Essays on Financial Information in Text

by

Gustaf Bellstam

B.S., University of Florida, 2009

M.S., Northeastern University, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Finance

2018

This thesis entitled:
Essays on Financial Information in Text
written by Gustaf Bellstam
has been approved for the Department of Finance

Prof. Diego Garca

Prof. J. Anthony Cookson

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Bellstam, Gustaf (Ph.D., Business Administration, Finance)

Essays on Financial Information in Text

Thesis directed by Prof. Diego Garca

In the first essay, I study how analysts' performance depends on their incentives and access to information using a regulatory shock and the textual content of analyst reports. My results focus on two aspects of performance, informativeness and bias. After incentives and access to information are reduced, analyst reports become less informative but also less biased. My identification strategy uses the Global Research Settlement as a shock that affected analysts at investment banks, but not other analysts, in a difference-in-difference design. I find that analyst reports become more similar to one another after the shock, an indication of less information content. Additionally, I find that text exhibits lower overall sentiment, which is driven by an increase in negativity, and that reports contain fewer markers of bias (less weaseling). The results highlight a trade-off with informativeness when regulating bias.

In the second essay, we develop a new measure of innovation using a textual analysis of analyst reports. Our text-based measure gives a useful description of innovation by mature firms with and without patenting and R&D. For non-patenting firms, the measure identifies firms that adopt novel technologies and innovative business practices (e.g., Walmarts cross-geography logistics). For patenting firms, the text-based measure strongly correlates with valuable patents, which likely capture true innovation. The text-based measure robustly forecasts greater firm performance and growth opportunities for up to four years, and these value implications hold just as strongly for non-patenting firms.

Acknowledgements

I am immensely grateful for help from my committee members, Diego Garca, J. Anthony Cookson, Yonca Ertimur, Jonathan Rogers, and Edward Van Wesep. I am also grateful to Jordan Martel and Katie Moon for many helpful suggestions. In addition, I would like to thank my co-authors on Chapter 2 of this dissertation; J. Anthony Cookson and Sanjai Bhagat.

All remaining errors are my responsibility.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Informativeness, and Bias: Evidence from Analyst Text

1.1 Introduction

There is widespread evidence that analysts provide useful information (Womack 1996), but also that improper incentives can lead to biased research output (see Michaely and Womack (1999), Merkley, Michaely and Pacelli 2017 and Cornaggia et al. 2016). For example, a series of scandals amid the dot-com bubble revealed that analysts produced unsubstantiated research to favor potential investment banking clients. These scandals led to a wave of regulations that disconnected analysts from investment bankers in terms of compensation and information, but did the regulations work to reduce analyst bias? Beyond recommendations, did these changes reduce analyst incentives or ability to provide valuable information? I provide novel insight into these questions by analyzing text in research reports, which contains indicators of both bias and informativeness, in the wake of a regulation change that influenced some, but not all, analysts'.

Between 2002 and 2003, a series of regulatory changes around the Global Research Settlement (hereafter collectively referred to as the GRS) aimed to reduce bias in financial analyst reports. These new regulations reduced analyst compensation and placed significant informational barriers between the analysts and investment bankers (Barber et al. 2006, Groysberg, Healy and Maber 2011)[1] . Moreover, the effects of these regulations were concentrated in investment banking firms (Barber et al. 2006). My empirical strategy exploits this differential effect on investment banking

[1] The GRS, NASD Rule 2711, NYSE Rule 472, and SOX 501 were during this period, all of which required extensive separation between investment banking activities and sell-side analysts. At the same time as the GRS, Regulation Analyst Certification was implemented and required a certification of truthfulness in analyst reports. Section 2 contains a more detailed description of The GRS and the related rule changes.

analysts in a difference-in-difference strategy that compares how the content of analyst reports changes after the GRS for investment-banking analysts versus the less affected, unaffiliated, analysts.

My focus is on the qualitative information contained in the text of analyst reports; an important channel for both informativeness and bias of analyst output. I find that, as a result of the GRS, reports written by investment banking analysts increased in similarity with prior reports relative to other analysts, an indication that less of the text in them is new information. In addition, investment banking analysts write more negative reports, a result consistent with a reduction in bias (less unsubstantiated optimistic reports) — a key goal of the regulation changes. I proceed by creating an alternative text-based measure of bias and show that it also goes down in response to the GRS. Overall, these findings are consistent with the argument made by the financial press that the regulatory goals of reduced conflicts have had unintended consequences for the informativeness of research reports. For example, Bob Pisani (a journalist with CNBC) argued in 2014 that analysts have either left for better jobs or reduced effort as a response to decreased compensation tied to the regulatory changes. He claims that "[t]he quality of research has declined as many of the brightest sell-side analysts have left for more lucrative jobs with hedge funds and other buy-side firms. Many of those remaining do little if any original research."

My empirical work uses a sample of 403,000 analyst reports. From this original sample, I construct textual measures of informativeness and bias. To proxy for (the lack of) informativeness, I compute the cosine similarity of an analyst report to other reports about the same firm from the prior three months. The intuition for this measure is that new information by necessity will look different from current information, in other words, it will have low similarity. This is motivated partly by the fact that a common way to produce an analyst report without information content is to maintain consensus (Hong, Kubik and Solomon 2000). As a sanity check, I plot the average magnitude of the stock market reaction during the report window against the decile rank of the **Similarity** score and find that there is an inverse relationship between **Similarity** and stock market reaction (see figure 1.4). To proxy for bias in an analyst report, I measure the extent to which

the analyst uses weasel language, a form of evasive language that is captured using Wikipedia's weasel tags (as in Cookson, Moon and Noh 2017). Intuitively, biased language should contain more evasive language when incorrect or slanted information is described as more accurate than it is. As an additional proxy for bias, I measure report sentiment (often referred to as "tone") and evaluate how positivity and negativity changes before and after the GRS.

After the GRS, banking analysts produce reports that are at least 0.3 standard deviations more similar to prior reports relative to non-banking analysts. Much of this effect comes from an increase in similarity with ones own reports. In related results, I show that, conditional on a big information event (an upgrade or a downgrade), this type of self-similarity has a dampening effect on the stock market reaction. The effect is distinct from traditional herding, which is often measured as temporal similarity between earnings forecasts, since it is driven by similarity with ones own output (a copy-paste effect). Overall, these results suggests that the analyst reports became less informative in response to regulation changes around the GRS.

Despite an overall reduction in informativeness, my finding that there is a shift away from optimism is consistent with conflicts of interest playing a smaller role after the GRS. However, since bias in the pre-period was a key driver that led to the GRS, it is possible that sentiment was exceptionally high in that particular period and any effects identified by studying variables based on sentiment, earnings forecasts, or recommendations could be driven by mean reversion (Ashenfelter, 1978). With this in mind, I create another measure of bias using a textual measure of the frequency of weasel words in analyst reports. Weasel words are used to create the impression that a meaningful statement has been made when in fact it has not. For example "people say" instead of a direct reference, or "many" instead of an available number. When an analyst is incentivized to produce research that differs from her underlying belief (i.e. biased), weasel statements are a natural consequence. I find that weaseling goes down by 0.3 standard deviations for banking analysts relative other analysts as a result of the GRS. In addition to the concerns about an Ashenfelter dip, my finding that the language exhibits fewer markers of bias via weasel words is useful because it sidesteps concerns apparent in prior studies that optimism and stock market reactions to optimism

could be jointly caused by new information.

Since Womack (1996)'s seminal evidence on analyst skill in forecasting stock price changes, an important literature has examined analyst skills and incentives, and how conflicts of interest can bias stock picking and forecasting outcomes (e.g., Hong, Kubik and Solomon 2000). Within this literature, studies of conflicts of interest in underwriting relationships are most closely related to my work. For example, previous work has shown a link between conflicts of interest and optimism in initiation reports and between conflicts and lower investment value of recommendations (Michaely and Womack 1999; Ertimur, Muslu and Zhang (2011b); Hirst, Koonce and Simko (1995)). Relative to this literature, my findings on the use of weasel words in biased reports are distinctive in that they show direct evidence of bias in the content of analyst reports. From this standpoint, the weaseling results give prospective guidance to readers of analyst reports to look for bias from an ex ante perspective, which would have been difficult to infer from prior work.

My findings also contribute to our understanding of the consequences of the Global Research Settlement and related rule changes. Previous studies of the the GRS have focused on easily-quantifiable measures, such as analyst recommendations and forecasts. Using these quantitative measures, the literature has shown a shift toward more sophisticated analysis following the GRS, such as the use of intrinsic value estimates (Barniv et al., 2009; Chen and Chen, 2009) or an increase in the precision of the ratings scale (Kadan et al., 2009), but also a departure of top talent (Guan, Lu and Wong, 2013). Despite some evidence that points to a reduction in conflicts, these changes have meant that the quantitative content of analyst reports (recommendations) have less overall market impact after the GRS. My focus on how the GRS changed the content of analyst reports allows me to evaluate this trade-off between bias and informativeness. My results show that – beyond quantitative assessments of analysts – markers of informativeness and bias in reports are both reduced after the GRS, and that qualitative assessments of informativeness have market impact beyond what others have shown for forecasts and recommendations. This finding is an important indication that the content of analyst reports contains valuable information, which is becoming more relevant as text processing becomes commonplace among sophisticated investors

(i.e., the digitization of EDGAR).

My study also relates to the broader literature on text and financial markets. Textual analysis has grown both because analyzing financial texts provides new data and evidence on difficult to measure outcomes (e.g., Hoberg and Phillips (2016), Bellstam, Bhagat and Cookson 2016, Popadak 2013a), but also because textual content matters for financial market outcomes (Tetlock (2007), Loughran and McDonald 2011, García 2013a). Within this broader literature, my work most closely relates to studies of analyst text, which has generally concerned itself with whether analysts produce new information. Early studies in this literature have considered how sentiment (often referred to as tone) affects market reaction to reports (e.g., Twedt and Rees 2012; Huang, Zang and Zheng 2014; Chen, Nagar and Schoenfeld 2015). My approach of developing textual measures beyond sentiment (similarity and weaseling) for information production is related to recent work by Huang et al. (2017), who employ a topic modeling approach to decompose information in analyst reports into discovery and interpretation. My analysis of the response of these measures to the GRS complements previous findings by showing evidence on how incentives for analysts relate to informativeness and bias in the text of their reports. I find that, although bias in reports goes down following the GRS, there is also a reduction in informativeness — an unintended consequence of the regulation change.

The paper progresses as follows: section 2 provides additional detail on the Global Research Settlement and associated regulations. In section 3, I describe the data, which is followed by a description of the empirical setting in section 4 and a discussion of results in section 5. In section 6, I check whether results are robust to different measurement windows, various classification schemes for weaseling, and to excluding the Reg FD period. I conclude in section 7.

1.2 The Global Research Settlement and Other Regulation

Between 2002 and 2003 there were four regulation changes as well as the announcement and finalization of the Global Research Settlement, all of which aimed to disconnect the content of analyst reports from sources of bias and corruption. The GRS was the outcome of a lawsuit

against 10 large investment banks. As a result of the settlement, investment banks were fined heavily and required to separate investment bankers from analysts. Shortly after the GRS, two self-regulating organizations, the NASD and the NYSE, imposed rules that were broadly similar to the requirements imposed by the GRS, but applied to all members. The other two rule changes during the period were Regulation Analyst Certification, which required analysts to certify their statements as true, and SOX section 501 which made some of the requirements from NYSE and NASD into law.

The GRS, NASD Rule 2711, and NYSE Rule 472 all required extensive separation between investment banking activities and sell-side analysts. The rules prohibited any direct link between investment bankers and analyst compensation and required informational firewalls between bankers and analysts. Specifically, the rules required a reporting line for research staff that is separate from investment banking; they required a dedicated legal staff for research; and they prohibited three-way meetings between investment bankers, investors, and analysts. Rules also prohibited bankers from influencing analysts in other ways. For example, bankers are not allowed to retaliate against analysts for unfavorable reports nor are they allowed to use research analysts for any sales or marketing efforts. In addition, there were disclosure requirements regarding investment banking relationships, compensation, and ownership or directorship in the subject company. Brokerages were also required to disclose the percentage of buy, hold, and sell recommendations.[2]

Despite the similarities between the GRS and the NASD and NYSE rules, there were some differences. In particular, the GRS was stricter and required the sanctioned firms to physically separate investment bankers from analysts. Sanctioned banks were also required to pay for, and make available to their clients, third party research. They were fined heavily and were required to have separate legal staff for their analyst business.[3]

[2] One finding in Kadan et al. (2009) is that the percentage of buy recommendations fell in favor of more hold and sell recommendations and that this made upgrades more informative.

[3] The effect of global settlement and related regulation on investment banking and analysts are discussed at length in "Frequently Asked Questions About Separation of Research and Investment Banking," available at https://media2.mofo.com/documents/frequently-asked-questions-about-separation-of-research-and-investment-banking.pdf. Kadan et al. (2009) and Corwin, Larocque and Stegemoller (2017) both contain good summaries.

To summarize, the rules had a number of provisions that plausibly affect analyst output. They required extensive extra disclosures (the effects of which I will try to isolate in my analysis), they changed how analysts are compensated (a key channel in my paper), and they affected analyst's access to information by erecting informational firewalls.

1.3 Data

In this section I describe the data. I use analyst text data from Investext accessed via Thomson One, analyst forecasts and recommendations from IBES, Investment Banking data from SDC Platinum, and stock market data from CRSP.

1.3.1 Analyst Report Text Sample

I start by selecting a sample of firms using the criterion that the firm must have been a member of the S&P500 at some point during the sample period, from 1990 to 2012, this leaves me with an initial sample of 797 firms. There are a few reasons why I focus on firms in the S&P500. First, data collection of analyst reports is time intensive and it is necessary to limit the sample. In addition, S&P500 firms are almost always covered by analysts, making the coverage decision itself less important for the research design when using this particular sample. On the other hand, I need to be careful with interpreting the generalizability of my results.

Using the initial sample of 797 firms, I search Investext via Thomson One and download analyst reports from 1990 to 2012 for each firm that I can identify, this procedure leaves me with an initial sample of 807,309 analyst reports for 750 unique firms. After downloading the reports, I clean them of common stop words using the stop word lists provided by Bill McDonald.[4] Documents are then stemmed using the Porter stemming algorithm and put into a document term matrix which is used throughout the paper. A document term matrix is a matrix with documents as rows and words as columns, each cell represents the number of occurrences of word j in document

[4] The data is described in Bodnaruk, Loughran and McDonald, 2015; Loughran and McDonald, 2011 and can be downloaded from Bill McDonald's website, http://www3.nd.edu/~mcdonald/Word_Lists.html

i. I drop documents with under 100 words remaining after the cleaning or over 5,847 words (the 98th percentile). This is still a large range of report lengths, with some report shorter than one page and some as long as 10 pages. With that in mind, I will control for report length in the main tables. After completing the cleaning described above and matching with CRSP identifiers, I end up with 669,555 reports for 747 firms. I also limit the sample to report where I can identify the analyst and brokerage firm, which leaves me with 445,056 reports. Finally I remove observations for which I do not have a valid similarity score (described in section 3.2) and end up with a sample of 403,000 reports, covering 698 firms, written by 2,208 analysts working for 133 brokerages.

Most tables in the paper describe difference-in-difference tests pre and post the GRS. For such tests, I use data from the pre-period and post-period only. The pre-period is 2000-2001 and the post-period is 2004-2005. I remove 2002 and 2003 as both years had multiple regulation changes. This leaves me with 1224,000 reports by 1452 analysts working at 76 brokerages that came out in 2000-2001 or 2004-2005. This limited data-set is described in table 1.2. The full data-set is used in graphs and tables that relates textual measures to returns.

1.3.2 Similarity

I use textual similarity as a measure of informativeness to argue that the GRS and related rule changes led to a decrease in the informativeness of reports. The measure I employ is based on the cosine similarity between analyst reports, a standard measure of similarity between two documents of text. Jurafsky and Martin (2017), the authors of one of the most popular textbooks about language processing, refer to it as "[b]y far the most common similarity metric." In this section, I describe how I calculate this measure.

After cleaning the text as described in section 3.1 to arrive at a useful document term matrix, I calculate the cosine similarity between the report in question and all reports about the same firm that came out in the prior three months (in robustness checks I use the prior month or the prior six months), not including the day of the report, and take their average. Equation 1 specifies how this is calculated. D is the time period (the prior three months not including today), K is the number

of reports in that time period, n is the number of words in the vocabulary, $W_{i,j}$ is the number of occurrences of word j in document i.

$$AvgSimilarity_{i,k} = \frac{1}{K} \sum_{k \in D} \frac{\sum_{j=1}^{n} W_{i,j} W_{k,j}}{\sqrt{\sum_{j=1}^{n} W_{i,j}^2} \sqrt{\sum_{j=1}^{n} W_{k,j}^2}} \tag{1.1}$$

The **Similarity, Other** and **Similarity, Self** measures are calculated in a similar way. **Similarity, Other** is the similarity of the report in question with all prior reports about the firm that were written by other analysts during the period. **Similarity, Self** is the opposite, i.e. the similarity between the report in question and all reports about the firm written by the same analyst during the period.

A key challenge in my paper is dealing with disclosure and form language — i.e. language that is not related to new information but rather things like legally mandated disclosures and company or analyst specific document templates. Brokerages will change their disclosure language over time which could contaminate my research design if the changes are related to the investment banking status of the brokerage. If there is an overall rise in the length of disclosure language over time that is sharper for investment banks than other brokerages, it could lead to spurious results in my settings. Disclosure could also be problematic because the regulation changes themselves made several specific requirements about disclosures, and this could drive the similarity results I document. I deal with this issue in a couple of ways, both by removing disclosure language and re-doing the analysis on that sample, and by controlling for similarity driven by the disclosure channel. As a control variable for disclosures, I create a similarity measure between an analyst report and other reports written by the same analyst but about different firms called **Self-Similarity, Other Firms**. Under the assumption that disclosure language will be the same or very similar for the same analyst at a given point in time across covered firms, this variable will capture the part of **Similarity** that is driven by disclosure language.

1.3.3 Sentiment

When I discuss how changing incentives led to a decrease in report bias, sentiment (often referred to as **tone** in the literature) is a key variable of analysis. This section describes what sentiment means and how it is calculated.

I use a common and simple way of calculating report sentiment, namely I count the number of positive and negative words and scale the difference by the report length. I also consider positivity and negativity separately by constructing measures that are the fraction of positive words in the report or the fraction of negative words in the report.

To create these measures, I need a list of words classified based on their positivity and negativity. Several such word lists exist and I use the one created by Loughran and McDonald (2011) that is available on Bill McDonald's website. The Loughran and McDonald word lists have been constructed specifically for financial report text and they argue that their lists are more appropriate to classify sentiment of financial text than other competing word lists.

1.3.4 Weaseling

Since bias in the pre-period was a key driver that led to the GRS, it is possible that sentiment was exceptionally high in that particular period and any effects identified by studying variables based on sentiment, earnings forecasts, or recommendations could be driven by mean reversion (Ashenfelter, 1978). With this in mind, I create a more direct measure of bias using weasel words. This exercise is also useful because it sidesteps concerns apparent in prior studies that optimism and stock market reactions to optimism could be jointly caused by new information.

The term "weasel word" is used by Wikipedia to describe language that is aimed at creating a false impression of accuracy. The concept is closely related to "linguistic hedges" in the natural language processing literature, which is defined as language which indicates that the speaker or writer is not backing up statements with facts. I use weaseling as a metric in evaluating how incentive changes relate to changes in textual bias under the assumption that weasel statements

are a natural consequence of trying to justify an incorrect (biased) analysis.

A key challenge in this endeavor is of course to classify analyst statements as weasels or not. Fortunately, Wikipedia's style guide recommends that editors avoid this type of language, and it asks editors to tag weasel statements that cannot be fixed on the spot. This has been recognized and evaluated as a classifier for linguistic hedges by Ganter and Strube (2009), who collect weasel tagged sentences from Wikipedia and argue that such a classifier is useful for a broad domain of language and text. Specifically, they find that the number of words and the distance from the weasel are strong classifiers for weasel language.

Inspired by the work of Ganter and Strube (2009), I download a fully tagged Wikipedia dump that contains all text data from Wikipedia, including all tags. I then collect all articles containing tags beginning in "{{weasel". At this point, I collect every sentence containing the tag. I then clean out tags that occur in the beginning of sentences as they often refer to the prior sentence. Similarly, I clean out tags that occur before the body of the article as these are often tags that apply to the entire document, inspection shows that these are often articles of poor quality with many issues. As a control sample, I collect sentences that occur three sentences after a weasel tag under the assumption that someone who finds and tags a weasel word would be likely to identify another one if it occurs shortly after.

Armed with weasel sentences and control sentences, I build a classifier to be applied to sentences in the analyst report text sample. Specifically, I reduce weasel and control sentences to a document term matrix where each sentence is a row (document). I then remove 100 classified sentences of each kind, i.e. 100 weasels and 100 controls. These 200 sentences make up the so called "held out" sample, which will be used to evaluate the out of sample fit of the classifier. Next, I fit a maximum entropy classifier to the sample without the held out documents and save the coefficients of the model. With only two groups, this amounts to predicting the weasel tag by using word counts in sentences with a logistic regression. This model is saved and applied to all sentences in the analyst text data. The measure of weaseling is then the proportion of sentences in each document classified as a weasel sentences using this method.

As a sanity check, I look at the performance of the weasel classifier within the already tagged held-out sample, i.e. an out of sample test. I find that it correctly classifies 40% of the weasel sentences as weasels and incorrectly classifies 10% of non-weasels as weasels. In other words 40% true positive and 10% false positive.

1.3.5 Classifying Investment Banks

I define an analyst employer as an investment bank if I can match the name of the brokerage from Investext with a name in SDC. Specifically, I download the lead manager field from the SDC Platinum equity issue and debt issue databases. Unlike the "manager parent" field, the "lead manager" field contains the lead underwriter with the current name as of the issue. Since name fields can differ between Investext and SDC Platinum, I use a semi-supervised matching strategy based on string matching followed by manual clean up of matches.

I initially match the first four characters of the brokerage names in my analyst sample with the first four characters of the names in the SDC sample in all cases when four character combinations are unique. I go over the results and clean out bad matches. As a second step, I use the same methodology for the first eight characters of the names and again clean out bad matches. As a last step, I match all the remaining names to their closest match using a fuzzy string matching algorithm that considers how many edits (substitutions, deletions, insertions) are required to go from one string to the next. Finally, I go through and clean out incorrect matches manually.

I use the investment banking classification described above because the regulation changes affected investment banks specifically and in many ways without regards to the relationship between the brokerage and the analyzed firm itself. The same investment banking classification has been used in prior studies, for example by Guan, Lu and Wong (2013) and Ertimur et al. (2007).[5] The literature on analyst conflicts of interest has used a few different methods to measure the severity of conflicts. Much of the early work on conflicts used affiliation between the investment bank and the

[5] Guan, Lu and Wong (2013) classify investment banking using SDC underwriting relationships and Nelson's Directory of Investment Research. Ertimur et al. (2007) use a classification based on the investment banking status of the brokerage house.

firm as an indicator for conflicted analysts. The idea is that conflicts are stronger when there is an existing relationship with the firm that is being analyzed.[6] A potential weakness of the affiliation classification scheme, as Ertimur et al. (2007) point out, is that conflicting incentives arise from the possibility of future investment banking business rather than past relationships. In my setting, the mechanism works via the status of the employer - i.e. employers that are investment banks are the ones supposed to follow these regulations.

1.4 Empirical Strategy

I use an identification strategy based on how the rules and regulations surrounding the GRS affected investment banks and other brokerages differentially.[7] I therefore classify each analyst in my sample based on whether their employer is an investment bank or not, and compare the output of investment banking analysts before and after the GRS with non-banking analysts using a difference-in-difference methodology.

The first main assumption is that the rules and regulation changes had a larger effect on investment banking analysts than non-banking analysts. The GRS explicitly focused on the conflict of interest between investment banking and analyst research, as did at least parts of each rule issued in this period. The second main assumption here is equal trends in outcomes between treatment groups, i.e. trends in similarity, weaseling, and sentiment for investment banking connected analysts and non-connected analysts.

[Figures 1.1-1.3]

As an initial check on the equal trends assumption, I plot the difference in outcome variables over time between the two groups. Figures 1.1-1.3 show these plots. In the cases of similarity and weaseling, trends are approximately equal. There is reason to be more suspicious of this assumption in the case of sentiment and its components because there is a sharp increase in sentiment among investment banking analysts leading up the the GRS. The specific concern is that since bias in the

[6] See Michaely and Womack (1999) and Lin and McNichols (1998).

[7] A similar strategy has been used to study other analyst outcomes by Guan, Lu and Wong (2013), Barber et al. (2006), Chen and Chen (2009), Kadan et al. (2009), and others.

pre-period was a key driver that led to the GRS, it is possible that sentiment was exceptionally high in that particular period and any effects identified by studying variables based on sentiment, earnings forecasts, or recommendations could be driven by mean reversion. This is part of the motivation for why I use weaseling to detect bias.

Policy changes do not occur in a vacuum and this is no different in my setting. In fact, the GRS and related regulations happened in response to a number of scandals and the regulation could plausibly have been anticipated by sell-side analysts. It is difficult to directly deal with this issue, so what I do is look at all the covariates for the treated and control groups just prior to the GRS.

Table 1.1 shows a covariate balance table for the treatment and control groups prior to the GRS. I merge each report with the average IBES information releases the day of the report.[8] In other words, I match this data based on firm and date. There are a few differences between investment banks and other brokerages in the pre-period. Most notably, annual forecasts are more negative on average for the non investment bank brokerages and investment banking analysts tend to cover slightly more firms. On the report level, investment banking reports tend to be significantly longer, something I deal with by controlling for length of report in all text related tables.

The differential timing of all the related regulation changes around the GRS present another problem. I deal with this by excluding the period of regulation changes. The wave of regulation changes came after a lawsuit by the attorney general of New York state against Merrill Lynch, where Merrill Lynch were accused of producing misleading research. This lawsuit was settled in the spring of 2002 (requiring Merrill Lynch to pay $100 million in fines without an admission of wrongdoing). The Merrill Lynch lawsuit was followed by another lawsuit that led to the GRS, affecting ten large Wall Street firms, and the regulation changes studied in this paper. In all regression tests in the paper, I exclude the period of regulation changes (2002-2003) and compare the difference between investment banking analysts and other analysts in each outcome variable

[8] I collapse IBES information by day and firm because I cannot reliably match reports in my sample to forecasts or recommendation changes in IBES.

Table 1.1: Pre-Period Covariate Balance

Note: **Earnings Announcement** is a dummy which is set to one if there is an earnings announcement the day of the report. **Earnings Forecast** is a similar dummy which is set to one if there is at least one new earnings forecast recorded in IBES the day of the report. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **SUE** is the standardized unexpected earnings measure, in other words it is the difference between the forecast and the surprise scaled by the standard deviation in this difference. **Boldness** is defined as the absolute difference between the earnings forecast in the report and the current consensus (the mean outstanding forecast for the period) scaled by the stock price of the firm, Q denotes the quarterly measure and A denotes annual. **Forecast** variables are calculated as **boldness** but are signed. **Experience** is the number of years since the analysts first showed up in IBES. **Breadth** is the number of firms covered by the analyst and **Coverage** is the number of firms covered by the brokerage house. **N per IBES date** are the number of reports per day when there is at least one report. **Word Count** and **Digits** are the counts of words and digits in the textual report, respectively. Returns are the CAR over the window, adjusted using the market model. Returns are scaled as basis points. All IBES variables are calculated by firm-day and merged with the text sample on that dimension. This means that any two reports issued about the same firm on the same day will have the same IBES data.

Variable	All	IB	Non-IB	Diff
IBES Measures				
Earnings Announcement	0.11	0.11	0.12	−0.00
SUE	0.08	0.07	0.09	−0.02
Earnings Forecast	0.45	0.45	0.44	0.02
Upgrade	0.06	0.06	0.06	0.00
Downgrade	0.07	0.07	0.07	0.00
Boldness (A)	0.17	0.16	0.17	−0.01
Boldness (Q)	0.05	0.05	0.04	0.01
Forecast (A)	−0.04	−0.02	−0.11	0.09***
Forecast (Q)	0.02	0.02	0.02	0.00
Log(Experience)	2.43	2.43	2.41	0.02
Log(Breadth)	2.79	2.80	2.76	0.04**
Log(Coverage)	6.67	6.68	6.67	0.01
N per IBES date	9.08	9.06	9.15	−0.10
Report Measures				
Log(Word Count)	6.86	6.92	6.67	0.24***
Log(Digits)	6.53	6.58	6.37	0.21***
Returns				
Return (0,0)	−17.92	−16.09	−23.57	7.48
Return (−1,1)	−29.14	−26.66	−36.82	10.2
Abs(Return (0,0))	311	309	316	−7.41

before and after the excluded period. Most such results are calculated using the two years prior to changes and the two years after the changes. As robustness, I have also looked at only the year prior and the year after, this is particularly important since regulation fair disclosure was implemented during the beginning of the pre-period in my main tests.

1.5 Results

1.5.1 Summary Statistics

Summary statistics are presented in table 1.2. This table shows the average of each of the text measures by investment banking association and in the pre regulation and post regulation periods.

The table shows that there are a total of 127,000 observations from 2000-2001 and 2004-2005, a little under three quarters of the observations are from investment banking analysts. We also see that overall similarity goes up between the two periods and that most of this increase is from the investment banking analysts. Sentiment of reports decreases for investment banking analysts, who, relative non-banking analysts, reduce the sentiment in their reports sharply. They similarly decrease their use of weasel language. These univariate results are also visualized in figures 1.1-1.3.

1.5.2 Informativeness

Further, if **Similarity** is a measure of how informative reports are, as I have conjectured, it should have a predictable relationship with stock market reactions to reports. Reports that contain less information should have a smaller market moving effect than reports with more information. Figure 1.4 plots the absolute value of the stock market reaction against the decile of the **Similarity** score.

The figure above illustrates the use of **Similarity** as a measure of (lack of) informativeness. The X-axis shows similarity deciles and the Y-axis shows the absolute value of the abnormal return in the report window. In the case of similarity with ones own reports (**Similarity, Self** in panel

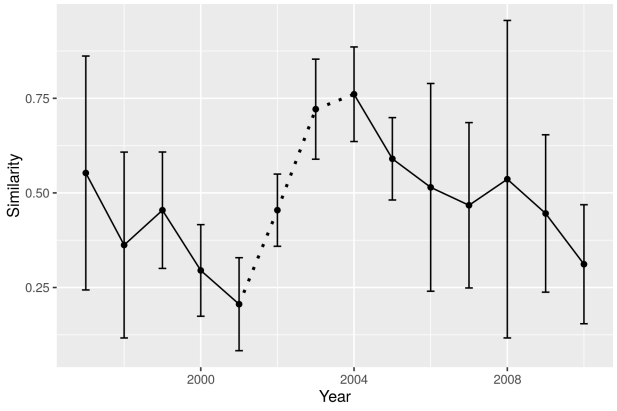
Table 1.2: Summary Statistics

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. IB Analysts are analysts who are employed by firms who provide investment banking services, Non-IB are other analysts. **Weaseling** is measured as the fraction of sentences in the document classified as weasels. **Sentiment** is measured as the number of positive words minus the number of negative words scaled by the length of the document. **Positivity** and **negativity** are defined analogously using either positive or negative words only. **N** represents the number of observations which are written by **N Analysts** different analysts from **N Brokerages** different brokerages. All measures (except counts) are centered at zero and normalized to have a standard deviation of 1.

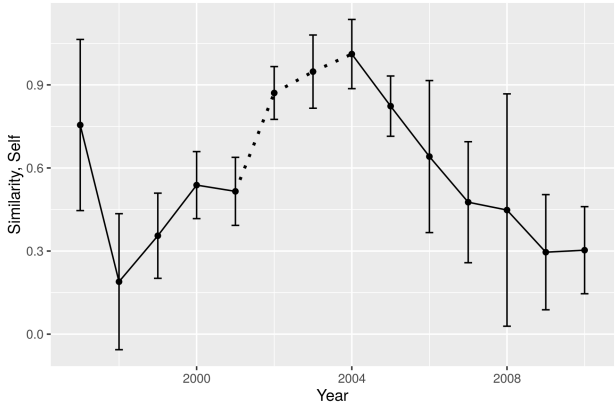
	Pre (2000-2001)			Post (2004-2005)			
Variable	IB	Non-IB	Diff	IB	Non-IB	Diff	DiD
Similarity Measures							
Similarity	-0.49	-0.74	0.25	0.47	-0.22	0.69	0.43***
Similarity, Other	-0.62	-0.78	0.16	0.50	-0.08	0.58	0.42***
Similarity, Self	-0.31	-0.81	0.50	0.45	-0.44	0.89	0.38***
Content Measures							
Sentiment	0.17	-0.15	0.32	-0.08	0.05	-0.13	-0.45***
Positivity	0.15	0.28	-0.13	-0.09	-0.07	-0.02	0.11*
Negativity	-0.08	0.53	-0.61	0.02	-0.15	0.17	0.78***
Weaseling	-0.08	-0.20	0.13	0.10	0.06	0.03	-0.09*
Observations							
N	32659	10837	21822	60272	20342	39930	18108
N Analysts	670	174	496	769	245	524	28.0
N Brokerages	44.0	13.0	31.0	38.0	16.0	22.0	-9.00

Figure 1.1: Difference-in-Difference - Similarity

Note: This figure shows how differences in **Similarity** between investment banking analysts and other analysts moves over time. **Similarity** is measured as the cosine similarity between the report and all other reports about the same firm that were issued in the prior three month. **Similarity, Self** is defined analogously, but using reports authored by the writer of the current report only.. **Similarity** is averaged by year for each of the two groups (investment banking analysts and non-investment banking analysts) and the difference is then plotted from 1998 to 2010. The dotted part of the plot indicates the period which I remove in the regression analysis that follows (the period of regulation change). All measures are centered at zero and normalized to have a standard deviation of 1.



[Similarity]



[Similarity, Self]

Figure 1.2: Difference-in-Difference - Weasel

Note: This figure shows how differences in **weaseling** between investment banking analysts and other analysts moves over time. **Weaseling** is measured as the fraction of **weasel sentences** in the document. The measure is then averaged by year for each of the two groups (investment banking analysts and non-investment banking analysts) and the difference is plotted from 1998 to 2010. The dotted part of the plot indicates the period which I remove in the regression analysis that follows (the period of regulation change). All measures are centered at zero and normalized to have a standard deviation of 1.

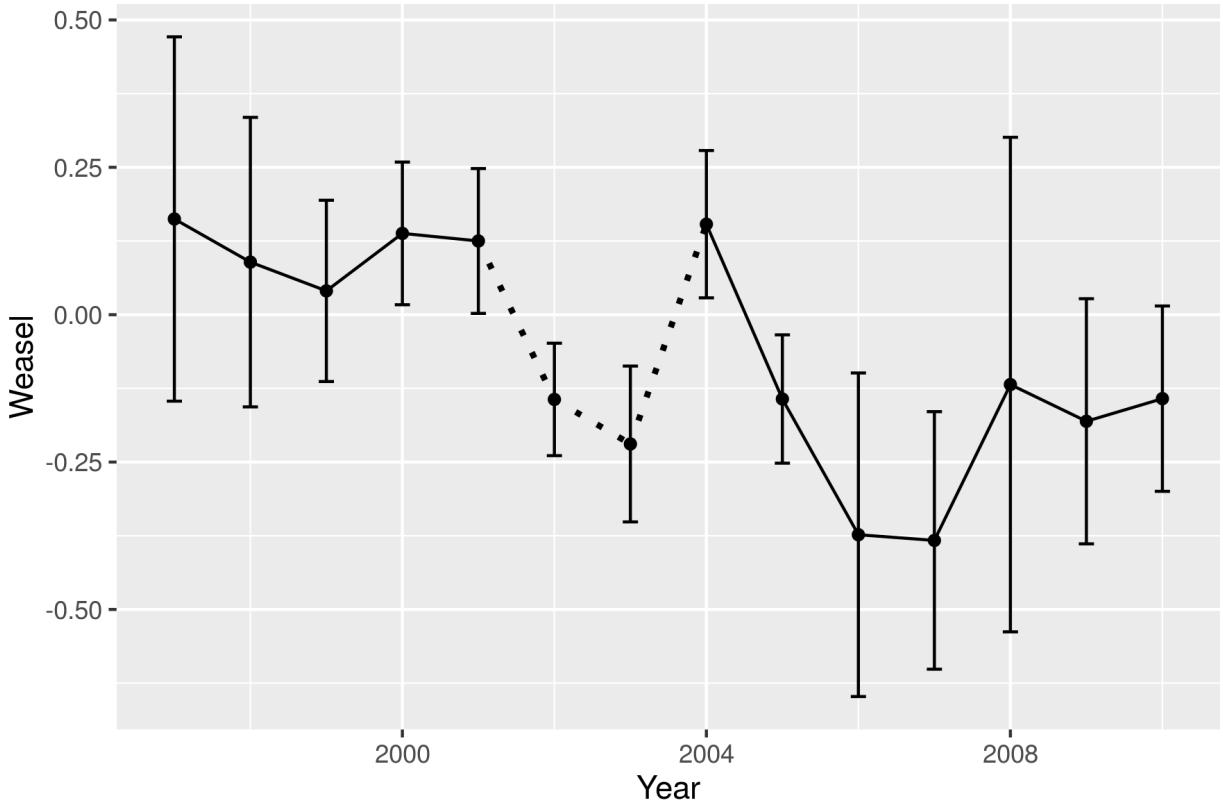
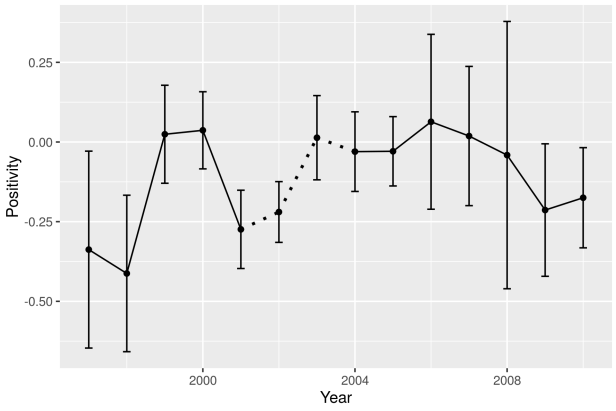
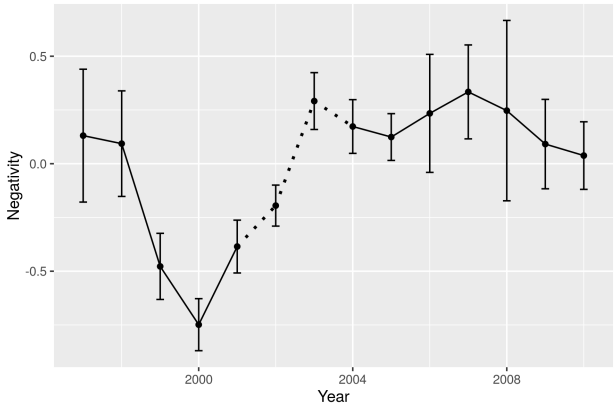


Figure 1.3: Difference-in-Difference - Sentiment

Note: This figure shows how differences in **positivity** and **negativity** between investment banking analysts and other analysts move over time. **Positivity** is measured as the fraction of positive words in the document, **negativity** is defined analogously. The measure is then averaged by year for each of the two groups (investment banking analysts and non-investment banking analysts) and the difference is plotted from 1998 to 2010. The dotted part of the plot indicates the period which I remove in the regression analysis that follows (the period of regulation change). All measures are centered at zero and normalized to have a standard deviation of 1.



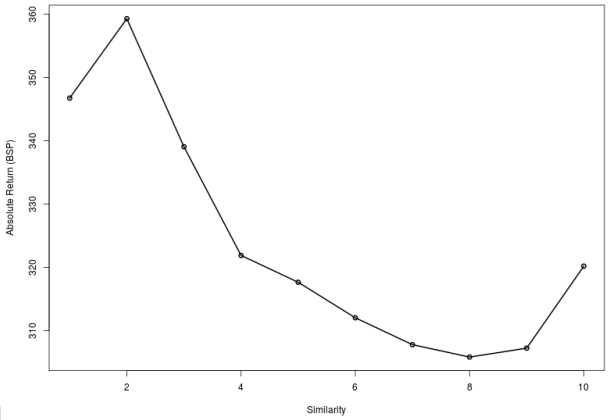
[Positivity]



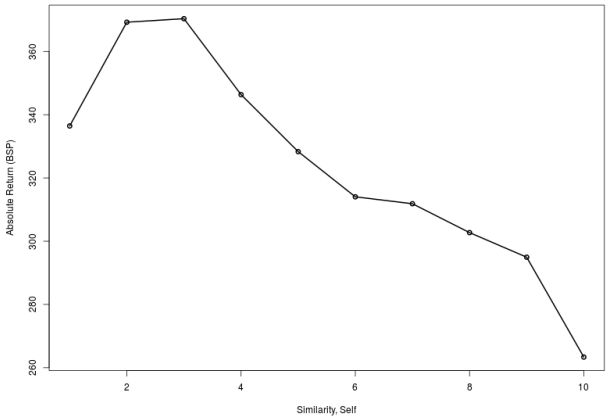
[Negativity]

Figure 1.4: Absolute CAR and Similarity

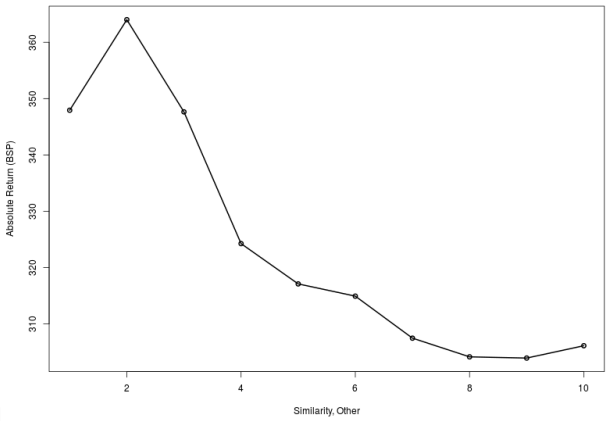
Note: This plot shows the relationship between the absolute value of the stock market reaction (in basis points) and **Similarity** in reports. **Similarity** is measured as the cosine similarity between the report and all other reports about the same firm that were issued in the prior three month. **Similarity, Self** is defined analogously, but using reports authored by the writer of the current report only. **Similarity, Other** uses reports authored by all analysts except the writer of the current report. All **Similarity** measures are centered at zero and normalized to have a standard deviation of 1.



[Similarity]



[Similarity, Self]



[Similarity, Other]

b), a move from the first to the tenth decile reduces the absolute value of the stock market reaction by almost 80 basis points. The other similarity measures show similar albeit weaker relationships with the size of the market reaction. Table 1.3 panel (a) shows the regression form of the same univariate relationship. A one standard deviation increase in **Similarity, Self** leads to 23 basis points lower stock market reactions.

Next, I dig further into the relationship between market reactions and **Similarity** scores. We see in table 1.5 that **Similarity** is related to the timing of firm events, as in the case of earnings announcements, and also with information events, as with upgrades and downgrades. In table 1.4, I control for earnings announcements, upgrades, downgrades, and surprises and ask how is **Similarity** related to the information content of upgrades and downgrades. I.e. conditional on an information event, what does **Similarity** say about the magnitude of the information during the event. If **Similarity** indeed is a measure of informativeness, I would expect to see that upgrades or downgrades with more information (less **Similarity**) would have a larger market reaction. Table 1.4 shows results consistent with this intuition. In columns 1-3, an upgrade with a one standard deviation higher **Similarity** score has a stock market reaction that is about 20 basis points smaller. For downgrades, the effect is a little bit bigger at 30-40 basis points. Columns 4-6 get at the same question but using the regular, signed, return measure. Here, we again see that downgrades with higher **Similarity** scores are followed with a smaller stock market hit than downgrades with low **Similarity** scores. For upgrades, the results point in a consistent direction but are not significant.

In the results that follow, I calculate the average cosine similarity with all reports about the same firm over the prior three months, not including the day of the report. As we will see shortly, analyst report informativeness is reduced significantly following the exogenous shock. If the GRS indeed had the unintended consequence of altering analyst incentives to provide and produce information, we should expect to see a difference in their output pre and post. Although the recommendation and forecast output prior to and after global settlement has been well studied, it is unclear what effect the regulation changes should have on the accompanied text for a couple of reasons. First of all, changes to analyst incentives in this period was driven partly by outrage

Table 1.3: Stock Market Reactions to Similarity - Univariate Results

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. Similarity measures are centered at zero and normalized to have a standard deviation of 1. The data is collapsed on firm-days to measure the reaction to the average similarity. **CAR** refers to the 3-day abnormal returns around the report using the one factor model. **Abs(CAR)** is the absolute value. Abnormal returns are winsorized at 1% and 99%. Errors are double clustered on firm and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:		
	Abs(CAR)		
	(1)	(2)	(3)
Similarity	-13.156^{***}		
	(4.196)		
Similarity, Self		-22.650^{***}	
		(4.342)	
Similarity, Other			-13.409^{***}
			(4.024)
Observations	382,943	382,943	382,943
Adjusted R^2	0.001	0.002	0.001
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 1.4: Stock Market Reactions to Similarity

Note: This table presents evidence related to the stock market response on the day of analyst reports. **Similarity, Self** is the similarity with prior reports that were written by the same analyst during the prior three months, the measure is standardized to have a deviation of 1 and mean of 0. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **CAR** refers to the 3-day abnormal returns around the report using the one factor model. **Abs(CAR)** is the absolute value. Abnormal returns are winsorized at 1% and 99%. Other controls are **Earnings Announcement, Earnings Forecast, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. All covariates are presented in table A.12. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Abs(CAR)			CAR		
	(1)	(2)	(3)	(4)	(5)	(6)
Similarity, Self	−5.865**	−3.467*	5.746***	0.279	1.452	1.143
	(2.591)	(2.034)	(1.729)	(1.914)	(2.161)	(2.145)
Upgrade	166.663***	145.656***	143.346***	290.040***	291.580***	290.576***
	(8.804)	(8.206)	(8.127)	(12.951)	(12.770)	(12.780)
Downgrade	253.163***	230.310***	226.188***	−406.978***	−403.450***	−403.893***
	(10.375)	(9.689)	(9.661)	(17.384)	(16.122)	(16.076)
Similarity, Self × Upgrade	−20.276***	−18.282***	−18.264***	−14.829	−12.177	−11.966
	(5.720)	(5.616)	(5.457)	(9.086)	(9.018)	(9.000)
Similarity, Self × Downgrade	−42.285***	−35.381***	−31.749***	49.403***	48.681***	48.840***
	(7.160)	(6.694)	(6.653)	(10.925)	(10.412)	(10.401)
Controls	X	X	X	X	X	X
Year FE			X			X
Analyst FE		X	X		X	X
Observations	382,943	382,943	382,943	382,943	382,943	382,943
Adjusted R^2	0.071	0.146	0.172	0.072	0.080	0.081
Note:						*p<0.1; **p<0.05; ***p<0.01

in the popular press. This outrage was often specifically targeted at investment bankers giving favorable recommendations to clients and prospective clients. Further, the GRS and associated rules mentioned recommendations and forecasts specifically, but discussed text only in terms of disclosures and certification (i.e. Reg AC).

There are also reasons to suspect that the informativeness of reports did change. During this period, analyst pay went down, and in particular, incentive pay went down (see Groysberg, Healy and Maber (2011)). While the objective of cutting the tie between investment bankers and analysts seems to have been at least partially met (i.e. the finding in Kadan et al. (2009) that they issue fewer buy recommendations and more sell recommendations), it is unclear to what extent such a change of incentives would change text output. The motivation for cutting the ties between bankers and compensation was to reduce incentives to misbehave, it is plausible that the changes also led to reduced information production.[9] If investment banking connected analysts lost incentives to produce or provide valuable information, we would expect their reports to be more similar to prior reports stemming from a reduced incentive to exert effort.

1.5.2.1 Determinants of Similarity

Before I get into the effect that the GRS had on the informativeness of analyst reports, I look at the determinants of similarity. Again, similarity is measured in three pieces. **Similarity**, which is the average similarity with all reports about the firm in the prior three months; **Similarity, Other**, which is the piece that is due to similarity with reports written by other analysts; and **Similarity, Self**, which is the piece due to similarity with reports by the same analyst. An increase in similarity in investment banking analysts reports could, as I will interpreted it, be due to a reduction in novel information. Another possibility is that investment banking analyst report similarity goes up because the regulation changes had exactly the intended effect. In such a story, investment banking analysts were producing reports in order to drive business to the investment

[9] A similar, but distinct, mechanism was found by Guan, Lu and Wong (2013), who show evidence of a brain drain in response to global settlement.

banking arm of their firm and were therefore producing reports that looked less like reports by other analysts. An increase in similarity may then reflect a shift away from this type of misleading output that would make reports in the post period look more like reports from other analysts.

Table 1.5 shows the relationship between similarity and the timing of these reports (i.e. if they were issued the same day as an announcement or a forecast update), as well as between similarity and upgrades, downgrades and report length.

The analyst output that has by far the largest impact on stock prices are upgrades and downgrades, presumably because such reports contain the most information. It is therefore comforting that upgrades and downgrades are both associated with lower similarity, consistent with my interpretation of dissimilarity as a measure of informativeness. Another interesting pattern is that reports that come out during an announcement or forecast event tend to be more similar to prior reports than reports issued in other time periods. This could simply be a reflection of reports issued in other time periods coming in response to information events. As an example, an announced merger would tend to be followed by analyst reports. These reports would likely look quite different from reports in the prior three months and this type of effect could drive the differences between announcement periods and other periods.

Upgrades have a somewhat stronger relationship with dissimilarity than downgrades, and both upgrades and downgrades are more important for the piece of similarity that is due to same analyst similarity.

1.5.2.2 Regulation Impact on Similarity

Next, I turn to the effect of the GRS on similarity by testing how similarity changed differentially between investment banking analysts (who were more affected by the GRS) and other analysts.

Indeed, I find that investment banking analysts greatly increase their similarity with prior reports following global settlement relative non-banking analysts. Table 1.6 shows that in the post period, similarity is up by 0.36-0.50 standard deviations for the connected analysts relative the

Table 1.5: Determinants of Similarity

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. Similarity measures are centered at zero and normalized to have a standard deviation of 1. **Earnings Announcement** is a dummy which is set to one if there is an earnings announcement the day of the report. **Earnings Forecast** is a similar dummy which is set to one if there is at least one new earnings forecast recorded in IBES the day of the report. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **SUE** is the standardized unexpected earnings measure, in other words it is the difference between the forecast and the surprise scaled by the standard deviation in this difference. **Boldness** is defined as the absolute difference between the earnings forecast in the report and the current consensus (the mean outstanding forecast for the period) scaled by the stock price of the firm, Q denotes the quarterly measure and A denotes annual. **Forecast** variables are calculated as **boldness** but are signed. **Experience** is the number of years since the analysts first showed up in IBES. **Breadth** is the number of firms covered by the analyst and **Coverage** is the number of firms covered by the brokerage house. **# Reports Today** are the number of reports during the day. **Word Count** and **Digits** are the counts of words and digits in the textual report, respectively. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Similarity		Similarity, Other		Similarity, Self	
	(1)	(2)	(3)	(4)	(5)	(6)
Earnings Announcement	0.111***	0.087***	0.084***	0.068***	0.062***	0.052***
	(0.012)	(0.014)	(0.012)	(0.014)	(0.011)	(0.012)
Earnings Surprise (SUE)	0.001	0.002	0.001	0.003	0.001	0.002*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Earnings Forecast	0.035***		0.046***		0.020***	
	(0.008)		(0.008)		(0.008)	
Upgrade	−0.064***	−0.066***	−0.051***	−0.054***	−0.070***	−0.062***
	(0.012)	(0.014)	(0.012)	(0.014)	(0.010)	(0.012)
Downgrade	−0.020**	−0.039***	−0.020**	−0.037***	−0.040***	−0.046***
	(0.010)	(0.011)	(0.010)	(0.012)	(0.011)	(0.012)
Log(Digits)	0.042***	0.015	0.031**	0.019	0.113***	0.074***
	(0.014)	(0.017)	(0.014)	(0.016)	(0.015)	(0.019)
Log(Words)	0.615***	0.703***	0.738***	0.794***	0.111***	0.204***
	(0.021)	(0.024)	(0.022)	(0.025)	(0.024)	(0.029)
# Reports Today	−0.001	−0.001*	−0.0004	−0.0003	−0.001*	−0.001***
	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0004)	(0.0004)
Log(Experience)		0.019*		0.013		0.005
		(0.010)		(0.010)		(0.010)
Log(Breadth)		−0.007		−0.011		−0.011
		(0.018)		(0.019)		(0.017)
Boldness Annual		0.005		0.004		0.012*
		(0.007)		(0.007)		(0.006)
Boldness Quarterly		−0.013		−0.013		−0.0001
		(0.011)		(0.012)		(0.006)
Year FE	X	X	X	X	X	X
Analyst FE × Firm FE	X	X	X	X	X	X
Observations	124,110	60,433	124,110	60,433	124,110	60,433
Adjusted R^2	0.688	0.701	0.700	0.710	0.705	0.721
Note:						*p<0.1; **p<0.05; ***p<0.01

non-connected analysts. In other words, they changed their textual output in a significant and specific way. Notice that this effect is robust to controlling for the analyst by analyzed-firm pair, so the effect is not simply a reflection of the brain drain found in Guan, Lu and Wong (2013).

I also consider an alternative measure of similarity that is based on reports in either a longer or a shorter window, the last month or the prior six months (see table A.8). Results remain similar both qualitatively and quantitatively.

1.5.2.3 Is Similarity Herding?

At this point, I have shown that overall similarity with prior reports increases as a response to the GRS and I have shown results that hint at an interpretation of dissimilarity as a measure of report informativeness (i.e. the relationship between similarity and recommendation changes in table 1.5 and the size of the market reaction in figure 2). This immediately looks similar to analyst herding, where analysts tend to herd around the same earnings forecasts (i.e. they issue similar earnings forecasts).

In table 1.7, I split the similarity measure into two pieces, one that is the similarity due to reports written by others and another that is the similarity due to reports written by the same analyst. Traditional herding is analogous to similarity with other analysts' reports (i.e. the part captured by **Similarity, Other**). Results show that there is a difference-in-difference increase in both types of similarity in response to the GRS, but that the increase is much sharper for **Similarity, Self**. This suggests a mechanism separate from the traditional herding explanations, potentially one driven by effort. If analysts reduce effort as a response to weakened incentives, a natural outlet for this reduction is to update reports less. Of course, another potential explanation for the larger magnitudes are that disclosure length increased more for investment bankers and this should be more apparent when comparing a report to another report written by the same analyst. I deal with this potential explanation next.

Table 1.6: Similarity with Prior Reports

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.4. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:				
	Similarity				
	(1)	(2)	(3)	(4)	(5)
IB × Post	0.363***	0.386***	0.388***	0.515***	0.500***
	(0.067)	(0.046)	(0.046)	(0.052)	(0.060)
Investment Bank	0.074	0.055	0.053		
	(0.054)	(0.034)	(0.033)		
Post	0.136**	0.153***			
	(0.069)	(0.050)			
Other Controls	X	X	X	X	X
Firm FE		X	X	X	
Year FE			X	X	X
Analyst FE				X	
Analyst FE × Firm FE					X
Observations	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.398	0.543	0.543	0.648	0.684
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1.7: Similarity with Prior Reports - Self vs Other

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.5. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

[Similarity with others vs similarity with self]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	1.008***	0.874***	0.886***	0.664***	0.382***	0.367***
	(0.071)	(0.066)	(0.070)	(0.072)	(0.061)	(0.065)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE	X	X		X	X	
Analyst FE × Firm FE			X			X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.635	0.656	0.709	0.491	0.622	0.695
Note:						*p<0.1; **p<0.05; ***p<0.01

[Controlling for time varying analyst-level similarity]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.355***	0.298***	0.279***	0.155	−0.010	0.003
	(0.088)	(0.079)	(0.082)	(0.120)	(0.112)	(0.116)
Self-Similarity, Other Firms	0.584***	0.560***	0.582***	0.384***	0.298***	0.347***
	(0.019)	(0.017)	(0.018)	(0.036)	(0.029)	(0.027)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE	X	X		X	X	
Analyst FE × Firm FE			X			X
Observations	30,654	30,654	30,654	30,654	30,654	30,654
Adjusted R^2	0.724	0.735	0.790	0.527	0.656	0.734
Note:						*p<0.1; **p<0.05; ***p<0.01

1.5.2.4 Is Similarity Disclosure and Form Language?

In this subsection, I distinguish similarity due to disclosure (more generally, form language) from similarity due to lack of new information. Form and disclosure language refers to parts of the reports that are not related to information production but that are either legal disclosures or other text written based on templates unrelated to information, such as headers and footers. To accomplish this, I create a similarity measure between an analyst report and other reports written by the same analyst but about different firms called **Self-Similarity, Other Firms** and use this new measures as a control for disclosures. Under the assumption that disclosure language will be the same or very similar for the same analyst across multiple firms, this variable will capture the part of **Similarity** that is driven by disclosure language. This exercise is only possible on the limited subset of analysts for which I have reports for multiple firms. There are relatively few such analysts since my sample is restricted to firms in the S&P500. The smaller sample for which I can measure self-similarity with other firms has a similar balance between investment banking analysts and other analysts, i.e. about three quarter investment banking analysts and one quarter other analysts.

Table 1.7, panel (b), show the results. Several interesting things happen. First of all, **Self-Similarity, Other Firms** is an important predictor of both **Similarity, Other** and **Similarity, Self**. This is a natural result since at least some disclosure language is legally mandated and there may be industry wide or economy wide information that is common between reports for different firms. The coefficient on **Similarity, Self** is larger than the coefficient on **Similarity, Other**, consistent with analyst reports by the same analyst containing identical headers and disclosures. More interestingly, controlling for disclosures and form language using **Self-Similarity, Other Firms** completely explains away the increase in **Similarity, Other** but not **Similarity, Self**. This suggests that investment banking analysts reduce their information production by editing their reports less over time than in the pre period. In the robustness section, I control for disclosures in an alternative way and find similar results.

1.5.2.5 Similarity and Confounding News

To rule out stories related to the timing between of analyst reports changing differentially between investment banking analysts and other analysts, I run my tests with firm-day fixed effects.

Table 1.8 shows results using the firm-day fixed effect. In this table, the effect is identified by comparing the treatment (banking analysts) and control (other analysts) groups before and after the period of regulation change on the same day for the same firm. This rules out a broad class of altenative stories that are either related to changing timing of reports or related to time-varying firm variables. Panel (b) adds the disclosure controls from the prior section. Results remain similar to results in prior tables, both in terms of coefficient sizes and in terms of significance.

1.5.2.6 Similarity and Banking Relationships

Here, I look at to what extent results are driven by existing banking relationships and the potential of future banking relationships. The idea is to differentiate between the information and incentive channels.

I run two different specifications in table 1.9. First, I look at whether a past bankning relationship between the analyst's employer and the analyzed firm extenuates the effect, panel (a) shows the results. I find that results are indeed stronger when there is a banking relationship between the analyst firm and the analyzed firm. Unfortunately, this does not help separate between the information and incentive channels since its possible that the analyst has had both a greater incentive shock for these firms, as they can no longer be compensated from investment banking revenues, and that they have less available information about these firm since there are now informational firewalls between the bankers and analysts.

To get at the channel, at least partially, I also run a specification where I interact the shock with whether or not the analyzed firm is in an industry that is reliant on investment banking services. Panel (b) shows these results. The idea here is that the incentive channel is similar between firms that are potential clients and firms that are past clients, but that the information

Table 1.8: Holding Information Constant

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.5. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

[Similarity with others vs similarity with self]

	Dependent variable:					
	Similarity, Self		Similarity		Similarity, Other	
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.963***	1.000***	0.555***	0.564***	0.444***	0.465***
	(0.074)	(0.086)	(0.054)	(0.060)	(0.052)	(0.063)
Other Controls	X	X	X	X	X	X
Firm-Day FE	X	X	X	X	X	X
Analyst FE	X		X		X	
Analyst FE × Firm FE		X		X		X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.730	0.755	0.762	0.775	0.767	0.775
Note:						*p<0.1; **p<0.05; ***p<0.01

[Controlling for time varying analyst-level similarity]

	Dependent variable:					
	Similarity, Self		Similarity		Similarity, Other	
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.406***	0.428***	0.096	0.113	0.175*	0.192
	(0.099)	(0.135)	(0.126)	(0.156)	(0.105)	(0.158)
Self-Similarity, Other Firms	0.577***	0.600***	0.467***	0.502***	0.364***	0.393***
	(0.034)	(0.037)	(0.025)	(0.028)	(0.035)	(0.039)
Other Controls	X	X	X	X	X	X
Firm-Day FE	X	X	X	X	X	X
Analyst FE	X		X		X	
Analyst FE × Firm FE		X		X		X
Observations	30,654	30,654	30,654	30,654	30,654	30,654
Adjusted R^2	0.802	0.824	0.813	0.827	0.798	0.806
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 1.9: Similarity with Prior Reports - Banking Relationship

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Relationship is a dummy equal to one if there is a relationship between the investment bank where the analyst works and the firm during the decade before the sample period, zero otherwise. An industry is classified as banking intensive if it has more than the median level of investment banking activity. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.5. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

[Existing Banking Relationship]

	Dependent variable:		
	Similarity, Self	Similarity	Similarity, Other
	(1)	(2)	(3)
IB × Post	0.850***	0.451***	0.302***
	(0.076)	(0.064)	(0.065)
IB × Post × Relationship	0.139**	0.189***	0.249***
	(0.055)	(0.054)	(0.046)
Other Controls	X	X	X
Year FE	X	X	X
Analyst FE × Firm FE	X	X	X
Observations	114,105	114,105	114,105
Adjusted R^2	0.709	0.684	0.695
Note:			*p<0.1; **p<0.05; ***p<0.01

[Banking Intensive Industries]

	Dependent variable:		
	Similarity, Self	Similarity	Similarity, Other
	(1)	(2)	(3)
IB × Post	0.955***	0.625***	0.475***
	(0.097)	(0.091)	(0.088)
IB × Post × IB Intensive	−0.098	−0.178**	−0.154**
	(0.082)	(0.083)	(0.073)
Other Controls	X	X	X
Year FE	X	X	X
Analyst FE × Firm FE	X	X	X
Observations	113,951	113,951	113,951
Adjusted R^2	0.709	0.684	0.695
Note:			*p<0.1; **p<0.05; ***p<0.01

channel is more important for past clients. My results are indicative of either a smaller drop or no difference in the change of informativeness for reports about potential clients.

Together, a larger drop in informativeness for past clients but not prospective clients hints at the information channel being more important than the incentive channel.

1.5.3 Bias

1.5.3.1 Sentiment

Prior to the GRS, investment banking analysts were accused of producing biased research in order to help their investment banking business. It has been shown that the GRS led to changed recommendations in a way consistent with reduced bias in recommendations (i.e. fewer buy recommendations and more hold and sell recommendations) but this does not necessarily imply a change in the textual content.

However, if analysts respond to altered incentives by changing their writing in response to the GRS, I would expect optimism to go down for investment banking analysts relative other analysts since their incentives to be over-optimistic have gone down. Similarly I would expect negativity to go up in the post settlement era. Results from this exercise are shown in table 1.10.

Indeed we see that the sentiment of communication changed drastically between the pre and post settlement eras. Sentiment drops by around .4 standard deviations for investment banking analysts relative non-banking analysts between the pre and post eras. This is driven primarily by an increase in negativity, consistent with reduced incentives for investment banking analysts to be overly optimistic.

1.5.3.2 Weasel Words

Weasel words, also called anonymous authority, are words which hedge the meaning of a statement, for example “some people think,” “scholars argue,” or “it is generally accepted.” Wikipedia describes these words and phrases as “aimed at creating an impression that a specific or meaningful statement has been made, when instead only a vague or ambiguous claim has actually been

Table 1.10: Change in Sentiment

Note: **Sentiment** is measured as the number of positive words minus the number of negative words scaled by the length of the document. **Positivity** and **negativity** are defined analogously using either positive or negative words only. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Sentiment measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.9. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Sentiment		Positivity		Negativity	
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	−0.467***	−0.434***	−0.011	0.019	0.668***	0.652***
	(0.087)	(0.098)	(0.064)	(0.078)	(0.100)	(0.115)
Other Controls	X	X	X	X	X	X
Firm FE	X		X		X	
Year FE	X	X	X	X	X	X
Analyst FE	X		X		X	
Analyst FE × Firm FE		X		X		X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.291	0.358	0.331	0.375	0.299	0.355
Note:						*p<0.1; **p<0.05; ***p<0.01

communicated." Weasel terms are therefore useful tools when writing reports with information that is consciously biased and I will interpret their reduction in usage in response to the GRS as a reduction in bias.

In table 1.11, we see that banking analysts reduce their use of weaseling words between the pre and post period relative other analysts by about 0.3 standard deviations. Given the discussion above, banking analysts reduced use of weasel terms is consistent with reduced conflicts of interests.

1.5.3.3 Bias and Confounding News

To rule out stories related to the timing between of analyst reports changing differentially between investment banking analysts and other analysts, I run my tests with firm-day fixed effects.

[Table 1.12, 1.13]

Tables 1.12 and 1.13 show results on sentiment and weaseling using the firm-day fixed effect. In this table, the effect is identified by comparing the treatment (banking analysts) and control (other analysts) groups before and after the period of regulation change on the same day for the same firm. This rules out a broad class of altenative stories that are either related to changing timing of reports or related to time-varying firm variables. Results remain close to results in prior tables, both in terms of coefficient sizes and in terms of significance.

1.6 Robustness

1.6.1 Disclosures

The regulations studied in this paper required additional disclosures. It's therefore important to check if the increase in similarity is driven by this. I have taken this into account in the main analysis by controlling for what I call Self-Similarity, Other Firms (see section 5.2.4). In this section, I redo the analysis using a different method to deal with disclosure and form language, namely by building a maximum entropy classifier for disclosure language by randomly selecting 2,000 sentences from the entire corpus and hand classifying them as either disclosure or not. The

Table 1.11: Weaseling

Note: **Weaseling** is measured as the fraction of sentences in the document classified as weasels. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. The weasel measure is centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.10. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:				
	Weasel Sentences				
	(1)	(2)	(3)	(4)	(5)
IB × Post	−0.030	−0.296***	−0.299***	−0.320***	−0.266***
	(0.081)	(0.084)	(0.084)	(0.080)	(0.091)
Investment Bank	0.067				
	(0.052)				
Post	0.073	0.425***			
	(0.073)	(0.056)			
Other Controls	X	X	X	X	X
Firm FE				X	
Year FE			X	X	X
Analyst FE		X	X	X	
Analyst FE × Firm FE					X
Observations	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.064	0.412	0.413	0.425	0.458
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1.12: Change in Sentiment - Holding Information Constant

Note: **Sentiment** is measured as the number of positive words minus the number of negative words scaled by the length of the document. **Positivity** and **negativity** are defined analogously using either positive or negative words only. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Sentiment measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.9. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Sentiment		Positivity		Negativity	
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	-0.326^{***}	-0.326^{***}	0.003	0.016	0.492^{***}	0.492^{***}
	(0.063)	(0.080)	(0.082)	(0.082)	(0.086)	(0.086)
Other Controls	X	X	X	X	X	X
Firm-Day FE	X	X	X	X	X	X
Analyst FE	X		X		X	
Analyst FE × Firm FE		X		X		X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.573	0.573	0.439	0.468	0.548	0.548
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 1.13: Weaseling - Holding Information Constant

Note: **Weaseling** is measured as the fraction of sentences in the document classified as weasels. **Post** is a dummy equal to one for 2004-2005 and zero for 2000-2001. **IB** is a dummy equal to one if the analyst's employer is an investment bank. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. The weasel measure is centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.10. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:	
	Weaseling	
	(1)	(2)
IB × Post	-0.639^{***}	-0.639^{***}
	(0.099)	(0.099)
Other Controls	X	X
Firm-Day FE	X	X
Analyst FE	X	
Analyst FE × Firm FE		X
Observations	114,105	114,105
Adjusted R^2	0.535	0.535
Note:	*p<0.1; **p<0.05; ***p<0.01	

classifier uses the document term matrix to predict whether a sentence is disclosure or not. I fit the model to 1,800 sentences, 250 of which are disclosure sentences. The held-out sample, i.e. the remaining 200 sentences are evenly split between disclosure and not and used to test the accuracy of the prediction model. The out of sample prediction accuracy is 55% true positive and 3% false positive.

Armed with the disclosure classifier, I do two different robustness tests. First, I use the proportion of each report that is disclosure as a control variable in the regression (not reported but virtually identical to table 1.6). Second, I cut out all sentences that are classified as disclosure sentences and rerun the main analysis using this purged corpus. Table A.1 show results from this robustness exercise. Results are very similar to the results of the disclosure exercise in section 5.2.4.

1.6.2 Weasel Classification

My main measure of weaseling is explained in section 3.3 and builds upon a maximum entropy classifier using training data from Wikipedia. As an alternative, I consider a more simplistic measure of weaseling - one based on a word count approach. In this alternative approach, I collect all weasel tagged sentences from a dump of Wikipedia, and rank words based on their occurrence. I then use the most common words from this exercise as a word-list that I use to count number of weasel words in analyst reports. The measure is then the fraction of weasel words in the document. As usual, the measure is standardized to have mean 0 and variance 1.

[Table A.11]

Results in table A.11 are similar to the main weasel result presented in table A.11, especially when I control for analyst and firm fixed effects. With such controls, results are qualitatively and quantitatively similar.

1.6.3 Regulation Fair Disclosure

Regulation fair disclosure was ratified by the SEC in October, 2000, and required companies to disseminate information to all investors simultaneously. This effected the value of analyst output

since analysts with favorable access to management lost potentially useful information. Since 2000 is part of the pre-period in my study, I re-run results excluding all data prior to regulation fair disclosure. In the tables below, I remove data up to and including October, 2000, and run a difference and difference study using November, 2000, to the end of 2001 as the pre-period.

[Table A.2]

Table A.2 shows the results after excluding the pre regulation fair disclosure period. Results are qualitatively and quantitatively similar to results presented earlier.

1.6.4 Alternative Measures of Similarity

I consider an alternative measure of similarity in this section to make sure the results are not driven by the choice in measure. Here, I use the Jensen-Shannon Divergence (JSD), which, intuitively, is the mutual information between two documents when the documents are viewed as probability distributions. In this case, a larger value of the measure means the documents are further away from each other, or less similar. So the effects of the shock on the JSD will have the opposite sign from the effects on similarity.

[Table A.7]

Table A.7 shows the results. The JSD measure is scaled by it's standard deviation to be comparable with other tables. What I see here is a reduction in the self-JSD of about one standard deviation as a result of the regulatory shock. This is consistent with what I found using the cosine similarity measure.

1.7 Conclusion

Regulatory changes surrounding the Global Research Settlement brought about important changes to the market for security analysts. Guan, Lu and Wong (2013) has found that all-star analysts left their jobs at investment banks as a result of the regulation changes, a channel through which information production is reduced. My paper provides complimentary evidence. I find that even for analysts who remain at the same investment bank before and after the change, weakened

incentives and reduced access to investment bankers led to less informative reports as measured by greater similarity between them.

In addition, my textual measures provide a novel take on the nature of bias in analyst reports by examining the extent of markers of bias in the analyst language by using so-called "weasel words". Consistent with the motivation behind the GRS to reduce bias among analysts, the content of analyst reports contains fewer markers of bias. Taken together, my findings highlight an important trade-off between informativeness and bias.

Chapter 2

A Text-Based Analysis of Corporate Innovation

Co-authored with Sanjai Bhagat and J. Anthony Cookson

2.1 Introduction

Innovation has long been thought to play a central role both for economic growth and short-term fluctuations (Schumpeter, 1939; Kuznets and Murphy, 1966; Nordhaus, 1969). Owing to its fundamental importance, innovation has attracted significant academic attention (e.g., Hall, 1990; Bhagat and Welch, 1995; Brown et al., 2009; Cohen et al., 2013). Nevertheless, our empirical understanding of innovation is incomplete because existing innovation proxies – typically, R&D intensity or outcomes related to patenting – do not fully capture the nature and scope of innovative output.

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

[1] As a measure of innovation, patents have a number of additional well-known weaknesses. For example, not all innovations are put under patent protection or can be put under patent protection (Moser, 2012; Hall et al., 2014), and some patents are filed for defensive reasons (e.g., see work on 'patent trolls' by Tucker, 2014, and Cohen, Gurun

To help bridge this gap, we propose a new measure of corporate innovation derived from textual descriptions of firm activities by financial analysts. Our measure encapsulates a broad notion of innovative processes, products, and systems, which well describes innovation in mature firms – i.e., firms in the S&P500. Innovation in mature firms has been sparsely studied despite these firms comprising the most valuable corporations in the economy. One reason for this lack of academic attention is because mature firm innovation involves much more than developing and introducing new products. By offering a measure of innovation beyond products, our analysis provides a useful first step toward understanding mature firm innovation.

We construct the text-based innovation measure using topic modeling tools that have been recently introduced to the finance literature (Israelsen, 2014; Goldsmith-Pinkham et al., 2016; Hoberg and Lewis, 2017; Lowry 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 count-based word-list techniques. The fitted LDA also provides an intensity with which each analyst report discusses each topic, which is the centerpiece of our innovation measure.

Our main measure is derived from a fitted LDA model that allows for 15 distinct topics to a corpus of 665,714 analyst reports of 703 firms that were in the S&P500 during 1990-2012. From this fitted topic model, we compute the Kullback-Liebler divergence of each topic from the language used in a mainstream textbook on innovation, and we select the topic that has the lowest divergence.

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

Figure 2.1: 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.

[Low Patent Industries]

[High Patent Industries]

Firm	Date	Excerpt
GOOGLE	2009-07-31	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	2002-02-04	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.

Beyond this selection criterion, the selected topic stands out as a reliable innovation proxy, both qualitatively and quantitatively. Qualitatively, the words in the innovation topic are also words that analysts should use to describe innovations (e.g., service, system, technology, product, solution). Quantitatively, the topic correlates strongly with patenting and R&D intensity among patenting firms. Beyond basic correlations, all of our findings using the text-based measure are robust to controlling for patenting, implying that the correlation with patenting does not drive our findings.

For studying innovation in mature firms, an important advantage of our text-based innovation measure is that it can be computed for firms that do not patent and do not use R&D, which provides a reliable basis for comparing mature firms' innovation to one another. Even within our sample of 703 firms from the S&P500, 329 firms have zero R&D and 219 firms have zero patents for the entire sample period (1990-2010). To illustrate that the measure is useful for non-patenting firms, we present tangible examples of content from analyst reports for non-patenting firms that score high on our measure. One such example, which highlights the value of our approach is Walmart. Walmart did not use patent protection in the early 1990s, but it has always been innovative with respect to how it organizes its cross-geography logistics (e.g., placement of warehouses and shipping logistics between locations). Taking an excerpt from a May 1993 analyst report (more detail in Figure 2.1), Walmart was described as "at the leading edge of retail store technology," very broadly in terms of tracking inventory, procurement and theft prevention. Our topic analysis captures this language, and as a result, we correctly classify Walmart as one of the most innovative companies in 1993, even though this was a time period when Walmart did not use patents at all.

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

Turning to corporate valuation implications of text-based innovation, higher innovation forecasts an increase in future operating performance, and an increase in measured growth opportunities embedded in Tobin's Q, results that are robust to firm fixed effects. Consistent with the nature of innovations that generate persistent improved performance and opportunities for growth, we find that both operating performance and Tobin's Q are significantly greater for up to four years after an increase in text-based innovation. Importantly, the valuation implications of innovation are similar for both patenting and non-patenting firms, providing further evidence that our measure extends in a useful manner beyond the set of firms that use patenting and R&D.

Even among the set of patenting firms, the text-based innovation measure provides useful additional information on innovation. We find that our text-based measure strongly correlates with the Kogan et al. (2017) patent valuation measure within the set of firms that patent. In this way, our text-based approach distinguishes true innovation captured by valuable patents from patenting outcomes that are not as valuable.

Because the text-based innovation measure applies across many contexts, the measure captures whether a company has an innovative system or platform. Indeed, the text of the topic does not reflect language surrounding specific products, but the systematic use of technology to enhance

revenue and decrease costs. This idea of innovative systems has been conceptually identified as important (see Egan, 2013), but traditional measurements have not captured this idea quantitatively. Empirically, we find that innovative firms are more acquisitive, especially of smaller firms, which is consistent with the incentives of a firm with an innovative system to acquire smaller firms as components to their revenue-generating system.

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

Our work contributes to an emerging line of research that draws a distinction between patenting measures and innovation (e.g., Kogan et al., 2017; 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 in mature firms studied in our paper is distinct from trade secrets, both kinds of innovation extend beyond the set of patenting firms. As both innovation in mature firms and non-patenting firms' innovative activities are understudied, we expect significant interest in approaches like ours to extend the analysis of innovation to new subsamples and types of innovation.

Beyond offering a useful measure of innovation, our work is part of a growing literature within finance and accounting that makes use of text descriptions to study important aspects of corporate

behavior. Recent text-based analyses in corporate finance have examined linkages between firms and industries, the value of corporate culture, product market fluidity, financial constraints, and the information content in IPO prospectuses (e.g., Hanley and Hoberg, 2010; Popadak, 2013b; Hoberg et al., 2014; Hoberg and Maksimovic, 2015; Agarwal et al., 2016). At the same time, the asset pricing literature has employed kindred text-analysis procedures to measure sentiment and other asset pricing risks and anomalies (Edmans et al., 2007; Garcia, 2013b; Dougal et al., 2012; Israelsen, 2014; Cohen et al., 2016). Within the broader literature on text analysis in finance, our work is most closely related to the growing set of papers that use Latent Dirchlet Allocation (Jegadeesh and Wu, 2017; Ganglmair and Wardlaw, 2017; Goldsmith-Pinkham et al., 2016; Hoberg and Lewis, 2017). Although there has been significant interest among finance scholars in text analysis in general and LDA in particular, our analysis is the first to systematically use a text analysis to construct a measure of innovation.[2]

In another vein, our use of the text of analyst reports relates to the study of the behavior and impact of analysts more broadly. Much of this work has focused on quantitative aspects of analyst reports (Loh and Mian, 2006), what information analysts actually produce (Swem, 2014), or the influence of analyst coverage on the real decisions of investors or firms (e.g., see analyst coverage tests in Cohen and Frazzini, 2008). Some of this work has shown how analyst coverage influences the innovativeness of firms (He and Tian, 2013), but none of this work has examined the information from the text of analyst reports as it relates to innovation. In this sense, our contribution is related to Asquith, Mikhail and Au (2005a), Huang, Zang and Zheng (2014), and Huang et al. (2015) who provide evidence, in a different context, that investors pay attention to the textual elements of analyst reports, rather than just the quantitative analyst forecasts. Our analysis suggests a new reason for investors to pay attention to the text of analyst reports: valuable information on firm innovation.

[2] Even related work on innovation using text analysis has not constructed a similar measure of innovation. Specifically, Fresard et al. (2017) studies how innovation and vertical integration relate to one another while making use of text analysis, but the text-analysis component of their work is confined to vertical relatedness rather than innovation. Their innovative outcomes are the more standard R&D intensity and patenting outcomes from the literature.

The remainder of the paper is organized as follows. Section 2.2 describes our data sources and sampling scope. Section 2.3 details how we construct our measure, and presents evidence on its time-series and cross-sectional properties. Section 2.4 presents the main results linking our text-based measure of innovation to firm performance and value. Section 2.5 presents an application of our measure to M&A activity. The final section concludes with a summary of future research directions.

2.2 Data

We begin with a sample of mature firms that were a member of the S&P500 at some point between 1990 and 2012. This initial sample contains 797 firms. To obtain the set of analyst reports these firms, we download analyst reports from Investext via Thomson One for the years 1990 to 2012, which provides an initial sample of 807,309 analyst reports for 750 unique S&P500 firms searchable in Thomson One.

After downloading the reports, we remove common stopwords (e.g., words commonly used in text without contextual meaning like "the", "that", "an") from the reports using a standard stopword list.[3] Prior to any textual analysis, we use a standard algorithm to stem the words contained in the analyst reports (i.e., group words into the same root as in "technolog" captures "technology" and "technological," among other related terms). To focus on a homogenous set of analyst reports, we drop reports with under 100 words remaining after the cleaning or over 5,847 words (the 98th percentile). After processing the text and matching with Compustat identifies, we obtain a final sample of 665,714 reports on which we base our textual analysis.

We combine the pure textual data from Thomson One with sentiment word lists (Loughran and McDonald 2011 and Bodnaruk, Loughran and McDonald 2015) as an integral part of our textual 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.

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

After constructing the main text sample, we calculate the measure we call the 'innovation' measure (as described in the Section 2.3.2) and aggregate it to the firm-year level before matching with accounting data from Compustat and patent data up to 2010 from Noah Stoffman's website (Kogan et al., 2017). The final sample has 6,200 observations from 703 unique firms for the period 1990-2010.

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

2.3 Text-Based Measure of Innovation

In this section, we describe how we construct the text-based measure of innovation using the Latent Dirchlet Allocation (LDA) method of Blei 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&P500 firms. The measure has desirable time series and cross-sectional properties for a measure of innovation.

As we described in the introduction, LDA has a number of advantages over naive word-list techniques (e.g., Loughran and McDonald, 2011). For our purposes, the most important advantage is that LDA accurately reflects context of the word usage, whereas a naive word-list textual analysis does not. As we show in the Appendix, Tables B.15 and B.16, the word-list measure exhibits slightly weaker valuation implications, and is not as robustly related to valuable patents as the more accurate LDA-based measure. This is to be expected because the LDA methodology is better equipped at getting the context of innovative language correct.

2.3.1 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. In addition to containing innovation-relevant information, the language of analyst reports has relatively common textual structure (i.e., similar word usage, jargon, specificity, and topics covered) relative to media reports about the firm, or even disclosures by the firm itself. This feature of analyst reports is convenient from the standpoint of our topic modeling approach described in the next subsection, which assumes that each report is built from a common set of latent topics.[4]

A potential concern regarding building the measure from analyst reports is that analysts cannot describe innovative activities that they cannot observe. Thus, our measure of innovation can only reflect publicly-observable information about the firm that was either disclosed by the firm or inferred by the analyst. As a result, our innovation measure will tend to reflect realized innovations (and their trajectory) in a similar manner to how patents reflect the realization of technical innovation. At the same time, the text-based innovation measure will also capture innovation activities beyond patents, which includes some trade secrets (e.g., although Coke's secret formula is a trade secret, the value of this secret is well-known to analysts). Beyond trade secrets, there are myriad ways for a firm to be innovative without filing for a patent or investing in R&D (e.g., see the Walmart example from the introduction and Figure 2.1). We expect that our analysis of the text of analyst reports reveals these innovative activities. Indeed, our measure identifies high-innovation firms and industries that do not patent or use R&D, which suggests that existing proxies overlook

[4] On the other hand, analysts are likely to be less informed about the firm's innovation activities than the firm's managers. This increases the noise in our innovation text measure, and biases the coefficients of our innovation text measure (in the regressions) towards zero. To the extent that we find our innovation text measure as statistically significant, it would be even more so if we could reduce this source of noise.

an important subset of innovative activites (see the discussion in Sections 2.3.3 and 2.3.4).[5]

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

There is a growing literature that shows that the qualitative nature of analyst text contains useful information. In one of the earliest contributions in this vein, Asquith, Mikhail and Au (2005a) hand classify a limited sample of analyst reports into various categories and show that some categories have investment value. More recently, authors have worked on parsing the text of analyst reports in a more systematic fashion. Using a sample of initiation reports, Twedt and Rees (2012) show that, controlling for recommendation changes and other factors, the tone of reports has an associated stock market reaction. Huang, Zang and Zheng (2014) is the first large sample study of text in analyst reports. Using a sample that overlaps our sample, they find results consistent with Twedt and Rees. Specifically, they find a stock market reaction associated with the tone of reports of between 1.5% and 3.5% (2-day CAR) for reports in the top quintile relative to those in the bottom quintile. They also show that the tone of more qualitative topics (those with few uses of "$" or "%") are more important, a strong indication that the qualitative and descriptive portions of the analyst text are a valuable source of new information.

Based on existing studies of analyst text, it is clear that the qualitative aspects of analyst

[5] Relative to the patenting measures of innovation, one notable limitation of the text-based measure is that it is observed at the firm-year level, which prevents within-firm, cross-sector analyses of innovation. From this standpoint, the text-based measure of innovation would not be useful to describe product market positioning, nor evaluate the determinants of innovation-level valuation. These potential concerns apply to other widely-used text-based measures of product competition (e.g., Hoberg et al., 2014; Hoberg and Phillips, 2016), and they are consistent with our interpretation of the text-based innovation measure as a measure of systems innovation.

text contain value-relevant information about the firm. This fact suggests that the analyst reports will provide insight into innovation, which is a critical resource that helps firms generate value. With this understanding of the qualitative content of analyst reports, we now turn to describing how we measure innovation using the analyst text.

2.3.2 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).

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

To construct our innovation measure, we estimate a LDA model with $K = 15$ topics using the 665,714 analyst reports as the underlying corpus of documents.[6] Fitting this LDA model gives the

[6] We experimented with other numbers of topics. Fitted LDA models with fewer topics tended to work similarly well (the model with $K = 10$ delivers all of the quantitative insights we report in the main text), whereas models fit with a greater pre-specified number of topics exhibit redundancy (i.e., multiple topics about the same essential idea). Although the number of topics is the only degree of freedom we have in fitting a LDA model, the extensive literature on LDA does not offer standardized guidance on how to select the appropriate number of topics because the appropriate number of topics depends on the application. Some applications of LDA have optimized an objective function to obtain an optimal number of topics in their context. For example, Goldsmith-Pinkham et al. (2016) maximize saliency of topics from one another, and other authors have estimated Hierarchical Dirchlet Process models (HDP-LDA), which obtains a likelihood-maximizing number of topics (Teh et al., 2006). Our objective is to select the number of topics to capture a general notion of innovation to apply across different contexts. Automated routines

15 topics – each a frequency distribution over words – that best fit the context of the analyst reports. To identify the topic that most accurately captures innovation, we compute each topic's statistical distance from the word frequencies used in a popular textbook on innovation, and select the topic with the word distribution that has the smallest statistical distance from the innovation textbook's word distribution.[7] Specifically, we compute the Kullback-Liebler (KL) divergence of each topic's word distribution from the source text on innovation, similar to Lowry et al. (2016). In our context, the KL divergence is useful because it is a measure of the expected information loss from using the topic distribution to proxy for the distribution of words in the textbook. Thus, selecting the topic with the lowest KL divergence is equivalent to picking the most informative topic about the source text. Figure 2.2 presents a summary of these KL divergence calculations, together with bootstrapped 95% confidence bands for the innovation topic and the average of the other topics. Using this method, the innovation topic is significantly more informative about textbook innovation than the typical topic written by analysts. To argue that this lower KL divergence is because of innovation rather than general finance language, the second panel of Figure 2.2 presents a placebo exercise in which the source text is a standard corporate finance textbook (Welch's "Corporate Finance: An Introduction"). Unlike the comparison to the innovation textbook, the innovation topic exhibits a similar KL divergence to other topics.

The measure also appears to intuitively measure the factors that describe innovative companies. For example, Figure 2.3 presents the topic distribution across words in the form of a word cloud (Appendix Figure B.3 provides word frequencies for the 10 most common words in the topic). When writing about this topic, analysts most frequently use words such as **revenue, growth, services, network, market**, and **technology**. Beyond the contextual word usage, we show that firms that have high values of this measure have the hallmarks of innovative firms.

Before using the loadings as a measure of innovation, it is helpful to refine the measure

that seek to maximize a likelihood function will tend to overfit by selecting a larger number of topics that adapt to different contexts. Thus, automated routines will tend to lead to topics that are too granular to capture a broad notion of innovation.

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

Figure 2.2: Selecting the Innovation Topic – Kullback-Liebler Divergence from an Innovation Textbook

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

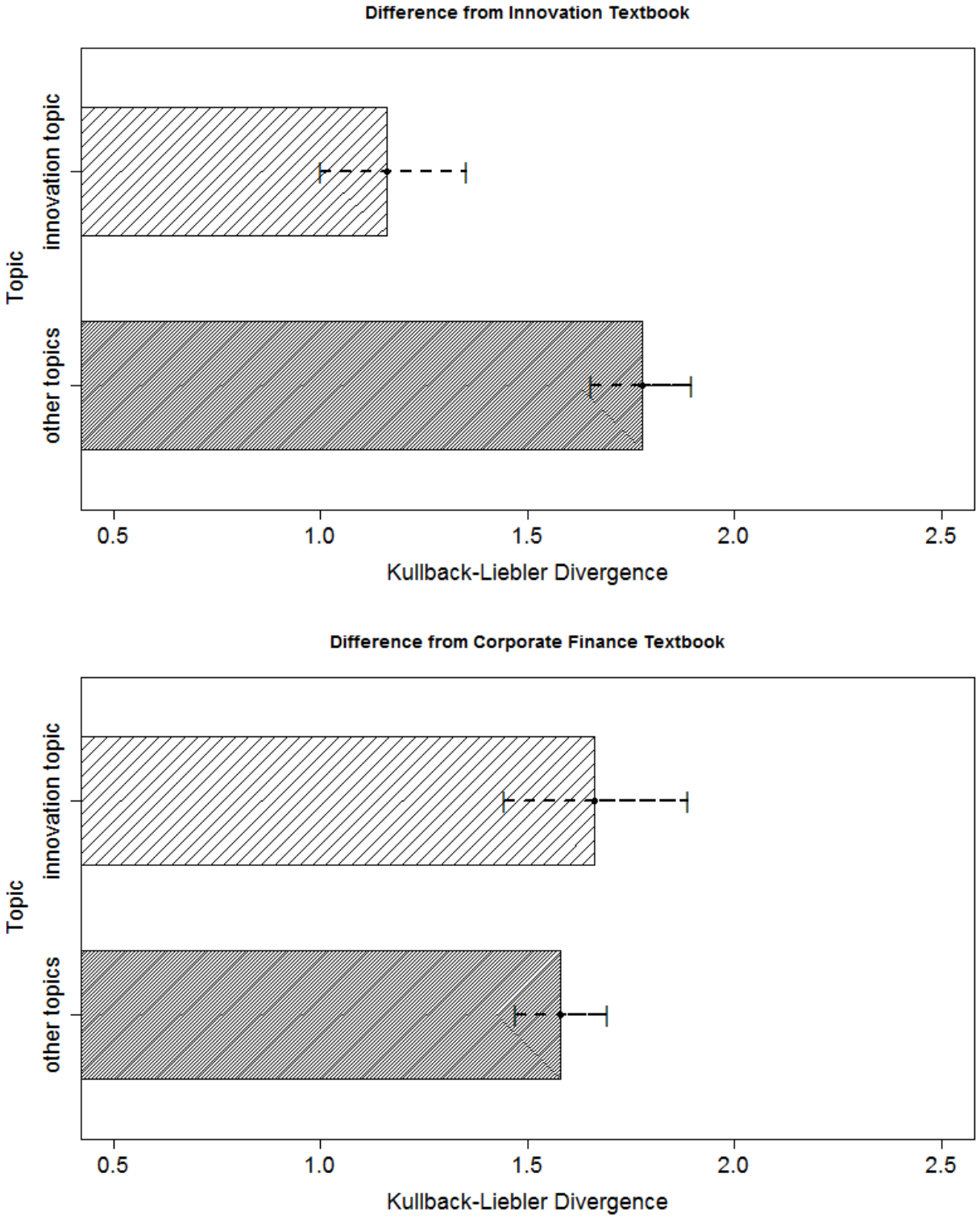


Figure 2.3: Text-Based Innovation Measure: Word Cloud

Note: This word cloud describes the frequency distribution of words used in the 'innovation' topic. The topic itself is from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P500 firms. We set the number of topics in the fitted LDA model to be 15, then select the topic (out of these 15) for which the distribution of words in the topic is closest to an innovation textbook (Tidd et al., 2005).

to account for analysts who write about the innovative activities of the firm in a negative or neutral tone. Specifically, if an analyst is talking with neutral or ambivalent sentiment about the company, it is less likely that the strong loading on the 'innovation' topic reflects stronger innovation by the company. We address this source of noise by focusing on the analyst reports that have relatively strong positive sentiment (i.e., those in the top quartile of sentiment, measured by $\frac{\#positive\_words - \#negative\_words}{\#total\_words}$ from the word list in Loughran and McDonald, 2011). For analyst reports with sentiment below the 75th percentile, we set the topic loading at the analyst report level to be zero in the sentiment-adjusted topic measure. We aggregate this sentiment-adjusted topic measure to the firm-year level to construct our text-based measure of innovation, $innov\_text_{it}$. It is the content of the topic, rather than the screen on sentiment, that drives the properties of our measure. Indeed, the innovation topic loadings and the sentiment have a low correlation equal to 0.08. Thus, reports that load on the innovation topic are unlikely to merely reflect positivity about earnings or revenue. Further, as robustness exercises, we have constructed the measure without the sentiment screen, and we have also controlled explicitly for average sentiment. In each case, the main results are similar.

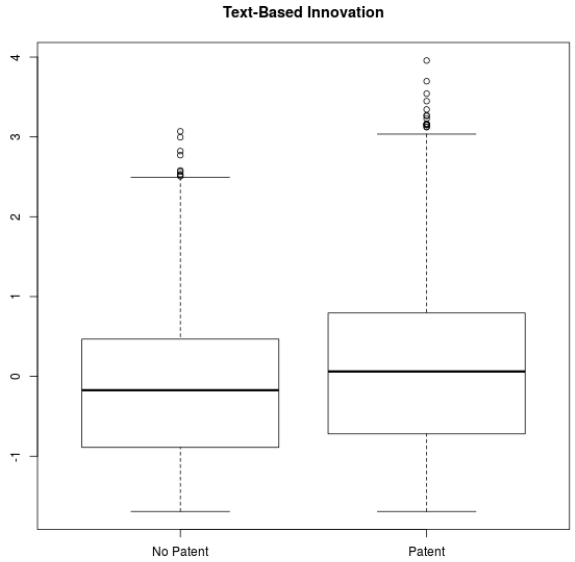
2.3.3 Comparison to Patenting Outcomes

An important advantage of the text-based innovation measure is that it captures innovative activities of firms that do not patent. In our sample of 703 mature firms in the S&P500, 219 firms have zero patenting throughout the full sample period (1990-2010). Although these firms do not patent, many are highly innovative. Panel (a) of Figure 2.4 presents side-by-side boxplots of our text-based innovation measure for patenting firms versus non-patenting firms. Although patenting firms have higher text-based innovation on average, the distribution of text-based innovation exhibits substantial overlap between non-patenting firms and patenting firms. Specific examples of highly innovative non-patenting firms are also consistent with this view.[8]

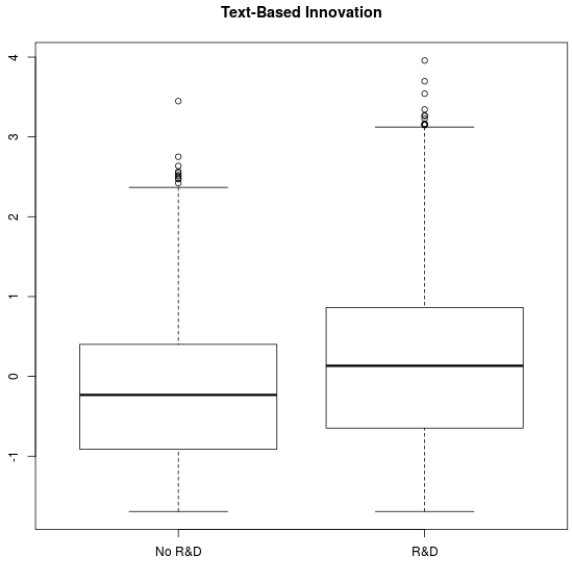
[8] The top three innovation firm-years among non-patenting firms in our sample highlight the ability of our measure to identify overlooked high innovation firms. First, in 1996, Shared Medical Systems Corporation produced information processing systems for the healthcare industry at a time when Internet technology was emerging, but was a non-patenting firm. Second, in 2000, BroadVision was non-patenting firm that was a software vendor for

Figure 2.4: Distribution of Text-Based Innovation

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



[With and Without Patents]



[With and Without R&D]

In columns 2 through 4 of Table 2.1, we present summary comparisons of text-based innovation for mature firms with and without patents. On average, patenting firms have higher text-based innovation than non-patenting firms by 0.27 standard deviations (0.20 sd at the firm level), a difference that is statistically significant at the one percent level, indicating a significant positive correlation between our text-based measure and whether a firm engages in patenting. Within the set of patenting firms, our text-based measure and patenting outcomes are also positively correlated. To this end, Figure 2.5 presents a graphical depiction of how the text-based measure fits patenting outcomes by plotting the log of patenting measures against decile bins of the text-based innovation measure. Regardless of the measure of patenting employed (counts, citations, or citations per patent), the text measure correlates strongly with patenting activity within the set of patenting firms.[9]

2.3.4 Comparison to Technology Development via R&D

The text-based innovation measure also captures innovative activities of firms that do not perform R&D. In our sample of 703 mature firms in the S&P500, 329 firms have zero R&D expenditures throughout the full sample period (1990-2010). Similar to the non-patenting firms, many non-R&D firms are highly innovative. Panel (b) of Figure 2.4 presents side-by-side boxplots of our text-based innovation measure for firms with and without R&D, which shows there is substantial overlap in the distribution of text-based innovation for firms with and without R&D.

In columns 5 through 7 of Table 2.1,[10] we present summary comparisons of text-based

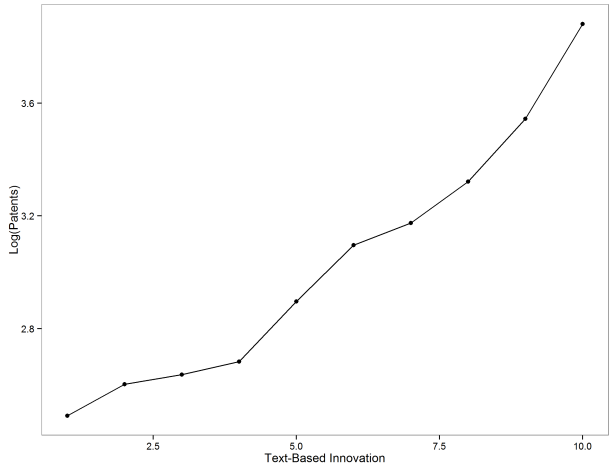
web applications that enhanced internal management systems of firms (HR, sales processing, online shopping, etc.). Finally, in 1994, Alltel Wireless was a wireless service provider that developed a large network of subscribers across much of the United States by adopting network technology manufactured by Lucent, Motorola, Nortel, Cisco, and Juniper Networks.

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

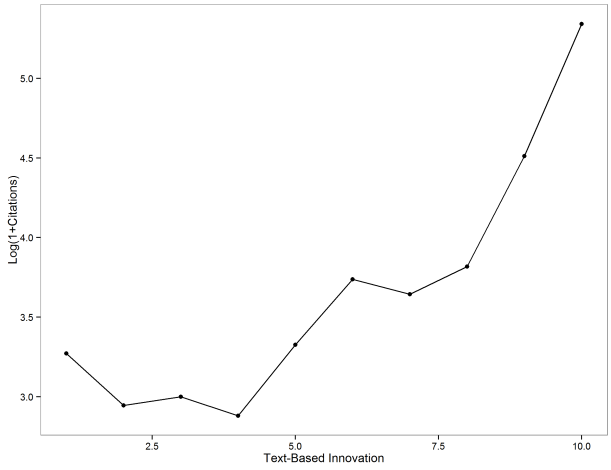
[10] The table presents comparisons of other characteristics as well, which are consistent with intuition about R&D and patenting. For example, there is a strong correlation between patenting and R&D expenditures. Both patenting

Figure 2.5: Relating Patent Counts and Patent Citations to the Text-Based Innovation Measure (Decile Bins)

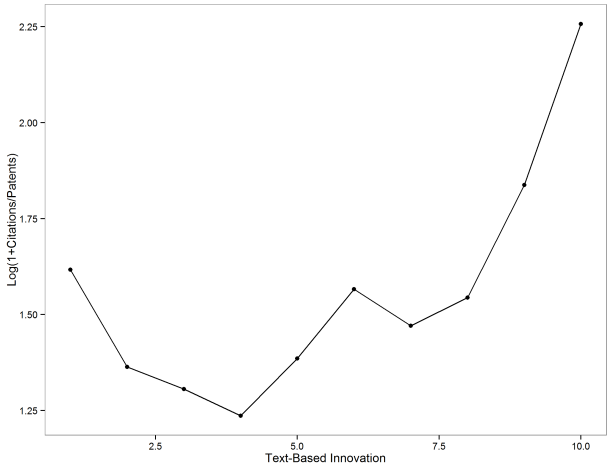
Note: This figure plots the relation between the text-based innovation measure and commonly-used patenting measures. In each panel, the text-based innovation measure is grouped into 10 deciles. Panel (a) presents the relation between text-based innovation and logged patent counts ($\log(1 + Patents)$),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



[Patent Counts]



[Patent Citations]



[Citations per Patent]

Table 2.1: Summary Statistics

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

[Summary Statistics]

Variable	All	Patents¿0	Patents=0	(5)-(6)	R&D¿0	R&D=0	(2)-(3)
Innovation Measures							
Text-Based Innovation	0.00	0.12	-0.16	0.27***	0.18	-0.21	0.39***
Patents	62.9	109	0.00	109***	118	1.71	116***
R&D/Assets	0.03	0.04	0.00	0.04***	0.05	0.00	0.05***
Performance Measures							
ROA	0.15	0.16	0.15	0.00	0.16	0.15	0.01
Log(Q)	0.55	0.61	0.47	0.15***	0.66	0.43	0.24***
Salesgrowth	0.09	0.08	0.10	-0.02	0.08	0.09	-0.01
Characteristics							
Log(Assets)	8.78	8.86	8.67	0.19**	8.75	8.81	-0.05
Asset Tangibility	0.36	0.30	0.43	-0.13***	0.27	0.46	-0.19***
Leverage	0.58	0.57	0.61	-0.04**	0.56	0.61	-0.05***
Log(Age)	3.18	3.20	3.15	0.06**	3.16	3.20	-0.04
Observations	6200	3586	2614		3268	2932	

[Firm-Level Summary Statistics]

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

innovation for firms with and without R&D. Firms with positive R&D expenditure have higher text-based innovation by 0.39 standard deviations (0.43 at the firm level), a difference that is statistically significant at the one percent level, indicating a significant positive correlation between our text-based measure and R&D expenditure. The time series and cross-industry correlations are also informative, both as a point of validation to the extent that the text-based measure is positively correlated with R&D intensity along these dimensions, but also to highlight specific industries and time periods in which text-based innovation is high and R&D intensity is low. Our interpretation of this section's results is that the text-based measure of innovation measures the adoption of technology, even in industries that have low R&D intensity.

In the time series (1990-2010), the text-based innovation captures the R&D boom and bust of the late 1990s and early 2000s, which is studied in Brown et al. (2009). Figure 2.6 presents the plot of the text-based measure of innovation over time (a value-weighted average across firms). For comparison, the time series of average R&D expenditures by year is also presented on the same plot. There is a strong relationship between these two series, which have a correlation of 0.51. This correlation suggests that our measure of innovation captures the macro-level trends in innovative activity well. In the cross-section, the text-based innovation measure also matches cross-industry differences in R&D expenditures well. Figure (2.7) 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 shows a significant relationship between R&D and the innovation measure at the industry level, which is also indicated by the correlation of 0.47.

Examining the fit industry-by-industry yields additional qualitative insight into what the text-based measure of innovation adds to existing proxies. Notably, industries with high text-based innovation and high R&D intensity tend to be industries in which it is more natural to develop technologies in-house (e.g, Electronic Equipment and Business Services). In contrast, the ill-fitting industries with high text-based innovation are industries in which the most innovative companies

and R&D firms have lower asset tangibility and lower leverage. In addition, R&D firms tend to be younger than non-R&D firms, and firms with patents tend to be older.

Figure 2.6: Time Series of Text-Based Innovation Measure and R&D (1990-2010)

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

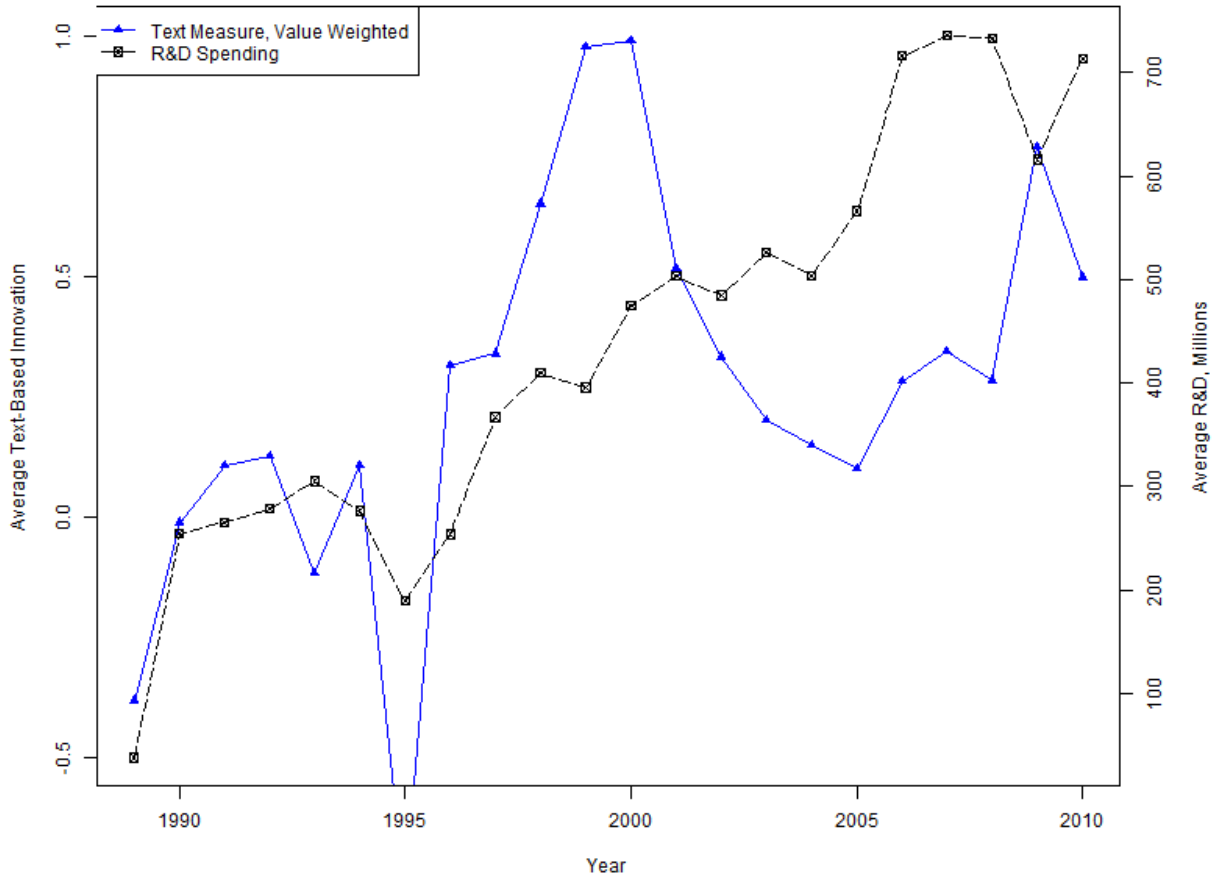
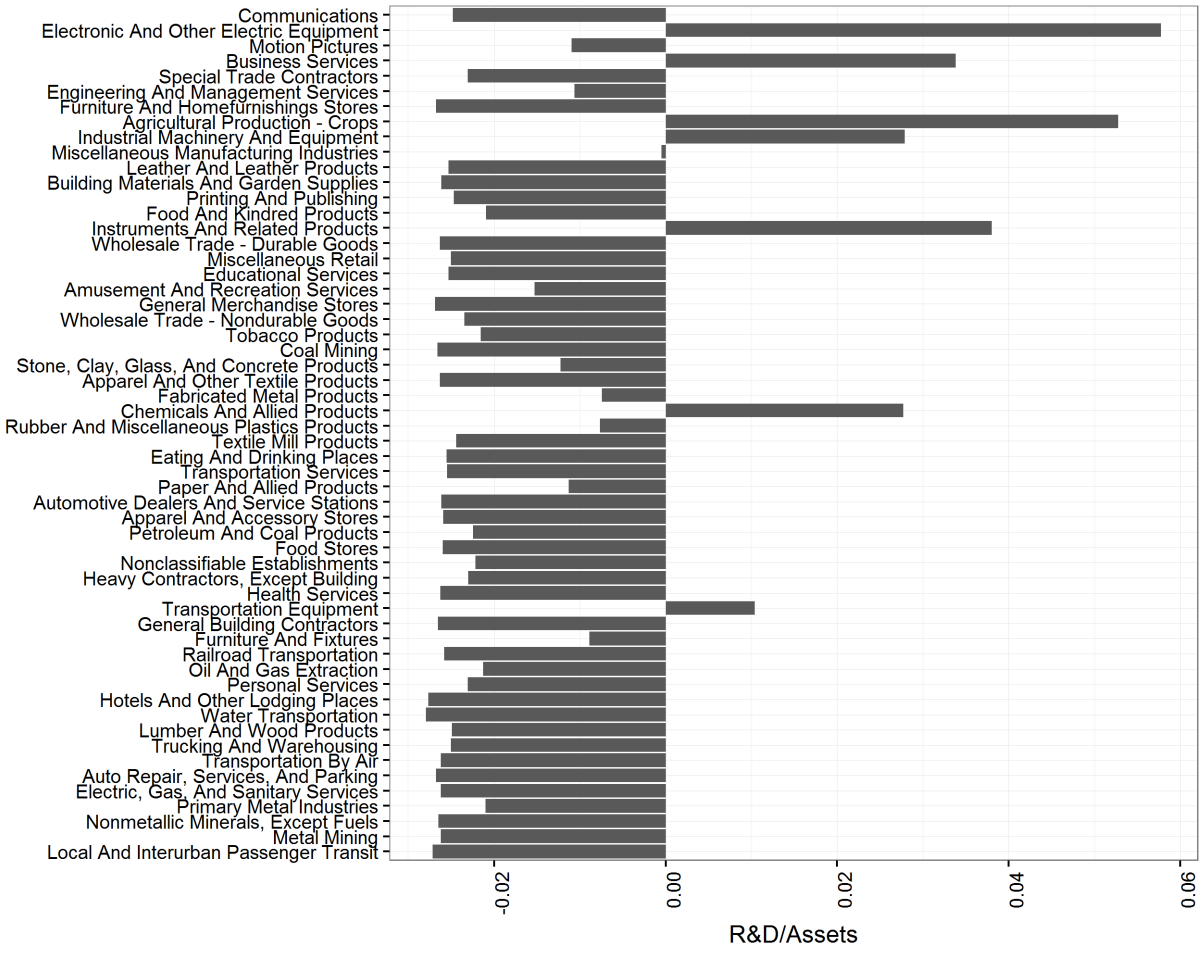


Figure 2.7: Cross-Industry Plot of R&D (1990-2004), Relationship to Text-Based Measure

Note: This figure provides a plot of R&D expenditures (demeaned by the average R&D/Assets) by industry covered in the sample of S&P500 firms. To show the relation between text-based innovation and R&D expenditures across industries, the industries in the plot are ordered from the highest value of text-based innovation to the lowest value. The correlation between R&D expenditures and the text-based measure across industries is 0.40, and statistically different from 0.



are skilled at technology adoption (e.g., Communications and Motion Pictures). These patterns suggest that the text-based measure is useful to identify firms that utilize technology to support a revenue generating system, and that the measure is most useful beyond standard measures when it reflects the firm's ability to adopt technology productively.

We have also estimated the relation between R&D intensity and the text-based measure more systematically in a panel data context (results presented in Appendix Table B.7). Even within narrowly-defined industries (4-digit SIC), there is a strong statistically significant link between R&D intensity and text-based innovation. The link between text-based innovation and R&D intensity persists after controlling for other firm-specific factors, and text-based innovation reliably forecasts R&D intensity one year ahead, even holding constant this year's R&D intensity. These within-industry findings are consistent with the text-based innovation measure capturing technology adoption decisions that are broader than the decision to develop technology via R&D expenditure.

2.4 Empirical Results

In this section, we use our text-based measure of innovation to evaluate the impact of innovation on various measures of performance (e.g., return on assets, Tobin's Q), and examine what the analyst text about innovative firms reveals about the value of innovation. We further examine the relation between our text-based measure and future values of patenting, and perform several robustness checks on our measure.

2.4.1 Innovation and Performance

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

2.4.1.1 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:

$$ROA_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \quad (2.1)$$

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 2.2 present the results of estimating equation (2.1). With industry and year dummies, there is a strong correlation between our text-based measure and the return on assets. A one standard deviation increase in the text measure is associated with a 0.9 percentage point increase in the return on assets in the following year. We find that this estimated effect is robust to including firm fixed effects, and thus, the within-firm variation in our text-based measure of innovation appears to be valuable in terms of generating abnormal operating performance. Moreover, we see that the text-based measure is more robustly associated with increases in operating performance than patent counts and R&D intensity. Patent counts are not significantly and positively correlated with operating earnings in any specification. Although R&D intensity is positively correlated with future operating performance and the magnitudes are similar to our measure, the statistical significance is lower and the result is not robust across specifications. Moreover, as our estimates using our text-based measure control for alternative measures of innovation, the findings imply that the innovations captured by our measure are valued beyond what existing measures of

innovation would predict.

A notable advantage of our text-based measure is that it can be computed for firms without patents, and thus, can help evaluate innovation for a broader set of firms than patenting firms. Panel (b) of Table 2.2 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 2.8 (a), we present a plot that summarizes the effect of innovation on operating performance for one through four years into the future. Consistent with how innovation should affect operating performance as a resource that earns the firm revenue, the effects are positive and significant for up to four years after a shock to innovation according to our measure, and these effects decay over time. By contrast, when we evaluate the effects of other measures of innovation over time, patents is unrelated to future operating performance, and the effect of R&D intensity decays much more rapidly over time (see Appendix Table B.14 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)

2.4.1.2 Growth Opportunities

Beyond the effects on operating performance, we expect innovation to have longer-term implications for the firm's growth opportunities. The intuition is that investors recognize an innovative firm when they see it, and rationally estimate an increase in the firm's future cash flows, thus enhancing its market valuation.

In line with this intuition, if text-based innovation is valuable in the same revealed preference sense, we should expect a significant effect on Tobin's Q because the market value will reflect this innovation premium. To evaluate this hypothesis, we examine the following specification:

Table 2.2: Performance of Firms and Text-Based Innovation (1990-2010)

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), 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 B.3). Variable definitions are presented in Table B.2. Standard errors that are double clustered on firm and year are reported in parentheses.

[Firm Performance]

	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.083^{***}	0.049^{***}	0.015^{***}	0.010^{**}
	(0.002)	(0.002)	(0.010)	(0.008)	(0.006)	(0.005)
$Log(Patents)_t$	0.002	−0.002	0.003	−0.027	−0.007	-0.015^{**}
	(0.003)	(0.003)	(0.015)	(0.022)	(0.005)	(0.007)
$Log(Citations)_t$	0.001	−0.0004	0.016^{*}	0.020^{*}	−0.003	0.002
	(0.002)	(0.001)	(0.009)	(0.010)	(0.003)	(0.004)
R&D/Assets $(Z)_t$	0.006	0.010^{**}	0.074^{***}	0.027	−0.001	−0.007
	(0.005)	(0.004)	(0.021)	(0.024)	(0.005)	(0.009)
Patenting Firm	0.009^{*}		0.037		0.00000	
	(0.006)		(0.031)		(0.009)	
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	0.436	0.674	0.577	0.771	0.099	0.159

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

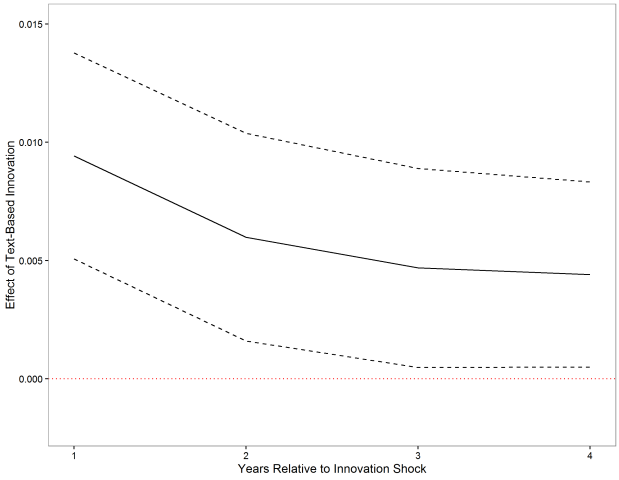
[Firm Performance - Patenting Firm Split]

	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^{**}	0.084^{***}	0.050^{***}	0.015^{***}	0.008
	(0.002)	(0.002)	(0.011)	(0.009)	(0.005)	(0.005)
× Non-Patenting Firm	0.010^{***}	0.006^{*}	0.078^{***}	0.046^{***}	0.017^{**}	0.020^{**}
	(0.004)	(0.003)	(0.016)	(0.014)	(0.008)	(0.009)
$Log(Patents)_t$	0.002	−0.002	0.003	−0.026	−0.007	-0.015^{**}
	(0.003)	(0.003)	(0.015)	(0.022)	(0.005)	(0.007)
$Log(Citations)_t$	0.001	−0.0004	0.016^{*}	0.020^{*}	−0.003	0.002
	(0.002)	(0.001)	(0.009)	(0.010)	(0.003)	(0.004)
R&D/Assets $(Z)_t$	0.006	0.010^{**}	0.074^{***}	0.027	−0.001	−0.007
	(0.005)	(0.004)	(0.021)	(0.024)	(0.005)	(0.008)
Patenting Firm	0.009		0.038		−0.0003	
	(0.006)		(0.031)		(0.009)	
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	0.436	0.674	0.577	0.771	0.099	0.159

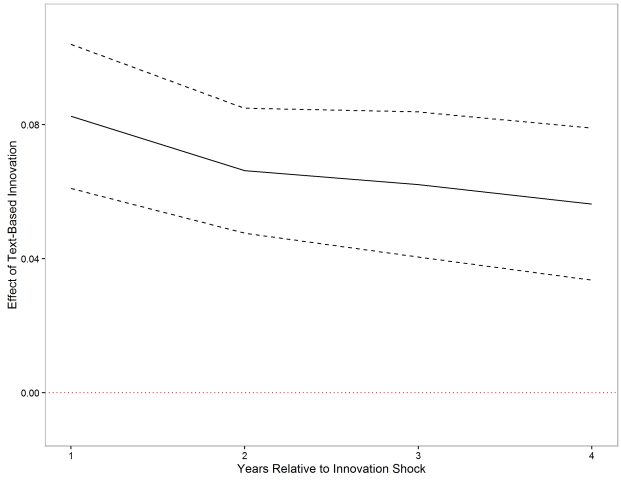
Note: *p<0.1; **p<0.05; ***p<0.01

Figure 2.8: Long Run Effects of Innovation on Performance – Forecasting ROA and Tobin's Q up to Four Years Out

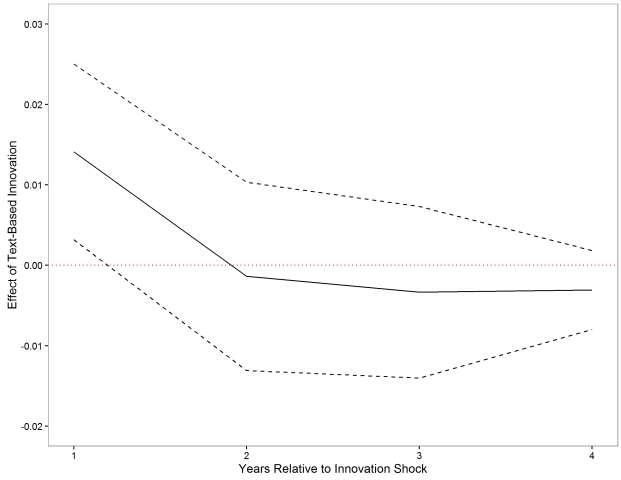
Note: These plots present the response in ROA, Q, and sales growth to a one standard deviation increase in the text-based measure of innovation. The X-axis represents the number of years ahead and the Y-axis is the beta estimate from appendix Table B.14. Dotted lines represent 95% confidence bands around the estimated effects.



[ROA]



[Q]



[Salesgrowth]

$$Q_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \quad (2.2)$$

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 2.2 present the results from estimating equation (2.2). 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 2.2 shows the results split by whether the firm uses patents. We see that an effect of innovation on Tobin's Q that is statistically indistinguishable between patenting and non-patenting firms. As with the results for operating performance, this finding highlights a notable advantage with our text-based innovation measure: it can be used for firms that do not use patents.

In Figure 2.8 (b), we present a plot that summarizes the effect of innovation on Q over time. Consistent with the idea that the market value of a firm reflects an innovation premium captured by our measure , the effects are positive and significant and these effects depreciate more slowly than the operating performance effects over time. For patents and R&D intensity, the effects over time are also persistent, but increase for some horizons (see Appendix Table B.14 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).

2.4.1.3 Growth in Sales

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

$$Salesgrowth_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \tag{2.3}$$

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 2.2 present the results from estimating equation (2.3). We find a statistically significant increase in sales for firms that have greater text-based innovation. As in the operating performance specifications, it is useful to compare the predictability of our text-based measure with R&D intensity and patent counts. Patent counts appear to be negatively associated with sales growth while there is no apparent relationship between sales growth and R&D intensity. Table 2.2 (b) shows the results split by whether the firm uses patents. Sales growth seems somewhat more associated with non-patenting firm innovation, though the results are not statistically different between non-patenting and patenting firms.

In Figure 2.8 (c), we present a plot that summarizes the effect of innovation on sales growth over time. Gains in sales growth are transitory, only occurring in the year following the increase in

innovation (see Appendix Table B.14 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.

2.4.1.4 Performance Results Using Rolling Window Version of the Measure

One concern with the innovation measure is the possibility of look-ahead bias in the performance regressions. Because we construct innovation topic from an LDA model fit on the entire sample period (1990-2010), a reader may be concerned that the innovation topic merely reflects factors that are eventually revealed to be valuable for firms, but that the information would not be viewed as innovation at the time of observation.

To address this potential concern, we reproduce the performance results using a 5-year rolling window version of text-based measure, which completely alleviates the look-ahead bias concern because the rolling window measure is based solely on past data. For example, in the rolling window version of the analysis, we construct the measure for a firm in 1995 using the topic loadings from a LDA model fit only using analyst reports from the previous five years (1990-1994).

Table 2.3 presents the performance results using the rolling window measure in place of the main measure. Results on operating performance and Tobin's Q are nearly identical in magnitude and statistical significance using the rolling window version, whereas the findings using sales growth are less robust (albeit the same sign and similar magnitude to the main result). These findings suggest that the relation between text-based innovation measure and firm performance reflects the value of true innovative activity rather than look-ahead bias.

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

In this subsection, we turn to examining the connection between the text-based innovation measure and patenting outcomes. Specifically, we examine the connection to standard patenting

Table 2.3: Performance of Firms and Text-Based Innovation – Rolling Window Version (1994-2010)

Note: This table presents OLS regressions that link the rolling window version of the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. The rolling window version of the text-measure is based on an LDA model of the 5 prior years of reports. Other innovation measures – log(patents), log(citations), 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 B.5). Variable definitions are presented in Table B.2. Standard errors that are double clustered on firm and year are reported in parentheses.

[Firm Performance]

	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.068^{***}	0.041^{***}	0.007	0.006
	(0.002)	(0.002)	(0.010)	(0.007)	(0.006)	(0.005)
$Log(Patents)_t$	0.003	−0.003	−0.010	-0.062^{***}	−0.005	−0.010
	(0.004)	(0.003)	(0.016)	(0.019)	(0.006)	(0.008)
$Log(Citations)_t$	0.0002	−0.001	0.027^{***}	0.031^{***}	−0.005	0.001
	(0.002)	(0.002)	(0.009)	(0.009)	(0.004)	(0.004)
R&D/Assets $(Z)_t$	0.006	0.011^{**}	0.084^{***}	0.030	0.001	−0.008
	(0.006)	(0.005)	(0.024)	(0.025)	(0.006)	(0.010)
Patenting Firm	0.012^{*}		0.040		0.006	
	(0.006)		(0.034)		(0.012)	
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.427	0.680	0.582	0.798	0.102	0.164

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

[Firm Performance - Patenting Firm Split]

	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^{**}	0.067^{***}	0.042^{***}	0.006	0.003
	(0.002)	(0.002)	(0.012)	(0.008)	(0.005)	(0.005)
× Non-Patenting Firm	0.012^{***}	0.008^{***}	0.069^{***}	0.036^{***}	0.011	0.015
	(0.004)	(0.003)	(0.014)	(0.011)	(0.014)	(0.016)
$Log(Patents)_t$	0.003	−0.003	−0.010	-0.062^{***}	−0.005	−0.010
	(0.004)	(0.003)	(0.016)	(0.019)	(0.006)	(0.008)
$Log(Citations)_t$	0.0003	−0.001	0.027^{***}	0.031^{***}	−0.005	0.001
	(0.002)	(0.002)	(0.009)	(0.009)	(0.004)	(0.004)
R&D/Assets $(Z)_t$	0.006	0.011^{**}	0.084^{***}	0.029	0.001	−0.008
	(0.006)	(0.005)	(0.024)	(0.025)	(0.006)	(0.010)
Patenting Firm	0.011^{*}		0.040		0.004	
	(0.006)		(0.036)		(0.012)	
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.427	0.680	0.582	0.798	0.102	0.164

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

outcomes (patent counts, citations, and impact), as well as the value of patents within the universe of firms that use patents (described in Kogan et al., 2017).

2.4.2.1 Patent Value Measures

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

$$Log(1 + PatentValue_{it+1}) = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \quad (2.4)$$

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

Panel (a) of Table 2.4 presents results from estimating equation 2.4 using the absolute dollar value measure of patent value. We find a robust relationship where text-based innovation is associated with meaningful increases in future patent values. This relationship holds after controlling for patent citations, a measure that is often used as a proxy for patent value, and beyond being robust to granular industry fixed effects and firm fixed effects, it is also robust to controlling for other time-varying firm characteristics. In panel (b), we report results using the Value per Patent measure, which show a similarly robust relationship between text-based innovation and future patenting values.

Table 2.4: Patent Value and Text-Based Innovation (1990-2010)

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

[Patent Value]

	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

[Value Per Patent]

	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

2.4.2.2 Text-Based Innovation and Patenting

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

$$Log(1 + \sum_{s=1}^{3} PatentingOutcome_{t+s}) \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \qquad (2.5)$$

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 2.5, we present the results from estimating equation (2.5). The text-based innovation measure is positively related to patent counts, citations, and citations per patent over the next three years. All of these estimates are statistically significant at better than the five percent level, and are robust to broad industry classifications (2-digit SIC).

The findings in this section indicate that our measure contributes valuable information, even within the set of firms that use patents to protect their innovations. Within the set of patenting firms, our text-based measure is strongly correlated with the most valuable patents, and it is a leading indicator of whether firms will patent in the coming years. Moreover, the text-based innovation measure can be computed using analyst reports in real time while patenting outcomes take longer (e.g., even counts of applications for eventually granted patents must wait for the patent to be granted or denied). Thus, our text-based measure is useful in providing a leading indicator for more traditional modes of innovative activity that take time to observe.

2.4.3 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

Table 2.5: Patents and Text-Based Innovation (1990-2010)

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

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

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

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

2.4.3.1 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:

$$Log(HHISimilarity_{i,t+1}) = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \quad (2.6)$$

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 (2.6) are presented in columns 1 and 2 of Table 2.6. Inconsistent with text-based innovation reflecting greater differentiation of the final product, we find no statistically significant relationship between our text-based measure of innovation and the Hoberg-Phillips HHI measure. We are cautious about over-interpreting a failure to reject, but note that the point estimate is small in magnitude, and opposite in sign from an innovation-as-differentiation

Figure 2.9: Valuable Patents (95th percentile)

Note: This is a list of patents on the 95th percentile of patent values ($80 million). Observations with only one patent grant during the day are shown.

Firm	Patent	Date	Title	Abstract
WASTE MANAGEMENT	4,927,317	1990-05-22	Apparatus for temporarily covering a large land area	A method for temporarily covering a large land area and an apparatus for suspending a flexible cover from a front loader bucket of an earth-moving vehicle.
COMPAQ	5,454,081	1995-09-26	Expansion bus type determination apparatus	A circuit that automatically detects whether an input/output expansion board is connected to an EISA system or an ISA system.
TEXACO	5,644,244	1997-07-01	Method for analyzing a petroleum stream	Methods are provided for determining a solids to liquids ratio in a flowing petroleum stream having an immiscible solids, oil and water flow.
3COM CORP	5,651,002	1997-07-22	Internetworking device with enhanced packet header translation and memory	An internetworking device providing enhanced packet header translation for translating the format of a header associated with a source network into a header format associated with a destination network of a different type than the source network.
ERICSSON	5,706,301	1998-01-06	Laser wavelength control system	A laser wavelength control system (20) stabilizes laser output wavelength. The control system includes a reflector/filter device (40) upon which laser radiation is incident for yielding both a filtered-transmitted signal (FS) and a reflected signal (RS).
HALLIBURTON	5,716,910	1998-02-10	Foamable drilling fluid and methods of use in well drilling operations	A foamable drilling fluid for use in well operations such as deep water offshore drilling where risers are not employed in returning the fluid to the surface mud pit.
ELECTRONIC DATA SYSTEMS	5,801,366	1998-09-01	Automated system and method for point-of-sale (POS) check processing	An automated check processing system includes an input device receiving checking account information and a check amount of a check provided for payment in a translation.
LILLY (ELI)	7,138,521	2006-11-21	Crystalline of N-[4-[2-(2-Amino-4,7-dihydro-4-oxo-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.

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 2.9.

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

$$Log(1 + ProductIntroductions_{i,t+s}) \quad = \quad \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \qquad (2.7)$$

where $ProductIntroductions_{i,t+1}$ is the count of firm i's product introductions (based on a textual analysis of firm press releases in Mukherjee 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 through 6 of Table 2.6 present results from estimating equation (2.7). 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,

Table 2.6: Product Differentiation and Product Announcements (1990-2010)

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

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

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

future patenting, and performance outcomes in a manner that is theoretically consistent with innovation. Thus, it is useful to dig into the nature of text-based innovation – specifically, the nature of valuable patents and the nature of highly-innovative firms outside of the set of patenting firms.

2.4.3.2 Contextual Examples of Systems Innovation in Mature Firms

Within the set of patenting firms, it is useful to examine the content of valuable innovations. Figure 2.9 presents a list of valuable patents in order of value starting at the 95th percentile of patent values. Most of these highly valuable patented innovations are not particular to a specific product, but rather reflect a valuable component or the patenting of a valuable process. In fact, only one patent in this list is directly related to a specific product – a vaccine. Other patents are either processes, components that can go in to one or several products, or components useful in the production process. Given that our measure appears to pick up on valuable patents with these characteristics that reflect innovative systems, these examples offer some insight into why we do not find a connection with product introductions or product differentiation.

Taking a step outside of the universe of patenting firms, we turn our attention to the retail sector in 1993, which our measure indicates as highly innovative, but nonetheless, is a low-patenting industry at the time. Figure 2.1 presents two excerpts from analyst reports of firms that are considered particularly innovative. These are firms that do not rely heavily on patents, but are considered innovative by the analyst. Consistent with our interpretation that the innovation we measure reflects innovative systems, the reports describe the firms as innovative in ways that are separate from bringing new products to market. For example, the analyst report about Walmart describes how Walmart "uses technology to improve productivity and at the same time reduce costs." The report describes several dimensions along which Walmart is innovative, and is an industry leader, in the way they use technology in their supply chain management and theft prevention. Because these innovations were not discovered using R&D expenditures and were not patented, our measure is in a unique position to capture this type of innovation, which is a common for mature firms like Walmart that have particularly innovative systems.

2.4.4 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)

2.4.4.1 Fitted Model Robustness

Table 2.7 presents robustness to the LDA model fit. First, in panel (a) we summarize the results of the 50-topic LDA robustness exercise. In our main specifications, we use relatively few topics to ensure that we capture the generality of the notion of innovation. If the 50-topic LDA has too many topics, the concern is that multiple topics could capture innovation in a similar way. To address this concern, we fit a topic model with 50 topics and identify the topic that is most similar to our main measure (Topic 6 from the 15-topic LDA). Two topics from the 50-topic model are highly correlated with our original topic, and the content of these topics is qualitatively similar (see Figure B.2). Table 2.7 (a) presents results with one of these two topics as the measure of innovation (using the other one makes no qualitative difference). We obtain results that are similar to Table 2.2(a) which suggests that the results in the paper are not driven by the choice of the number of topics.

Second, in panel (b), we address the concern that the other topics in the 15-topic LDA are correlated with our measure, and thus, drive the result for a more mechanical reason (e.g., an 'operating performance' topic emerges in the 15-topic LDA, see Figure B.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 2.7 indicate, the main results are qualitatively similar after controlling for other topic loadings, though in some cases, they become stronger.

Table 2.7: Robustness of LDA Model Fit

Note: The specifications and variable definitions for ROA, Q, and Salesgrowth are analogous to those in Table 2.2. 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.

[Firm Performance, K=50 (1990-2010)]

	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

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

	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

2.4.4.2 Robustness to Alternative Explanations

Table 2.8 presents robustness to three other alternative explanations. In particular, because construction of the measure relies on only the reports with high analyst sentiment, a reader may be concerned that the sentiment of the reports rather than their content is driving the relation of text-based innovation to the performance measures. Panel (a) of Table 2.8 presents the results controlling for analyst sentiment, which are similar to the main results.

In addition, given the words most prominently used in the innovation topic, a reader may have a separate concern that the LDA topic is merely a crude technique to approximate for whether analysts discuss the firm's revenue or growth prospects, unrelated to innovation. To address this issue, we construct word counts of analyst usage of the words "revenue" and "growth" to be used as controls in the specification. Panel (c) of Table 2.8 presents these results, which show that controlling for the relative word usage of "revenue" or "growth" does not explain the topic's relationship to firm performance. In panel (c) of Table 2.8, we conduct a similar exercise using words with the root "tech" in them. These word count controls indicate that the topic is not merely selecting the relative incidence of particular words, but consistent with the motivation to use LDA, our methodology seems to be picking up these words when they are used together contextually.

2.5 Innovation and Acquisition Activity

This section provides a useful application of the text-based innovation measure. Specifically, for the mature firms we study, we examine the relation between text-based innovation and subsequent acquisition activity. The results in this section are consistent with text-based innovation reflecting an innovative system that generates productive merger opportunities, and are difficult to reconcile with agency-based rationales for merging.

Table 2.8: Accounting for Alternative Explanations (1990-2010)

Note: The specifications and variable definitions for ROA, Q, and Salesgrowth are the same as in Table 2.2. 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.

[Controlling for average sentiment]

	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

[Controlling for words related to revenue and growth]

	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

[Controlling for "technology" words]

	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

2.5.1 Text-based Innovation and Acquisition Activity

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

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

$$Log(1 + \sum_{s=1}^{3} Acquisitions_{t+s}) = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \tag{2.8}$$

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

Columns 1 and 2 of Table 2.9 Panel (a) presents the results from estimating equation (2.8). Across specifications, we find evidence that greater measured innovation today is associated with greater acquisition activity over the next three years.[11] This estimated effect is robust to controlling for profitability and the market-to-book ratio, which proxy for free cash flow and relatively

[11] In the appendix (Table B.10), we present estimates from an analogous specification, but using a linear probability model for whether the firm engages in any acquisitions in the coming 3 years (rather than the number of acquisitions during that time). We find that higher innovation is associated with the extensive margin (i.e., the main result is not just an increase in acquisitions among the set of firms that would conduct M&A anyway).

overvalued equity. Thus, the relation between innovation and acquisitions is unlikely to be driven by agency-based explanations for abnormal M&A activity. For the mature firms we study, this finding suggests that innovation is complementary to acquisition activity. This finding corroborates the underlying intuition described by Fresard et al. (2017) in the context of vertical acquisitions, and extends this finding to mature firm innovation.

2.5.2 Text-based Innovation and Small Acquisitions

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

In columns 3 through 6 of Table 2.9 (a), we present the results for two splits of acquisition activity: large acquisitions (columns 3 and 4) and small acquisitions (columns 5 and 6). Across specifications, we find that greater text-based innovation is associated with significantly more small acquisitions in the future, but there is a much weaker relation between text-based innovation and large acquisitions.[12] Moreover, the positive relation between text-based innovation and small acquisitions is consistent with our overall interpretation that the text-based innovation measure captures innovative systems. This subsample finding is inconsistent with other prominent rationales

[12] In the appendix (Table B.12), we also relate merger announcement returns (CARs) to our text-based measure of innovation. For acquirers with high text-based innovation, we find that small acquisitions are viewed significantly more favorably than large acquisitions. Together with our finding that innovation is associated with more small acquisitions (but not large acquisitions), the CAR analysis suggests that text-based innovation generates synergies that are well recognized by investors, and that corporate actions inconsistent with this synergy view (i.e., attempting to acquire a large firm, potentially with integration risk) are viewed negatively by investors.

Table 2.9: Predicting Acquisition Activity Using the Text-Based Innovation Measure (1990-2010)

Note: The dependent variable in panel (a) is number of acquisitions completed in the next three years; this is the count of acquisition records from the SDC database which fall in the next three fiscal years. Panel (b) uses an alternative text-measure that is calculated without words that start with "merg" and "acqui". As in other tables, the text-based measure is standardized to have a mean of 0 and a standard deviation of 1. Other controls include log (patents), ROA, R&D intensity, log(assets), asset tangibility, leverage, log(age), log(Q), and a dummy for patenting firm. Full results are presented in the appendix (Table B.8). Variable definitions are presented in Table B.2. Standard errors that are double clustered on firm and year are reported in parentheses.

[Acquisition Count – Main Innovation Measure]

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

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

[Acquisition Count – Corpus Purged of Merger and Acquisition Words]

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

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

to merge (e.g., empire building), which would tend to lead to larger acquisitions.[13]

2.5.3 Purging the Innovation Measure of Acquisition-Specific Language

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

In Panel (b) of Table 2.9, we present the results on acquisition activity using this acquisitions-purged text-based measure of innovation. As is apparent from the results, the broad conclusions not only remain, but in some cases the magnitudes and statistical significance on the relation between text-based innovation and subsequent acquisition activity strengthens. In this way, these findings enhance our confidence that the underlying relation between our text-based innovation measure and acquisition activity reflects incentives faced by highly innovative firms rather than an artifact of the underlying textual descriptions.

2.6 Conclusions

In this paper, we have developed a useful new measure of corporate innovation based on a textual analysis of analyst reports. Our text-based innovation measure provides a useful description of innovation in mature firms without patents and with zero R&D expenditure. Such firms are common, even among our sample of 703 firms from the S&P500, there are 219 firms with no patents and 329 firms that had zero R&D expenditure for our entire sample period (1990-2012). Moreover, there is a substantial overlap between the distribution of innovation for patenting firms

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

and the distribution of innovation for non-patenting firms (similarly for R&D versus zero-R&D), which indicates that important innovative activities are overlooked by using patenting and R&D as proxies for innovation. Indeed, this view is confirmed by notable examples of firms that do not patent or use R&D, but are nonetheless identified as highly innovative by our measure (e.g., Walmart).

Beyond expanding the sample of innovative firms to study, our textual analysis provides a useful step toward understanding innovation in the spirit of Schumpeter (1934), who described five types of innovation: new products, new methods of production, new sources of supply, exploitation of new markets, and new ways to organize business. Patenting and R&D expenditure typically pertain to product innovation, and the literature's focus on these measures has left the other categories understudied. To take one example of how adopting this broader view (and measurement) of innovation is useful, recent research by Fresard et al. (2017) argues that firms with realized innovations are more likely to be acquired in a vertical merger because realized innovations are easier to commercialize than innovations in progress. The authors use patenting outcomes to proxy for realized innovation, and thus, their focus is primarily on innovation and commercialization of products. As our analysis shows, the text-based innovation measure captures important innovative activity in business systems, unrelated to products. This mode of innovation likely exhibits a different relation to corporate outcomes that have been linked to product innovations. In this vein, future research could use textual measures of innovation to examine the extent to which the lessons learned from studying product innovations translate into other types of corporate innovation.

Finally, although our analysis is applied to the text of analyst reports, our textual approach could be applied to other settings to identify complementary measures of innovation. Media articles, required firm disclosures (10Ks), and press releases may also contain information about firms' innovative activities. Recent work has considered some of these textual databases as a source of information on corporate innovation (e.g., see the analysis of product innovation in Mukherjee et al., 2016 using press releases), but given the available wealth of textual sources of information about firms, much more progress is possible. Our text-based innovation measure suggests that examining

these sources of textual information about firms is fertile ground for future research.

Bibliography

Agarwal, Sumit, Sudip Gupta, and Ryan D. Israelsen (2016) "Public and Private Information: Firm Disclosure, SEC Letters, and the JOBS Act," *Working Paper SSRN 2891089*.

Ashenfelter, Orley (1978) "Estimating the effect of training programs on earnings," *The Review of Economics and Statistics*, pp. 47–57.

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

Asquith, Paul, Michael B Mikhail, and Andrea S Au (2005b) "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," *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.

Barber, Brad M, Reuven Lehavy, Maureen McNichols, and Brett Trueman (2006) "Buys, holds, and sells: The distribution of investment banks stock ratings and the implications for the profitability of analysts recommendations," *Journal of accounting and Economics*, Vol. 41, No. 1, pp. 87–117.

Barniv, Ran, Ole-Kristian Hope, Mark J Myring, and Wayne B Thomas (2009) "Do analysts practice what they preach and should investors listen? Effects of recent regulations," *The Accounting Review*, Vol. 84, No. 4, pp. 1015–1039.

Bellstam, Gustaf, Sanjai Bhagat, and J. Anthony Cookson (2016) "A Text-Based Analysis of Corporate Innovation," *Working Paper*.

Bernstein, Shai (2015) "Does going public affect innovation?" *Journal of Finance*, Vol. 70, No. 4, pp. 1365–1403.

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," *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Bodnaruk, Andriy, Tim Loughran, and Bill McDonald (2015) "Using 10-K text to gauge financial constraints," *Journal of Financial and Quantitative Analysis*, Vol. 50, pp. 623–646.

Boni, Leslie (2005) "Analyzing the analysts after the Global Settlement," in *conference presentation, Brookings-Nomura Seminar, Brookings Institution, Washington, DC*.

Boni, Leslie and Kent Womack (2002) "Wall Street's credibility problem: misaligned incentives and dubious fixes?" *Brookings-Wharton papers on financial services*, Vol. 2002, No. 1, pp. 93–130.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson (2013) "Which News Moves Stock Prices? A Textual Analysis," *NBER Working Paper No. 18725*.

Bradley, Daniel, Sinan Gokkaya, and Xi Liu (2015a) "Before an analyst becomes an analyst: Does industry experience matter?" *Journal of Finance, Forthcoming*.

Bradley, Daniel, Incheol Kim, and Xuan Tian (2015b) "Do unions affect innovation?" *Management Science, Forthcoming*.

Bradshaw, Mark T (2009) "Analyst information processing, financial regulation, and academic research," *The Accounting Review*, Vol. 84, No. 4, pp. 1073–1083.

Bradshaw, Mark Thomas (2011) "Analysts' forecasts: What do we know after decades of work?" *Working Paper 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*, Vol. 64, No. 1, pp. 151–185.

Brown, Lawrence D, Andrew C Call, Michael B Clement, and Nathan Y Sharp (2015) "Inside the "Black Box" of Sell-Side Financial Analysts," *Journal of Accounting Research*, Vol. 53, No. 1, pp. 1–47.

Busse, Jeffrey A and T Clifton Green (2002) "Market efficiency in real time," *Journal of Financial Economics*, Vol. 65, No. 3, pp. 415–437.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2011) "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, Vol. 29, No. 2, pp. 238–249.

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

Chen, Chih-Ying and Peter F Chen (2009) "NASD Rule 2711 and changes in analysts' independence in making stock recommendations," *The Accounting Review*, Vol. 84, No. 4, pp. 1041–1071.

Chen, Jason V, Venky Nagar, and Jordan Schoenfeld (2015) "Information Dispersion in Financial Markets," *Ross School of Business Paper*, No. 1197.

Clement, Michael B and Senyo Y Tse (2005) "Financial analyst characteristics and herding behavior in forecasting," *The Journal of finance*, Vol. 60, No. 1, pp. 307–341.

Cohen, Lauren, Karl Diether, and Christopher Malloy (2013) "Misvaluing innovation," *Review of Financial Studies*, Vol. 26, No. 3, pp. 635–666.

Cohen, Lauren and Andrea Frazzini (2008) "Economic links and predictable returns," *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," *NBER Working Paper No. 20322*.

Cohen, Lauren, Christopher Malloy, and Quoc H. Nguyen (2016) "Lazy Prices," *Working Paper SSRN 1658471*.

Cookson, J. Anthony, S. Katie Moon, and Joonki Noh (2017) ""Some People Say," Evasive Language in Corporate Disclosures," *Working Paper*.

Cooper, Michael J., Anne Marie Knott, and Wenhao Yang (2015) "Measuring Innovation," *Working Paper SSRN 2631655*.

Cornaggia, Jess, Kimberly J Cornaggia, and Han Xia (2016) "Revolving doors on wall street," *Journal of Financial Economics*, Vol. 120, No. 2, pp. 400–419.

Corwin, Shane A, Stephannie A Larocque, and Mike A Stegemoller (2017) "Investment banking relationships and analyst affiliation bias: The impact of the global settlement on sanctioned and non-sanctioned banks," *Journal of Financial Economics*, Vol. 124, No. 3, pp. 614–631.

Da, Zhi and Xing Huang (2016) "Harnessing the wisdom of crowds," *Working Paper*.

Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher A Parsons (2012) "Journalists and the stock market," *Review of Financial Studies*, Vol. 25, No. 3, pp. 639–679.

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," *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 SSRN 2364096*.

Ertimur, Yonca, William J Mayew, and Stephen R Stubben (2011a) "Analyst reputation and the issuance of disaggregated earnings forecasts to I/B/E/S," *Review of Accounting Studies*, Vol. 16, No. 1, pp. 29–58.

Ertimur, Yonca, Volkan Muslu, and Frank Zhang (2011b) "Why are recommendations optimistic? Evidence from analysts' coverage initiations," *Review of Accounting Studies*, Vol. 16, No. 4, pp. 679–718.

Ertimur, Yonca, Jayanthi Sunder, and Shyam V Sunder (2007) "Measure for measure: The relation between forecast accuracy and recommendation profitability of analysts," *Journal of Accounting Research*, Vol. 45, No. 3, pp. 567–606.

Fresard, Laurent, Gerard Hoberg, and Gordon M. Phillips (2017) "Innovation Activities and the Incentives for Vertical Acquisitions and Integration," *Working Paper SSRN 2242425*.

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 (2015) "Patents Rights and Innovation by Small and Large Firms," *Working Paper SSRN 2694725*.

Ganglmair, Bernhard and Malcolm Wardlaw (2017) "Complexity, standardization, and the design of loan agreements," *Working Paper*.

Ganter, Viola and Michael Strube (2009) "Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 173–176, Association for Computational Linguistics.

García, Diego (2013a) "Sentiment during recessions," *The Journal of Finance*, Vol. 68, No. 3, pp. 1267–1300.

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

García, Diego (2017) "The kinks of financial journalism," *Working Paper*.

Goldsmith-Pinkham, Paul, Beverly Hirtle, and David Lucca (2016) "Parsing the Content of Bank Supervision," *Working Paper FRBNY Staff Report No. 770*.

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.

Griliches, Zvi ed. (1984) *R&D, Patents, and Productivity*: University of Chicago Press.

Griliches, Zvi (1998) *The search for R&D spillovers, R&D and Productivity: The Econometric Evidence*: University of Chicago Press.

Groysberg, Boris, Paul M Healy, and David A Maber (2011) "What Drives Sell-Side Analyst Compensation at High-Status Investment Banks?" *Journal of Accounting Research*, Vol. 49, No. 4, pp. 969–1000.

Guan, Yuyan, Hai Lu, and MH Franco Wong (2013) "Regulations and brain drain: Evidence from the Wall Street star analysts career choices," *Review of Accounting Studies forthcoming*.

Hall, Bronwyn H (1990) "The Impact 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," *NBER Working Paper No. 8408*.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg (2005) "Market Value and Patent Citations," *Rand Journal of Economics*, Vol. 36, No. 1, 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 (2016) "... and the Cross-Section of Expected Returns," *Review of Financial Studies*, Vol. 29, No. 1, pp. 5–68.

He, Jie 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 A, Yaron Levi, Ben Lourie, and Siew Hong Teoh (2017) "Decision Fatigue and Heuristic Analyst Forecasts," *Working Paper*.

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.

Hirshleifer, David and Tyler Shumway (2003) "Good day sunshine: Stock returns and the weather," *The Journal of Finance*, Vol. 58, No. 3, pp. 1009–1032.

Hirst, D Eric, Lisa Koonce, and Paul J Simko (1995) "Investor reactions to financial analysts' research reports," *Journal of Accounting Research*, pp. 335–351.

Hoberg, Gerard and Craig Lewis (2017) "Do Fraudulent Firms Produce Abnormal Disclosure?" *Journal of Corporate Finance, Forthcoming*.

Hoberg, Gerard and Vojislav Maksimovic (2015) "Redefining Financial Constraints: A Text-Based Analysis," *Review of Financial Studies*, Vol. 28, No. 5, pp. 1312–1352.

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

——— (2016) "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, Vol. 124, No. 5, pp. 1423–1465.

Hoberg, Gerard and Gordon M. Phillips (2016) "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, Vol. 124, No. 5, pp. 1423–1465.

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

Hong, Harrison, Jeffrey D Kubik, and Amit Solomon (2000) "Security analysts' career concerns and herding of earnings forecasts," *The Rand journal of economics*, pp. 121–144.

Hsu, Charles, Xi Li, Zhiming Ma, and Gordon M Phillips (2015) "Does Product Market Competition Influence Analyst Coverage and Analyst Career Success?" *Working Paper. Available at SSRN*.

Huang, Allen H, Reuven Lehavy, Amy Y Zang, and Rong Zheng (2017) "Analyst information discovery and interpretation roles: A topic modeling approach," *Management Science*.

Huang, Allen H, Amy Y Zang, and Rong Zheng (2014) "Evidence on the information content of text in analyst reports," *The Accounting Review*, Vol. 89, No. 6, pp. 2151–2180.

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

Huang, Allen, Amy Zang, and Rong Zheng (2014) "Evidence on the Information Content of Text in Analyst Reports," *The Accounting Review*, Vol. 89, No. 6, pp. 2151–2180.

Israelsen, Ryan D. (2014) "Tell It Like It Is: Disclosed Risks and Factor Portfolios," *Working Paper SSRN 2504522*.

Jegadeesh, Narasimhan, Joonghyuk Kim, Susan D Krische, and Charles Lee (2004) "Analyzing the analysts: When do recommendations add value?" *The journal of finance*, Vol. 59, No. 3, pp. 1083–1124.

Jegadeesh, Narasimhan and Di Wu (2017) "Deciphering Fedspeak: The Information Content of FOMC Meetings," *Working Paper SSRN 2939937*.

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

Jones, Charles I (2002) "Sources of US economic growth in a world of ideas," *American Economic Review*, Vol. 92, No. 1, pp. 220–239.

Jurafsky, Dan and James H Martin (2017) *Speech and language processing*, 3rd edition.

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

Kadan, Ohad, Leonardo Madureira, Rong Wang, and Tzachi Zach (2009) "Conflicts of interest and stock recommendations: The effects of the global settlement and related regulations," *The Review of Financial Studies*, Vol. 22, No. 10, pp. 4189–4217.

Kisgen, Darren J, Matthew Osborn, and Jonathan Reuter (2016) "Analyst Promotions within Credit Rating Agencies: Accuracy or Bias?"Technical report, National Bureau of Economic Research.

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 (2017) "Technological Innovation, Resource Allocation, and Growth," *Quarterly Journal of Economics, Forthcoming*.

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 (2015) "Search Based Peer Firms: Aggregating Investor Perceptions through Internet Co-Searches," *Journal of Financial Economics*, Vol. 116, No. 2, pp. 410–431.

Lin, Hsiou-wei and Maureen F McNichols (1998) "Underwriting relationships, analysts' earnings forecasts and investment recommendations," *Journal of Accounting and Economics*, Vol. 25, No. 1, pp. 101–127.

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

Lowry, Michelle, Roni Michaely, and Ekaterina Volkova (2016) "Information Revelation Through Regulatory Process: Interactions between the SEC and Companies Ahead of the IPO," *Working Paper SSRN 2802599*.

Mann, William (2016) "Creditor rights and innovation: Evidence from patent collateral," *Working Paper SSRN 2356015*.

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

McNichols, Maureen and Patricia C O'Brien (1997) "Self-selection and analyst coverage," *Journal of Accounting Research*, Vol. 35, pp. 167–199.

Mehran, Hamid and René M Stulz (2007) "The economics of conflicts of interest in financial institutions," *Journal of Financial Economics*, Vol. 85, No. 2, pp. 267–296.

Mensah, Yaw M and Rong Yang (2008) "An empirical evaluation of analysts' herding behavior following Regulation Fair Disclosure," *Journal of Accounting and Public Policy*, Vol. 27, No. 4, pp. 317–338.

Merkley, Kenneth, Roni Michaely, and Joseph Pacelli (2017) "Does the Scope of the Sell-Side Analyst Industry Matter? An Examination of Bias, Accuracy, and Information Content of Analyst Reports," *The Journal of Finance*, Vol. 72, No. 3, pp. 1285–1334.

Michaely, Roni and Kent L Womack (1999) "Conflict of interest and the credibility of underwriter analyst recommendations," *Review of financial studies*, Vol. 12, No. 4, pp. 653–686.

Mikhail, Michael B, Beverly R Walther, and Richard H Willis (2003) "The effect of experience on security analyst underreaction," *Journal of Accounting and Economics*, Vol. 35, No. 1, pp. 101–116.

Miller, George A. (1995) "WordNet: A Lexical Database for English," *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41.

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," *American Economic Review*, Vol. 98, No. 4, pp. 1370–1396.

Nordhaus, William D (1969) "An economic theory of technological change," *American Economic Review*, Vol. 59, No. 2, pp. 18–28.

Popadak, Jillian A (2013a) "A corporate culture channel: How increased shareholder governance reduces firm value," *Working Paper*.

Popadak, Jillian A. (2013b) "A Corporate Culture Channel: How Increased Shareholder Governance Reduces Firm Value," *Working Paper SSRN 2345384*.

Previts, Gary John, Robert J Bricker, Thomas R Robinson, and Stephen J Young (1994) "A content analysis of sell-side financial analyst company reports," *Accounting Horizons*, Vol. 8, No. 2, p. 55.

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

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

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

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

Solow, Robert M (1956) "A contribution to the theory of economic growth," *Quarterly Journal of Economics*, Vol. 70, No. 1, pp. 65–94.

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

Teh, Yee Whye, Michael I. Jordan, Mathew J. Beal, and David M. Blei (2006) "Hierarchical Dirchlet Process," *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581.

Tetlock, Paul C (2007) "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, Vol. 62, No. 3, pp. 1139–1168.

Tian, Xuan (2012) "The role of venture capital syndication in value creation for entrepreneurial firms," *Review of Finance*, Vol. 16, No. 1, pp. 245–283.

Tian, Xuan and Tracy Yue Wang (2014) "Tolerance for Failure and Corporate Innovation," *Review of Financial Studies*, Vol. 27, No. 1, pp. 211–255.

Tidd, Joe, John Bessant, and Keith Pavitt (2005) *Managing Innovation: Integrating Technological, Market and Organizational Change*: John Wiley & Sons.

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

Trueman, Brett (1994) "Analyst forecasts and herding behavior," *The Review of Financial Studies*, Vol. 7, No. 1, pp. 97–124.

Tucker, Catherine (2014) "Patent Trolls and Technology Diffusion: The Case of Medical Imaging," *Working Paper 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.

Welch, Ivo (2000) "Herding among security analysts," *Journal of Financial economics*, Vol. 58, No. 3, pp. 369–396.

Womack, Kent L (1996) "Do brokerage analysts' recommendations have investment value?" *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.

Appendix A

Appendix to Chapter 1

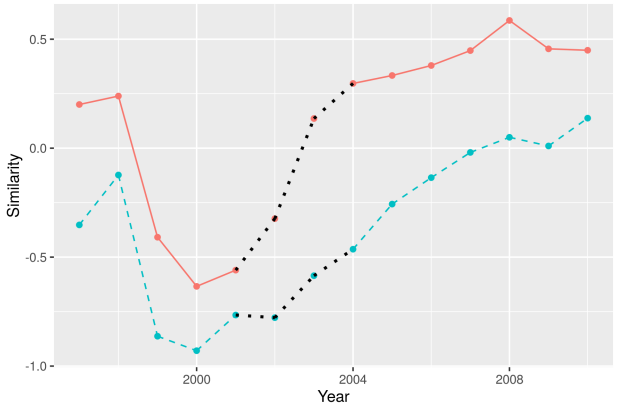
Figure A.1: Weasel Examples

Note: This figure presents a few examples of sentences from Wikipedia containing weasel tags. The sentences are collected from a Wikipedia dump that contains all text and all tags from Wikipedia on June 1, 2017. Editors at Wikipedia are asked to tag sentences as weasel sentences using a tag that starts "{{weasel" when such sentences are found. The tagged sentences are the basis for my **weaseling** measure. Weasel words are "words and phrases aimed at creating an impression that a specific or meaningful statement has been made".

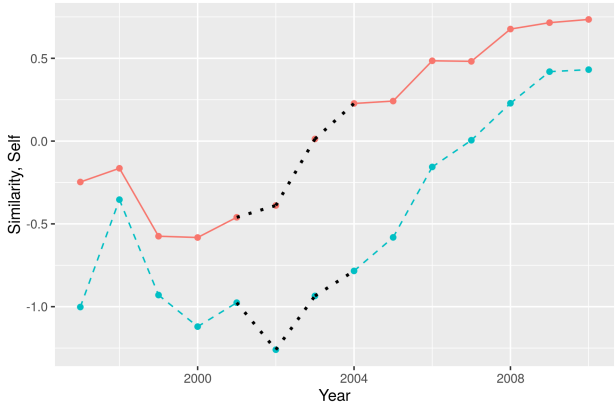
- It has been reported{{weasel inline—date=June 2016}} that the British often marched to a version believed to be about a man named [[Thomas Ditson]] of [[Billerica, Massachusetts]].
- Some people{{who—date=October 2015}}{{weasel-inline—date=October 2015}} suggest that digital pets are preferable for a number of reasons.
- "'Indrani Iriyagolle"' (c. 1933-2015) was a well-known figure{{weasel-inline—date=October 2015}} in Sri Lanka for rehabilitation, welfare, women's rights and humanitarian work.
- The Scheme, which is recognised as a leading example of [[public/private partnership]] in the UK{{weasel-inline—date=March 2016}}, is owned by its Members but is underpinned by a [[HM Treasury]] commitment to support Pool Re if ever it has insufficient funds to pay a legitimate claim.

Figure A.2: Equal Trends - Similarity

Note: This figure shows how the **Similarity** measure moves over time. **Similarity** is measured as the cosine similarity between the report and all other reports about the same firm that were issued in the prior three month. **Similarity, Self** is defined analogously, but using reports authored by the writer of the current report only. The two lines represent reports by investment banking analysts (red solid) and other analysts (teal dashed). **Similarity** is averaged by year for each of the two groups and then plotted from 1998 to 2010. The dotted part of the plot indicates the period which I remove in the regression analysis that follows (the period of regulation change). All are measures centered at zero and normalized to have a standard deviation of 1.



[Similarity]



[Similarity, Self]

Figure A.3: Equal Trends - Weasel

Note: This figure shows how **weaseling**, which is measured as the fraction of weasel sentences in the document, moves over time. The two lines represent reports by investment banking analysts (red solid) and other analysts (teal dashed). The measure is averaged by year for each of the two groups and then plotted from 1998 to 2010. The dotted part of the plot indicates the period which I remove in the regression analysis that follows (the period of regulation change). All measures are centered at zero and normalized to have a standard deviation of 1.

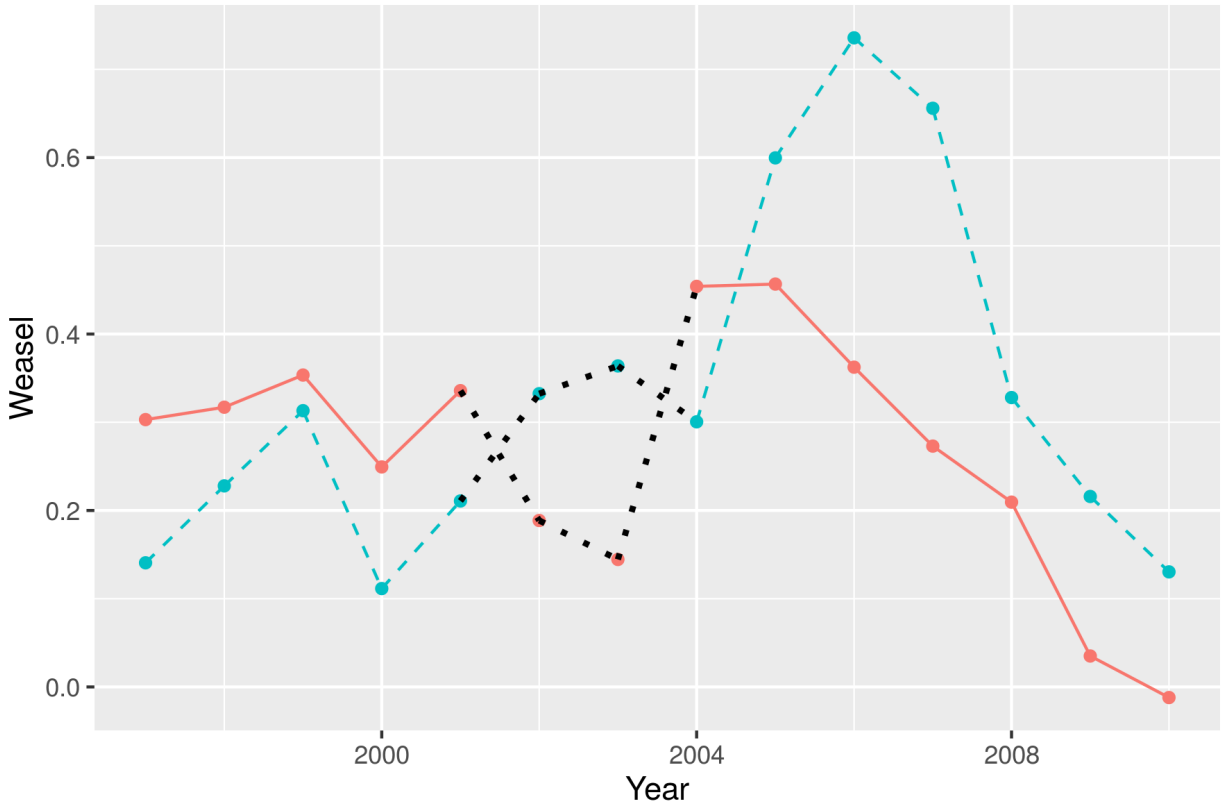
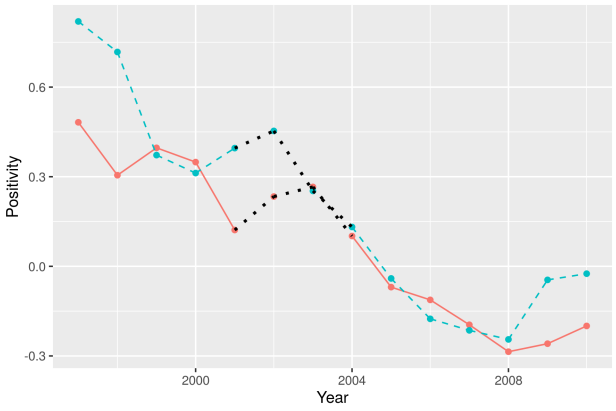
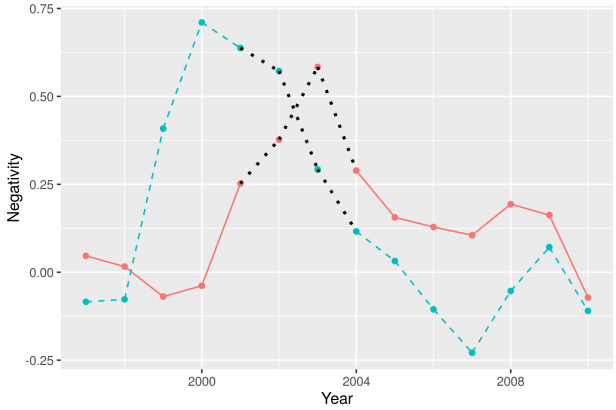


Figure A.4: Equal Trends - Sentiment

Note: This figure shows how **positivity** (panel a) and **negativity** (panel b) move over time. **Positivity** is measured as the fraction of positive words in the document, **negativity** is defined analogously. The two lines represent reports by investment banking analysts (red solid) and other analysts (teal dashed). The measure is averaged by year for each of the two groups and then plotted from 1998 to 2010. The dotted part of the plot indicates the period which I remove in the regression analysis that follows (the period of regulation change). All measures are centered at zero and normalized to have a standard deviation of 1.



[Positivity]



[Negativity]

Table A.1: Robustness to Disclosures

Note: In this table, similarity measures are calculated using a sample that has been purged of disclosure sentences using a disclosure classifier. *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.746***	0.558***	0.554***	0.139**	−0.114**	−0.105*
	(0.075)	(0.064)	(0.068)	(0.056)	(0.054)	(0.054)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE	X	X		X	X	
Analyst FE × Firm FE			X			X
Observations	110,153	110,153	110,153	110,153	110,153	110,153
Adjusted R^2	0.453	0.504	0.551	0.423	0.537	0.588
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.2: Robustness - Excluding RegFD

Note: Regulation Fair Disclosure was ratified by the SEC in October of 2000, in this table, I redo the analysis from earlier while limiting the pre-period to the period after regulation fair disclosure. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:				
	Similarity, Self	Positivity	Negativity	Sentiment	Weaseling
	(1)	(2)	(3)	(4)	(5)
IB × Post	0.909***	0.036	0.671***	−0.433***	−0.380***
	(0.080)	(0.079)	(0.148)	(0.133)	(0.097)
Other Controls	X	X	X	X	X
Year FE	X	X	X	X	X
Analyst FE × Firm FE	X	X	X	X	X
Observations	97,605	97,605	97,605	97,605	97,605
Adjusted R^2	0.717	0.371	0.362	0.360	0.494
Note:					*p<0.1; **p<0.05; ***p<0.01

Table A.3: Sanctioned in GRS

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. IB is a dummy set to one if the firm that employs the analyst is an investment bank. Sanctioned is a dummy set to one if the firm where the analyst works was sanctioned in GRS. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. All covariates are presented in table A.4. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Similarity, Self		Similarity		Similarity, Other	
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.809***	0.816***	0.603***	0.551***	0.488***	0.454***
	(0.073)	(0.076)	(0.070)	(0.073)	(0.074)	(0.079)
Sanctioned × Post	−0.169***	−0.187***	0.195***	0.133**	0.274***	0.230***
Other Controls	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Analyst FE	X		X		X	
Analyst FE × Firm FE		X		X		X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.657	0.709	0.590	0.684	0.623	0.695
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.4: Similarity with Prior Reports - All Covariates

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. **Earnings Announcement** is a dummy which is set to one if there is an earnings announcement the day of the report. **Earnings Forecast** is a similar dummy which is set to one if there is at least one new earnings forecast recorded in IBES the day of the report. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **Surprise** is the standardized unexpected earnings measure. **# IBES Reports** are the number of reports about the firm during the day. **Word Count** and **Digits** are the counts of words and digits in the textual report, respectively. Similarity measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:				
	Similarity				
	(1)	(2)	(3)	(4)	(5)
IB × Post	0.363***	0.386***	0.388***	0.515***	0.500***
	(0.067)	(0.046)	(0.046)	(0.052)	(0.060)
Investment Bank	0.074	0.055	0.053		
	(0.054)	(0.034)	(0.033)		
Post	0.136**	0.153***			
	(0.069)	(0.050)			
Earnings Announcement	0.200***	0.145***	0.145***	0.107***	0.108***
	(0.017)	(0.014)	(0.014)	(0.010)	(0.011)
Surprise	−0.002	−0.001	−0.001	−0.0004	0.0001
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Earnings Forecast	0.042**	0.027**	0.027**	0.033***	0.041***
	(0.020)	(0.011)	(0.011)	(0.008)	(0.008)
Upgrade	−0.091***	−0.078***	−0.078***	−0.060***	−0.058***
	(0.016)	(0.013)	(0.013)	(0.011)	(0.011)
Downgrade	−0.059***	−0.037***	−0.037***	−0.025***	−0.018*
	(0.017)	(0.011)	(0.011)	(0.009)	(0.010)
Log(Digits)	−0.040**	−0.042***	−0.043***	0.041***	0.041***
	(0.016)	(0.013)	(0.013)	(0.013)	(0.013)
Log(Word Count)	0.731***	0.721***	0.722***	0.602***	0.607***
	(0.026)	(0.020)	(0.020)	(0.020)	(0.021)
# IBES Reports	−0.003***	−0.002***	−0.002***	−0.001**	−0.001
	(0.001)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Firm FE		X	X	X	
Year FE			X	X	X
Analyst FE				X	
Analyst FE × Firm FE					X
Observations	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.398	0.543	0.543	0.648	0.684
Note:					*p<0.1; **p<0.05; ***p<0.01

Table A.5: Similarity with Prior Reports - Self vs Other

Note: **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. **Earnings Announcement** is a dummy which is set to one if there is an earnings announcement the day of the report. **Earnings Forecast** is a similar dummy which is set to one if there is at least one new earnings forecast recorded in IBES the day of the report. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **Surprise** is the standardized unexpected earnings measure. **# IBES Reports** are the number of reports about the firm during the day. **Word Count** and **Digits** are the counts of words and digits in the textual report, respectively. Similarity measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011). Panel (b) shows results with **Self-Similarity, Other Firms** as a control for disclosure language.

[Similarity with others vs similarity with self]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	1.008***	0.874***	0.886***	0.664***	0.382***	0.367***
	(0.071)	(0.066)	(0.070)	(0.072)	(0.061)	(0.065)
Earnings Announcement		0.077***	0.062***		0.109***	0.093***
		(0.010)	(0.010)		(0.011)	(0.011)
Surprise		−0.0004	0.001		−0.002	0.0002
		(0.001)	(0.001)		(0.002)	(0.001)
Earnings Forecast		0.038***	0.026***		0.059***	0.051***
		(0.009)	(0.008)		(0.011)	(0.008)
Upgrade		−0.066***	−0.068***		−0.054***	−0.050***
		(0.010)	(0.009)		(0.012)	(0.011)
Downgrade		−0.039***	−0.036***		−0.035***	−0.018*
		(0.009)	(0.009)		(0.011)	(0.010)
Log(Digits)		0.124***	0.107***		0.038***	0.028**
		(0.014)	(0.013)		(0.014)	(0.014)
Log(Word Count)		0.105***	0.105***		0.716***	0.733***
		(0.027)	(0.024)		(0.023)	(0.022)
# IBES Reports		−0.001**	−0.001*		−0.0002	−0.0004
		(0.0004)	(0.0003)		(0.0004)	(0.0004)
Year FE	X	X	X	X	X	X
Analyst FE	X	X		X	X	
Analyst FE × Firm FE			X			X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.635	0.656	0.709	0.491	0.622	0.695

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

[Controlling for time varying analyst-level similarity]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.355***	0.298***	0.279***	0.155	−0.010	0.003
	(0.088)	(0.079)	(0.082)	(0.120)	(0.112)	(0.116)
Self-Similarity, Other Firms	0.584***	0.560***	0.582***	0.384***	0.298***	0.347***
	(0.019)	(0.017)	(0.018)	(0.036)	(0.029)	(0.027)
Earnings Announcement		0.035***	0.019		0.085***	0.060***
		(0.013)	(0.012)		(0.015)	(0.017)
Surprise		−0.004	−0.002		0.001	0.002
		(0.003)	(0.002)		(0.003)	(0.003)
Earnings Forecast		0.024**	0.004		0.048***	0.039***
		(0.012)	(0.009)		(0.018)	(0.011)
Upgrade		−0.017	−0.031**		−0.031*	−0.016
		(0.016)	(0.015)		(0.018)	(0.017)
Downgrade		−0.038**	−0.031*		−0.037**	−0.020
		(0.017)	(0.018)		(0.019)	(0.018)
Log(Digits)		0.092***	0.071***		0.059***	0.033
		(0.015)	(0.014)		(0.020)	(0.021)
Log(Word Count)		0.082***	0.094***		0.684***	0.716***
		(0.025)	(0.021)		(0.037)	(0.034)
# IBES Reports		−0.001	−0.0003		−0.001	−0.001
		(0.001)	(0.0005)		(0.001)	(0.001)
Year FE	X	X	X	X	X	X
Analyst FE	X	X		X	X	
Analyst FE × Firm FE			X			X
Observations	30,654	30,654	30,654	30,654	30,654	30,654
Adjusted R^2	0.724	0.735	0.790	0.527	0.656	0.734

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

Table A.6: Similarity with Prior Reports - Weasel and Sentiment Controls

Note: *Similarity* is the average cosine similarity with reports issued about the firm in the prior three months. **Similarity, Self** is the similarity with prior reports that were written by the same analyst and **Similarity, Other** is the similarity with prior reports written by other analysts. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

[Similarity with others vs similarity with self]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.971***	0.839***	0.855***	0.681***	0.385***	0.365***
	(0.070)	(0.066)	(0.070)	(0.070)	(0.061)	(0.066)
Weaseling	−0.074***	−0.064***	−0.057***	−0.007	−0.021**	−0.019**
	(0.006)	(0.007)	(0.007)	(0.009)	(0.008)	(0.008)
Sentiment	−0.023***	−0.033***	−0.038***	0.045***	0.017**	0.006
	(0.007)	(0.007)	(0.008)	(0.009)	(0.007)	(0.008)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE		X			X	
Analyst FE × Firm FE			X			X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.639	0.660	0.711	0.493	0.622	0.695

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

[Disclosure Control]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.364***	0.304***	0.285***	0.173	0.003	0.012
	(0.087)	(0.078)	(0.080)	(0.113)	(0.107)	(0.113)
Weaseling	−0.029***	−0.030***	−0.028***	0.029*	−0.004	0.003
	(0.007)	(0.008)	(0.008)	(0.015)	(0.013)	(0.010)
Sentiment	0.026***	0.018**	0.024***	0.073***	0.041***	0.039***
	(0.008)	(0.008)	(0.007)	(0.011)	(0.009)	(0.010)
Self-Similarity, Other Firms	0.583***	0.558***	0.583***	0.399***	0.304***	0.355***
	(0.019)	(0.017)	(0.018)	(0.036)	(0.029)	(0.027)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE	X	X		X	X	
Analyst FE × Firm FE			X			X
Observations	30,654	30,654	30,654	30,654	30,654	30,654
Adjusted R^2	0.725	0.736	0.791	0.532	0.658	0.735
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.7: Similarity with Prior Reports - Self vs Other - JSD

Note: *JSD* is the average square root Jensen-Shannon Divergence with reports issued about the firm in the prior three months. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

[Similarity with others vs similarity with self]

	Dependent variable:					
	JSD, Self			JSD, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	−1.094***	−0.961***	−0.971***	−0.734***	−0.405***	−0.368***
	(0.073)	(0.069)	(0.078)	(0.062)	(0.048)	(0.052)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE		X			X	
Analyst FE × Firm FE			X			X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.682	0.702	0.757	0.601	0.763	0.811
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.8: Similarity with Prior Reports - Different Windows

Note: This table shows results using a **Similarity** measure calculated using either a shorter or a longer window. *Similarity* is the average cosine similarity with reports issued about the firm in the prior month (panel a) or prior six months (panel b). Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Similarity measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

[One Month Similarity]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	0.960***	0.863***	0.874***	0.573***	0.350***	0.333***
	(0.078)	(0.073)	(0.075)	(0.093)	(0.079)	(0.093)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE		X			X	
Analyst FE × Firm FE			X			X
Observations	67,542	67,542	67,542	67,542	67,542	67,542
Adjusted R^2	0.573	0.586	0.631	0.435	0.545	0.613
Note:						*p<0.1; **p<0.05; ***p<0.01

[Six Month Similarity]

	Dependent variable:					
	Similarity, Self			Similarity, Other		
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	1.031***	0.891***	0.896***	0.655***	0.367***	0.345***
	(0.070)	(0.065)	(0.069)	(0.070)	(0.059)	(0.065)
Other Controls		X	X		X	X
Year FE	X	X	X	X	X	X
Analyst FE		X			X	
Analyst FE × Firm FE			X			X
Observations	119,823	119,823	119,823	119,823	119,823	119,823
Adjusted R^2	0.665	0.688	0.744	0.508	0.642	0.717
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.9: Change in Sentiment

Note: **Sentiment** is measured as the number of positive words minus the number of negative words scaled by the length of the document. **Positivity** and **negativity** are defined analogously using either positive or negative words only. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Sentiment measures are centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Sentiment		Positivity		Negativity	
	(1)	(2)	(3)	(4)	(5)	(6)
IB × Post	−0.467***	−0.434***	−0.011	0.019	0.668***	0.652***
	(0.087)	(0.098)	(0.064)	(0.078)	(0.100)	(0.115)
Earnings Announcement	0.028	0.026	0.046***	0.048***	0.009	0.015
	(0.019)	(0.018)	(0.017)	(0.018)	(0.018)	(0.017)
Surprise	0.017***	0.017***	0.007**	0.007**	−0.017***	−0.016***
	(0.006)	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)
Earnings Forecast	−0.025*	−0.015	0.070***	0.079***	0.113***	0.108***
	(0.013)	(0.013)	(0.012)	(0.012)	(0.014)	(0.014)
Upgrade	0.160***	0.161***	0.187***	0.184***	−0.028*	−0.033*
	(0.018)	(0.018)	(0.015)	(0.016)	(0.017)	(0.017)
Downgrade	−0.309***	−0.283***	−0.135***	−0.119***	0.303***	0.283***
	(0.021)	(0.021)	(0.015)	(0.015)	(0.021)	(0.021)
Log(Digits)	−0.017	−0.014	−0.061***	−0.053**	−0.042***	−0.037**
	(0.015)	(0.016)	(0.022)	(0.024)	(0.015)	(0.015)
Log(Word Count)	0.138***	0.134***	0.406***	0.402***	0.242***	0.244***
	(0.034)	(0.036)	(0.042)	(0.045)	(0.026)	(0.027)
# IBES Reports	0.0002	0.0003	0.003***	0.003***	0.003***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Firm FE	X		X		X	
Year FE	X	X	X	X	X	X
Analyst FE	X		X		X	
Analyst FE × Firm FE		X		X		X
Observations	114,105	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.291	0.358	0.331	0.375	0.299	0.355

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

Table A.10: Weaseling - All Covariates

Note: **Weaseling** is measured as the fraction of sentences in the document classified as weasels. **Earnings Announcement** is a dummy which is set to one if there is an earnings announcement the day of the report. **Earnings Forecast** is a similar dummy which is set to one if there is at least one new earnings forecast recorded in IBES the day of the report. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **Surprise** is the standardized unexpected earnings measure. **# IBES Reports** are the number of reports about the firm during the day. **Word Count** and **Digits** are the counts of words and digits in the textual report, respectively. The weasel measure is centered at zero and normalized to have a standard deviation of 1. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:				
	Weasel Sentences				
	(1)	(2)	(3)	(4)	(5)
IB × Post	−0.030	−0.296***	−0.299***	−0.320***	−0.266***
	(0.081)	(0.084)	(0.084)	(0.080)	(0.091)
Investment Bank	0.067				
	(0.052)				
Post	0.073	0.425***			
	(0.073)	(0.056)			
Earnings Announcement	−0.092***	−0.095***	−0.094***	−0.101***	−0.111***
	(0.020)	(0.012)	(0.012)	(0.011)	(0.012)
Surprise	−0.0002	−0.002	−0.001	−0.001	−0.001
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Earnings Forecast	−0.077***	−0.053***	−0.052***	−0.063***	−0.064***
	(0.015)	(0.009)	(0.009)	(0.008)	(0.008)
Upgrade	0.055***	0.048***	0.050***	0.039***	0.039***
	(0.016)	(0.012)	(0.012)	(0.011)	(0.011)
Downgrade	0.119***	0.093***	0.095***	0.085***	0.072***
	(0.017)	(0.011)	(0.011)	(0.011)	(0.011)
Log(Digits)	−0.264***	−0.290***	−0.290***	−0.288***	−0.288***
	(0.021)	(0.016)	(0.016)	(0.015)	(0.016)
Log(Word Count)	0.461***	0.465***	0.464***	0.461***	0.464***
	(0.041)	(0.027)	(0.027)	(0.027)	(0.028)
# IBES Reports	−0.003***	−0.003***	−0.004***	−0.004***	−0.003***
	(0.001)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Constant	−1.536***				
	(0.198)				
Other Controls	X	X	X	X	X
Firm FE				X	
Year FE			X	X	X
Analyst FE		X	X	X	
Analyst FE × Firm FE					X
Observations	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.064	0.412	0.413	0.425	0.458
Note:					*p<0.1; **p<0.05; ***p<0.01

Table A.11: Robustness to Weasel Classification

Note: **Weaseling** is measured as the fraction of words in the document classified as weasels using a list of weasel words. The measure is standardized to have 0 mean and 1 standard deviation. Other controls are **Earnings Announcement, Earnings Forecast, Upgrade, Downgrade, SUE, Word Count**, and **Digit Count**. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:				
	Weasel Sentences				
	(1)	(2)	(3)	(4)	(5)
IB × Post	0.151	−0.331***	−0.334***	−0.332***	−0.278***
	(0.102)	(0.085)	(0.084)	(0.080)	(0.092)
Investment Bank	0.032				
	(0.090)				
Other Controls	X	X	X	X	X
Firm FE				X	
Year FE			X	X	X
Analyst FE		X	X	X	
Analyst FE × Firm FE					X
Observations	114,105	114,105	114,105	114,105	114,105
Adjusted R^2	0.084	0.328	0.329	0.349	0.386
Note:					*p<0.1; **p<0.05; ***p<0.01

Table A.12: Stock Market Reactions - All Covariates

Note: This table presents evidence related to the stock market response on the day of analyst reports. **Similarity, Self** is the similarity with prior reports that were written by the same analyst, the measure is standardized to have a deviation of 1 and mean of 0. **Upgrade** and **Downgrade** are the fraction of reports issued during the same day that contained either an upgrade or a downgrade. **CAR** refers to the 3-day abnormal returns around the report using the one factor model. **Abs(CAR)** is the absolute value. Abnormal returns are winsorized at 1% and 99%. Other controls are **Earnings Announcement, Earnings Forecast, SUE, Boldness**, **Forecast**, **Word Count**, and **Digit Count**. Errors are triple clustered on firm, analyst, and date (see Cameron, Gelbach and Miller, 2011).

	Dependent variable:					
	Abs(CAR)			CAR		
	(1)	(2)	(3)	(4)	(5)	(6)
Similarity, Self	−5.865**	−3.467*	5.746***	0.279	1.452	1.143
	(2.591)	(2.034)	(1.729)	(1.914)	(2.161)	(2.145)
Upgrade	166.663***	145.656***	143.346***	290.040***	291.580***	290.576***
	(8.804)	(8.206)	(8.127)	(12.951)	(12.770)	(12.780)
Downgrade	253.163***	230.310***	226.188***	−406.978***	−403.450***	−403.893***
	(10.375)	(9.689)	(9.661)	(17.384)	(16.122)	(16.076)
Similarity, Self × Upgrade	−20.276***	−18.282***	−18.264***	−14.829	−12.177	−11.966
	(5.720)	(5.616)	(5.457)	(9.086)	(9.018)	(9.000)
Similarity, Self × Downgrade	−42.285***	−35.381***	−31.749***	49.403***	48.681***	48.840***
	(7.160)	(6.694)	(6.653)	(10.925)	(10.412)	(10.401)
Earnings Announcement	41.535***	44.277***	44.990***	7.826	12.758	12.891
	(7.276)	(6.245)	(6.001)	(13.425)	(13.128)	(13.129)
Surprise	−0.895	−0.450	−0.039	30.243***	30.396***	30.261***
	(1.745)	(1.546)	(1.513)	(6.768)	(6.676)	(6.680)
Earnings Forecast	2.155	16.848***	20.351***	−6.304	−7.658	−8.074
	(4.367)	(3.851)	(3.775)	(5.145)	(5.073)	(5.092)
Log(Digits)	10.008***	4.135**	−8.168***	1.210	−0.212	1.769
	(2.448)	(1.924)	(1.548)	(1.779)	(2.023)	(2.002)
Log(Word Count)	−47.430***	−28.039***	6.576**	7.453**	9.163**	5.754
	(4.374)	(3.901)	(3.033)	(3.217)	(3.853)	(3.842)
# IBES Reports	2.146***	2.086***	2.289***	−0.595**	−0.780***	−0.765***
	(0.211)	(0.194)	(0.192)	(0.262)	(0.264)	(0.266)
Year FE			X			X
Analyst FE		X	X		X	X
Observations	382,943	382,943	382,943	382,943	382,943	382,943
Adjusted R^2	0.071	0.146	0.172	0.072	0.080	0.081
Note:						*p<0.1; **p<0.05; ***p<0.01

Appendix B

Appendix to Chapter 2

B.1 Additional Detail on LDA

B.2 Additional Tables and Full Results

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

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

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

As a robustness check, we also recreate the measure based on the underlying skewed distributions (using both raw loadings, and IHS transformed loadings) and use these alternative measures in our main specifications. Table B.13 presents the main results using these alternative measures.

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

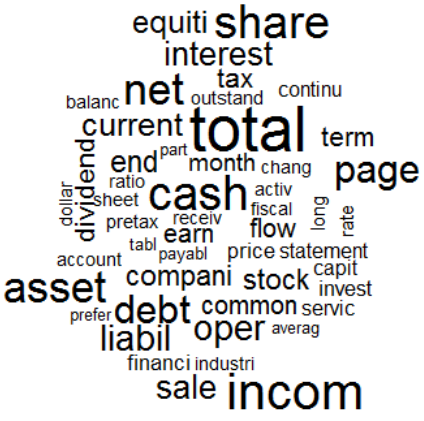
Note: This table presents the t-statistics and adjusted R-squared on the linear relationship between a firm's patent applications and the loadings of each of the 15 topics from the fitted LDA model. Topic 6 is the innovation topic that we use for our text-based measure of innovation. This topic explains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

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

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

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

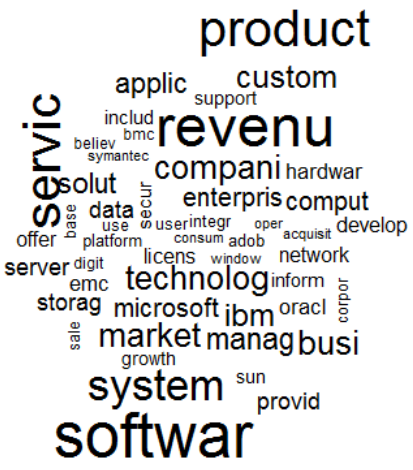
[Underperforming Benchmark Topic ($t = -8.678$)]



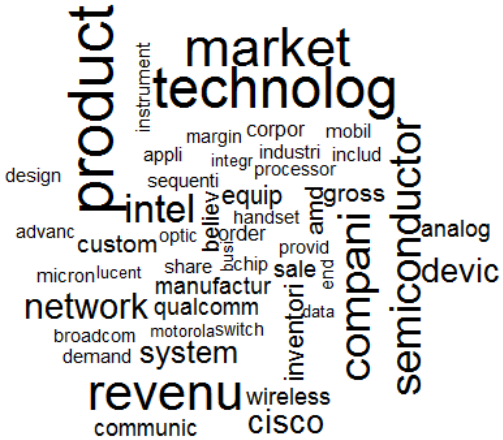
[Operating Performance Topic ($t = 8.697$)]

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

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



[First Innovation Topic]



[Second Innovation Topic]

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

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

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

Table B.2: Variable Definitions

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

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

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

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. Sales growth is defined as the percentage growth in sales between year t and year t+1 (in decimal form). The text-based innovation measure is converted to a Z-score for ease of interpretability. Patents is the count of granted patents which were applied for during the year. Citations are the forward citations of the patents 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.

[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^{***}	0.083^{***}	0.049^{***}	0.015^{***}	0.010^{**}
	(0.002)	(0.002)	(0.010)	(0.008)	(0.006)	(0.005)
$Log(Patents)_t$	0.002	-0.002	0.003	-0.027	-0.007	-0.015^{**}
	(0.003)	(0.003)	(0.015)	(0.022)	(0.005)	(0.007)
$Log(Citations)_t$	0.001	-0.0004	0.016^{*}	0.020^{*}	-0.003	0.002
	(0.002)	(0.001)	(0.009)	(0.010)	(0.003)	(0.004)
$R\&D/Assets\ (Z)_t$	0.006	0.010^{**}	0.074^{***}	0.027	-0.001	-0.007
	(0.005)	(0.004)	(0.021)	(0.024)	(0.005)	(0.009)
$Log(Assets)_t$	-0.001	-0.027^{***}	-0.030^{*}	-0.208^{***}	0.002	-0.071^{***}
	(0.003)	(0.005)	(0.016)	(0.023)	(0.005)	(0.018)
$Asset\ Tangibility_t$	0.103^{***}	0.055^{**}	0.171^{*}	-0.048	-0.060^{*}	-0.308^{***}
	(0.017)	(0.022)	(0.091)	(0.109)	(0.036)	(0.106)
$Leverage_t$	-0.008	-0.008	-0.126	-0.127^{*}	-0.086^{***}	-0.055
	(0.021)	(0.020)	(0.084)	(0.070)	(0.029)	(0.041)
$Log(Age)_t$	0.002	-0.004	-0.079^{**}	-0.149	-0.025^{**}	-0.022
	(0.007)	(0.018)	(0.032)	(0.114)	(0.010)	(0.050)
$Cash/Assets_t$	0.102^{***}	0.038	0.983^{***}	0.389^{***}	0.057	0.014
	(0.030)	(0.030)	(0.123)	(0.091)	(0.050)	(0.056)
Patenting Firm	0.009^{*}		0.037		0.00000	
	(0.006)		(0.031)		(0.009)	
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	0.436	0.674	0.577	0.771	0.099	0.159

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

Table B.4: Full Results on Performance of Innovative Firms (1990-2010)

[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^{**}	0.084^{***}	0.050^{***}	0.015^{***}	0.008
	(0.002)	(0.002)	(0.011)	(0.009)	(0.005)	(0.005)
× Non-Patenting Firm	0.010^{***}	0.006^{*}	0.078^{***}	0.046^{***}	0.017^{**}	0.020^{**}
	(0.004)	(0.003)	(0.016)	(0.014)	(0.008)	(0.009)
$Log(Patents)_t$	0.002	−0.002	0.003	−0.026	−0.007	-0.015^{**}
	(0.003)	(0.003)	(0.015)	(0.022)	(0.005)	(0.007)
$Log(Citations)_t$	0.001	−0.0004	0.016^{*}	0.020^{*}	−0.003	0.002
	(0.002)	(0.001)	(0.009)	(0.010)	(0.003)	(0.004)
$R\&D/Assets\ (Z)_t$	0.006	0.010^{**}	0.074^{***}	0.027	−0.001	−0.007
	(0.005)	(0.004)	(0.021)	(0.024)	(0.005)	(0.008)
$Log(Assets)_t$	−0.001	-0.027^{***}	-0.030^{*}	-0.208^{***}	0.002	-0.071^{***}
	(0.003)	(0.005)	(0.016)	(0.023)	(0.005)	(0.018)
Asset $Tangibility_t$	0.103^{***}	0.055^{**}	0.171^{*}	−0.048	-0.060^{*}	-0.308^{***}
	(0.017)	(0.022)	(0.091)	(0.109)	(0.036)	(0.106)
$Leverage_t$	−0.008	−0.008	−0.126	-0.127^{*}	-0.086^{***}	−0.056
	(0.021)	(0.020)	(0.084)	(0.070)	(0.029)	(0.041)
$Log(Age)_t$	0.002	−0.004	-0.079^{**}	−0.149	-0.025^{**}	−0.024
	(0.007)	(0.018)	(0.032)	(0.114)	(0.010)	(0.050)
$Cash/Assets_t$	0.103^{***}	0.038	0.983^{***}	0.389^{***}	0.057	0.014
	(0.030)	(0.030)	(0.123)	(0.091)	(0.050)	(0.056)
Patenting Firm	0.009		0.038		−0.0003	
	(0.006)		(0.031)		(0.009)	
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	0.436	0.674	0.577	0.771	0.099	0.159

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

Table B.5: Full Results on Performance of Innovative Firms - Rolling Measure (1990-2010)

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. Sales growth is defined as the percentage growth in sales between year t and year t+1 (in decimal form). The text-based innovation measure is converted to a Z-score for ease of interpretability. Patents is the count of granted patents which were applied for during the year. Citations are the forward citations of the patents 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.

[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^{***}	0.068^{***}	0.041^{***}	0.007	0.006
	(0.002)	(0.002)	(0.010)	(0.007)	(0.006)	(0.005)
$Log(Patents)_t$	0.003	-0.003	-0.010	-0.062^{***}	-0.005	-0.010
	(0.004)	(0.003)	(0.016)	(0.019)	(0.006)	(0.008)
$Log(Citations)_t$	0.0002	-0.001	0.027^{***}	0.031^{***}	-0.005	0.001
	(0.002)	(0.002)	(0.009)	(0.009)	(0.004)	(0.004)
$R\&D/Assets\ (Z)_t$	0.006	0.011^{**}	0.084^{***}	0.030	0.001	-0.008
	(0.006)	(0.005)	(0.024)	(0.025)	(0.006)	(0.010)
$Log(Assets)_t$	0.0003	-0.031^{***}	-0.024	-0.246^{***}	0.002	-0.109^{***}
	(0.003)	(0.005)	(0.019)	(0.028)	(0.006)	(0.022)
Asset $Tangibility_t$	0.101^{***}	0.041^{*}	0.188^{*}	-0.126	-0.056	-0.378^{***}
	(0.020)	(0.024)	(0.106)	(0.109)	(0.040)	(0.134)
$Leverage_t$	-0.0004	0.0002	-0.145	-0.156^{**}	-0.076^{**}	-0.047
	(0.025)	(0.023)	(0.100)	(0.073)	(0.037)	(0.052)
$Log(Age)_t$	0.001	0.011	-0.079^{**}	-0.176	-0.029^{***}	-0.030
	(0.007)	(0.021)	(0.035)	(0.123)	(0.011)	(0.069)
$Cash/Assets_t$	0.106^{***}	0.040	0.967^{***}	0.341^{***}	0.088	0.018
	(0.033)	(0.029)	(0.137)	(0.079)	(0.054)	(0.067)
Patenting Firm	0.012^{*}		0.040		0.006	
	(0.006)		(0.034)		(0.012)	
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.427	0.680	0.582	0.798	0.102	0.164
Note:						*p<0.1; **p<0.05; ***p<0.01

Table B.6: Full Results on Performance of Innovative Firms - Rolling Measure (1990-2010)

[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^{**}	0.067^{***}	0.042^{***}	0.006	0.003
	(0.002)	(0.002)	(0.012)	(0.008)	(0.005)	(0.005)
× Non-Patenting Firm	0.012^{***}	0.008^{***}	0.069^{***}	0.036^{***}	0.011	0.015
	(0.004)	(0.003)	(0.014)	(0.011)	(0.014)	(0.016)
$Log(Patents)_t$	0.003	−0.003	−0.010	-0.062^{***}	−0.005	−0.010
	(0.004)	(0.003)	(0.016)	(0.019)	(0.006)	(0.008)
$Log(Citations)_t$	0.0003	−0.001	0.027^{***}	0.031^{***}	−0.005	0.001
	(0.002)	(0.002)	(0.009)	(0.009)	(0.004)	(0.004)
$R\&D/Assets\ (Z)_t$	0.006	0.011^{**}	0.084^{***}	0.029	0.001	−0.008
	(0.006)	(0.005)	(0.024)	(0.025)	(0.006)	(0.010)
$Log(Assets)_t$	0.0003	-0.031^{***}	−0.024	-0.246^{***}	0.002	-0.109^{***}
	(0.003)	(0.005)	(0.019)	(0.028)	(0.006)	(0.022)
$Asset\ Tangibility_t$	0.101^{***}	0.041^{*}	0.188^{*}	−0.126	−0.056	-0.378^{***}
	(0.020)	(0.024)	(0.106)	(0.109)	(0.040)	(0.134)
$Leverage_t$	−0.0005	0.0001	−0.145	-0.156^{**}	-0.076^{**}	−0.048
	(0.025)	(0.023)	(0.100)	(0.073)	(0.037)	(0.052)
$Log(Age)_t$	0.001	0.011	-0.079^{**}	−0.175	-0.029^{***}	−0.032
	(0.007)	(0.021)	(0.035)	(0.123)	(0.011)	(0.069)
$Cash/Assets_t$	0.106^{***}	0.040	0.967^{***}	0.341^{***}	0.088	0.018
	(0.033)	(0.029)	(0.136)	(0.079)	(0.054)	(0.068)
Patenting Firm	0.011^{*}		0.040		0.004	
	(0.006)		(0.036)		(0.012)	
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	4,898	4,898	4,793	4,793	4,902	4,902
Adjusted R^2	0.427	0.680	0.582	0.798	0.102	0.164

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

Table B.7: Text-Based Innovation and R&D Expenses (1990-2010)

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

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

Table B.8: Full Results on Predicting Acquisition Activity (1990-2010)

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

[Acquisition Count]

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

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

Table B.9: Full Results on Predicting Acquisition Activity (1990-2010)

[Acquisition Count - Purged Corpus]

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

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

Table B.10: Predicting Acquisition Activity - LPM (1990-2010)

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

[Linear Probability Model – Main Innovation Measure]

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

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

Table B.11: Predicting Acquisition Activity - LPM (1990-2010)

[Linear Probability Model- Corpus Purged of Merger and Acquisition Words]

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

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

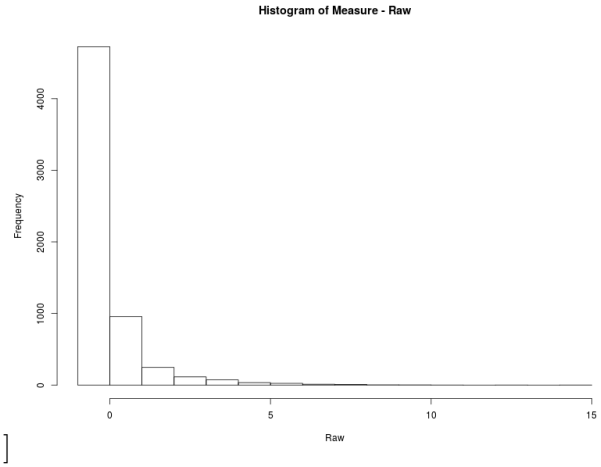
Table B.12: Relation of Text-Based Innovation to Merger Announcement CARs (-1 to +1 day)

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

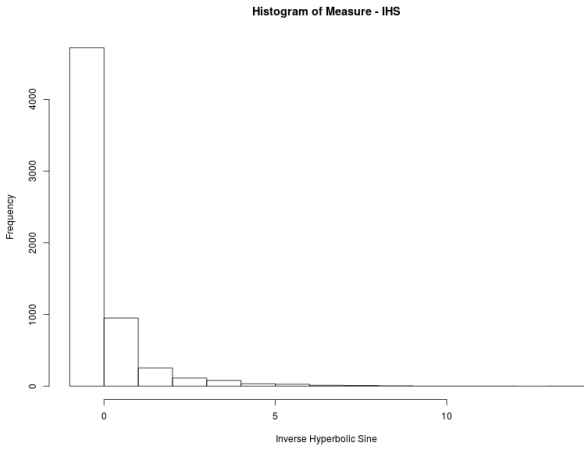
	Dependent variable:			
	CAR			
	(1)	(2)	(3)	(4)
Text-Innovation $(Z)_t$	-0.001	-0.010^{*}	-0.010^{**}	-0.012^{*}
	(0.001)	(0.006)	(0.005)	(0.006)
Text-Innovation $(Z)_t \times$ SmallAcq		0.009^{*}	0.010^{**}	0.012^{**}
		(0.005)	(0.005)	(0.006)
SmallAcq		-0.005	-0.005	-0.005
		(0.005)	(0.005)	(0.005)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

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

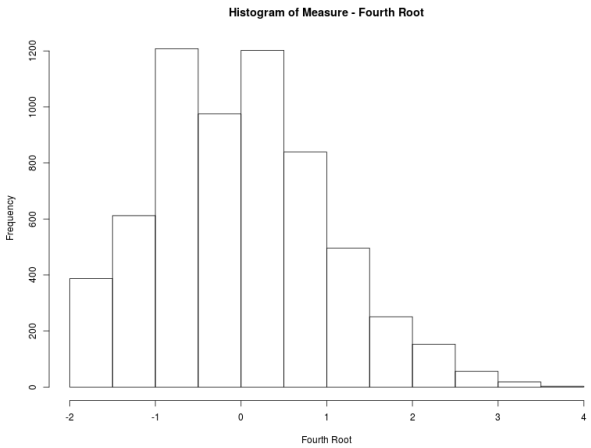
Note: This figure presents sample histograms of the topic loadings used as a basis for the text-based innovation measure. In panel (a), the measure is the raw mean of the innovation topic loading for positive analyst reports about the firm over the fiscal year. Panel (b) uses the inverse hyperbolic sine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



[Raw Topic Loadings]



[Inverse Hyperbolic Sine of Topic Loadings]



[Fourth Root of Topic Loadings]

The sign and significance of the main results are broadly consistent with our main measure, but the estimates tend to be more precise and stable using our primary measure, which confirms the rationale for using the less skewed measure in the first place.

B.4 Long-Term Dynamics

B.5 Word-List Measure versus Latent Dirchlet Allocation

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

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

Table B.13: Results Using Alternative Scaling of the Topic Loading to Construct the Text-Based Innovation Measure

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

[Main Results Using Raw Topic Loadings]

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

[Main Results Using Inverse Hyperbolic Sine of Topic Loadings]

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

Table B.14: Long-Term Tobin's Q, ROA, and Salesgrowth Using the Text-Based Innovation Measure

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

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

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

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

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

We take the resulting word list and measure its similarity with each of our analyst reports by counting how many innovation words each document contains and scaling it by the document length.[3] For consistency with our main LDA-based measure, we aggregate the word-list measure

[1] 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

[2] 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'.

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

across analyst reports written about the same firm in the same fiscal year for positive reports only (sentiment above the 75th percentile). Tables B.15 and B.16 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:,

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

and

$$Patenting_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \epsilon_{it} \tag{B.2}$$

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 B.15 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 B.16 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 B.15: Patent Value, Word-List Measure (1990-2010)

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

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

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

Table B.16: Patent Value, Word-List Measure (1990-2010)

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

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