different splits of acquisition activity: big acquisitions (columns 3 and 4) and small acquisitions (columns 5 and 6). Consistent with similar sample splits in the literature (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 big otherwise. Across these specifications, we find that greater text-based innovation is associated with significantly more small acquisitions in the future, whereas we find a much weaker relationship to larger acquisitions. The linear probability model estimates in Panel (b) show a similar relationship between large and small acquisitions as in Panel (a), alleviating the concern that the result is driven by the functional form of the main specification.[10]

Relating to the mechanisms we outlined above, the broad finding relating text-based innovation to overall acquisitiveness suggests that innovation-to-synergy tends to be more empirically relevant than innovation as a substitution for acquisitions. This finding is consistent with recent findings by Fresard et al. (2016) who use more standard measures of R&D and patenting to measure innovation, and link it to merger activity, especially in the context of vertical mergers. Moreover, the fact that these findings are concentrated in acquisitions of small targets is consistent with our overall interpretation that the text-based innovation measure captures innovative systems, which are enhanced by the acquisitions of small firms (often bringing complementary components into the acquiring firm).

Alternatively, if innovation leads to excess free cash flow (through greater ROA, as in Table 2 columns (1) and (2)) or greater market-to-book (as in Table 2 columns (3) and (4)), and thus overvalued equity, an increase in acquisitions would be natural from the standpoint of empire building motives (e.g., see Harford et al., 2012). Aside from providing evidence that our text-based measure reflects a firm's innovative system, the fact that we find that the acquisitions are focused on smaller

[10] We have also conducted an analysis of merger announcement returns (CARs) as they relate to our text-based measure of innovation. According to this merger CAR analysis, small acquisitions (<5% of acquirer value) by firms with high text-based innovation are viewed significantly more favorably than relatively larger acquisitions by firms with high text-based innovation. The difference in announcement return is approximately 1.5% of firm value for a firm with text-based innovation one standard deviation above the mean. Together with the fact that the increase in acquisitions is concentrated among the small acquisitions, this finding suggests that text-based innovation generates synergies that are well recognized by the market, 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. See the results in Table A.6.

firms where the potential synergies are greater suggests that these motives are less likely.

6 Conclusions

We propose a new measure of corporate innovative activity based on textual analysis of analyst reports. There are several benefits of this text-based measure of innovation. First, it allows inclusion and measurement of non-patented innovation. Second, it is not subject to the problems inherent in the use of the production function to measure the impact of innovation. Third, it is not subject to the concerns of using the stock market's response to announcements of a patent award as a measure of the value of the innovation. Finally, our text-based measure is less susceptible to endogeneity concerns compared to measures constructed from managerial disclosures, such as annual reports, since analysts are outsiders vis-à-vis firm managers.

We analyze 665,714 analyst reports of 703 firms that were in the S&P 500 for some period during 1990-2012, and construct a text-based measure of innovation using state-of-the-art topic modeling techniques (Latent Dirchlet Allocation). Our measure of innovation exhibits sensible time-series and cross-sectional properties, as well as offering useful within firm variation in the level of innovation. Our results offer fresh perspective on extant measures of innovation. In particular, R&D intensity is well forecasted using our measure, but patenting outcomes are less robustly related to innovation as we measure it. These findings suggest that our new measure of innovation has much to add to existing measures, and should be viewed as complementary to existing tools to evaluate corporate innovation.

Turning to value implications, we also document a significant and positive relation between the text-based innovation measure and (a) the firm's future operating performance, and (b) the firm's future growth opportunities as proxied by market to book value of assets. Taken together, these results suggest our text-based measure of innovation is a valid measure of corporate innovation, and there is a significant and positive relation between innovation and firm performance, and innovation and firm value.

References

Asquith, Paul, Michael B. Mikhail, and Andrea S. Au (2005) “Information content of equity analyst reports,” *Journal of Financial Economics*, Vol. 75, No. 2, pp. 245 – 282.

Atanassov, Julian (2013) “Do Hostile Takeovers Stifle Innovation? Evidence from Antitakeover Legislation and Corporate Patenting,” *The Journal of Finance*, Vol. 68, No. 3, pp. 1097–1131.

Baier, Scott L, Gerald P Dwyer, and Robert Tamura (2006) “How important are capital and total factor productivity for economic growth?” *Economic Inquiry*, Vol. 44, No. 1, pp. 23–49.

Bhagat, Sanjai, Ming Dong, David Hirshleifer, and Robert Noah (2005) “Do tender offers create value? New methods and evidence,” *Journal of Financial Economics*, Vol. 76, No. 1, pp. 3–60.

Bhagat, Sanjai and Ivo Welch (1995) “Corporate research & development investments international comparisons,” *Journal of Accounting and Economics*, Vol. 19, No. 2, pp. 443–470.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003) “Latent dirichlet allocation,” *the Journal of machine Learning research*, Vol. 3, pp. 993–1022.

Bodnaruk, Andriy, Tim Loughran, and Bill McDonald (2013) “Using 10-k text to gauge financial constraints,” *Available at SSRN 2331544*.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson (2013) “Which News Moves Stock Prices? A Textual Analysis,” No. 18725.

Bradley, Daniel, Incheol Kim, and Xuan Tian (Forthcoming) “Do unions affect innovation?” Ph.D. dissertation.

Bradshaw, Mark Thomas (2011) “Analysts’ forecasts: what do we know after decades of work?” *Available at SSRN 1880339*.

Brown, James R., Steven M. Fazzari, and Bruce C. Petersen (2009) “Financing Innovation and Growth: Cash Flow, External Equity and the 1990s R&D Boom,” *Journal of Finance*.

Caskurlu, Tolga (2015) “Effects of Patent Rights On Industry Structure and R&D,” *Working Paper*.

Cohen, Lauren and Andrea Frazzini (2008) “Economic links and predictable returns,” *The Journal of Finance*, Vol. 63, No. 4, pp. 1977–2011.

Cohen, Lauren, Umit Gurun, and Scott Duke Kominers (2014) “Patent Trolls: Evidence from Targeted Firms.”

Cooper, Michael J., Anne Marie Knott, and Wenhao Yang (2015) “Measuring Innovation.”

Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher A Parsons (2012) “Journalists and the stock market,” *Review of Financial Studies*, p. hhr133.

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

Edmans, Alex, Diego Garcia, and Øyvind Norli (2007) “Sports sentiment and stock returns,” *The Journal of Finance*, Vol. 62, No. 4, pp. 1967–1998.

Egan, Edward J. (2013) “How Start-Up Firms Innovate: Technology Strategy, Commercialization Strategy, and their Relationship,” *Working Paper*.

Fresard, Laurent, Gerard Hoberg, and Gordon M. Phillips (2016) “Innovation Activities and the Incentives for Vertical Acquisitions and Integration,” *Working Paper*.

Fried, Dov and Dan Givoly (1982) “Financial analysts’ forecasts of earnings: A better surrogate for market expectations,” *Journal of Accounting and Economics*, Vol. 4, No. 2, pp. 85–107.

Galasso, Alberto and Mark Schankerman (2016) “Patents Rights and Innovation by Small and Large Firms,” *Working Paper*.

Ganglmair, Bernhard and Malcolm Wardlaw (2015) “Measuring Contract Completeness: A Text Based Analysis of Loan Agreements.”

Garcia, Diego (2013) “Sentiment during recessions,” *The Journal of Finance*, Vol. 68, No. 3, pp. 1267–1300.

Goldsmith-Pinkham, Paul, Beverly Hirtle, and David Lucca (2016) “Parsing the Content of Bank Supervision,” *Working Paper*.

Griliches, Zvi (1980) *New Developments in Productivity Measurement*, Chap. Returns to Research and Development Expenditures in the Private Sector, pp. 419–462: University of Chicago Press.

——— (1984) “R&D, Patents, and Productivity.”

——— (1998) “The search for R&D spillovers, R&D and Productivity: The Econometric Evidence.”

Hall, Bronwyn H (1990) “TheImpact of Corporate Restructuring On Industrial Research and Development,” *Brookings papers on economic activity: Microeconomics*.

Hall, Bronwyn H., Christian Helmers, Mark Rogers, and Vania Sena (2013) “The Importance (or not) of Patents to UK Firms,” *Oxford Economic Papers*, pp. 603 – 629.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg (2001) “The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools,” No. 8498.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg (2005) “Market Value and Patent Citations,” *The RAND Journal of Economics*, Vol. 36, No. 1, pp. pp. 16–38.

Hall, Bronwyn H, Jacques Mairesse, and Pierre Mohnen (2010) “Measuring the Returns to R&D,” *Handbook of the Economics of Innovation*, Vol. 2, pp. 1033–1082.

Hall, Bronwyn, Christian Helmers, Mark Rogers, and Vania Sena (2014) “The Choice between Formal and Informal Intellectual Property: A Review,” *Journal of Economic Literature*, Vol. 52, No. 2, pp. 375–423.

Hanley, Kathleen Weiss and Gerard Hoberg (2010) “The information content of IPO prospectuses,” *Review of Financial Studies*, Vol. 23, No. 7, pp. 2821–2864.

Harford, Jarrad, Mark Humphery-Jenner, and Ronan Powell (2012) “The sources of value destruction in acquisitions by entrenched managers,” *Journal of Financial Economics*, Vol. 106, pp. 247–261.

Harvey, Campbell R., Yan Liu, and Heqing Zhu (2015) "... and the Cross-Section of Expected Returns," *Review of Financial Studies*.

He, Jie (Jack) and Xuan Tian (2013) "The dark side of analyst coverage: The case of innovation," *Journal of Financial Economics*, Vol. 109, No. 3, pp. 856 – 878.

Hellwig, Martin and Andreas Irmen (2001) "Endogenous Technical Change in a Competitive Economy," *Journal of Economic Theory*, Vol. 101, No. 1, pp. 1 – 39.

Hirshleifer, David, Angie Low, and Siew Hong Teoh (2012) "Are Overconfident CEOs Better Innovators?" *The Journal of Finance*, Vol. 67, No. 4, pp. 1457–1498.

Hoberg, Gerard and Craig Lewis (2015) "Do Fraudulent Firms Produce Abnormal Disclosure?" *Vanderbilt Owen Graduate School of Management Research Paper No. 2298302; Robert H. Smith School Research Paper*.

Hoberg, Gerard and Gordon Phillips (2010) "Real and financial industry booms and busts," *The Journal of Finance*, Vol. 65, No. 1, pp. 45–86.

Hoberg, Gerard and Gordon M. Phillips (2016) "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala (2014) "Product market threats, payouts, and financial flexibility," *The Journal of Finance*, Vol. 69, No. 1, pp. 293–324.

Huang, Allen, Reuven Lehavy, Amy Zang, and Rong Zheng (2015) "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Working Paper, Available at SSRN 2409482*.

Huang, Allen, Amy Zang, and Rong Zheng (2012) "Evidence on the Information Content of Text in Analyst Reports," *The Accounting Review*.

Jegadesh, Narasimhan and Di Wu (2015) "Deciphering Fedspeak: The Information Content of FOMC Meetings," *Working Paper, December 2015*.

Jiang, J.J. and D.W. Conrath (1997) "Semantic similarity based on corpus statistics and lexical taxonomy," pp. 19–33.

Jones, Charles I (2002) "Sources of US economic growth in a world of ideas," *American Economic Review*, pp. 220–239.

Jurafsky, Daniel and James H. Martin (2009) "Speech and Language Processing, 2nd edition,": Pearson Education Inc.

Knott, Anne Marie (2008) "R&D/returns causality: Absorptive capacity or organizational IQ," *Management Science*, Vol. 54, No. 12, pp. 2054–2067.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman (2012) "Technological Innovation, Resource Allocation, and Growth," No. 17769.

Kuznets, Simon and John Thomas Murphy (1966) *Modern economic growth: Rate, structure, and spread*, Vol. 2: Yale University Press New Haven.

Lee, Charles M.C., Paul Ma, and Charles C.Y. Wang (2014) "Search Based Peer Firms: Aggregating Investor Perceptions through Internet Co-Searches," *Working Paper*.

Loh, Roger K and G Mujtaba Mian (2006) "Do accurate earnings forecasts facilitate superior investment recommendations?" *Journal of Financial Economics*, Vol. 80, No. 2, pp. 455–483.

Loughran, Tim and Bill McDonald (2011a) "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, Vol. 66, No. 1, pp. 35–65.

——— (2011b) "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, Vol. 66, No. 1, pp. 35–65.

Mann, William (2016) "Creditor rights and innovation: Evidence from patent collateral," *Working Paper*.

Manso, Gustavo (2011) "Motivating Innovation," *The Journal of Finance*, Vol. 66, No. 5, pp. 1823–1860.

Moser, Petra (2012) "Innovation without Patents: Evidence from World's Fairs," *Journal of Law and Economics*, Vol. 55, No. 1, pp. 43 – 74.

Mukherjee, Abhiroop, Manpreet Singh, and Alminas Zaldokas (2016) "Do Corporate Taxes Hinder Innovation?," *Journal of Financial Economics, Forthcoming*.

Nicholas, Tom (2008) "Does Innovation Cause Stock Market Runups? Evidence from the Great Crash," *The American Economic Review*, Vol. 98, No. 4, pp. pp. 1370–1396.

Nordhaus, William D (1969) "An economic theory of technological change," *The American Economic Review*, pp. 18–28.

Popadak, Jillian A. (2013) "A Corporate Culture Channel: How Increased Shareholder Governance Reduces Firm Value."

Saidi, Farzad and Alminas Zaldokas (2016) "Patents as Substitutes for Relationships," *Working Paper, Available at SSRN 2735987*.

Scherer, F. M. (1965) "Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions," *The American Economic Review*, Vol. 55, No. 5, pp. pp. 1097–1125.

Schumpeter, Joseph Alois (1939) *Business cycles: a theoretical, historical, and statistical analysis of the capitalist process*: McGraw-Hill New York.

Solow, Robert M (1956) "A contribution to the theory of economic growth," *The quarterly journal of economics*, pp. 65–94.

Swem, Nathan (2014) "Information in Financial Markets: Who Gets It First?" *Available at SSRN 2437733*.

Tian, Xuan (2011) "The role of venture capital syndication in value creation for entrepreneurial firms," *Review of Finance*, p. rfr019.

Tian, Xuan and Tracy Yue Wang (2011) "Tolerance for Failure and Corporate Innovation," *Review of Financial Studies*.

Trajtenberg, Manuel (1990) "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *The RAND Journal of Economics*, Vol. 21, No. 1, pp. pp. 172–187.

Tucker, Catherine (2014) "Patent Trolls and Technology Diffusion: The Case of Medical Imaging," *Available at SSRN 1976593*.

Twedt, Brady and Lynn Rees (2012) "Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports," *Journal of Accounting and Public Policy*, Vol. 31, No. 1, pp. 1–21.

University, Princeton (2010) "About WordNet."

Womack, Kent L (1996) "Do brokerage analysts' recommendations have investment value?" *The journal of finance*, Vol. 51, No. 1, pp. 137–167.

Yim, Soojin (2013) "The Acquisitiveness of Youth: CEO Age and Acquisition Behavior," *Jounral of Financial Economics*, Vol. 108, No. 1, pp. 250–273.

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: Text-Based Innovation Measure: Word Cloud

Note: This word cloud describes the frequency distribution of words used in the ’Product Revenue’ topic. The topic itself is from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P 500 firms. We set the number of topics in the fitted LDA model to be 15, then selected the topic (out of these 15) for which the topic loadings had the strongest correlation with patent counts. To avoid an overfitting interpretation, our tests that utilize this measure either forecast future values of patents or patent citations, or predict other measures of innovation and performance (i.e., R&D expenditures).

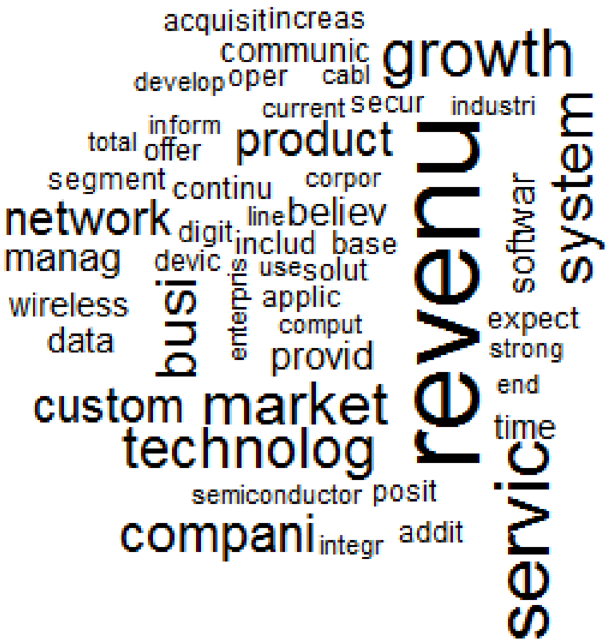
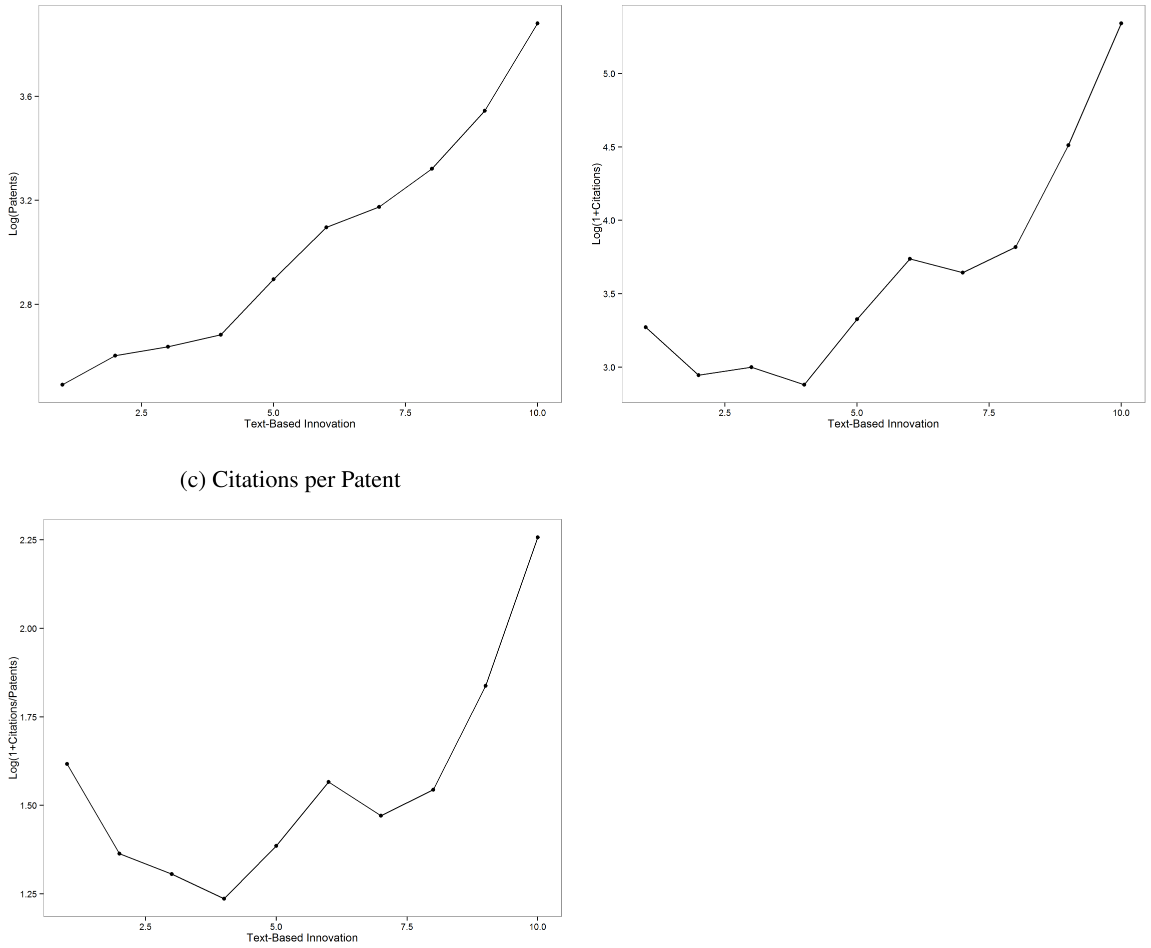


Figure 3: Relating Patent Counts and Patent Citations to the Text-Based Innovation Measure (Decile Bins)

Note: This is a description of the text-based innovation measure and how it relates to the commonly-used patenting measures. In each panel, the text-based innovation measure is grouped into 10 deciles. Panel (a) presents the relationship to logged patent counts ($\log(1+Patents)$), Panel (b) presents the relationship to patent citations ($\log(1+Citations)$), and Panel (c) presents the relationship to citations per patent ($\log\left(1+\frac{Citations}{Patent}\right)$) .



(a) Patent Counts

(b) Patent Citations

(c) Citations per Patent

Figure 4: 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 are strongly correlated with one another, with a correlation of 0.58, which is statistically different from zero.

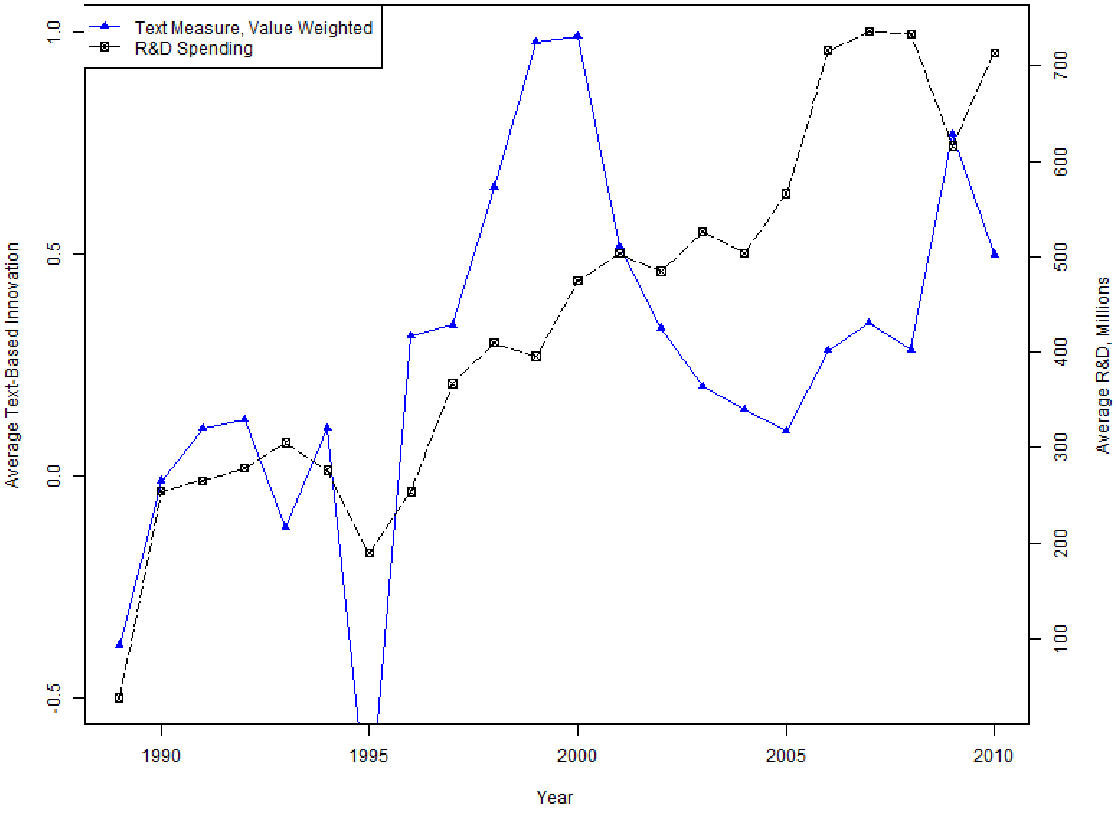


Figure 5: Cross-Industry Plot of R&D (1990-2004), Relationship to Text-Based Measure

Note: This figure provides a plot of R&D expenditures (average R&D/Assets) by industry covered in the sample of S&P500 firms. To show the relationship to the text-based measure of innovation, 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.

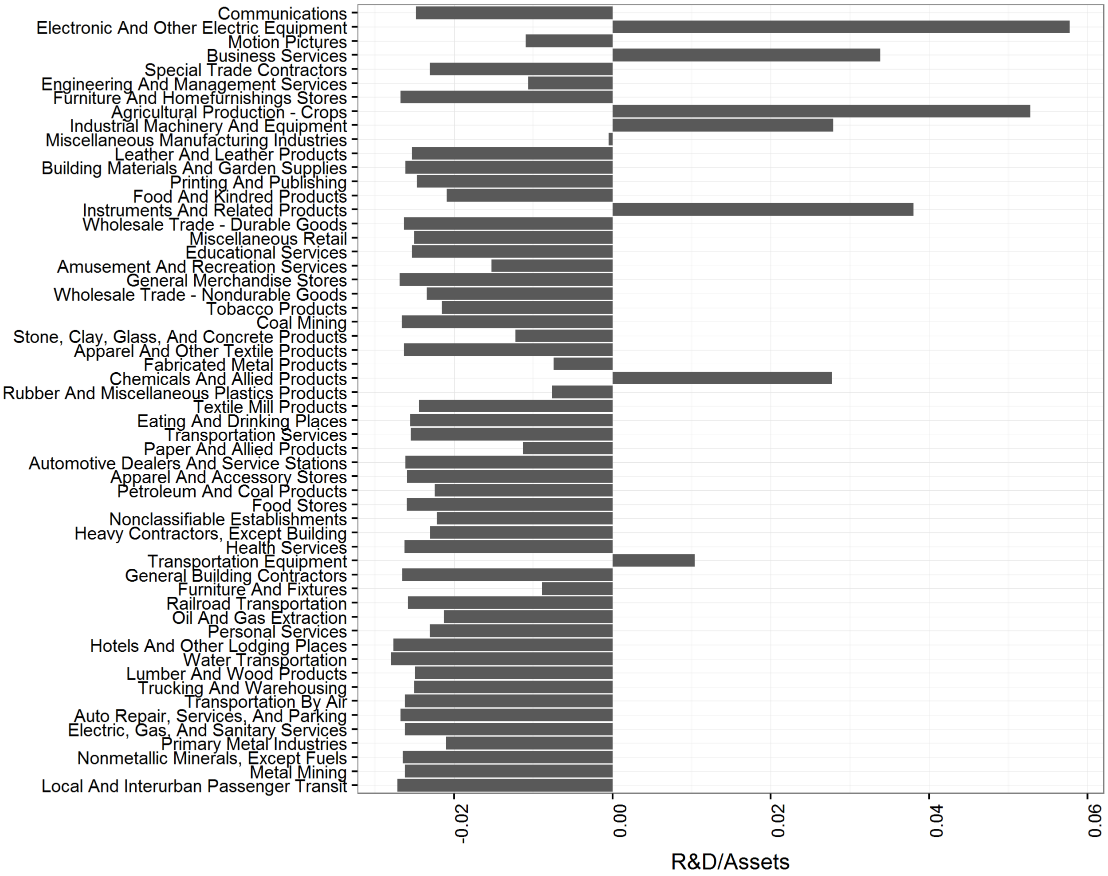
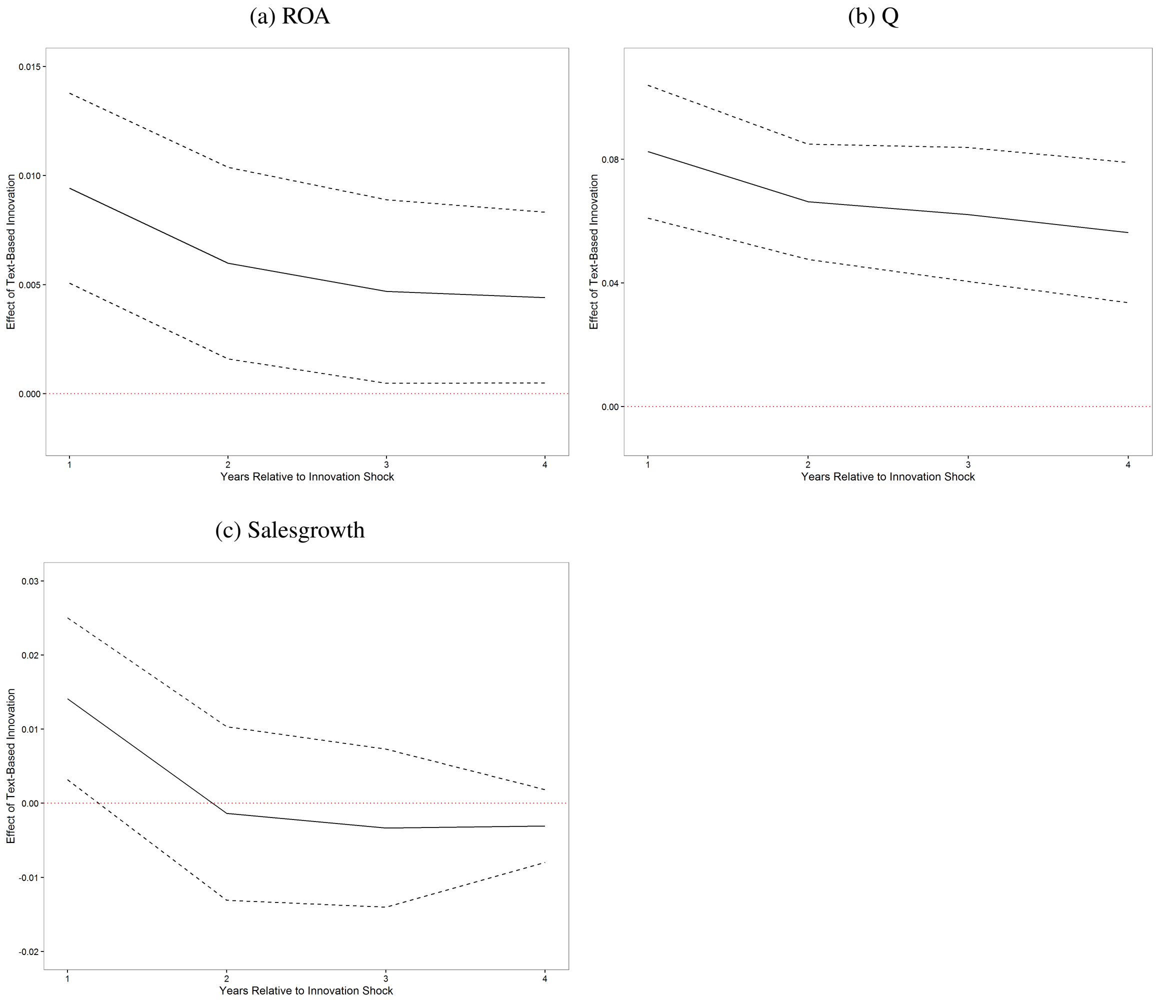


Figure 6: Long Run Effects of Innovation on Performance – Forecasting ROA and Tobin's Q up to Four Years Out

Note: These plots present the response in future operating performance and Q 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.7. Dotted lines represent 95% confidence bands around the estimated effects.


(a) ROA

(b) Q

(c) Salesgrowth

Figure 7: 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 the mean of the ’product revenue’ topic loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed measure and convert it to a Z-score. Patents is the count of granted patents which were applied for during the year. Cumulative patents is the count of all patents granted with application years before the fiscal 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. 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).

Statistic	N	Mean	St. Dev.	Min	Max
Text-Based Innovation	6,201	0.000	1.000	−1.691	3.957
Patents	6,201	62.918	235.367	0	5,063
R&D/Assets	6,201	0.026	0.045	0.000	0.605
Log (Assets)	6,201	8.778	1.255	5.242	13.590
ROA	6,201	0.154	0.088	−0.670	0.905
Asset Tangibility	6,201	0.359	0.231	0.002	0.970
Leverage	6,201	0.585	0.196	0.032	1.899
Log(Age)	6,201	3.019	0.414	0.693	3.584

Table 2: Performance of Firms and Text-Based Innovation (1989-2010)

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)$_t$						
× 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)
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: Patent Value and Text-Based Innovation (1989-2010)

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. (2012)). 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 4: Patents and Text-Based Innovation (1989-2010)

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)
Log(1 + Patents)$_t$		1.018^{***}		0.961^{***}		-0.065^{***}
		(0.060)		(0.107)		(0.025)
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)
Return on Assets$_t$		-0.100		-0.532		0.759^{**}
		(0.320)		(0.963)		(0.351)
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 5: Product Differentiation and Product Announcements (1989-2010)

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)-(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:					
	$\text{Log(Total Similarity)}_{t+1}$		$\text{Log}(1 + \text{Products})_t$		$\text{Log}(1 + \text{Products})_{t+1}$	
	(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)
$\text{Log}(1 + \text{Patents})_t$	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)
$\text{Log(Total Assets)}_t$	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)
$\text{Asset Tangibility}_t$	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)
$\text{Return on Assets}_t$	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 6: 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 (1989-2010)

	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) Rolling Window Measure (1994-2010)

	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)$_t$	0.009***		0.071***		0.007	
	(0.002)		(0.010)		(0.006)	
× Patenting Firm		0.009***		0.071***		0.005
		(0.002)		(0.012)		(0.005)
× Non-Patenting Firm		0.012***		0.072***		0.011
		(0.005)		(0.014)		(0.014)
Controls, Industry FE, 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	0.428	0.580	0.580	0.102	0.102
Note:				*p<0.1; **p<0.05; ***p<0.01		

(c) Controlling for other topics, K=15 (1989-2010)

	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 7: Accounting for Alternative Explanations (1989-2010)

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 8: Predicting Acquisition Activity Using the Text-Based Innovation Measure (1989-2010)

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 indicator variable that is set to 1 if there is an acquisition in the next year. 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

	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 $(Z)_t$	0.089***	0.035**	0.005	0.010**	0.088***	0.031**
	(0.018)	(0.015)	(0.005)	(0.005)	(0.018)	(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	6,201	6,201	6,201	6,201	6,201	6,201
Adjusted R^2	0.310	0.393	0.117	0.133	0.317	0.412
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Linear Probability Model

	Dependent variable:					
	$\text{I(Acquisition)}_{t+1}$		$\text{I(Big Acquisition)}_{t+1}$		$\text{I(Small Acquisition)}_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.040***	0.017**	0.002	0.004	0.039***	0.014*
	(0.008)	(0.008)	(0.003)	(0.004)	(0.008)	(0.007)
Other Controls		X		X		X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,075	6,075	6,075	6,075	6,075	6,075
Adjusted R^2	0.166	0.201	0.033	0.039	0.173	0.216
Note:						*p<0.1; **p<0.05; ***p<0.01

Appendix to:

A Text-Based Analysis of Corporate Innovation

A Appendix Tables and Figures

A.1 Additional Detail on LDA

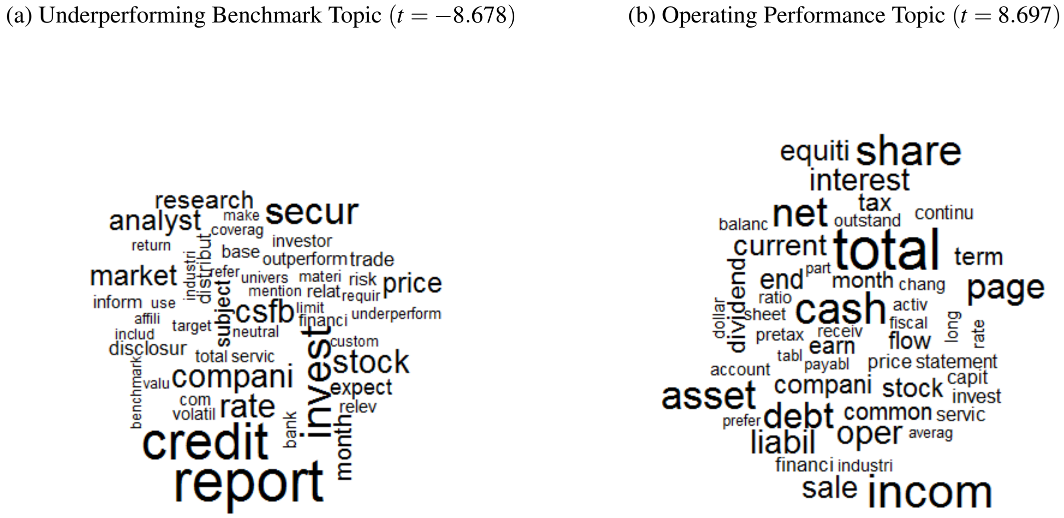
Table A.1: Fit of Patenting Outcomes to Loadings for Every Topic in the 15-Topic LDA

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 product revenue 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 intuitively matches the notion of innovation. 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 describes 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 Product Revenue topic, these topics are computed from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P 500 firms. We set the number of topics in the fitted LDA model to be 15.

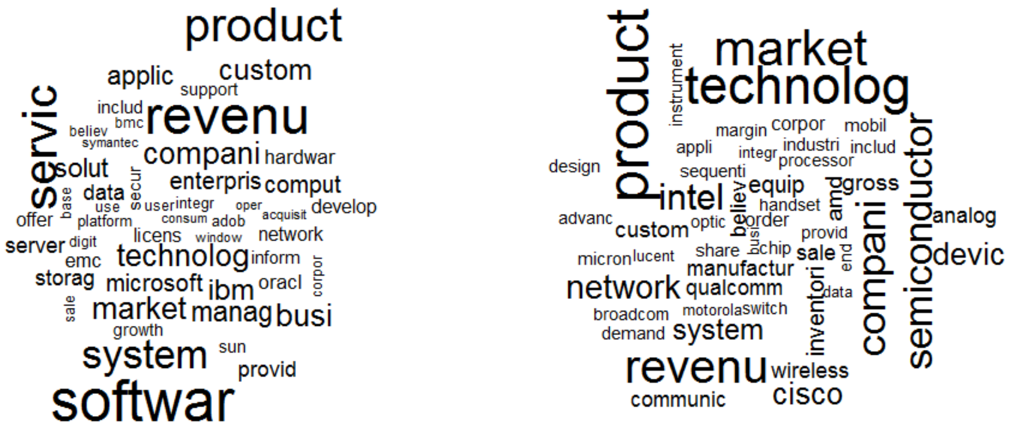


(a) Underperforming Benchmark Topic ($t = -8.678$)

(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 Product Revenue Toopic form the 15-topic LDA. As with the Product Revenue 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&P 500 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 ’Product Revenue’ 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&P 500 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 loadings had the strongest correlation with patent counts. To avoid an overfitting interpretation, our tests that utilize this measure either forecast future values of patents or patent citations, or predict other measures of innovation and performance (i.e., R&D expenditures).

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.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. (2012))*
Products	Product Announcements	*The count of product amouncements in which the stock return exceeded the 75th percentile in Mukherjee et al. (2016).*

Table A.3: Full Results on Performance of Innovative Firms (1989-2010)

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 the mean of the 'product revenue' topic loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed 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). 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.3: Full Results on Performance of Innovative Firms (1989-2010)

(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 (1989-2010)

Note: The dependent variable is the ratio of R&D expenses to total assets. The text-based innovation measure is the mean of the 'product revenue' topic loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed 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). 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 (1989-2010)

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 indicator variable that is set to 1 if there is an acquisition in the next year. The text-based innovation measure is the mean of the 'product revenue' topic loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed measure and convert it to a Z-score. 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 $(Z)_t$	0.089^{***}	0.035^{**}	0.005	0.010^{**}	0.088^{***}	0.031^{**}
	(0.018)	(0.015)	(0.005)	(0.005)	(0.018)	(0.014)
Log(Patents)_t		0.056^{***}		0.001		0.057^{***}
		(0.017)		(0.005)		(0.016)
ROA_t		0.603^{**}		−0.009		0.608^{**}
		(0.260)		(0.066)		(0.254)
$\text{R\&D/Assets}_t$		−0.734		-0.491^{***}		−0.471
		(0.683)		(0.161)		(0.708)
Log(Assets)_t		0.174^{***}		-0.022^{***}		0.188^{***}
		(0.024)		(0.006)		(0.024)
$\text{Asset Tangibility}_t$		-0.367^{***}		-0.082^{***}		-0.302^{**}
		(0.116)		(0.030)		(0.118)
Leverage_t		-0.531^{***}		-0.071^{**}		-0.487^{***}
		(0.079)		(0.029)		(0.076)
Log(Age)_t		0.050		−0.015		0.063
		(0.055)		(0.013)		(0.053)
Log(Q)_t		0.189^{***}		-0.033^{***}		0.214^{***}
		(0.044)		(0.011)		(0.043)
Patenting Firm		-0.084^{**}		−0.006		-0.073^{**}
		(0.036)		(0.014)		(0.036)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,201	6,201	6,201	6,201	6,201	6,201
Adjusted R^2	0.310	0.393	0.117	0.133	0.317	0.412

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table A.5: Full Results on Predicting Acquisition Activity (1989-2010)

(b) Linear Probability Model

	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)$_t$	0.040***	0.017**	0.002	0.004	0.039***	0.014*
	(0.008)	(0.008)	(0.003)	(0.004)	(0.008)	(0.007)
Log(Patents)$_t$		0.014**		−0.0004		0.016**
		(0.007)		(0.003)		(0.007)
ROA$_t$		0.239*		0.035		0.216*
		(0.132)		(0.055)		(0.129)
R&D/Assets$_t$		−0.426		−0.269***		−0.301
		(0.264)		(0.102)		(0.285)
Log(Assets)$_t$		0.069***		−0.014***		0.078***
		(0.012)		(0.003)		(0.011)
Asset Tangibility$_t$		−0.200***		−0.041**		−0.169***
		(0.062)		(0.018)		(0.061)
Leverage$_t$		−0.272***		−0.031		−0.259***
		(0.044)		(0.022)		(0.038)
Log(Age)$_t$		0.049*		−0.010*		0.056**
		(0.026)		(0.006)		(0.024)
Log(Q)$_t$		0.074***		−0.019**		0.087***
		(0.020)		(0.009)		(0.020)
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,075	6,075	6,075	6,075	6,075	6,075
Adjusted R^2	0.166	0.201	0.033	0.039	0.173	0.216

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

Table A.6: 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)$_t$ × SmallAcq		0.015^{***}	0.015^{***}	0.016^{***}
		(0.003)	(0.003)	(0.005)
Text-Innovation (Z)$_t$	-0.003^{**}	-0.016^{***}	-0.016^{***}	-0.018^{***}
	(0.001)	(0.003)	(0.003)	(0.005)
SmallAcq		−0.005	−0.005	−0.005
		(0.004)	(0.005)	(0.005)
Log(Patents)$_t$			−0.001	−0.001
			(0.001)	(0.001)
Return on Assets$_t$			0.014	0.038
			(0.018)	(0.029)
R&D/Assets$_t$			0.034	0.033
			(0.031)	(0.075)
Log(Assets)$_t$			0.002	0.003
			(0.002)	(0.004)
Asset Tangibility$_t$			−0.008	−0.033
			(0.011)	(0.023)
Leverage$_t$			0.002	0.008
			(0.010)	(0.013)
Log(Age)$_t$			−0.004	−0.023
			(0.004)	(0.020)
Log(Q)$_t$			−0.004	-0.010^{*}
			(0.004)	(0.006)
Cash/Assets$_t$			−0.002	−0.014
			(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	0.066	0.073	0.073	0.158

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

A.3 Long-Term Dynamics

Table A.7: 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 the mean of the 'product revenue' topic loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed measure and convert it to a Z-score. All firms that have at least one patent during the sample period (1989-2004) 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.4 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.[11] 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),[12] and restrict attention to the top 1% words of most related words. Specifically, we use the Jiang

[11] 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

[12] The synset for 'innovation' is defined as 'a creation (a new device or process) resulting from study and experimentation'. The synset for 'innovate' is defined as 'bring something new to an environment'.

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.[13] 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.8 and A.9 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:,

$$Performance_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}_{it}'\Gamma + \varepsilon_{it} \qquad (10)$$

and

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

$$Patenting_{it+1} \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{11}$$

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.8 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.9 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.8: Patent Value, Word-List Measure (1989-2010)

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 the mean of the innovation word-list loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed 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.9: Patent Value, Word-List Measure (1989-2010)

Note: The dependent variable is a patent value measure. The first four columns aggregates the value of all patents granted during the year, scaled by patent count in columns 3-4. Columns 5-6 uses the citation weighted patents over the next three years as the measure of patent value. We use patent value data from Kogan et al. (2012) 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. The text-based innovation measure is the mean of the innovation word-list loading for positive analyst reports about the firm over the fiscal year. We take the fourth root of this highly skewed measure and convert it to mean zero and unit variance. 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	