Norwegian School of Economics
Bergen, Spring 2019

The Complexity of Readability

The Effect of Industry Complexity on Annual Report Readability

August Joachim Holst Strømnes and Eirik Fjelltveit Skagseth

Supervisor: Maximilian Rohrer

Master thesis, Economics and Business Administration

Major: Finance / Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

This thesis is written as a part of the Master of Science in Economics and Business Administration at the Norwegian School of Economics.

We would like to thank our supervisor, Maximilian Rohrer for his support, guidance, and feedback of our thesis. Moreover, we would like to thank him for providing us with the annual report text files which our analysis relies upon. We would also like to thank the Norwegian School of Economics for providing an extensive and interesting Master of Science program.

Norwegian School of Economics

Bergen, June 2019

Eirik Fjelltveit Skagseth

August Joachim Holst Strømnes

Abstract

We challenge the empirical relationship between annual report readability and subsequent stock return volatility, used to validate readability measures' ability to reflect the effective communication of valuation-relevant information. We establish that vocabulary most indicative of higher and lower readability scores are words specific to selected industries, and that both readability measures and stock return volatility hold strong time trends. When controlling for the unobserved time-varying heterogeneity across industries we find that both the magnitude and the statistical significance of the association between readability and volatility mitigates. Overall, the results support the notion that this association reflects the underlying complexity of the firm's business rather than the effective communication of valuation-relevant information.

Keywords – Readability, textual analysis, Mandatory disclosure

Contents

List of Figures

List of Tables

1 Introduction

The federal securities laws require all U.S. publicly traded firms to provide annual reports of the company's business and financial condition to the Securities and Exchange Commission (SEC) on Form 10-K. Thus, the 10-K is an essential source of information for investors when evaluating a specific firm. However, to fully benefit from the disclosed information, it is important that the investors actually understand the presented information. With the plain English mandate of 1998, the SEC ratified that companies must provide disclosures following plain English rules, emphasizing the importance of providing more readable disclosures that are easier to understand. As a tool in judging the level of compliance with the plain English mandate, the SEC has considered the use of traditional readability measures like the Fog Index (Cox, 2007).

In recent years a discussion in the literature has emerged on what readability measure best captures the comprehension of financial disclosures. Loughran and McDonald (2014), and Bonsall IV et al. (2017) both find problems with existing readability measures such as the Fog Index and present the alternatives, the file size of the 10-K filing and the Bog Index, respectively. To validate their new measure, Loughran and McDonald (2014) show that less readable 10-Ks are significantly associated with higher stock return volatility subsequent to the filing date. The assumption behind this test is that less readable disclosures will result in more ambiguity in validation in the period after the filing, as reflected by a higher stock return volatility. Then, if a readability measure has a positive association with stock return volatility, that measure is a valid proxy for readability.

Another notion from the literature is the difficulty of disentangling the role of firm-level complexity from readability. Loughran and McDonald (2016) point out that managers might produce less readable disclosures simply because they describe a more complex business operation or situation. Describing a complex reality is distinct from reporting a message in an easy versus a complex way. Thus, if not properly controlling for complexity, researches might draw an incorrect inference when assessing the impact of disclosure readability on capital market outcomes.

Following this notion, we establish that vocabulary most indicative of higher and lower

readability scores are words specific to selected industries. Second, we find that both the readability measures and stock price volatility show clear trends over the years. Third, we find that the magnitude and the statistical significance of the association between readability measures and subsequent volatility mitigates when we introduce controls for time-varying industry heterogeneity.

This thesis investigates commonly used readability measures' ability to reflect the effective communication of valuation-relevant information in 10-k filings. The readability measures we investigate include the Fog Index, the Bog Index, File Size, and Total Words. We conduct this analysis in two stages. First, we apply the framework introduced by Taddy (2013) to investigate what lexical features of the 10-K filings are associated with higher levels of the readability measures. If the readability measures primarily reflect the effectiveness of communication, we expect facets of readability to be reflected in the output(i.e., long words, complex words). In contrast to this, we find that vocabulary associated with higher or lower levels of the readability measures is related to specific industries. For example, 10-Ks containing sequences of two adjacent words such as *gene therapy*, *rare disease*, or *medicine product* are associated with higher levels of the Bog Index, indicating a less readable document. 10-Ks containing phrases such as *restaurant company* or *restaurant manager* are associated with lower levels of the Bog Index. Thus, vocabulary related to the pharmaceutical industry is associated with less readable 10-Ks, and vocabulary related to the restaurant industry is associated with more readable 10-Ks. Following the notion of the close link between complexity and readability in the literature, these findings indicate that variation in the readability measures is related to differences in the industry level of complexity.

Second, based on the previous findings, we investigate whether the association between annual report readability and stock price volatility is robust to time-varying industry effects. Both Loughran and McDonald (2014) and Bonsall IV et al. (2017) find that less readable annual reports are associated with higher subsequent stock price volatility (controlling for other variables, including the historical level of volatility). To account for unobserved heterogeneity between firms and industry, they additionally include Fama and French (1997) industry fixed effects and firm fixed effects, respectively.

Second, based on the previous findings, we hypothesize that the association between

annual report readability and stock price volatility is driven by time-varying industry effects. Both Loughran and McDonald (2014) and Bonsall IV et al. (2017) find that less readable annual reports are associated with higher subsequent stock price volatility (controlling for other variables, including the historical level of volatility). Additionally, they include industry fixed effects and firm fixed effects, respectively.

However, we consider that these methods do not sufficiently take into account the unobserved heterogeneity between firms, as they are limited to account for unobserved heterogeneity that is constant over time. Both volatility in the period after the filing date and the readability measures show clear time trends. Loughran and McDonald (2014) and Bonsall IV et al. (2017) include year fixed effects to account for the general trends in the economy. However, this approach does not consider that trends might differ between industries.

Hoberg and Phillips (2016) introduce two new sets of industry classifications based on the similarity between firms business descriptions within 10-Ks. They show that these industries better capture which firms that are exposed to high competition and are better at detecting firm rivals than other existing classifications. By applying these industry classifications to the regression models of Bonsall IV et al. (2017) and Loughran and McDonald (2014), we account for unobserved time-varying industry effects in two distinct ways. In the first model, we include industry-year dummy variables to account for the unobserved time-varying heterogeneity. In the second model, we include a control for the uncertainty among peer firms, as reflected by the mean post-filing volatility of firms in the same industry.

The results of both tests indicate that much of the association between annual report readability and subsequent stock volatility can be explained by time-varying industry characteristics. For example, the coefficient of the Bog Index holds a value of 0.032 and is significant on the 99% level when using the methods of Bonsall IV et al. (2017). However, when including additional controls for unobserved time-varying industry effects, the magnitude of association is more than halved, and the coefficient is no longer statistically significant. The only readability measure that remains significantly associated with subsequent stock volatility is Total Words, but the magnitude of this association is considerably reduced. The coefficient drops from 0.023 to 0.011 when including additional

controls. Following the assumption of Loughran and McDonald (2014), that more effective communication of valuation-relevant information leads to less ambiguity in validation in the period after the filing, our results imply that the current readability measures are not able to distinguish the level of effective communication between annual reports.

The results presented in this thesis suggest that unobserved factors, such as the industry level of complexity, are influencing the readability measures. Thus, we urge caution when interpreting the impact of disclosure readability on various capital market outcomes, as time-varying industry effects might lead to incorrect inference.

The remainder of this thesis is organized as follows: Section 2 presents the readability measures and reviews the literature. Section 3 describes the data collection and presents summary statistics. Sections 4 and 5 present methodologies used in this thesis and results. Section 6 concludes.

2 Background

In this section, we provide the definitions of four commonly used readability measures in financial and accounting research. Next, we present the Plain English Mandate and review the relevant literature.

2.1 Readability Measures

2.1.1 The Fog Index

The Fog Index was developed in 1952 by Robert Gunning (1952) and is a common readability measure in accounting and finance research (Bonsall IV et al., 2017). It consists of the sum of two components that are scaled by 0.4, and its index score is meant to reflect the years of education that is needed to comprehend a text at first reading. A Fog index of 6 means that a 6th grader should be able to understand its contents, and a score >18 indicates that a text is practically unreadable. The Fog Inde is defined as:

$$Fog\ Index = 0.4(words\ \ per\ \ sentence\ +\ percentage\ \ complex\ \ words)$$

where a *complex word* is a word with 3 or more syllables. The Fog Index proposes that all else equal, longer sentences and longer words make documents harder to read (Li, 2008).

2.1.2 Total Words and File Size

Loughran and McDonald (2014) argue that annual reports are less differentiable by the writing style, and readability should, therefore, reflect how hard it is for the reader to assimilate valuation-relevant information. The presumption is that as more text and longer documents require more parsing and filtering by the reader, longer documents seem more deterring and harder to read. Loughran and McDonald (2014) propose the file size (in megabytes) of the unaltered 10-K file uploaded to the SEC's EDGAR database as a

readability measure.

Under this same presumption, the total number of words in a 10-K document is a similar measure of readability. Several papers use the number of words as a measure of readability (see Li (2008), Miller (2010), Lawrence (2013)). The number of words is defined as the count of words left in the text after non-text attributes of the 10-K filing have been removed. This includes removing tables, numbers and markup language (XML, HTML).

2.1.3 The Bog Index

Bonsall IV et al. (2017) introduce the Bog Index as a readability measure of financial disclosures. The index is constructed to penalize documents based on violations of a broad set of plain English attributes as outlined in the *Plain English Handbook* (SEC, 1998). The Bog Index scores a text document based on the sum of three multifaceted components, where the first component, *Sentence Bog*, is a function of the average sentence length. *Word Bog* penalizes words and phrases that violate different plain English attributes, in addition to assigning penalties for each word based on its difficulty. The final component, *Pep*, rewards a document for having an interesting writing style that facilitates reading interpretation. It is defined as:

$$Bog\ Index = Sentence\ Bog + Word\ Bog - Pep,$$

where a higher Bog Index indicates a less readable document. The range of the Bog Index goes from zero to potentially over a thousand, but generally, a score above 70 is considered difficult to read (Editor Software, 2014). The Bog Index is developed by the company *Editor Software* and is part of the linguistics software *StyleWriter*.

2.2 The Plain English Mandate

In 1969 the U.S. Securities and Exchange Commission (SEC) published *The Wheat Report*, investigating the disclosure provisions of the current legislature. The report noted that the

average investor was unable to readily understand firm's prospectuses and recommended that firms should make their writing easier to comprehend (SEC, 1969).

As the legislation at the time did not warrant sufficiently readable disclosures, the SEC implemented the plain English mandate in 1998, further ratifying the importance of efficient communication between firm and investor. Plain English can be summarized as a way of writing that facilitates communication through the use of layout, content, words, and sentences. In an accompanying guide *A plain English Handbook*, the SEC lay out several benefits of more readable disclosures. These include that investors will make better-informed decisions when buying and selling investments, brokers and investment advisers can make better recommendations to clients when the relevant documents are quicker and easier to comprehend, and companies can save the cost of having to further explain their statements to investors (SEC, 1998). One way the SEC has considered in judging the level of compliance with plain English, is using readability measures (Cox, 2007).

2.3 Readability in Finance and Accounting Literature

Although the applications of readability measures have a long history in finance and accounting, as Loughran and McDonald (2016) note, much of the research prior to Li (2008) has suffered from small sample sizes and problematic methods. Jones and Shoemaker (1994) review 32 prior studies on readability in finance and accounting, of which 26 focus on annual report disclosures. They argue that these studies have little predictive validity, due to their lack of face validity and small sample sizes. Nine of the studies investigated by Jones and Shoemaker (1994) use less than 30 reports, and only two studies use samples larger than 100. Similarly small sample sizes extend to later papers such as Courtis (1998) and Clatworthy and Jones (2001) that investigate the association between readability profitability in 60 companies in Hong Kong and the U.K., respectively.

Li (2008) is considered the first to apply readability measures to financial disclosures with a sample of more than 50.000 firm-year observations. Applying the Fog Index and the number of words in a 10-K filing, Li finds that poor performing firms tend to have

annual reports that are harder to read. He also finds that firms with more readable annual reports tend to have more persistent positive earnings. Bloomfield (2008) interprets Li's results as that business in a complex business environment might require longer and more complex explanations to explain this reality in their annual report. Later, in support of this interpretation, Guay et al. (2016) find that some firms with harder to read financial statements will try to mitigate this negative impact on the information environment by offering voluntary disclosures.

Following Li (2008), several papers include the Fog Index as a readability measure. One line of research focuses on investor reactions to disclosure readability. For instance, Biddle et al. (2009) find that more readable disclosures are associated with less over- and underinvestment. On the effects of readability in regards to the investor size, Miller (2010) find that more complex 10-K filings are associated with reduced trading activity and lower consensus for smaller investors. Lawrence (2013) finds that the individual's shareholdings increase with clearer and more concise financial disclosures, and the effect mitigates with higher frequency trading and more financially literate investors. In addition, Lawrence (2013) documents that higher individual investor returns are associated with lower Fog Index and document length, such that clearer and more concise disclosures can mitigate the individual's information disadvantage. Lehavy et al. (2011) find that investors demand greater amounts of analyst services as the disclosures get longer and less readable. Lehavy et al. (2011) also find that analysts take longer time to provide coverage, more analysts provide coverage, and that their estimates are more dispersed in response to a less readable 10-K.

More recent studies include the 10-K file size and the Bog Index as readability measures. Investigating the relationship between 10-K readability and creditors, Ertugrul et al. (2017) find that annual report readability, as measured by file size, increases the perceived risk of information to creditors, leading to higher costs for external financing. Further, Bonsall and Miller (2017) find that less readable annual reports are linked to less favorable bond ratings, more frequent and pronounced bond rating agency disagreement, and higher costs of debt capital.

With the introduction of the file size of the 10-K filing by Loughran and McDonald (2014) and the Bog Index by Bonsall IV et al. (2017), a discussion of what measure is more

accurate in reflecting the comprehension of financial disclosures emerges. For instance, Loughran and McDonald (2014) show that a large portion of the words classified as complex by the Fog Index are well understood by investors and analysts. They argue that terms such as corporation, company, and management are all multisyllable words that appear very frequently in annual reports but do not make the annual reports harder to read. In addition, Loughran and McDonald (2014) find that the components of the Fog Index, *Average words per sentence* and *Percentage complex words* are negatively correlated and only *Average sentence length* is significant in explaining post-filing stock volatility. They note that it is unlikely that these components both measure readability, and that *Percentage complex words* contributes noise to the data.

Further, Loughran and McDonald (2014) emphasize that the parsing procedure needed to calculate other readability measures makes replication challenging. As researchers use varying parsing algorithms this could lead to differing readability scores of the same 10-K file. Loughran and McDonald (2014) highlight that disclosure file size is easy to determine and escapes sources of measurement errors caused by textual parsing or calculation and thus recommend its further use in research.

Furthering the discussion of how to best capture the comprehension of financial disclosures, Bonsall IV et al. (2017) discuss problems with using file size as a readability measure. Their main concern with the file size is that it contains elements that are unrelated to the 10-K filing text, such as compensation contracts, supplier/customer agreements, bond indentures in addition to pictures, PDF attachments, and HTML/ XML syntax. Bonsall IV et al. (2017) document that these elements have increased the past years, outnumbering the 10-K text itself in terms of megabytes. The authors argue that this may lead to over/under rejection when using the file size, depending on the influence of non-textual components.

Another important notion from the literature is the difficulty of disentangling annual report readability from firm complexity. Li (2008) notes that firms with more complex operations and financial situations are more likely to have more complicated annual reports. Loughran and McDonald (2014) further highlights this and points out that once the complexity of the firm is properly controlled for, the link between the readability measures and subsequent stock volatility, earnings surprise and analyst dispersion might

disappear. Bonsall IV et al. (2017) also acknowledge this ambiguity of the readability measures. The authors note on the possibility that the Bog Index is correlated with another unobservable variable that is correlated with the outcome variable examined (e.g., complexity), and that readability does not affect capital market outcomes such as stock volatility.

3 Data

In this section, we present our data collection and sample creation. We subsequently report descriptive statistics for the key variables in our sample.

3.1 Data Collection

We obtain 97 417 complete annual report text files from our thesis' supervisor. The annual reports are from U.S. firms and consists of document filings of type *10-K, 10-K405, 10-KT, 10KSB, 10KSB40*, and *10KT405*. Each filing has a CRSP permanent number (PERMNO) match. Amendments are not included. This text data had been parsed to remove HTML tags, headings, and tables, leaving the text, punctuation, and numbers. Following Loughran and McDonald (2014) and Bonsall IV et al. (2017), we look at the entire 10-K document to assess the readability of each document [1].

We collect share price, shares outstanding, SIC codes, book value of equity, book value of assets, and associated exchange for each firm from the CRSP and COMPUSTAT merged database. We obtain pre-calculated Bog Index for each annual report from Brian P. Miller's website [2], referenced in Bonsall IV et al. (2017). We collect market return and the risk-free rate from Kenneth French's website[3]. Lastly, text-based industry classifications are obtained from the Hoberg and Phillips data library[4], referenced in Hoberg and Phillips (2016).

3.2 Sample Creation

Of the initial sample, we apply a similar data screening procedure to that of Loughran and McDonald (2014). We remove the observation if the filing date is less than 180 days

[1] Other researchers such as Feldman et al. (2010) and Li (2010a) focus solely on the management discussion and analysis(MDA). However, Jones and Shoemaker (1994) find mixed empirical evidence of whether readability varies in the different sections of the annual report.

[2] *https://kelley.iu.edu/bpm/activities/bogindex.html*

[3] *https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html*

[4] *http://hobergphillips.tuck.dartmouth.edu/industryclass.htm*

from the prior filing (dropping 4 observations). We only include stocks with at least 2000 words(dropping 1623 observations). We drop observations that are not from common stocks, defined as having share codes 10 or 11 (dropping 8678 observations), and we limit the sample to only contain firms with a stock price greater than 3$ to avoid market microstructure effects(dropping 14491 observations). We also require a minimum of 10 days of stock price data for days [6, 28] and at least 60 observations for days [-252, -6], relative to the 10-K filing date. In addition, we exclude annual reports that do not contain sentences between 3 and 50 words. This filter is applied to remove documents that do not contain actual 10-k text.

Table 3.1 denotes the full sample screening procedure, from the original 97 417 to 70 106 observations. Our sample year period from 1996 to 2016 differ from that of Loughran and McDonald (2014) and Bonsall IV et al. (2017), which include the period from 1994 to 2011[5]. With the available data, the sample size is similar to that of Loughran and McDonald (2014) (66 707) and Bonsall IV et al. (2017)(66 173). When we subsequently include the text-based industry classifications from Hoberg and Phillips (2016), we note the change in sample size.

Table 3.1: Sample Creation

This table summarizes the effect of the filters on the original 10-K sample.
RMSE is the root mean square error from a market model regression for days [6, 28] following the 10-K filing.

	Dropped	Sample Size
10-K files 1996 to 2016		97 417
Drop if file date $<$ 180 days from prior filing	4	97 413
Drop if number of words $<$ 2 000	1 623	95 790
Reported on CRSP as ordinary common equity	8 678	87 112
Price on filing date minus one $\geq$ \$3	14 491	72 621
Post-filing date market model RMSE for days [6, 28]	151	72 470
At least 60 days' data available for market model estimates from event days [-252, -6]	102	72 368
Returns for days 0–1 in event period	8	72 360
Book-to-market COMPUSTAT data available and book value $>$ 0	1 756	70 604
Match with BOG index	495	70 109
Contains sentences of length 3-50 words	3	70 106

[5]a few large firms started electronic filing in 1994. Electronic filing was required of all firms, with minor exceptions, beginning in 1996(Loughran and McDonald, 2016)

3.3 Descriptive Results

Table 3.2 present the descriptive statistics for our main variables. A detailed description of these variables are presented in A1. Despite some differences in sample, the results show great resemblance to those of Loughran and McDonald (2014) and Bonsall IV et al. (2017). The mean Fog Index of annual reports is 19.6. According to the intended interpretation, this would indicate that an investor would require 19.6 years of education to comprehend the average annual report at first reading. However, we do not believe that an investor is required a Ph.D. level of education to comprehend the information of the average annual report. These statistics are likely an artifact of the application of the Fog Index outside of its intended use. Table 3.2 further displays a low variation in The Fog Index. Less than one unit of the Fog Index separates the 25th and the 75th percentile, meaning that the knowledge gain of less than one year's worth of studies is what separates half of the observations in terms of readability.

The mean of the Bog Index is 82.9, similarly indicating that the average annual report is not easy to comprehend. As with the Fog Index, the Bog Index has limited variation. Half of the observations are in the interval between 78 and 88, further indicating that the readability of annual reports is mostly similar and mostly hard to comprehend.

Table 3.2 further shows that the two quantity based readability measures, File size and Total words, have a positive skew with a low mean, but with several large observations. We follow Loughran and McDonald (2014) and transform these variables by the natural logarithm in further regressions. The same logarithmic transformation is performed to the variables Market capitalization and Book-to-market.

To illustrate the time variations in the variables we split the sample in two. Table 3.3 reports a strong trend in our two quantity-based measures of readability, file size and total words. The mean of File size is more than 14 times the size in the later subsample, compared to the earlier. Bonsall IV et al. (2017) show that this growth is mainly driven by the implementation of HTML and XML code in the filings, in addition to pictures, PDFs and more extensive use of tables. The amount of disclosed text also increases substantially between the two samples, where the mean of total words doubles between the

Table 3.2: Descriptive statistics

This table reports summary statistics for the key variables in our sample. In subsequent regressions variables *File size, Total Words, Market capitalization and Book-to-market* are log-transformed. See Appendix A1 for detailed variable descriptions.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Bog Index	70,106	82.951	7.621	47	78	88	140
Fog Index	70,106	19.605	0.894	14.935	19.033	20.190	23.740
File size	70,106	4.525	9.122	0.033	0.323	3.333	414.522
Total Words	70,106	26.568	16.750	2.003	14.145	35.831	251.727
Post-filing RMSE	70,106	2.487	2.111	0.000	1.202	3.088	106.960
Pre-filing alpha	70,106	0.046	0.253	−4.660	−0.081	0.142	6.963
Pre-filing RMSE	70,106	2.936	1.888	0.000	1.692	3.702	97.718
Abs(abnormal return)	70,106	0.033	0.044	0.000	0.008	0.041	1.740
Market Capitalization	70,106	3,512.658	16,653.430	0.654	103.676	1,481.017	638,654.200
Book-to-Market	70,106	0.665	0.745	0.0001	0.301	0.842	66.049
NASDAQ dummy	70,106	0.597	0.490	0	0	1	1

two subsamples. Our other two readability measures also show a positive trend through the sample period. The mean Fog index is 19.4 in the earlier period, compared to 19.8 in the later period, whereas the mean Bog index increase from 80.9 to 85.4 in the later period. Overall, this indicates that annual reports have become harder to read over the last twenty years. Opposed to the readability measures, the dependent variable in later regressions, post-filing RMSE, shows a negative trend through our sample period. The mean post-filing RMSE decreases from 3.0 in our early period to 1.9 in the later period.

Table 3.3: Mean statistics

This table displays mean descriptive statistics for the key variables in our sample. In subsequent regressions variables *File size, Total Words, Market capitalization and Book-to-market* are log-transformed. See Appendix A1 for detailed variable descriptions.

Statistic	1996 -2005	2006-2016
Bog Index	80.896	85.428
Fog Index	19.424	19.823
File size	0.641	9.206
Total Words	18.411	36.397
Post-filing RMSE	2.986	1.885
Pre-filing alpha	0.067	0.019
Pre-filing RMSE	3.360	2.426
Abs(abnormal return)	0.035	0.030
Market capitalization	2,373.822	4,884.806
Book-to-Market	0.643	0.690
NASDAQ dummy	0.619	0.57
N	38 310	31 796

Figure 3.1: Time trend in readability measures and post-filing RMSE

This figure shows the trend in each of the readability measures and post-filing RMSE by year over our entire sample period. For ease of interpretation, the mean values by year are standardized with a mean zero and standard deviation of one.

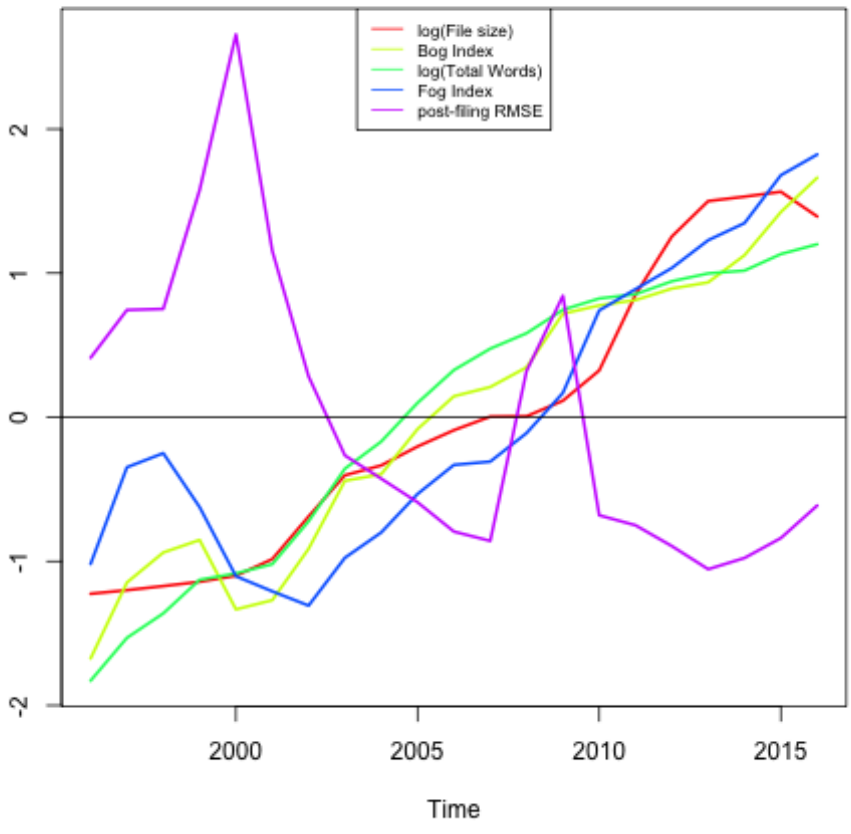

Figure 3.1 further illustrates the time trends of the readability measures and post-filing RMSE. The post-filing RMSE peaked during the Dotcom crash in the early 2000s and the financial crisis in 2008, but in general, show a negative trend throughout the sample period. The Fog Index and Bog Index show a modest decline following the introduction of the plain English mandate in 1998, but since the early 2000s both show a positive trend. Similarly, the quantitative measures, log(File size) and log(Total words) show a positive trend throughout the sample period.

4 What Affects the Readability Measures?

This section is structured as follows: First we present the main research question of this thesis. Second, we describe the methods we use to investigate this question.Then, we present the results.

4.1 Research Question

This thesis aims to answer the following research question:

Do readability measures reflect the effective communication of valuation-relevant information in annual reports?

The general consensus is that readability refers to the ease with which a reader can process and comprehend written text (Bonsall IV et al., 2017). However, the exact definition varies with the context of its application. In the context of financial disclosures, Loughran and McDonald (2014) define readability as the effective communication of valuation-relevant information. They advocate that more readable disclosures would produce less ambiguity in valuation, and validate their readability measures by showing that disclosures rated as less readable, are associated with higher subsequent stock return volatility.

There are, however, alternative explanations to these findings. By using such archival-based capital market test to validate readability measures, one is essentially testing two things at the same time. First, that the readability measures capture the concept of readability, and second, that readability is related to the outcome variable of interest (i.e, subsequent stock volatility). For instance, if the readability measures are correlated with another unobserved variable (e.g., complexity) that is correlated with the outcome variable examined, one could falsely conclude that the readability measure is a valid proxy for readability.

Following the notion of the possible close link between readability and firm complexity, we want to further investigate what determines a higher or lower readability score. It is possible that that unobserved factors, such as firm complexity, is more decisive in

determining the given readability score than the level of plain English violations, complex words or amount of disclosure. If such unknown variables correlate with both volatility and readability, they could be the cause of the previously discovered relation between these variables.

To answer our research question, we start by investigating whether the readability measures mainly reflect the concept of effective communication, or if there are other common features among the documents rated as less readable.

4.2 Methodology

Taddy (2013) introduces a framework to investigates the relationship between text data and a variable associated with the said text. We use this framework to determine which lexical features of annual reports are associated with high and low readability scores, as measured by four different readability measures. More specifically, Taddy (2013) calculates the population average effect a given word (or combination of words) has on an associated variable. This effect is represented by a word loading, where a high positive loading for a word indicates that documents containing this word are associated with higher levels of the associated variable.

We apply the methods of Taddy (2013) on our full sample of 70 106 annual reports and treat the four different readability measures as variables associated with the said text. Instead of investigating loadings for single words (unigrams), we focus on bigrams, defined as sequences of two adjacent words in a text document, as this speaks more specific to the context compared to unigrams (Taddy, 2013). For example, the bigrams *annual_ report* or *wall_ street* speak more specifically to the context of which it is used, compared to the unigrams *street, report, annual*, and *wall*. In Appendix A2 we present a more detailed description of our data cleaning steps and the implementation of Taddy (2013) on our sample.

4.3 Results

In this section, we present the results from using the framework of Taddy (2013) to investigate the relationship between the text of 70 106 annual reports and corresponding readability measures. We expect that if the readability measures mainly reflect the level of effective communication and not some other feature, facets of readability should be reflected in the word lists of bigrams. Examples of this could be that bigrams containing more complex words, longer words (where easier synonyms exist), abstract words and legal terms should get high positive loadings and bigrams that are more easy to comprehend get negative loadings.

Table 4.1 presents the 40 bigrams with the top positive and negative loadings related to the Bog Index. The bigrams with high positive and negative loading are less differentiated by their complexity, but rather reflect patterns of industry-specific vocabulary. For example, among the top positive loadings related to the Bog Index, there are the bigrams *gene_therapies*, *clinical_holding*, and *rare_disease*, related to healthcare and pharmaceutical industries. The top negative loadings include *restaurant_company*, *line_bank*, and several other bigrams related to the restaurant and banking industries.

These patterns indicate that the annual reports of firms in specific industries are measured as less readable by the Bog Index. However, describing a complex business or situation is distinct from reporting a situation in an easy versus a complicated way. For instance, a firm in the pharmaceutical industry might produce an annual report in accordance with the plain English rules. Still, describing the technical aspects of their operation might require complex language, leading to a higher Bog Index. At the same time, a firm in the restaurant industry might produce a disclosure with many violations of the plain English rules. However, due to the lower complexity of their operations, less technical language is needed. Thus, the Bog Index might falsely classify the disclosure of the pharmaceutical company as less readable than the disclosure of the restaurant company. This highlights how differences in the industry level of complexity might lead to biased readability measures.

Table 4.1: Top 40 Bigrams with Positive and Negative Loadings related to the Bog Index

The table displays the 40 bigrams with the top positive and negative loadings related to the Bog Index, using Taddy (2013). High positive loadings indicate that documents using these bigrams are associated with higher levels of the Bog Index, and are thus considered less readable. Similarly, a negative loading is associated with a more readable document, as measured by the Bog Index. The loadings is calcualted using the text of 70 106 annual reports from 1996-2016, and the associated Bog Index of each annual report. The wordlist consist of bigrams of stemmed words, using the Porter stemmer.

	negative bigrams	Loadings	Positive bigrams	Loadings
1	restaur_compani	-0.1664	gene_therapi	0.1578
2	page_registr	-0.1639	serious_life-threaten	0.1565
3	restaur_manag	-0.1580	licens_collabor	0.1562
4	open_restaur	-0.1542	prioriti_review	0.1560
5	line_bank	-0.1537	acceler_approv	0.1554
6	counti_market	-0.1527	candid_delay	0.1545
7	bank_main	-0.1526	clinic_hold	0.1545
8	one_restaur	-0.1501	candid_approv	0.1544
9	restaur_offer	-0.1494	rare_diseas	0.1544
10	presid_merchandis	-0.1485	product_candid	0.1542
11	main_bank	-0.1476	candid_obtain	0.1541
12	automat_teller	-0.1469	candid_manufactur	0.1540
13	compani_restaur	-0.1459	avail_therapi	0.1540
14	page_proxi	-0.1456	well-control_clinic	0.1534
15	addit_restaur	-0.1454	secondari_endpoint	0.1529
16	store_featur	-0.1453	medicin_product	0.1525
17	restaur_open	-0.1451	safeti_toler	0.1525
18	merchandis_manag	-0.1450	grant_orphan	0.1524
19	restaur_sale	-0.1446	trial_site	0.1523
20	take_stock	-0.1442	primari_endpoint	0.1523
21	store_rang	-0.1438	candid_receiv	0.1523
22	apart_complex	-0.1429	submiss_ind	0.1521
23	restaur_general	-0.1428	addit_preclin	0.1519
24	bif_fdic	-0.1417	candid_depend	0.1519
25	bank_open	-0.1415	design_drug	0.1518
26	presid_store	-0.1415	candid_expect	0.1516
27	new_restaur	-0.1415	commerci_licens	0.1516
28	restaur_also	-0.1414	review_nda	0.1516
29	director_page	-0.1412	protocol_detail	0.1514
30	sound_feder	-0.1411	collabor_licens	0.1513
31	exist_restaur	-0.1410	safer_effect	0.1512
32	compens_page	-0.1410	candid_develop	0.1511
33	food_item	-0.1407	central_procedur	0.1510
34	ice_cream	-0.1406	trial_protocol	0.1510
35	out-of-st_bank	-0.1406	candid_commerci	0.1510
36	three_store	-0.1401	approv_label	0.1510
37	main_offic	-0.1394	nasdaq_biotechnolog	0.1508
38	fund_bif	-0.1392	medicin_agenc	0.1508
39	store_compani	-0.1387	adequ_well-control	0.1507
40	store_personnel	-0.1382	price_reimburs	0.1504

In Appendix A3 we present word list of bigrams with the top positive and negative loadings related to the Fog Index, log(Total Words), and log(File size), respectively. Similar to the word list presented in Table 4.1, the Fog Index word list shows that vocabulary related to specific industries are associated with less readable disclosures. Among the top

negative loadings, we find the bigrams *steel_ manufacturer* and *steel_ product*, indicating that annual reports of firms in the steel industry are linked to higher readability, as measured by the Fog index. In the top positive loadings, there are bigrams such as *homeowner_ insurance*, *physician_ service*, and *surgery_ center*. These bigrams indicate that firms within the healthcare and insurance industries are generally considered less readable by the Fog Index. Thus, with the same argumentation as above, we argue that the high loadings of industry-specific terminology signals that the Fog Index is affected by the industry level of complexity.

Industry-specific vocabulary is also prominent among the bigrams with the highest association to longer 10-K texts. Vocabulary related to the financial industry such as *credit_ derivative*, *default_ swap* and *credit_ spread* are all among the bigrams with the higher positive loadings. Considering the role of credit default swaps in the financial crisis, the aforementioned bigrams might also signal that financial companies had to produce longer annual reports to describe their situation following the financial crisis. This notion is supported by the fact that the bigram *loss_ billion* is among the bigrams with the highest positive loading. Further, vocabulary related to the energy industry, such as *wholesale_ energy*, *power_ price*, and *energy_ supplier*, are also associated with longer annual reports.

In the bigrams associated with smaller file size we find *non-it_ system* and *year_ y2k* associated with the IT-industry, but most words are not limited to a single industry. Among the bigrams associated with larger file size, we find *volcker_ rule*, *basel_ iii*, and *implement_ dodd-frank*. These bigrams relate to government regulation of the banking industry in the period after the financial crisis. Industries affected by these regulations might need to include details or attachments that affect the file size.

Taddy (2013) further shows that the word loadings can be used to predict the associated variable for additional text documents. Using the word loadings the from the initial set of documents and the word counts of the new documents he produces a single information-preserving score for each document. This sufficient reduction (SR) projection is then used to predict the level of the associated variable for the additional documents, in a forward regression.

To validate whether the word lists of bigrams are suited to reflect what lexical features that are associated with higher and lower levels of readability, we test whether the word lists can be used to predict the readability score of documents. Following Taddy (2013) we use the sufficient reduction projection constructed from the bigram loadings in a forward regression. We find that the sufficient reduction projections are significant in predicting the readability score of documents. Thus, indicating that the word lists of bigrams are in fact reflecting features related to the readability measures. Appendix A4 provide a more detailed description of the forward regression and interpretation of the results.

Together these findings signal that the readability measures are affected by features of the firm distinct from the effective communication of valuation-relevant information. We find that industry-specific vocabulary correlates with the readability measures. Following the notion of the close link between complexity and readability in the financial disclosure literature, we interpret these findings as that much of the variation in the readability measures is due to differences in the industry level of complexity.

5 Impact of Readability Measures in Regressions of Volatility

This section is organized as follows: First, we present our hypothesis and the intuition for this. Second, we describe the models used in previous research and present the models we use. Third, we present results from using both sets of models and compare the results.

5.1 Hypothesis Development

As the previous tests show that the readability measures are affected by the industry level of complexity, we challenge the empirical link between readability and stock price volatility found by Bonsall IV et al. (2017) and Loughran and McDonald (2014). It is possible that the association between the variables is a result of the unobserved industry effects correlating with both the volatility and the readability measures.

Both Loughran and McDonald (2014) and Bonsall IV et al. (2017) validate their readability measures ability to reflect the level of effective communication by testing if higher levels of their readability measures are associated with higher stock return volatility in the month following the filing date of the 10-K. The intuition behind this test builds on the assumption that less readable 10-Ks are harder to understand and lead to ambiguity in validation of the associated stock. To capture the uncertainty in the information environment attributable to readability they use the subsequent stock return volatility. Loughran and McDonald (2014) argue that the volatility of returns immediately surrounding the filing date is affected by both the information signal and its uncertainty, but that the uncertainty component are more likely to persist in the period following the filing. Thus, they use use the root mean square error (RMSE) from a market model estimated using trading days [6, 28] relative to the 10-K file date. Both Loughran and McDonald (2014) and Bonsall IV et al. (2017) find that less readable annual reports are associated with higher subsequent stock price volatility (controlling for other variables, including the historical level of volatility). To account for unobserved heterogeneity between firms and

industry, they additionally include Fama and French (1997) industry fixed effects and firm fixed effects, respectively.

However, we consider that these methods do not sufficiently take into account the unobserved heterogeneity between industries. For example, Hoberg and Phillips (2016) argue that the Standard Industrial Classification (SIC) does not accurately represent a firm's rivals. They explain that such industry classifications do not adequately reclassify firms over time as the product market evolves, and the SIC do not accommodate innovations that create entirely new product markets.

In addition, industry fixed effects are limited to account for unobserved heterogeneity that is constant over time. As illustrated by Figure 3.1 both volatility in the period after the filing date and the readability measures show clear time trends. Loughran and McDonald (2014) and Bonsall IV et al. (2017) include year fixed effects to account for the general trend in the economy. However, this approach does not consider that trends might differ between industries. As Matsa (2010) points out, industries react independently to demand shocks and implications of new legislation. Further, Cetorelli and Strahan (2006) highlight the impact of technological changes that drive both structure and regulations within industries. Thus, we hypothesize the following:

The association between annual report readability and subsequent stock return volatility is driven by time-varying industry effects.

5.2 Methodology

To test this hypothesis, we implement regression analyses, following Loughran and McDonald (2014) and Bonsall IV et al. (2017). We first describe the models Loughran and McDonald (2014) and Bonsall IV et al. (2017) use, before we present the models including controls for time-varying industry characteristics.

5.2.1 Methods used by Loughran and McDonald (2014) and Bonsall IV et al. (2017)

To isolate the effect of readability on subsequent stock volatility, Loughran and McDonald (2014) use a regression model, controlling for other firm characteristics linked to volatility. In addition, they include industry and year dummy variables to account for fixed industry and year effects. Loughran and McDonald (2014) use the following regression:

$$\sigma_{i,j,t} = \alpha + \lambda X_{i,j,t} + \beta' Y_{i,j,t} + \phi_j + \gamma_t + \epsilon_{i,j,t}, \tag{5.1}$$

where $\sigma_{i,j,t}$ is the RMSE for trading days [6,28] relative to the filing date for firm i in industry j in year t. $X_{i,j,t}$ is one of the four readability measures of the annual report of firm i in year t and $Y_{i,j,t}$ is a vector of firm-specific characteristics, including the pre-filing RMSE, of firm i in year t. ϕ_j and γ_t represent the Fama and French (1997) industry fixed effects and year fixed effects, respectively.

The regression model used by Bonsall IV et al. (2017) differs from that of Loughran and McDonald (2014) by replacing the industry fixed effects (ϕ_j) with firm fixed effects (f_i) to mitigate endogeneity concerns raised by Li (2010b). Bonsall IV et al. (2017) use the following regression model:

$$\sigma_{i,t} = \alpha + \lambda X_{i,t} + \beta' Y_{i,t} + f_i + \gamma_t + \epsilon_{i,j,t}, \tag{5.2}$$

5.2.2 Introducing Additional Controls

Following their criticism of commonly used industry classifications, Hoberg and Phillips (2016) introduce two new sets of industries based on product similarity between firms, the Fixed Industry Classifications (FIC) and the Text-Based Network Industries (TNIC). The FIC are analogous to commonly used industry classifications such as SIC, in that industries exist as groups of similar firms. However, FIC are distinct in that firms are grouped as industries based on word similarities in the product descriptions of annual

reports, such that FIC are annually updated with the release of new 10-K documents.

TNIC differ from other industry classifications in that instead of grouping firms by similarity, each firm has its own distinct set of peers. Firms are classified as being peers if the product descriptions in their 10-K filings have words similarities above a given threshold. This allows a firm to have different competitors in different years, similarly to FIC, and relaxes the assumption of transitivity[6]. A more detailed description of FIC and TNIC is presented in Appendix A5.

5.2.2.1 Industry-Year effects - FIC Model

Through their website, Hoberg and Phillips provide FIC with different granularity. They recommend using the FIC with 300 different industries as these are most analogous to the three-digit SIC codes and argue that such a division is best suited to explain firm-level data. Thus, we include the 300 industries classification in our further analysis.

In addition, to address the concern that industries may react independently to changes in their environment over time(i.e, some industries are more exposed to oil price shocks or new industry specific legislation) we include industry-year fixed effects, in addition to the firm fixed effects. Hence, we look at variation within a firm across time, adjusting for within firm changes at other firms in the same industry and year. We propose the following model:

$$\sigma_{i,j,t} = \alpha + \lambda X_{i,j,t} + \beta' Y_{i,j,t} + f_i + \delta_{j,t} + \epsilon_{i,j,t}, \tag{5.3}$$

where $\sigma_{i,j,t}$ is the RMSE for trading days [6,28] relative to the filing date for firm i in industry j in year t. $X_{i,j,t}$ is one of the four readability measures for firm i, in industry j, in year t. $Y_{i,j,t}$ are a vector of the same six firm-specific control variables as used by Loughran and McDonald (2014) and Bonsall IV et al. (2017). Following Bonsall IV et al. (2017) we also include firm fixed effects (f_i). Instead of the fixed year effects used by both Loughran

[6]To illustrate how transitivity may restrict fixed industry classification, suppose firms A and B both consider firm C as a rival. If C have products similar to both those of A and B, but the products of A and B are not similar, then A and B may not compete against each other, as they may serve different product segments.

and McDonald (2014) and Bonsall IV et al. (2017) we include fixed industry-by-year effects $\delta_{j,t}$ using the FIC by Hoberg and Phillips (2016) for industry classification.

5.2.2.2 TNIC Peer Average Model

In an additional regression model, we investigate whether the association between readability and subsequent volatility is robust the additional control of subsequent volatility among a firm's peers. The intuition behind this is that the variation in firms volatility presumed to be explained by readability measures could be due to circumstances affecting the industry as a whole. More precisely, we test whether higher levels of volatility following a 10-K filing are associated with less readable disclosures when we control for the uncertainty surrounding similar firms in the same year, as reflected by the stock price volatility subsequent to their filing.

For this model, we apply the text-based network classifications by Hoberg and Phillips (2016). As this classification assign each firm a unique set of peers for each year, it can not account for industry fixed effects by including industry dummies. However, in an earlier version of their 2016 paper, Hoberg and Phillips (2010) propose a simple industry-averaging method, analog to using industry-year fixed effects. The proposed method involves taking the average of the given firm characteristic (the dependent variable) within each industry in each year and use this average as a single additional control variable. Compared to using industry-year fixed effects, this method substantially reduces the degrees of freedom used with considerably fewer dummy variables accounting for the time effects. More degrees of freedom leaves more information for the parameter estimates, which is beneficial when small changes in the significance levels are detrimental to the results. Further, the method allow us to exclude the firm itself from the industry average, as opposed to fixed effects methods, where the values of the firm itself is included in the average. We calculate the new variable as follows:

$$\overline{\sigma}_{j-i,t} = \frac{1}{n_{j-i,t}} \sum_{j=1}^{n_{j-i,t}} \sigma_{j-i,t}, \tag{5.4}$$

where $\overline{\sigma}_{j-i,t}$ represents the average post-filing RMSE among peers classified to be in the

industry j of firm i in year t, excluding firm i from the average. In later regression we refer to this variable as *industry volatility*. $n_{j-i,t}$ represent the number of firms in the industry j of firm i in year t, excluding firm i, and $\sigma_{j-i,t}$ represent the post-filing RMSE for each firm in industry j of firm i, excluding firm i, calculated in the month following the filing date of each firm in year t.

Gormley and Matsa (2013) argue that using the simple industry mean as a control variable instead of including fixed effect, could lead to biased estimators. Thus, when interpreting the results we will have the model limitations in mind, and emphasize the findings accordingly. We propose the following regression model:

$$\sigma_{i,t} = \alpha + \lambda X_{i,t} + \overline{\sigma}_{j-i,t} + \beta' Y_{i,t} + f_i + \gamma_t + \epsilon_{i,t}, \tag{5.5}$$

where $\sigma_{i,t}$ is the RMSE for trading days [6,28] relative to the filing date. $X_{i,t}$ is one of the four readability measures for firm i in year t. $Y_{i,t}$ is a vector of the same six firm-year specific control variables as used by Loughran and McDonald (2014) and Bonsall IV et al. (2017), including the pre-filing RMSE. Following Bonsall IV et al. (2017) we also include firm (f_i) and year (γ_t) fixed effects. The average post-filing RMSE among peers of firm i in year t is represented by $\overline{\sigma}_{j-i,t}$. To ensure the industry average among a firm's peers is representative, we impose the restriction that each firm must have at least five peers to classify as an industry and be included in the sample ($n_{j-i,t} \geq 5$).

5.3 Results

To add validity to our further analysis we begin by replicating the results of Bonsall IV et al. (2017) and Loughran and McDonald (2014). We subsequently present the results of our additional regression models that apply controls for time-varying industry effects.

5.3.1 Replicating Previous Literature

We calculate the Fog Index, log(File Size), and log(Total Words) for all documents in the sample of 70,106 10-K files. The Bog Index data is pre-calculated. As noted, our sample years extend that of Loughran and McDonald (2014) and Bonsall IV et al. (2017) to include data from 2012-2016, but drop the years 1994-1995. Using the same models we therefore expect some minor deviations relative to their results.

The results of the test replicating Bonsall IV et al. (2017) is reported in Table 5.1, and the results of the test replicating Loughran and McDonald (2014) is reported in Appendix Table A6.1. For ease of interpretation, all regression variables are standardized with a mean of zero and a standard deviation of one. In addition to an intercept, firm and year dummies, each regression has the following firm-specific control variables:

- *Pre-filing alpha*, the alpha from a market model estimated for the period prior to the 10-K filing date.
- *Pre-filing RMSE*, The root mean squared error from a market model estimated for the period prior to the 10-K filing.
- *Abs(abnormal return)*, The absolute value of the two-day buy-and-hold abnormal return from the filing date to the next date of trading.
- *log(Market capitalization*, the natural logarithm of each firms market capitalization the day prior to the filing.
- *log(Book-to-Market)*, the natural logarithm of the book-to-market ratio calculated from data reported prior to the filing date.
- *NASDAQ dummy*, dummy variable equal to 1 if the firms stock is listed on NASDAQ, zero otherwise.

A more detailed variable description is presented in Appendix A1.

Column (1) of Table 5.1 shows the results from a regression on Post-filing RMSE considering the above-mentioned controls. Five out of six control variables are significant at the 1%

level. However, we find that with our sample, that book to market is not significantly related to post-filing RMSE. The results imply that larger firms (market capitalization), firms tilted toward value (i.e., high book-to-market ratio), and firms with better pre-filing performance are less volatile subsequent to annual report filings. Firms with higher pre-filing volatility(RMSE), larger absolute abnormal return on the filing date and firms listed on NASDAQ, all else equal, are more volatile in the days following a 10-K filing. The R^2 of the first regression is 62.0%

In columns (2), (3), (4), and (5) the readability measures the Bog Index, log(File Size), log(Total Words), and the Fog index are included as explanatory variables, respectively. Consistent with Loughran and McDonald (2014) and Bonsall IV et al. (2017), all four measures of readability are positively associated with future stock return volatility, and have similar levels of significance and magnitude of association. The Bog Index is the only of the four measures that is significant at the 1% level(t-stat of 2.82). log(Total Words)(t-stat of 2.15) and the Fog Index(t-stat of 2.05) is significant at the 5% level, whereas log(File Size)(t-stat of 1.81) is significant at the 10% level.

Table 5.1: Post-Filing RMSE and Readability Measures Following Bonsall IV et al. (2017)

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression includes an intercept, calendar year dummies and firm dummies in addition to the presented control variables. For ease of interpretation, all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 70 106 firm-year observations during 1996-2016.

	Dependent variable:				
	Post-filing RMSE				
	(1)	(2)	(3)	(4)	(5)
Bog Index		0.032^{***} t = 2.862			
Log(File Size)			0.023^{*} t = 1.808		
Log(Total Words)				0.019^{**} t = 2.149	
Fog Index					0.010^{**} t = 2.046
Pre-filing alpha	-0.050^{***} t = −2.722	-0.049^{***} t = −2.715	-0.049^{***} t = −2.711	-0.049^{***} t = −2.708	-0.049^{***} t = −2.709
Pre-filing RMSE	0.367^{***} t = 14.507	0.366^{***} t = 14.273	0.367^{***} t = 14.476	0.366^{***} t = 14.275	0.367^{***} t = 14.494
Abs(abnormal return)	0.100^{***} t = 13.605	0.100^{***} t = 13.569	0.100^{***} t = 13.619	0.100^{***} t = 13.641	0.100^{***} t = 13.605
Log(market capitalization)	-0.113^{***} t = −2.708	-0.114^{***} t = −2.732	-0.116^{***} t = −2.783	-0.115^{***} t = −2.771	-0.114^{***} t = −2.708
Log(book-to-market)	−0.029 t = −1.414	−0.030 t = −1.460	−0.029 t = −1.438	−0.029 t = −1.436	−0.029 t = −1.417
NASDAQ dummy	0.030^{***} t = 3.001	0.030^{***} t = 3.036	0.030^{***} t = 3.006	0.030^{***} t = 3.028	0.030^{***} t = 3.000
Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	70,106	70,106	70,106	70,106	70,106
R^2	0.620	0.621	0.621	0.621	0.620
Note:					*p<0.1; **p<0.05; ***p<0.01

Comparing the regression results of the two tests show no dramatic changes in the coefficient of the readability measures. However, we note that the statistical significance of all readability measures increases compared to the results in Table 5.1. Both the Bog Index and the Fog Index are in this case significant at the 1% level. The disclosure quantity measures log(File size) and log(Total Words) are both significant at the 5% level and show a greater association with post-filing RMSE than when firm fixed effects are used.

To highlight the time component of the variables, we split the sample into two parts. We observe that when performing the previous tests on both subsamples, the Bog index is still significant in explaining the post-filing volatility in the earlier sample, but as Table 5.2 report, the coefficients of all four readability measures are not significant in the later sample. There could be several causes for this. Changes in firm characteristics in the latter years could capture the effect that was previously picked up by the readability measures. An alternative explanation is that by reducing the sample size, there is not sufficient variation in the data to pick up the small effect of annual report readability on the subsequent stock volatility. Independent of the explanation behind this result, a researcher with only data available for the last ten years would conclude that readability has no effect on stock return volatility. A similar regression for the earlier subsample is reported in Appendix Table A6.2.

Table 5.2: Post-Filing RMSE and Readability Measures Following Bonsall IV et al. (2017) for the Years 2006 to 2016

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression includes an intercept, calendar year dummies and firm dummies in addition to the same 6 control variables presented in Table 5.1. For ease of interpretation, all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 31 796 firm-year observations during 2006-2016.

	Dependent variable:			
	Post-filing RMSE			
	(1)	(2)	(3)	(4)
Bog Index	0.014 t = 0.692			
Log(File Size)		0.006 t = 0.297		
Log(Total Words)			0.002 t = 0.171	
Fog Index				−0.006 t = −0.415
Controls	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	31,796	31,796	31,796	31,796
Adjusted R^2	0.532	0.532	0.532	0.532
Note:				*p<0.1; **p<0.05; ***p<0.01

5.3.2 Testing Robustness

In the following we test our hypothesis by estimating the regression models from sections 5.2.2.1 and 5.2.2.2.

5.3.2.1 FIC

We merge our original data sample with the Hoberg and Phillips (2016) FIC industry data. This reduces the sample size from 70,106 to 66,213 firm-year observations. We report regression results with this reduced sample following Bonsall IV et al. (2017) in Appendix Table A6.3. We note that the Fog Index is no longer significant in explaining post-filing RMSE in this sample which emphasizes the inconsistency of the readability measures in different samples, as displayed in Table 5.2.

Table 5.3 reports the results for the regression model with fixed industry-year effects. From column (1) we observe that the coefficient of the Bog Index is no longer statistically significant at the 10% level and the magnitude of association with post-filing RMSE is halved compared to using the methodology of Bonsall IV et al. (2017). In column (2) log(File size) is an explanatory variable. We note that the corresponding coefficient is reduced to less than a quarter of its value when using the methodology of Bonsall IV et al. (2017) and the coefficient is statistically indistinguishable from zero. Column (4) shows that the Fog Index is not significantly associated with the post-filing RMSE. However, this is due to the reduced sample size, as the Fog index is also insignificant when using the method of Bonsall IV et al. (2017) for the same sample. Thus, we can not assume the non-significance of the Fog Index is due to the additional controls in our model.

The results in column (3) stand out from the other three regressions. In this regression, we study the association between log(Total Words) and the subsequent stock volatility. We observe that the magnitude of association is reduced by almost 50% compared to when using the methodology of Bonsall IV et al. (2017). A one standard deviation increase in log(Total Words) is associated with a 1.1% standard deviation increase in the subsequent stock return volatility. However, the coefficient is still statistically distinguishable from zero at the 5% level. Thus, indicating a positive association between the total amount of

disclosed 10-K text, and the subsequent stock volatility.

Overall, the results support the hypothesis. The findings indicate that time-varying industry characteristics are driving much of the association between the readability of annual reports and subsequent stock volatility in prior literature. The already limited economic magnitude of the readability measures in explaining subsequent volatility is further reduced when controlling for time-varying industry characteristics, and for the majority of the readability measures, the association is statistically insignificant.

Table 5.3: Post-Filing RMSE and Readability Measures with Industry-Year Fixed Effects (FIC)

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression includes an intercept, Hoberg and Phillips (2016) industry-year dummies and firm dummies in addition to the control variables presented in Table 5.1. For ease of interpretation, all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 66,213 firm-year observations during 1996-2016.

	Dependent variable:			
	Post-filing RMSE			
	(1)	(2)	(3)	(4)
Bog Index	0.016			
	t = 1.625			
Log(File size)		0.006		
		t = 0.617		
Log(Total Words)			0.013**	
			t = 2.317	
Fog Index				0.006
				t = 1.209
Controls	Yes	Yes	Yes	Yes
industry-year Fixed Effects	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Observations	66,213	66,213	66,213	66,213
Adjusted R^2	0.589	0.589	0.589	0.589
Note:				*p<0.1; **p<0.05; ***p<0.01

5.3.2.2 TNIC

To estimate the TNIC regression model using the simple averaging method presented in section 5.2.2.2, we first merge or data with the Hoberg and Phillips (2016) TNIC data. We require each firm to have at least 5 peers to be included in the sample, to mitigate a single firm's volatility in distorting the peer mean. This further reduces the data sample to 49 678 firm-year observations. We note that the Fog index is non-significant in explaining

post-filing RMSE with a regression following Bonsall IV et al. (2017) for this sample, while the other three readability measures remain significant. This regression is reported in Appendix Table A6.4.

Table 5.4 displays the regression results from using post-filing RMSE as the dependent variable, where the Industry Volatility (mean post-filing RMSE among the company's peers) is included as an additional control variable. The coefficient of the industry volatility variable is significant in all four regressions(t-stat > 8.5), indicating that the industry volatility in a year is associated with the volatility for individual firms. Further, the results show that the Bog Index and File size are not significantly related to the post-filing RMSE when inducing industry-volatility as an additional control. As the Fog Index is not significant due to the change in sample size, we can not conclude on the effects of the additional control variable for this readability measure. Total words is the only readability measure that retains significance with the added industry control, although with a reduced coefficient. Overall the results suggest that the association between readability and volatility is not robust to the inclusion of time-varying industry controls.

The results from the regression reported in Table 5.4 are resemblant to the results of the FIC-regression reported in Table 5.3. Although they use different methods to account for the impact industries, the results of both regressions indicate that the association between the readability measures and stock price volatility is not robust to controls for time-varying industry effects.

Table 5.4: Post-Filing RMSE and Readability Measures with Industry Peer RMSE Control Variable (TNIC)

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression includes an intercept, calendar year dummies and firm dummies in addition to the same 6 control variables as presented in table 5.1. For ease of interpretation, all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 49 678 firm-year observations during 1996-2016.

	Dependent variable:			
	Post-filing RMSE			
	(1)	(2)	(3)	(4)
Bog Index	0.010			
	t = 1.037			
Log(File Size)		0.015		
		t = 1.362		
Log(Total Words)			0.011**	
			t = 2.179	
Fog Index				−0.001
				t = −0.200
Industry volatility	0.292***	0.292***	0.292***	0.292***
	t = 8.592	t = 8.506	t = 8.567	t = 8.536
Controls	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	49,678	49,678	49,678	49,678
Adjusted R^2	0.593	0.593	0.593	0.593
Note:				*p<0.1; **p<0.05; ***p<0.01

6 Conclusion

In this thesis we investigate whether readability measures reflect the effective communication of valuation-relevant information in annual reports. We find that vocabulary reflecting higher and lower readability scores in 10-K filings consists of industry-specific words, suggesting that much of the variation in the readability measures is due to differences in the industry level of complexity, rather than the level of effective communication. Based on this result we hypothesize that the previously found empirical relationship between annual report readability and subsequent stock return volatility is driven by time-varying industry effects. This association has in previous literature been used to validate readability measures' ability to reflect the effective communication of valuation-relevant information in financial disclosures (Loughran and McDonald, 2014). In two distinct tests, we find that the association between readability measures and volatility diminishes when we include controls for time-varying industry effects. These results support our hypothesis, and further highlight the readability measures inability to reflect the effective communication of valuation-relevant information.

The results are consistent with the notion from previous literature, that the association between 10-K readability and subsequent stock return volatility could be a result of an unobserved variable (e.g., complexity) that correlates with both volatility and the readability measures (see e.g., Loughran and McDonald (2014); Loughran and McDonald (2016); Li (2010b)). Our findings suggest that one or more such variables are related to industry-specific changes over time. Causes of such time-varying industry heterogeneity include industry-level shocks to demand and ramifications of new legislation as discussed by Matsa (2010), in addition to technological advances as considered by Cetorelli and Strahan (2006).

With the inclusion of time-varying industry controls, the Bog Index and File Size are no longer significantly associated to the post-filing volatility. Similarly, the association between Fog Index and post-filing volatility is insignificant in these tests. However, this effect is due to the reduction in sample size, and can not be ascribed to the inclusion of industry controls. Still, this highlights the weakness of the association between Fog Index

and post-filing volatility. Total words in the 10-K filing is the only readability measure that retains significance in the link to stock price volatility with the inclusion of new industry controls. This result is consistent with the notion from Loughran and McDonald (2014), that firms trying to obfuscate earnings-relevant information are more likely to hide this data in longer documents than with complex language. However, the already limited economic magnitude of the association is further reduced. A one standard deviation increase in log(Total Words) is associated with an increase in subsequent volatility that is 1.9% of subsequent volatility's standard deviation without controls for time-varying industry effects. With these controls the economic significance declines to 1.3% using FIC industry-year fixed effects and 1.1% using TNIC peer mean volatility.

Viewed collectively, the findings of this thesis indicate that the current readability measures are affected by unobserved variables related to time-varying industry complexity. The implications of this relate to the interpretations of readability measures and highlight the limitations of their use in financial documents. When applied to 10-K filings, the readability measures are not objective indicators of the communication of information, as they display a bias to certain industries. Based on this, we advice against using readability measures as a tool in gauging plain English compliance, as previously discussed by the SEC. At the very least our findings urge caution when interpreting the impact of disclosure readability on various market outcomes, as time-varying industry effects might lead to incorrect inference.

References

Biddle, G. C., Hilary, G., and Verdi, R. S. (2009). How does financial reporting quality relate to investment efficiency? *Journal of accounting and economics*, 48(2-3):112–131.

Bloomfield, R. (2008). Discussion of "annual report readability, current earnings, and earnings persistence". *Journal of Accounting and Economics*, 45(2-3):248–252.

Bonsall, S. B. and Miller, B. P. (2017). The impact of narrative disclosure readability on bond ratings and the cost of debt. *Review of Accounting Studies*, 22(2):608–643.

Bonsall IV, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357.

Cetorelli, N. and Strahan, P. E. (2006). Finance as a barrier to entry: Bank competition and industry structure in local us markets. *The Journal of Finance*, 61(1):437–461.

Clatworthy, M. and Jones, M. J. (2001). The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal*, 14(3):311–326.

Courtis, J. K. (1998). Annual report readability variability: tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal*, 11(4):459–472.

Cox, C. (2007). Speech by SEC Chairman: Closing Remarks to the Second Annual Corporate Governance Summit. Closing Remarks by SEC Chairman Christipher Cox, to the Second Annual Corporate Governance Summit at the USC Marshall School of Business, Los Angeles, California, March 23, 2007. Available at https://www.sec.gov/news/speech/2007/spch032307cc.htm [accessed: 2019/05/05].

Editor Software (2014). *Can You Read Me Now? How to Use the New BOG Readability Formula to Write for Your Target Audience!* Editor Software UK ltd, Dursley. https://ia800500.us.archive.org/1/items/TheNewBogIndexReadabilityFormula/bog-readability-ebook.pdf.

Ertugrul, M., Lei, J., Qiu, J., and Wan, C. (2017). Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis*, 52(2):811–836.

Fama, E. F. and French, K. R. (1997). Industry costs of equity. *Journal of financial economics*, 43(2):153–193.

Fama, E. F. and French, K. R. (2001). Disappearing dividends: changing firm characteristics or lower propensity to pay? *Journal of Financial economics*, 60(1):3–43.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 15(4):915–953.

Gormley, T. A. and Matsa, D. A. (2013). Common errors: How to (and not to) control for unobserved heterogeneity. *The Review of Financial Studies*, 27(2):617–661.

Guay, W., Samuels, D., and Taylor, D. (2016). Guiding through the fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics*, 62(2-3):234–269.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill, New York.

Hoberg, G. and Phillips, G. (2010). Text-based network industries and endogenous product differentiation. NBER Working Paper No. 15991, 1050 Massachusetts Avenue, Cambridge, MA 0238.

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

Jones, M. J. and Shoemaker, P. A. (1994). Accounting narratives: A review of empirical studies of content and readability. *Journal of accounting literature*, 13:142.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, 56(1):130–147.

Lehavy, R., Li, F., and Merkley, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review*, 86(3):1087–1115.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, 45(2-3):221–247.

Li, F. (2010a). Managers' self-serving attribution bias and corporate financial policies. Working paper, University of Michigan, Available at SSRN 1639005.

Li, F. (2010b). Textual analysis of corporate disclosures: a survey of the literature. *Journal of accounting literature*, 29:143–165.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Matsa, D. A. (2010). Capital structure as a strategic variable: Evidence from collective bargaining. *The Journal of Finance*, 65(3):1197–1232.

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review*, 85(6):2107–2143.

SEC (1969). *Disclosure to Investors: A Reappraisal of Federal Administrative Policies Under the'33 and'34 Acts*. Commerce Clearing House.

Securities and Exchange Commission (1998). *A plain English handbook: How to create*

clear SEC disclosure documents. US Securities and Exchange Commission, Washington, DC. http://www.sec.gov/pdf/handbook.pdf.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. et al. (2015). Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Appendix

A1 Variable definitions

Table A1.1: Variable definitions

Variable Names	Definitions
Readability Measures:	
Fog Index	The Gunning (1952) Fog Index, equal to 0.4*(average number of words per sentence+percent of complex words). High values of the Fog Index imply less readable text. In our analysis we use the "quanteda" package from the Comprehensive R Archive Network (CRAN) to calculate the Fog Index.
Log(File size)	The natural logarithm of the file size in megabytes of the SEC EDGAR "complete submission text file" for the 10-K filing.
Bog Index	A proprietary measure of readability created by Editor Software's plain English software, StyleWriter. The formula is based on several plain English factors such as sentence length, passive voice, weak verbs, overused words, complex words, and jargon. Higher values of the index imply lower readability (Bonsall IV et al., 2017). The Bog Index for each 10-K filing is collected from Brian P. Miller's website.
Log(Total Words)	The number of words (in 000's) contained in the complete 10-K filing from EDGAR (Bonsall IV et al., 2017). The natural logarithm rather than the raw number of words is used because of the skewness in the number of words across firms and some extreme values. In our analysis we utilize the "ngram" package from the Comprehensive R Archive Network (CRAN) to calculate the number of words in the parsed 10-K files.
Dependent variable	
Post-filing RMSE	The root mean squared error from a market model multiplied by 100. The model is estimated using trading days[6,28] relative to the 10-K file date, where a minimum of 10 observations are required to be included in the sample.
Other control variables	

(Continued)

Variable Definitions - *Continued*

Pre-filing alpha	The alpha from a market model using trading days [252, 6]. At least 60 observations of daily returns must be available to be included in the sample.
Pre-filing RMSE	The root mean squared error from a market model multiplied by 100. The model is estimated using trading days[-257,-6] relative to the 10-K filing date, where a minimum of 60 observations of daily returns must be available to be included in the sample.
Abs(filing period abnormal return)	The absolute value of the filing date excess return. The buy-and-hold return period is measured from the filing date(day 0) through day +1 minus the buy-and-hold return of the CRSP value-weighted index over the same 2-day period.
Log(Market capitalization)	The natural logarithm of the CRSP stock price times shares outstanding on the day prior to the 10k-K filing date(in $ millions).
Log(Book-to-market)	The natural log of book-to-market, following Fama and French (2001) using data from both COMPUSTAT (book value from most recent year prior to filing date) and CRSP (market value of equity). Firms with negative book value are removed from the sample.
NASDAQ dummy	Dummy variable set to one if the firm is listed on NASDAQ at the time of the 10-K filing, else zero.
Year fixed effects	Year dummies
Industry fixed effects	Fama and French (1997) 48 industry dummies
Firm fixed effects	Firm dummies
Industry-year fixed effects	Industry-year dummies using Hoberg and Phillips (2016) 300 FIC.
Industry volatility	Calculated as the mean post-filing RMSE among firms classified as peers by the Hoberg and Phillips (2016) TNIC in a given year. Requiring that we have pricing data available for at least 5 peers to be included in the sample.

A2 Multinomial Inverse Regression

In this section we present the framework for sentiment-preserving dimension reduction of text-data introduced by Taddy (2013). We apply this framework to investigate what bigrams are most associated with higher and lower readability scores.

First, clean the text data of the 10-K files. In this cleaning we remove punctuation, numbers and we convert all remaining text to lowercase. We further remove a set of stop words (e.g. but, and, to, is) that occur very frequently in English text. Lastly we stem the words with the Porter stemmer. This treats words that have the same meaning as the same word (e.g. decline, declines and declined become declin), and ensures that variations of the same word is not counted separately.

The next data preparation step is to parse clean text into informative language tokens. One way of doing this is to represent each document as a word-count vector, referred to as a "bag-of-words" representation. A more refined representation as described in Taddy (2013) is to represent each token as a count of bigrams. A bigram is defined as a sequence of two adjacent words, and can therefore provide more information to the context of its use compared to single words (unigrams) (Taddy, 2013). For example, we believe that the bigram "annual report" is more informative than "annual" and "report" viewed separately.

We create bigrams for each text document so that each document is represented as a given bigram count of $\mathrm{x}_i = [x_{i1} \ ... \ x_{ip}]$' for each p bigram in the vocabulary. In order to avoid terms that are rare and firm-specific we limit the number of tokens to those that appear in at least one percent of all documents, ending up with a vocabulary of 129.857 bigrams.

For each document, we have a sentiment variable y_i that is the score of each readability measure. A generic regression for $y_i|\mathrm{x}_i$ is computationally unfeasible for the large count of tokens, so it is necessary to simplify x_i. Taddy (2013) uses the multinomial distribution of x_i, implied by the exchangeability of token-counts, as basis for further dimension reduction. Taddy (2013) further propose an inverse regression approach wherein the inverse conditional distribution for text sentiment is used to obtain low dimensional document scores that preserve information relevant to the sentiment variable y_i.

As an example, let $\mathrm{x}_y = \sum_{i:y_i=y}$ for all $y \in Y$, the support for an ordered discrete sentiment variable. The sentiment variable is in this case the different readability measures. A basic multinomial inverse regression model is then

$$x_y \sim MN(q_y, m_y) \quad \text{with} \quad q_y = \frac{exp[a_j + y\varphi_j]}{\sum_{l=1}^{p} exp[a_l + y\varphi_l]} \text{ for } j = 1, .., p, \; y \in Y$$

Where each MN is a p-dimensional multinomial distribution with size $m_y = \sum_{i:y_i=y} m_i$ and probabilities $q_y = [q_{y1}, ..., q_{yp}]'$ that are a linear function of y through a logistic link.

The next step of the framework is obtaining the token coefficients . We apply the *dmr* function in R, from the *Distrom* package made by Taddy. This function returns the list of tokens with loadings, shown in A3. These loadings are obtained from fat tailed, sparsity-inducing Laplace priors. These Laplace priors are chosen from full regularization paths, to account for uncertainty about the appropriate level of variable-specific penalties.

The following step of the framework is estimating sufficient reduction (SR) scores from these token coefficients. With these parameters, Taddy (2013) shows that a sufficient reduction score for frequencies $\mathrm{f}_i = \frac{x_i}{m_i}$ is then:

$$z_i = \varphi' f_i \Rightarrow y_i \perp\!\!\!\perp x_i, m_i | z_i.$$

Given this sufficient reduction, z_i, full x_i is ignored and the model for further determining the text-sentiment relationship becomes a univariate regression problem, making the model sufficiently computationally efficient. We implement this regression in A4.1 to verify that the z-scores relate to the readability measures scoring of documents. For implementing the Taddy (2013) MNIR model we use the *srproj* function from the *textir* package for R introduced by Taddy (2013).

A3 Word lists, calculated following Taddy (2013)

Table A3.1: Top 40 Bigrams with Positive and Negative Loadings related to the Fog Index

The table displays the 40 bigrams with the top positive and negative loadings related to the Fog Index, using Taddy (2013). High positive loadings indicate that documents using the these bigrams are associated with higher levels of the Fog Index, and are thus considered less readable. Similarly, a negative loading is associated with a more readable document, as measured by the Fog Index. The loadings are calculated using the text of 70 106 annual reports from 1996-2016, and the associated Fog Index. The word list consist of bigrams of stemmed words, using the Porter stemmer.

	negative.bigrams	Loadings	Positive.bigrams	Loadings.1
1	u_s	-1.4735	peopl_unit	1.6238
2	total_acr	-1.1489	block_busi	1.6069
3	food_inc	-1.1246	physician_servic	1.5788
4	santa_fe	-1.1173	physician_group	1.4876
5	capac_ton	-1.1168	reinsur_insur	1.4835
6	apart_complex	-1.1022	cede_compani	1.4440
7	shall_mean	-1.0862	compani_reinsur	1.4214
8	robert_m	-1.0691	catastroph_reinsur	1.4172
9	execut_deliv	-1.0628	variabl_annuiti	1.4168
10	steel_manufactur	-1.0524	assum_reinsur	1.3944
11	store_expand	-1.0400	includ_reinsur	1.3845
12	store_fixtur	-1.0325	homeown_insur	1.3843
13	season_merchandis	-1.0200	reinsur_may	1.3766
14	store_rang	-1.0027	physician_practic	1.3744
15	r_brand	-1.0019	reinsur_loss	1.3732
16	steel_produc	-0.9996	facult_reinsur	1.3530
17	averag_squar	-0.9883	individu_life	1.3507
18	store_size	-0.9878	reinsur_program	1.3488
19	style_color	-0.9875	accid_year	1.3430
20	three_store	-0.9857	individu_famili	1.3426
21	shall_constitut	-0.9838	reserv_unpaid	1.3194
22	equival_begin	-0.9833	use_reinsur	1.3188
23	discount_store	-0.9805	surgeri_center	1.3161
24	store_averag	-0.9789	hipaa_privaci	1.3141
25	merchandis_price	-0.9774	reinsur_protect	1.3103
26	open_store	-0.9772	insur_segment	1.3040
27	depart_specialti	-0.9751	reinsur_market	1.2977
28	mill_oper	-0.9715	practic_medicin	1.2961
29	averag_store	-0.9699	reduc_medicar	1.2904
30	mortgag_million	-0.9691	agreement_reinsur	1.2899
31	bond_seri	-0.9672	reduct_medicar	1.2895
32	per_store	-0.9635	reinsur_busi	1.2890
33	secur_parti	-0.9626	certain_reinsur	1.2840
34	steel_product	-0.9613	physician_provid	1.2774
35	supermarket_chain	-0.9606	insur_exposur	1.2753
36	presid_merchandis	-0.9598	econom_clinic	1.2722
37	two_store	-0.9594	develop_pattern	1.2674
38	store_featur	-0.9593	underwrit_price	1.2627
39	store_end	-0.9583	term_reinsur	1.2623
40	merchandis_manufactur	-0.9553	support_insur	1.2620

Table A3.2: Top 40 Bigrams with Positive and Negative Loadings related to log(Total words)

The table displays the 40 bigrams with the top positive and negative loadings related log(Total words), using Taddy (2013). High positive loadings indicate that documents using the these bigrams are generally longer. Similarly, a negative loading is associated with shorter documents. The loadings are calculated using the text of 70 106 annual reports from 1996-2016, and the associated word count. The word list consist of bigrams of stemmed words, using the Porter stemmer.

	negative.bigrams	Loadings	Positive.bigrams	Loadings.1
1	page_registr	-2.6527	combin_note	3.6294
2	report_sharehold	-2.3959	credit_deriv	2.8903
3	compens_page	-2.3809	block_busi	2.8740
4	sharehold_fiscal	-2.3400	default_swap	2.7068
5	director_page	-2.3002	loss_billion	2.6607
6	page_proxi	-2.2524	variabl_annuiti	2.6178
7	analysi_page	-2.2220	state_louisiana	2.6116
8	page_compani	-2.2049	collater_post	2.5883
9	sharehold_year	-2.2001	credit_default	2.5537
10	report_stockhold	-2.1977	econom_capit	2.5504
11	caption_select	-2.1920	note_combin	2.5267
12	thereon_incorpor	-2.1648	counterparti_exposur	2.5259
13	page_caption	-2.1552	energi_suppli	2.5096
14	refer_page	-2.1428	account_valu	2.4751
15	data_page	-2.0984	net_deriv	2.4537
16	exhibit_inform	-2.0878	commerci_mortgage-back	2.4478
17	contain_page	-2.0552	deriv_counterparti	2.3887
18	caption_quarter	-2.0527	wholesal_energi	2.3860
19	transact_page	-1.9953	market-bas_rate	2.3670
20	sharehold_incorpor	-1.9755	credit_spread	2.3623
21	llp_date	-1.9649	credit_protect	2.3616
22	page_exhibit	-1.9597	collater_counterparti	2.3420
23	oper_page	-1.9294	nonperform_risk	2.2992
24	page_page	-1.9281	counterparti_net	2.2953
25	refer_select	-1.9231	non-perform_risk	2.2884
26	expressli_incorpor	-1.9120	deriv_associ	2.2882
27	inform_page	-1.8983	risk_committe	2.2774
28	busi_registr	-1.8855	credit_event	2.2700
29	page_annual	-1.8584	net_agreement	2.2698
30	registr_believ	-1.8576	energi_sale	2.2674
31	registr_annual	-1.8533	loss_mitig	2.2620
32	page_incorpor	-1.8213	power_price	2.2584
33	caption_common	-1.8208	expens_billion	2.2463
34	stockhold_fiscal	-1.8191	herein_addit	2.2357
35	februari_appear	-1.8052	emiss_allow	2.2352
36	financi_summari	-1.7921	chief_risk	2.2303
37	registr_busi	-1.7772	incom_billion	2.2272
38	stockhold_year	-1.7694	coal-fir_generat	2.2246
39	registr_also	-1.7658	manag_framework	2.2205
40	forth_page	-1.7330	energi_capac	2.2052

Table A3.3: Top 40 Bigrams with Positive and Negative Loadings related to log(File Size)

The table displays the 40 bigrams with the top positive and negative loadings related log(File size), using Taddy (2013). High positive loadings indicate that documents using the these bigrams generally have larger 10-K filings measured in megabytes. Similarly, a negative loading is associated with smaller filings. The loadings are calculated using the text of 70 106 annual reports from 1996-2016, and the associated file size. The word list consist of bigrams of stemmed words, using the Porter stemmer.

	negative bigrams	Loadings	Positive bigrams	Loadings
1	primari_fulli	-1.4415	volcker_rule	1.6224
2	compliant_compani	-1.3912	combin_note	1.4145
3	miscalcul_caus	-1.3786	basel_iii	1.3604
4	year_compliant	-1.3498	conserv_buffer	1.2729
5	text_compani	-1.3445	equiti_tier	1.2457
6	failur_miscalcul	-1.3440	capit_conserv	1.2347
7	year_y2k	-1.3406	implement_dodd-frank	1.1613
8	marwick_llp	-1.3384	stabil_oversight	1.1573
9	digit_rather	-1.3329	enforc_master	1.1389
10	accept_audit	-1.3196	implement_basel	1.1380
11	rather_four	-1.2906	section_dodd-frank	1.1255
12	non-it_system	-1.2879	tdr_loan	1.1191
13	y2k_issu	-1.2871	requir_dodd-frank	1.1169
14	year_problem	-1.2802	bureau_cfpb	1.1165
15	temporari_inabl	-1.2772	pursuant_dodd-frank	1.0949
16	calcul_primari	-1.2737	iii_capit	1.0885
17	becom_year	-1.2706	credit_impair	1.0576
18	window_nt	-1.2699	oversight_council	1.0519
19	address_year	-1.2674	subject_master	1.0489
20	non-inform_technolog	-1.2663	regul_dodd-frank	1.0358
21	two_digit	-1.2649	provis_dodd-frank	1.0292
22	readi_compani	-1.2297	addit_dodd-frank	1.0213
23	r_brand	-1.2175	cyber_attack	1.0174
24	primari_earn	-1.2120	disclosur_dispos	1.0167
25	c_text	-1.2072	evid_credit	1.0141
26	remedi_year	-1.1913	result_dodd-frank	1.0089
27	year_complianc	-1.1901	incom_entireti	1.0055
28	four_digit	-1.1843	restructur_tdrs	0.9997
29	thing_temporari	-1.1681	disclosur_offset	0.9895
30	text_follow	-1.1652	dodd-frank_act	0.9848
31	program_written	-1.1610	gaap_reclassifi	0.9837
32	transact_send	-1.1558	basel_committe	0.9814
33	date_field	-1.1456	collater_post	0.9765
34	internet_intranet	-1.1401	carryforward_similar	0.9725
35	waterhous_llp	-1.1339	cross-refer_disclosur	0.9675
36	year_readi	-1.1282	reclassifi_entireti	0.9655
37	rather_year	-1.1256	financi_reform	0.9617
38	written_use	-1.1252	net_arrang	0.9585
39	opinion_express	-1.1191	supervis_basel	0.9573
40	regard_year	-1.1174	signific_vie	0.9568

A4 Forward regressions

Table A4.1: Forward Regressions for Word Lists SR-Scores

The table summarize the results of a forward regression for each readability measure. Coefficients and t-stats for Z are averages across the five regressions from a five fold cross validation. The only control variable in regression (1),(3),(5) and (7) is the number of bigrams in the particular document. Regression (2),(4),(6) and (8) additionally includes year and firm fixed effects, with clustered standard errors by firm and year.

	Dependent variable:							
	Bog Index		Fog Index		log(Total Words)		log(File Size)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Z	0.51***	0.60***	0.62***	0.95***	0.36***	0.61***	0.66***	0.06**
	(75.86)	(8.62)	(84.04)	(17.25)	(71.48)	(24.12)	(94.19)	(1.92)
Control for length	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Year Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Note:								*p<0.1; **p<0.05; ***p<0.01

To test the relationship between the lists of bigrams and the readability measures, we implement a forward regression following Taddy (2013). In this regression we test whether the SR-scores based on the bigram loadings are related to the observed readability scores. For each simulation the data is divided in to 5 random groups. Each of these groups are then used as an out-of-sample testing set, while the four other groups are used as training data.

First, for each training sample the loadings for each bigram are calculated with the *dmr* function referenced in Taddy et al. (2015), using the readability score and bigram count vector for each document. These bigram loadings are then matched with the bigram count per document in the test sample to create SR-scores for each test sample. Finally, the observed readability scores of each document in the test sample is implemented as the dependent variable in a regression with the SR-scores as a explanatory variable. If a document contains many bigrams with high positive loadings, this will give a higher SR-score and vice versa. If these loadings are correctly weighted this regression will show a positive association between Z-scores and the observed readability measures. Thus, the validation of the bigram lists is based on their ability to produce z-scores that can predict the observed readability score. We implement the following regression for each readability measure:

$$Readability\ Score = \beta_0 + \beta z_i + m_i + \epsilon_i$$

where z_i is the SR-scores for each document, and m_i is a control for the length of the document.

Table A4 shows the relationship between z-scores and the readability measures. The z-scores for all readability measures are significantly linked to their respective observed readability scores, and this is robust to the inclusion of firm and year fixed effects. The economic significance is also considerable. For instance, one standard deviation increase in the z-score reflecting the Bog Index leads to an increase corresponding to 60% of the standard deviation in the observed Bog Index, when including firm and year fixed effects. The exception to this the file size when including firm and year fixed effects as shown in column (8), where much of the economic association between the Z-score and observed file size is removed when looking at the effect within a firm and year. The association is however still positive and significant at the 5% level. The results from this test suggest that the word lists are related to the observed readability scores.

A5 Hoberg Phillips industry classification

To determine the similarity between two firms, Hoberg and Phillips (2016) use the pairwise cosine similarity between the firms, based on what words they use in their product description. Consider the collection W of unique words used in the union of product descriptions[7]. The vocabulary of each firm i can then be represented by a W-vector P_i, where each element of the vector holds the value 1 if the word is present in the firms product description, and 0 if it is not. The vector is then normalized as follows:

$$V_i = \frac{P_i}{\sqrt{P_i \cdot P_i}} \qquad \forall i, j$$

The normalized vectors of all firms within a given year are then collected in a $N_t \times W$ matrix Q_t, where N_t is the number of firms in year t. Each row in Q_t now contain the normalized vector V_i for the year t. To construct the firm-to-firm network representation of industries Hoberg and Phillips (2016) then calculate the product cosine similarity of firm i and j as follows:

$$Product\,Cosine\,Similarity_{i,j} = (V_i \cdot V_j)$$

This results in an $N_t \times N_T$ matrix M_t for each year, where the entry in row i and column j represent the cosine similarity between firms i and j. A higher cosine similarity indicate that the firms use more of the same words in their product description.

A5.1 Fixed industry classifications(FIC)

Their next step is then to use the M_t matrices of similarities to classify each firm to the best fitting industry, but first they need to specify the set of industries to be hold fixed for all years must be created. This is done by taking the subsample of N single-segment

[7]For Hoberg and Phillips (2016) W is 61,146 unique nouns and proper nouns in 1996 and 55,605 in 2008.

firms [8] from the first year of full data. The industry classification is then set to have N industries, before computing the pairwise cosine similarities $I_{j,k}$ for each unique pair of industries j and k.

To reduce the set of industries Hoberg and Phillips (2016) take the maximum pairwise industry similarity as follows:

$$\max_{j,k,\, j \neq k} I_{j,k}$$

The two industries with the highest similarity is then combined, reducing the number of industries by one. The procedure is then repeated until the desired amount of industries is reached, recalculating the industry similarities each time.

When the set of industries are complete, Hoberg and Phillips (2016) start classifying the remaining firms to the industries, one year at a time. to do this, the pairwise cosine similarity between the respective firm and each industry is calculated. The firm is then assigned to the industry of which it holds the highest similarity to.

A5.2 Text-based Network Industry Classifications (TNIC)

Hoberg and Phillips (2016) also propose a additional industry classification scheme, where each firm has its own set of rivals. This classification scheme relaxes the notion of transitivity, namely that if firms B and C are in firm A's industry, then firms B and C are also in the same industry.

Instead of assigning each firm to a set of predefined industries, each firm is assigned a unique set of peers within each year. To determine which firms that should be classified as peers, the same pairwise similarity score of product descriptions that is used to determine the fixed industry classifications is used. Firms is classified as peers if the pairwise similarities with a given firm is above a given threshold. This threshold is set so that the coarseness of the TNIC matches the coarseness of three digit SIC codes. For example, the likelihood of two random firms in the CRSP/COMPUSTAT universe being in the same

[8] multiple-segment firms are identified using the Compustat segment database

three SIC industry is 2.05%. Thus, the threshold of which firms are classified as peers are set such that the likelihood of two random firms being deemed as peers is also 2.05%.

A6 Complementary Regressions Results

Table A6.1 present the results form using the method of Loughran and McDonald (2014) with Fama and French (1997) 48-industry dummies on the full sampler of 70 106 firm-year observations. The results are analog to the results of Table 5.1

Table A6.2 present the regression results from using the method of Bonsall IV et al. (2017) on the subsample of the first 10 years of observations in our sample. Comparable to Table 5.2.

Table A6.3 present the regression results from using the method of Bonsall IV et al. (2017) on the same subsample of 66 213 firm-year observations as Table 5.3.

Table A6.4 present the regression results from using the method of Bonsall IV et al. (2017) on the same subsample of 49 678 firm-year observations as Table 5.4.

Table A6.1: Post-Filing RMSE and Readability Measures Following Loughran and McDonald (2014)

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression include an intercept, calender year dummies, and Fama and French (1997) 48-industry dummies in addition to the presented control variables. For ease of interpretation all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and industry. All regressions include 70 106 firm-year observations during 1996-2016.

	Dependent variable:				
	Post-filing RMSE				
	(1)	(2)	(3)	(4)	(5)
Bog Index		0.032^{***} t = 3.977			
Log(File Size)			0.028^{**} t = 2.451		
Log(Total Words)				0.027^{**} t = 2.212	
Fog Index					0.010^{***} t = 2.929
Pre-filing alpha	-0.081^{***} t = −3.814	-0.080^{***} t = −3.802	-0.081^{***} t = −3.789	-0.080^{***} t = −3.723	-0.081^{***} t = −3.802
Pre-filing RMSE	0.459^{***} t = 12.863	0.455^{***} t = 12.553	0.458^{***} t = 12.779	0.454^{***} t = 12.298	0.458^{***} t = 12.825
Abs(abnormal return)	0.111^{***} t = 11.925	0.110^{***} t = 11.809	0.111^{***} t = 11.954	0.110^{***} t = 11.881	0.111^{***} t = 11.918
Log(market capitalization)	-0.079^{***} t = −4.798	-0.084^{***} t = −4.899	-0.085^{***} t = −5.115	-0.084^{***} t = −4.815	-0.081^{***} t = −4.864
Log(book-to-market)	-0.048^{**} t = −2.538	-0.049^{***} t = −2.608	-0.050^{***} t = −2.593	-0.049^{**} t = −2.568	-0.048^{**} t = −2.535
NASDAQ dummy	0.046^{***} t = 3.282	0.046^{***} t = 3.347	0.046^{***} t = 3.283	0.045^{***} t = 3.345	0.045^{***} t = 3.272
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	70,106	70,106	70,106	70,106	70,106
R^2	0.444	0.445	0.444	0.445	0.444
Note:					*p<0.1; **p<0.05; ***p<0.01

Table A6.2: Post-Filing RMSE and Readability Measures Following Bonsall IV et al. (2017) for the Years 1996 to 2005

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression include an intercept, calender year dummies and firm dummies in addition to the presented control variables. For ease of interpretation all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 38 310 firm-year observations during 1996-2005.

	Dependent variable:			
	Post-filing RMSE			
	(1)	(2)	(3)	(4)
Bog Index	0.024*			
	t = 1.663			
log(File Size)		−0.009		
		t = −0.933		
log(Total Words)			0.005	
			t = 0.616	
Fog Index				0.011
				t = 1.450
Pre-filing alpha	−0.043*	−0.043*	−0.043*	−0.043*
	t = −1.718	t = −1.723	t = −1.715	t = −1.713
Pre-filing RMSE	0.309***	0.310***	0.309***	0.310***
	t = 7.425	t = 7.435	t = 7.394	t = 7.416
Abs(abnormal return)	0.101***	0.101***	0.101***	0.101***
	t = 11.156	t = 11.161	t = 11.131	t = 11.162
Log(Market capitalization)	−0.070	−0.070	−0.071	−0.071
	t = −1.156	t = −1.166	t = −1.181	t = −1.167
Log(book-to-market)	−0.045*	−0.044*	−0.045*	−0.044*
	t = −1.876	t = −1.856	t = −1.875	t = −1.867
NASDAQ dummy	0.041**	0.040**	0.041**	0.041**
	t = 2.416	t = 2.370	t = 2.387	t = 2.399
Firm Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	38,310	38,310	38,310	38,310
Adjusted R^2	0.521	0.520	0.520	0.520
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A6.3: Post-Filing RMSE and Readability Measures Following Bonsall IV et al. (2017) on the Same Subsample as Table 5.3

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression include an intercept, calender year dummies and firm dummies in addition to the presented control variables. For ease of interpretation all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 66,213 firm-year observations during 1996-2016.

	Dependent variable:			
	Post-filing RMSE			
	(1)	(2)	(3)	(4)
Bog Index	0.031^{***}			
	t = 2.690			
Log(File Size)		0.027^{**}		
		t = 2.041		
Log(Total Words)			0.023^{**}	
			t = 2.566	
Fog Index				0.007
				t = 1.491
Pre-filing alpha	-0.045^{**}	-0.045^{**}	-0.045^{**}	-0.045^{**}
	t = −2.451	t = −2.447	t = −2.445	t = −2.449
Pre-filing RMSE	0.373^{***}	0.374^{***}	0.373^{***}	0.374^{***}
	t = 13.172	t = 13.339	t = 13.163	t = 13.356
Abs(abnormal return)	0.102^{***}	0.102^{***}	0.102^{***}	0.102^{***}
	t = 14.625	t = 14.675	t = 14.681	t = 14.666
Log(market capitalization)	-0.113^{***}	-0.115^{***}	-0.114^{***}	-0.113^{***}
	t = −3.066	t = −3.132	t = −3.120	t = −3.036
Log(book-to-market)	−0.033	−0.032	−0.033	−0.032
	t = −1.470	t = −1.463	t = −1.461	t = −1.435
NASDAQ dummy	0.027^{***}	0.027^{***}	0.027^{***}	0.027^{***}
	t = 2.624	t = 2.614	t = 2.637	t = 2.605
Firm Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	66,213	66,213	66,213	66,213
Adjusted R^2	0.561	0.561	0.561	0.561
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A6.4: Post-Filing RMSE and Readability Measures Following Bonsall IV et al. (2017) on the Same Subsample as Table 5.4

The dependent variable in each regression is the market model RMSE for trading days [6,28]. See Appendix for detailed variable definitions. Each regression include an intercept, calender year dummies and firm dummies in addition to the presented control variables. For ease of interpretation all regression variables are standardized with a mean of zero and standard deviation of one. t-statistics are displayed under each coefficient, with standard errors clustered by year and firm. All regressions include 49 678 firm-year observations during 1996-2016.

	Dependent variable:			
	Post-filing RMSE			
	(1)	(2)	(3)	(4)
Bog Index	0.026**			
	t = 1.963			
Log(File Size)		0.031**		
		t = 2.314		
Log(Total Words)			0.020**	
			t = 2.377	
Fog Index				−0.001
				t = −0.187
Pre-filing alpha	−0.046**	−0.046**	−0.046**	−0.046**
	t = −2.339	t = −2.335	t = −2.334	t = −2.353
Pre-filing RMSE	0.384***	0.384***	0.383***	0.385***
	t = 10.542	t = 10.642	t = 10.492	t = 10.656
Abs(abnormal return)	0.098***	0.098***	0.098***	0.098***
	t = 13.100	t = 13.138	t = 13.127	t = 13.110
Log(market capitalization)	−0.078*	−0.080**	−0.079*	−0.077*
	t = −1.910	t = −1.969	t = −1.949	t = −1.886
Log(Book-to-Market)	−0.040*	−0.040*	−0.040*	−0.039*
	t = −1.793	t = −1.818	t = −1.801	t = −1.773
NASDAQ dummy	0.020*	0.020*	0.020*	0.020*
	t = 1.856	t = 1.827	t = 1.862	t = 1.835
Firm Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	49,678	49,678	49,678	49,678
Adjusted R^2	0.579	0.579	0.579	0.579
Note:				*p<0.1; **p<0.05; ***p<0.01