# Annual Report Readability, Current Earnings, and Earnings Persistence *

Feng Li

Ross School of Business, University of Michigan

701 Tappan St., Ann Arbor, MI 48109

Phone: (734)936-2771

Email: Feng@umich.edu

First Draft: July 2005
This Draft: September 15, 2006

*The paper was previously titled "Annual report readability, earnings, and stock returns". I acknowledge the financial support of the Harry Jones Endowment for Research on Earnings Quality at the Ross School of Business, University of Michigan. I thank Rob Bloomfield, Daniel Cohen, Ilia Dichev, Scott Richardson, Doug Skinner, Suraj Srinivasan, Franco Wong, and the participants of the Hosmer Lunch Workshop at the University of Michigan for comments. The comments and suggestions of Bob Holthausen (the editor) and an anonymous referee greatly improved the paper. All errors remain my own.

**Abstract**

This paper examines the relationship between annual report readability and firm performance and earnings persistence. This is motivated by the Securities and Exchange Commission's plain English disclosure regulations that attempt to make corporate disclosures easier to read for ordinary investors. I measure the readability of public company annual reports using both the Fog Index from computational linguistics and the length of the document. I find that the annual reports of firms with lower earnings are harder to read (i.e., they have higher Fog and are longer). Moreover, the positive earnings of firms with annual reports that are easier to read are more persistent. This suggests that managers may be opportunistically choosing the readability of annual reports to hide adverse information from investors.

# 1 Introduction

Ever since the passage of the Securities Act of 1933, the Securities and Exchange Commission ("SEC") has made consistent efforts to make the disclosure documents of public companies more readable(Firtel (1999)). The most recent of these efforts is the plain English disclosure rules adopted by the SEC on January 22, 1998. The underlying argument for the plain English disclosure regulation is that (1) firms could use vague language and format in disclosure to hide adverse information, and (2) the average investors may not understand complex documents and this could result in capital market inefficiency.

The relevance of the regulation, however, is not straightforward. First, there are other information sources (such as financial analyst reports) for investors. Depending on whether the different information sources are complements or substitutes, annual report readability may or may not be relevant. Second, some critics contend that disclosure should primarily be geared towards sophisticated investors due to the complicated nature of technical and financial information (Firtel (1999)). Finally, to the extent that the marginal investors are sophisticated and understand complex disclosure, stock price may not be distorted even if complicated language and format are used.

This paper attempts to provide the first large sample evidence on the readability and other lexical features of corporate disclosure. More specifically, I ask the following questions: Is there a relation between a company's annual report readability and its current performance? What is the implication of disclosure readability for future performance and earnings persistence?

If disclosure readability is strategically used by managers to hide adverse information, a relationship between firm performance and readability would be expected. This management opportunism story argues that managers have incentives to obfuscate information when the current performance is bad (Bloomfield (2002)). Given that the annual report contains detailed financial numbers on historical firm performance, however, the marginal benefit of using disclosure readability to hide current poor performance seems small. Hence, I also examine the implication of disclosure readability for future performance. In particular, I

examine whether the positive earnings of firms with more complex annual reports are less persistent and whether the negative earnings of these firms are more persistent in the next several years.

I empirically measure annual report readability using two variables. The first variable is the Fog Index from computational linguistics based on syntactical textual features (such as words per sentence and syllables per word) in the 10-K filing. The intuition is that, everything else equal, more syllables per word or more words per sentence make a document harder to read. The second measure of readability is the annual report length based on the intuition that longer documents are more deterring and require higher information costs. Using a sample with more than 50,000 firm-years, I find that firms with lower earnings tend to file annual reports that are more difficult to read; an increase (decrease) in earnings from the previous year also results in annual reports that are easier (more difficult) to read compared with the previous year's reports. This effect holds after controlling for other firm and industry specific factors. However, although this effect is statistically significant, the economic magnitude is small.

I find that annual report readability is related to earnings persistence. Firms with more complicated annual reports have a lower persistence of earnings when they are profitable. The effect is significant both economically and statistically. An inter-quartile change in readability has a similar impact on positive earnings persistence comparable to the effect of an inter-quartile change in the absolute amount of accruals.

Other lexical features of the annual reports also have systematic associations with earnings persistence, confirming the findings based on readability. For profitable firms, longer annual reports are associated with lower earnings persistence, higher frequency of causation words (such as "because") in the MD&A section is associated with less persistent earnings, more positive emotion words (relative to negative emotion words) are associated with more persistent earnings, and higher frequency of future tense verbs (relative to past/present tense verbs) indicates lower earnings persistence. On the other hand, loss firms with more positive emotion words (relative to negative emotion words) in their MD&A have less persistent

earnings.

Taken together, the evidence in this paper suggests that, consistent with the motivation behind the plain English disclosure regulation of the Securities and Exchange Commission, managers may be opportunistically choosing the readability of annual reports to hide adverse information from investors.

Certain caveats are in order. My empirical measures of annual report readability capture only part of the many requirements of the SEC plain English rules. Hence, it does not speak to other parts of the regulation such as formatting of the prospectuses. Due to data availability, my sample period includes only the post-1994 years. It is possible that the cross-sectional variation in annual report readability may be smaller during this period, resulting in lower powered tests.

This paper contributes to the literature in the three ways. First, there is extensive research on the determinants and consequences of accounting choices (Fields, Lys, and Vincent (2001)) and quality of disclosure (Healy and Palepu (2001)). But there is no large-sample study of the determinants and consequences of annual report readability and other lexical properties even though the SEC has been advocating disclosure based on more plain English (SEC (1998)). This paper is the first large-sample study to examine the cross-sectional variation in annual report readability and its implications for current earnings and earnings persistence. It extends the strategic reporting literature (e.g., Schrand and Walther (2000)) by showing that disclosure readability and lexical features may be used strategically by managers. This lends further support to the "incomplete revelation hypothesis" in Bloomfield (2002)) and sheds light on the relevance of the plain English disclosure regulation.

Second, much of the empirical literature on corporate disclosure quality has focused on the determinants and consequences of the amount of disclosure (e.g., Miller (2002)). Most papers have small sample sizes, as the disclosure is in general manually coded. Annual report readability and lexical features capture the characteristics, rather than the content, of disclosure. To the extent that more complicated annual reports increase the information processing cost for investors and hence have lower disclosure quality, this paper provides a

new empirical measure of disclosure quality that can be studied in a large sample.

Finally, there is extensive research on earnings quality (see Dechow and Schrand (2004) for a comprehensive review). But prior research in general does not study the association between firm disclosure quality and earnings quality.[1] While many papers explicitly link firm performance with disclosure quality (e.g., Lang and Lundholm (1993)) and other papers use earnings quality as a proxy for disclosure quality (e.g., Francis, LaFond, Olsson, and Schipper (2005)), few examine the implication of disclosure quality for future earnings. This paper extends this literature by showing that the quality of disclosure is correlated with earnings persistence and contains information about earnings quality.

The remainder of the paper proceeds as follows. I discuss prior literature and hypotheses in Section 2 and empirical measures of annual report readability in Section 3. I present the basic empirical findings on readability in Section 4 and other lexical properties of annual report in Section 5. I explore some additional empirical tests in Section 6. Section 7 concludes.

# 2 Literature and Hypotheses

The SEC has continually attempted to make public company prospectus more readable and understandable. In several Securities Act Releases after the 1933 Securities Act, it encouraged greater clarity in the disclosure documents with an emphasis on not compromising full and fair disclosure (Firtel (1999)). In 1967, the SEC constituted an internal study group to examine and make recommendations for improving its disclosure regime. This resulted in the 1969 "*Wheat Report*". Among other findings, *the Wheat Report* noted that the average investor could not readily understand the complicated prospectuses and therefore recommended that companies avoid unnecessarily complex, lengthy or verbose writing.

In October 1998, the SEC adopted new plain English disclosure rules that require the

[1] One exception is Francis, Nanda, and Olsson (2005), who examine the relation between voluntary disclosure and accrual quality.

usage of plain English in the drafting and format of all prospectuses in registered public offerings by domestic and foreign issuers. The SEC's Investor Ed Office published and posted on its website "A plain English handbook, how to create clear SEC disclosure documents" which provides practical tips for disclosure documents. For instance, when drafting the front and back cover pages, the summary and risk factors sections, an issuer must comply with the following six basic principles: short sentences; definite, concrete, everyday language; active voice; tabular presentation or bullet lists for complex material, whenever possible; no legal jargon or highly technical business terms; and no double negatives. More recently, the SEC has taken several steps in making the disclosure of mutual funds more readable (Glassman (2005)).

## 2.1 Literature

Given the importance of the plain English disclosure regulation, surprisingly, there is little large sample empirical evidence on its relevance. Jones and Shoemaker (1994) reviewed 32 studies in the fields of accounting, business communication, and management which study the readability of annual report narratives (26 studies), tax law (3 studies), or accounting textbook (3 studies).

Most studies try to assess the reading ease of the annual report and its components. For instance, Smith and Smith (1971) study the readability of the financial statements footnotes of Fortune 50 companies and conclude that the readability level of the notes is restrictive. Healy (1977) studies the reading ease of the footnotes to the financial statements of 50 New Zealand firms. Lebar (1982) studies the Forms 10-Ks, annual reports, and press release by 10 NYSE firms in 1978 and compares the differences in topics and information between them. The general conclusion from these studies is that corporate annual reports are very difficult to read and may be classified as technical literature which risks "being inaccessible to a large proportion of private lay shareholders" (Jones and Shoemaker (1994)). Some studies also specifically investigate whether annual reports have become more difficult to read over time (e.g., Soper and Dolphin (1964); Barnett and Leoffler (1979)) and the evidence is rather

mixed (Jones and Shoemaker (1994)).

Other studies examine the association between readability and other variables, including the identity of the external auditor (Smith and Smith (1971) and Barnett and Leoffler (1979)) and corporate profitability (Courtis (1986); Baker and Kare (1992); Subramanian, Insley, and Blackwell (1993)). The evidence is again mixed and inconclusive. For instance, Courtis (1986) finds no strong correlation between readability and net profits and return on capital. However, Subramanian, Insley, and Blackwell (1993) find annual reports of profitable firms are significantly easier to read than those of poor performers.

The sample sizes of the previous studies, however, are very small. Only two of the thirty-two studies reviewed by Jones and Shoemaker (1994) have a sample size slightly larger than 100. Among the sixteen papers examined in Table I of Clatworthy and Jones (2001), fourteen have a sample size of 50 or smaller and the largest sample size is 120.

In this paper, I extend this literature using a large sample with a particular focus on the association between annual report readability and firm performance, future earnings, and earnings persistence.

## 2.2 The Implications of Annual Report Readability

### 2.2.1 Current Performance

As a motivation for the plain English disclosures regulation, the SEC argues that disclosures easier to understand can better inform investors (SEC (1998)). While the SEC may be more worried about boilerplate legalese, perhaps what's more relevant to investors is possible management obfuscation of information through complex disclosures. The maintained assumption of the managerial obfuscation argument (i.e., management is less forthcoming in disclosing information when the firm is performing poorly) is the "incomplete revelation hypothesis". Because the information that is more costly to process is perhaps less completely reflected in market prices (Grossman and Stiglitz (1980) and Bloomfield (2002)), managers may want to strategically hide bad information through less transparent disclosure. In particular, Bloomfield (2002) argues that managers make many decisions motivated, at least

partly, by a desire to make it harder for investors to uncover information that the managers do not want to affect the firms' stock prices. Therefore, by increasing the processing cost of adverse information, managers hope that it is not reflected in stock prices or reflected in prices with a delay. Current empirical evidence seems to support the strategic reporting and incomplete revelation hypotheses: managers announce *pro forma* earnings numbers that emphasize improvements relative to their own strategically chosen benchmarks, while making it more difficult for investors to observe other measures of performance (Schrand and Walther (2000)); the special items recognized as a line item on the income statement are also less persistent than those only disclosed in the footnotes (Riedl and Srinivasan (2005)). The managerial obfuscation story thus predicts a negative relation between firm current performance and annual report complexity.

However, this hypothesized relation between disclosure readability and a firm's current performance may not be significant. First, corporate annual reports contain a lot of financial information about current and historical performance. Hence, the benefit to the managers of making the annual reports harder to read in order to hide adverse information about current performance seems small. Second, if the good current earnings is (partially) due to strategic manipulation, then managers may not necessarily want to make the annual reports easier to read when the reported earnings is "good".

For these reasons, the relation between annual report readability and current performance is not clear-cut and the benefit of managerial strategic reporting using annual report readability is more likely to lie in hiding or delaying *future* adverse information. Therefore, I further examine the implication of annual report readability for future performance with a particular focus on earnings persistence.

### 2.2.2 Future Performance

The intuition on the relation between disclosure quality and a firm's current performance can be extended to future performance. Opportunistic managers may have incentives to make the annual report harder to read, if good earnings of this year are not persistent or if poor

earnings are very persistent. On the other hand, firms with better future performance may want to disclose information more transparently to lower the information processing cost and distinguish themselves from the "lemons". In other words, to the extent that complicated annual reports can hide the transitory nature of the good news or the permanent nature of the bad news by increasing investors' information processing cost, the management obfuscation hypothesis predicts that the profits (losses) of firms with more complex annual reports are less (more) persistent.

Most prior studies on disclosure either examine the relation between disclosure quality and firm performance (e.g., Lang and Lundholm (1993)) or use earnings quality as a proxy for disclosure quality (e.g, Francis, LaFond, Olsson, and Schipper (2005) and Cohen (2005)). A few papers study the relation between disclosure quality and earnings quality: Francis, Nanda, and Olsson (2005) find a positive relationship between voluntary disclosure quality and the accruals quality; Riedl and Srinivasan (2005) examine the implication for earnings persistence of whether special items are recognized as a line item on the income statement or only disclosed in the footnotes. I extend this literature by examining the implication of disclosure readability for earnings persistence.

# 3 Data and Empirical Measures of Annual Report Readability

## 3.1 Sample

I collect my sample as follows: (1) I start with the intersection of CRSP-COMPUSTAT firm-years. (2) I then manually match GVKEY (from COMPUSTAT) and PERMNO (from CRSP) with the Central Index Key (CIK) used by SEC online Edgar system. Firms without matching CIK are dropped. (3) I download from Edgar the 10-K filing for every remaining firm-year. Those firm-years that do not have electronic 10-K filings on Edgar are then

excluded.[2] (4) For each 10-K file, all the heading items, paragraphs that have fewer than one line, and tables are deleted and those 10-K filings that have less than 3000 words or 100 lines of remaining texts are dropped. The calculation of the annual report readability is based on the remaining material. Details of these steps are presented in Appendix 1. Notice that it is important to delete the tables and financial statements in this step, since the readability indices are designed for text rather than numbers or tables. (5) Finally, firm-years that have operating earnings (scaled by book value assets) greater than 1 or less than -1 are deleted from the sample. This yields a sample of 55,719 firm-years with annual report filing date between 1994 to 2004. Since most of the firms have December fiscal year end, my sample mainly covers fiscal years 1993 to 2003.

## 3.2 The Readability Measures

I use two statistics to measure the annual report readability. The first is the Fog Index from computational linguistics. The Fog index, developed by Robert Gunning, is a well known and simple formula for measuring readability. Assuming that the text is well formed and logical, it captures text complexity as a function of syllables per word and words per sentence. [3] The index indicates the number of years of formal education a reader of average intelligence would need to read the text once and understand that piece of writing with its word sentence workload. It is calculated in the following way:

$$Fog = (words\_per\_sentence + percent\_of\_complex\_words) * 0.4, \quad (1)$$

[2]SEC has electronic Edgar filing available online from 1994.

[3]There are two other popular measures of readability: the Kincaid index and the Flesch Reading Ease Index. The Kincaid Index, also referred to as the Flesch-Kincaid formula and calculated as (11.8 * syllables_per_word) + (0.39 * words_per_sentence) - 15.59, rates text on U.S. grade school level. So a score of 8.0 means that the document can be understood by an eighth grader. The Flesch Reading Ease which rates text on a 100 point scale and is calculated as 206.835 - (1.015 * words per sentence) - (84.6 * syllables per word). The higher the Flesch Reading Ease, the easier is the text. The empirical results based on the Kincaid Index and the Flesch Index are similar to those based on Fog index and are therefore unreported. For more information about the readability measures, see http://www.plainlanguage.com/Resources/readability.html.

where complex words are defined as words with three syllables or more. The relation between Fog and reading ease is as follows: $FOG >= 18$ means the text is unreadable; 14-18 (difficult); 12-14 (ideal); 10-12 (acceptable); and 8-10 (childish).

The second measure I use to capture annual report readability is the length of the document. Because the information processing cost of longer documents is presumably higher, everything else equal, longer documents seem more deterring and more difficult to read. Therefore, the length of annual report could be strategically used by managers to make annual report less transparent and hide adverse information from investors. The SEC has consistently suggest companies avoid lengthy sentences and documents (SEC (1998)). Practitioners also use lengthy document as an example of bad and complex disclosure (e.g., Barker (2002)). There are pros and cons of using the length of a document as a measure of disclosure complexity. The advantage is that it is easy to calculate and understand. Compared with the readability indices, the disadvantage of the document length as a measure of readability is that it is more likely to be correlated with the amount of the disclosure. I define the length of annual report as:

$$Length = log(NWords), \tag{2}$$

where $NWords$ is the number of words in the document. The natural logarithm rather than the raw number of words is used because of the skewness in the number of words across firms and the few extreme values.

I use the Lingua::EN:Fathom package of PERL language to analyze the raw 10-K files and calculate the $Fog$ and $Length$. The Appendix gives the details of the calculation.[4] This program has been used in various fields including information science and business communication. Examples include Collins-Thompson and Callan (2005) and Muresan, Cole, Smith, Liu, and Belkin (2006).

To check the validity of the PERL program in calculating $Fog$, I first compare the numbers reported in this study with those from other studies. Smith and Smith (1971) manually calculate the Flesch Reading Index of some randomly selected footnotes of the

[4]For more information, see http://search.cpan.org/dist/Lingua-EN-Fathom/lib/Lingua/EN/Fathom.pm.

50 biggest Fortune companies. The mean of the Flesch Index per their calculation is 23.49 (Table II of Smith and Smith (1971)). For my sample, the mean and median of the Flesch Index calculated using PERL program are 24.44 and 24.63, which are pretty close to their manually calculated number.

A second way of checking the validity of the calculation is to compare it with manual calculation or other computer program using the same text. MS WORD can report the Kincaid Index. However, for unexplained reasons, Microsoft's version of the Kincaid Index does not score above grade 12, although the original formula scored up to a graduate school level. Since any grade level above 12 will be reported as grade 12, documents at a graduate school reading level will be reported as grade 12 - a measurement error of about 7 grades.[5] For this reason, I randomly select 3 paragraphs from 10 annual reports and count the number of words per sentence and syllables per word manually. The difference between the results from the manual calculation and the PERL programs is smaller than 5% in most cases, confirming the validity of the program.

Overall, I believe that the programs used in this paper should measure the readability reasonably well. There is no reason to believe that any measurement error is systematic and biases the results.

One concern regarding using syntactical features such as the Fog Index to measure readability is that they may not reflect actual comprehension difficulty. However, this concern is more problematic if researchers want to assess the *absolute* level of readability (Jones and Shoemaker (1994)). The focus here is on the relative readability of the annual reports in a cross-section. Hence, while still a caveat, the concern is less worrisome.

I calculate $Fog$ and $Length$ for both the whole annual report and sub-sections of the file. In particular, I focus on two sub-sections: the MD&A (Management Discussions and

[5]There is evidence that previous research relying on Word's Flesch-Kincaid formula seriously underestimates a document's grade level and all of the research done on health materials using Word's Flesch-Kincaid is seriously flawed. Several readability researchers have contacted Microsoft about this problem, but the company has neither acknowledged the problem nor fixed it. See the Reader Feedback section of this page: http://www.wats.ca/resources/determiningreadability/1.

Analysis) and the Notes to the financial statements. The MD&A section contains the discussion by managers of past performance and future outlook and Notes to financial statements have detailed assumptions behind the reported financial numbers. Details of extracting the sections electronically are presented in Appendix 2. Companies use different formats in their annual reports and this electronic extracting of MD&A and Notes are certainly not perfect. However, tests based on 50 randomly selected annual reports show that the algorithms can do a very reasonable job. I require the MD&A section to have at least 100 words and the Notes section to have at least 1000 words to be included in the analysis.

## 3.3 Summary Statistics

Table 1 Panel A presents the summary statistics of the sample. Overall, the annual reports of public companies are very difficult to read. The mean and median Fog Index of the whole annual report are 19.4 and 19.2 respectively, which are "unreadable" according to the usual interpretation of the index. The mean (median) *Length* is 10.08 (10.05) and this translates into a mean (median) of 31,034 (23,122) words. To provide a benchmark, I check the readability index for the articles from *the Wall Street Journal*. I download all the Editorials from the June 2005 issues of *the Wall Street Journal*. On average, these Editorials have a Fog of 15.2 and are much shorter, suggesting they are much easier to read than a typical annual report.

The standard deviation and the inter-quartile range of the $Fog$ ($Length$) of the 10-K filings in my sample are 1.4 (1.4) and 0.7 (0.9) respectively. This variation seems substantial. For instance, the difference in the Fog index between Reader's Digest and the TIME magazine is about 2.[6] The variation in year-by-year change in $Fog$ and $Length$ is not small either. The standard deviation of the change in Fog Index is 1.46 and that of $Length$ is 0.66. The 25th and the 75th percentile of year-to-year change in Fog are -0.59 and 0.65 respectively.

Panel A also presents the readability of the MD&A and the notes to the financial statements. The MD&A section of the annual report is much easier to read than the document

[6]Source: http://en.wikipedia.org/wiki/Fog_index#Typical_Gunning-Fog_indices_of_selected_magazines.

as a whole, with the mean (median) Fog index being 18.23 (17.98). Moreover, the variation in the MD&A readability is much bigger than the whole annual report with the standard deviation of $Fog$ being 2.55 and the inter-quartile range being about 2.8. The Notes to the financial statements have a mean Fog of 18.96 and a median of 18.83 and are slightly easier to read than the annual report as a whole. The variation is also comparable to that of the whole annual report. The median number of words of the whole annual report, the MD&A section, and the Notes section are 23122, 3325, and 6135 respectively.

Figure 1 A plots the median level of $Fog$ and $Length$ of the annual reports for the sample firms over time.[7] Interestingly, there is an obvious drop in $Fog$ in the years immediately after 1999, suggesting that the plain English disclosure regulation of 1998 might make companies take efforts to make their annual reports more readable. However, this trend reverses dramatically after 2002 and the annual reports filed by public firms seem to become even more difficult to read compared with the pre-1998 years. It would be interesting to get a longer time-series of data to examine whether this is related to the Sarbanes-Oxley Act regulation. In contrast, the $Length$ of the annual reports experienced a steady increase over time.

Figure 1 B and Figure 1 C plot the median level of $Fog$ and $Length$ of the MD&A and the Notes sections. The drop in year 2000 of the readability of the whole annual report observed in Figure 1 A primarily comes from the MD&A section, but not the Notes to financial statements. Both the MD&A and the Notes sections experienced dramatic increase in $Fog$ in 2003 and 2004.

Panel B of Table 1 presents the Pearson correlations of $Fog$ and $Length$ of the annual reports with some firm characteristics. There is a significant correlation between $Fog$ and $length$ of the whole annual reports with a Pearson correlation coefficient of 0.377. The $Fog$ of the Notes to the financial statements is also positively correlated with its $Length$ (Pearson correlation coefficient 0.383). However, the $Fog$ of the MD&A section has a negative association with its $Length$ (Pearson correlation coefficient of -0.189).

[7] The same graph (unreported) based on a constant sample, defined as firms with at least 8 years of data between 1994 and 2004, shows that the same time-series pattern is also seen in a constant sample.

There is strong correlation (economically and statistically) between the readability of MD&A section, Notes to financial statements, and the annual report as a whole. The Pearson correlation coefficient between Fog of the whole annual report and the MD&A Fog (the Notes Fog) is 0.368 (0.599). The correlation coefficient of MD&A Fog and the Notes Fog is 0.227.

Overall, bigger firms tend to have annual reports that are more difficult to read, as evidenced by the correlation coefficient of 0.007 and 0.263 between $Fog$ and $Length$ and firm size. Growth firms (firms with higher market-to-book ratio) tend to have annual reports with higher $Fog$, with the Pearson correlation coefficients between market-to-book and Fog being 0.014, but growth firms do not appear to differ in length from low market-to-book firms (Pearson coefficient between $Length$ and market-to-book of -0.006 and statistically insignificant).

From Panel C of Table 1, the five 2-digit SIC industries with the highest annual report Fog are Insurance Agents (2-digit SIC code 64), Health Services (80), Insurance Carriers (63), Electric and Gas (49), and Building Construction (15); the five industries with the lowest Fog are Stone, Clay, Glass, and Concrete Products (32), Transportation by Air (45), Leather and Leather Products (31), Apparel and Accessory Stores (56), and Food and Kindred Products(20). Firms in Communications (48), Insurance Carrier (63), Hotel (70), Petroleum Refining (29), and Electric, Gas, and Sanitory Services (49) have the longest annual reports. In addition, financial companies tend to have longer MD&A section, as Security and Commodity Brokers (62), Insurance Carriers (63), and Depository Institutions (60) are among the five industries with the longest MD&A section.

Panel D shows the persistence of annual report readability for firms in the first and fifth quintiles of $Fog$ and $Length$. Every year, firms are sorted into five quintiles based on $Fog$ or $Length$. For firms in the first and fifth quintiles, I track their readability level in the next three years. For instance, there are 11,479 (100%) firm-years in the fifth quintile of Fog in year 0. In the next year, 44.60% of these firms still remain in the fifth quintile, 24.57% switch to quintile 4, 14.17% are in quintile 3, 10.00% are in quintile 2 and 6.65% go to quintile 1.

Overall, there seems to be some time-series variation in annual report readability. Of the firms in the fifth quintile of Fog in year 0, only about 61% stay in quintiles 4 and 5 and the rest belong to the first three quintiles in year 3. Unreported results indicate similar persistence in the readability of MD&A section and the Notes section.

## 3.4 Determinants of annual report readability

This section discusses the (non-strategic) determinants of annual report readability. I explore the determinants of annual report readability in a multivariate regression setting. The implicit assumption here is that the relation between firm performance and readability is strategic.[8] Ex ante, there are many factors that might affect annual report readability non-strategically. It is important to empirically document the determinants and control for them in my later empirical tests. The factors examined here include the following variables:

- size: Size captures many aspects of a firm's operation and business environment. For instance, the accounting literature has used firm size to proxy for a firm's political cost (e.g., Watts and Zimmerman (1986)). Hence, I include $SIZE$, defined as logarithm of the market value of equity at the end of the fiscal year, as a variable to explain annual report readability. Ex ante, I expect bigger firms to have longer and more complex annual reports.

- Market-to-book: High market-to-book firms are different from low market-to-book firms in many aspects, including the investment opportunity set and growth potential. Market-to-book ratio ($MTB$), defined as the market value of equity plus book value of liability and divided by the book value of total assets at the end of the fiscal year end, is included as a potential determinant of annual report readability. Growth firms

[8] A relation between performance and readability is certainly consistent with the managerial obfuscation story. However, with the current research design in this paper, it is difficult to separate it from a simple association story. In Section 6.1, I attempt to provide some preliminary evidence to distinguish between them.

may have more complex and uncertain business models and thus more complex annual reports.

- Firm age: Old firms may exhibit different annual report readability because there is less information asymmetry and information uncertainty for these firms. If investors are more familiar with and have more precise information about the business models of older firms, then annual reports of older firms should be simpler and more readable. I proxy for firm age using the number of years since a firm shows up in CRSP monthly stock return files ($AGE$).

- Special items: Firms with significant amount of special items are likely to experience some unusual events. $SI$, defined as the amount of special items scaled by book value of assets, is included as a potential determinant of annual report readability. Everything else equal, I expect firms with lower special items (i.e., more negative special items) to have more complex annual reports.

- Volatility of business or operations: Communications to investors by firms with more volatile business environment are presumably more complicated. I use firm-specific stock return volatility ($RET\_VOL$, measured as the standard deviation of the monthly stock returns in the last year) and earnings volatility ($EARN\_VOL$, measured as the standard deviation of the operating earnings in the last five fiscal years) to capture the volatility of business.

- Complexity of operations: Firms with more complex operations are likely to have more complex annual reports. To measure the complexity of business and operations, I use the logarithm of the number of business segments ($NBSEG$) and the logarithm of the number of geographic segments ($NGSEG$) from Compustat segment files at the end of a fiscal year.

- Financial complexity: Firms which have more complex financial situations are also likely to have more complicated annual reports. I use the logarithm of the number of

non-missing items in Compustat as a proxy for financial complexity ($NITEMS$). The underlying assumption is that if a firm needs to report more items in their financial statements, it's more complex financially.

- Firm events: Unusual firm events may require extra and more detailed disclosures. I create two dummies, $MA$ and $SEO$, to capture firm-year specific merger-and-acquisition and seasoned equity offering events. $MA$ is set to 1 in a year if a company appears as an acquirer in this calendar year in SDC Platinum M&A database and 0 otherwise; $SEO$ is set to 1 in a year if a company has a common equity offering in the secondary market according to the SDC Global New Issues database and 0 otherwise.

- Incorporation state: Finally, firms that are incorporated in Delaware have different corporate laws, investor protections, are more likely to receive takeover bids and be acquired, and are valued higher than similar firms incorporated elsewhere (Daines (2001)). Therefore, I include a Delaware incorporation dummy to check whether Delaware firms have different annual reports in terms of readability.

In addition, I include year and industry fixed effects as potential determinants of the readability.[9]

Table 2 shows the regression results of $Fog$ and $Length$ on their potential determinants. Since the readability of annual report is likely to be correlated within industries, the standard errors are clustered at two-digit SIC industry level. In column [1] of Table 2 Panel

[9]As an alternative specification, I drop the year dummies and include the accumulated CRSP value-weighted stock market returns in the last twelve months in the regression to examine the effect of macro economic conditions. I also drop the industry fixed effects and examine two industry-specific variables as potential determinants of annual report readability: the Herfindahl Index and a high-tech industry dummy defined by the American Electronics Association. In addition, firms facing more litigation risks may therefore want to write their annual reports more rigorously and end up with annual reports that are harder to read (Bencivenga (1997)). I therefore construct an industry-specific litigation risk using the Securities Class Action Clearinghouse Database from the Stanford Law School. The untabulated results show that the aggregate stock returns and the litigation risk are both positively related to readability, but the Herfindahl index and the high-tech dummies do not have explanatory power for annual report readability.

A, the Fog index of the whole annual report is regressed on the variables with year and industry fixed effects. Bigger firms, firms with more volatile business, firms with merger and acquisition activities, and firms incorporated outside of Delaware have more complex annual reports, as evidenced by the positive and significant coefficients on $SIZE$, $RET\_VOL$, $EARN\_VOL$, $MA$, and $DLW$.[10] On the other hand, $AGE$, $SI$, $NGSEG$, and $SEO$ are negatively associated with Fog, suggesting that younger firms, firms with more negative special items, firms with fewer geographic segments, and firms that are not issuing new equity have more complex annual reports. The counter-intuitive result is the negative coefficient on $NGSEG$, suggesting firms with more geographic segments tend to have less complicated annual reports. The explanatory power of all the variables combined together, however, is pretty small, as evidenced by the 8% adjusted R-squared in the regression and half of this explanatory power comes from industry dummies.

Column [2] reports the determinants of the MD&A $Fog$. Unlike the results based on the readability of the whole annual report, $SIZE$ and $AGE$ are not significantly related to MD&A readability, whereas $MTB$ is positively associated with it. This is perhaps because growth opportunities are harder to describe than assets-in-place in management discussion. Another interesting difference is that the association between $SI$, $RET\_VOL$, and $EARN\_VOL$ and readability are much stronger for MD&A than the whole document. For instance, the coefficient on $EARN\_VOL$ is 0.822 (t-statistic 5.68) in column [5], while the coefficient is 0.182 (t-value 2.20) in column [3]. This suggests that more negative special items and more volatile business environment are harder to explain in the MD&A section.

In column [3], the dependent variable is the Fog of the Notes to the financial statements. The negative and significant coefficient on $SIZE$ suggests that smaller firms tend to have more complicated Notes. Compared with the MD&A section, $MTB$ is only marginally related to the Notes Fog (coefficient of 0.012 with a t-statistic of 1.86). The amount of special items is not associated with the Notes readability. When a firm is involved in M&A

[10] $NBSEG$ is positively related to $Fog$ if industry fixed effects are not included and becomes insignificant if industry dummies are controlled.

transactions, the Notes to financial statements become more complex, as indicated by the positive coefficient (0.059 with a t-statistic of 2.44) on $MA$. Surprisingly, the negative coefficient on $NITEMS$ indicates that firms with more non-missing Compustat items have simpler annual reports, suggesting that $NITEMS$ may not capture firm financial complexity well.

Panel B of Table 2 shows the regressions of annual report length on potential determinants. The determinants of the length of the whole annual report, the MD&A section, and the Notes to the financial statements are quite similar. Bigger firms, low market-to-book firms, younger firms, firms with very negative special items, firms with high return and earnings volatility, firms involved in M&A transactions, and Delaware firms have longer annual reports. Not surprisingly, firm size is the single most important factor in explaining the length of annual reports.

# 4 Empirical Results

## 4.1 Current Earnings and Annual Report Readability

I first check the relation between firm performance and annual report readability (i.e., $Fog$ and $Length$). Table 3 shows the results of regressing $Fog$ and $Length$ on earnings (scaled by book value of assets) using both level (Panel A) and change (Panel B) specifications. In all the regressions, the variables used in Table 2 as determinants of annual report readability are included as control variables. The results without these control variables are not reported but are of similar magnitude and statistical significance. Year and industry fixed effects are also included in all the regressions. All the standard errors are clustered at industry level to control for within-industry correlation of annual report readability.

The negative coefficients on earnings indicate that firms with higher earnings have annual reports that are easier to read (i.e., lower Fog and shorter). In columns [1] and [3] of Panel A, the coefficients on earnings are -0.458 (t-statistic -4.44) and -0.508 (t-statistic -12.93) when it is used to explain the Fog and Length of the whole annual report. Replacing the

earnings level with a profit/loss dummy, which equals one if a company reports a profit and zero otherwise, gives similar results: The coefficients on the dummy are -0.163 (t-statistic -3.95) and -0.184 (t-statistic -17.61) in columns [2] and [4] of Panel B. The results indicate that the annual reports of loss firms are harder to read than those of profit firms.

The negative relation between firm performance and annual report $Fog$ and $Length$ also holds in a change specification. Firms that experience an increase in earnings tend to write their annual report in a more readable way than last year. In Panel B of Table 3, when the control variables and fixed effects are included, year-to-year change in earnings is negatively related to change in $Fog$ and $Length$ (columns [1] and [3]). Columns (2) and (4) show that, on average, the change in $Fog$ ($Length$) of firms with an increase in earnings is 0.094 (0.053) lower than those with a decrease in earnings.

Separating the annual report into sections shows that the relation between earnings and $Fog$ mainly comes from the MD&A section. In column [5] of Table 3 Panel A, the coefficient on earnings is -1.66 (t-statistic -8.38), more than three times the coefficient in column [1] when the Fog of the whole annual report is used. On the other hand, while the Fog of the Notes to the financial statements is negatively associated with earnings, the coefficient on earnings is much smaller: -0.185 (t-statistic -2.53) in column [9] and -0.257 (t-statistic -3.67) in column [11], which are less than half of the coefficients in columns [1] to [4].

However, the relation between earnings and $Length$ comes more from the Notes section than the MD&A section. Splitting annual report into MD&A and Notes to the financial statements shows that the Notes (coefficient on earnings is -0.551 with t=-5.80) is more negatively correlated with earnings than MD&A (coefficient -0.284 and t=-4.93). This suggests that length of Notes are more likely to be used as a strategic deterrence to investors. The change specification further confirms that the negative relation between firm performance and readability is stronger in the MD&A section and weaker in the Notes to financial statements.

The incremental R-squared of earnings in explaining $Fog$ and $Length$, however, is trivial. Comparing column [1] of Table 3 Panel A with column [1] of Table 2 Panel A reveals

that adding current earnings increases the R-squared by 0.00. This suggests that economic performance is not a first-order determinant of annual report readability. To gauge the economic size of the effects, I do the following calculation. On average, increasing a firm's earnings from .00 (25th percentile of the sample) to 0.11 (75th percentile) will lead to a decrease in Fog Index of about 0.05. This is small compared to the variation of Fog in the sample (Table 1). Put differently, the annual reports of firms at 25th percentile of earnings have about 0.13 more syllables per word or about 0.13% more complex words than those of firms at 75th percentile. The Fog Index (Length) of loss firms is higher than that of profit firms by 0.16 (0.184), which is also small.

To summarize, I find that firms with better performance have annual reports that are harder to read. The effects are statistically significant, but the economic magnitude seems small. This is consistent with the marginal benefit of making annual reports more complex to hide poor current performance is small.

## 4.2 Earnings Persistence and Annual Report Readability

In this section, I examine the implication of annual report readability for earnings persistence. Management opportunism suggests that when annual reports are harder to read, good news may be more transitory and bad news may be more persistent.

I find that, indeed, the positive earnings of firms with more "foggy" or longer annual reports are less persistent. Table 4 Panel A presents the regression results of one-year and two-year ahead earnings on this year's earnings, Fog, and their interaction using a sample of all firm-years with positive earnings.[11] The interaction term captures the change in earnings persistence as annual report readability changes. In all the regressions, the variables that are potential determinants of readability (i.e., $SIZE$, $MTB$, $AGE$, $SI$, $RET\_VOL$, $EARN\_VOL$, $NBSEG$, $NGSEG$, $NITEMS$, $SEO$, $MA$, and $DLW$) and their interactions with earnings are included as control variables. In addition, the absolute amount of

[11] I also checked the three-year and four-year ahead earnings. The results are similar but statistically weaker.

accruals ($ABSACC$) and a dividend dummy ($DIV$, which equals one if a company pays dividend and zero otherwise) and their interactions with earnings are also included, because Sloan (1996) documented a negative relation between the absolute amount of accruals and earnings persistence and Skinner (2004) found a positive association between dividend and earnings persistence. The results without the control variables are similar and not reported.

In all cases, the interaction term is negative. For instance, in columns [1] and [2] of Table 4 Panel A, where the $Fog$ of the whole annual report is used to explain year $t+1$ and $t+2$ earnings persistence, the interaction term coefficients are -0.028 (t=-3.74 with the standard errors clustered at industry-level) and -0.041 (t=-2.95). This means that, as Fog of the whole annual report goes up (i.e., annual reports become harder to read), the earnings persistence becomes smaller for profitable firms.

To gauge the economic significance, I compare the impact of annual report readability on earnings persistence with that of accruals. Everything else equal, for an inter-quartile increase in Fog (an increase from 18.44 to 20.16), the one-year ahead earnings persistence of profitable firms goes down by 0.05 (calculated as -0.028 * (20.16-18.44), where -0.028 is from Column [1] of Table 4 Panel A.) and the two-year ahead earnings persistence goes down by 0.07 (calculated as -0.041 * (20.16-18.44), where -0.041 is from Column [2] of Table 4 Panel A.) Untabulated results also indicate that, on average, firms with Fog Index greater than 18 has an earnings persistence lower than those with Fog Index less than 14 by 0.12. An inter-quartile increase in the absolute amount of accruals, on the other hand, will lower the earnings persistence by about 0.05. This suggests that the Fog Index has economically significant implications for the persistence of earnings of profitable firms.

Focusing on the readability of the MD&A section and the Notes to financial statements (columns [5] to [12]) shows that the Fog of both sections are negatively related to earnings persistence. The effect of MD&A Fog is slightly smaller. However, the cross-sectional variation in the MD&A Fog is also bigger (Table 1) and the overall economic effect of MD&A readability on earnings persistence is comparable to the readability of the whole annual report.

Panel B of Table 4 documents the negative relation between annual report length and the earnings persistence of profit firms. In column [1] of Table 4 Panel A, where one-year ahead earnings is regressed on current earnings and its interaction with annual report length using positive earnings sample, the coefficient on the interaction term has a coefficient of -0.06 (t=-2.97). The effect of annual report length on earnings persistence is economically big: an increase of length from 9.63 (the 25th percentile from Table 6 Panel A) to 10.52 (the 75th percentile) implies an earnings persistence lower by 0.05. The length of Notes is more associated with earnings persistence than the MD&A length: In column [5] of Table 4 Panel B, the coefficient on the interaction of earnings with Notes length is -0.025 (t=-2.87); in column (3), the coefficient on the interaction of earnings with MD&A length is -0.019 (t=-1.23). Overall, it seems that the length of annual report is negatively related to performance and earnings persistence, and this effect is stronger in Notes than MD&A section.

Panel C of Table 4 includes the readability of the whole annual report, the MD&A section and the Notes to the financial statements in one regression to examine which part of the annual report has the biggest impact on earnings persistence. The results indicate that the Fog of the whole document, MD&A, and Notes are all negatively related to one-year ahead and two-year ahead earnings persistence, but only the effect of MD&A Fog is statistically significant (see columns [1] and [2] of Table 4 Panel C), suggesting that managerial strategic disclosure may come more from the MD&A section. However, the insignificant coefficients could be due to high correlations among the Fog of the sections. For instance, in column [1], the coefficient on the interaction of MD&A Fog and earnings is -0.010 (t=-1.78) and that of the Notes Fog is -0.015 (t=-1.66). Hence, although marginally insignificant, the effect of Notes Fog on earnings persistence is comparable to that of MD&A Fog in economic magnitude. Overall, it seems that the readability of both the MD&A section and the Notes section contain information about earnings persistence.

On the other hand, I find little evidence that the annual report readability affects the persistence of losses. As can be seen from Table 5 Panel A, the Fog Index of annual reports has no impact on the persistence of losses. Both the coefficient magnitude and the t-statistics

of the interaction term of Fog and earnings are small. Evidence from *Length* (Panel B) is similar, with the exception that the length of the whole annual report is negatively correlated with the persistence in two-year ahead earnings.

In summary, the evidence here is consistent with firms using more complicated language in their annual reports to present good news that are less persistent. On the other hand, I do not find significant evidence that firms make their annual report more difficult to read to hide more persistent bad news.[12]

# 5 Beyond Readability: Additional Lexical Features of Annual Reports

In this section, I analyze other lexical properties of annual reports and provide preliminary evidence on their implications for firm performance and earnings persistence.[13] Managerial strategic disclosure is just one of the possible explanations for my findings. For instance, perhaps poor and less persistent earnings are inherently more difficult to present and this

[12]The results are robust empirically. First, one concern may be that some firm characteristics drive both the annual report readability and earnings persistence. To rule this out, I construct a panel data set by keeping firms with at least 10 years of data. I then redo the tests by adding firm dummies in the regressions and the results still hold. Second, I include earnings-squared as an additional explanatory variable to control for possible non-linearity in earnings persistence and unreported results show that the results are slightly stronger both statistically and economically. Third, Dechow and Ge (2005) find that the low persistence of earnings in low accrual firms is primarily driven by balance sheet adjustments relating to special items. Therefore, I further examine whether unusual events related to special items are driving the empirical findings using a sub-sample of firm-years which have a zero amount of special items. Unreported results based on this sub-sample are similar to the main results. Finally, firms with poor current or future performance are more likely to use more sophisticated language in disclosure to avoid potential lawsuits (Bencivenga (1997)). However, my main results come from the profitable firms. Untabulated results show that more than 90% of these firms still report a profit in the next year and more than 80% of them remain profitable every year in the next one to four years. It seems unlikely that litigation is a first-order concern for these firms.

[13]I thank the referee for suggesting the analysis and the software package to me.

leads to more "foggy" and longer annual reports. One approach to mitigate this concern is to go beyond readability and examine other features of the annual reports. If the findings on the association of annual report readability and firm performance and earnings persistence are due to the strategic behavior of managers, there should be other lexical features of the annual reports that are related to earnings persistence.[14]

In particular, I focus on five categories of writing styles of the MD&A section: the relative frequency of self-reference words, exclusive words, causation words, positive emotion words, and future tense verbs. Psychology research shows that words that reflect how people are expressing themselves can often be more informative than what they are expressing (Undeutsch (1967), Pennebaker and King (1999), Pennebaker, Mehl, and Nierderhoffer (2003), and Shapiro (1989)) and that liars and truth-tellers communicate in qualitatively different ways.[15]

More specifically, Newman, Pennebaker, Berry, and Richards (2003) found that when people tell the truth, they are more likely to use 1st person singular pronouns and they also use more exclusive words like "except", "but", "without", and "excluding". They argue that words such as this indicate that a person is making a distinction between what they did do and what they didnt do and liars have a problem with such complex ideas. Therefore, the first two measures that I examine are the percentages of self-reference and exclusive words in the MD&A section of the annual report. I focus on the MD&A section instead of the whole annual report, because it is closer to the documents typically analyzed by psychologists in

[14]Davis, Piger, and Sedor (2005) document a positive (negative) association between optimistic (pessimistic) language usage and future firm performance and a significant incremental market response to optimistic and pessimistic language usage in earnings press releases.

[15]A real-life example is provided by Newman, Pennebaker, Berry, and Richards (2003): In 1994, Susan Smith appeared on television claiming that her two young children had been kidnaped at gunpoint. Eventually, authorities discovered she had drowned her children in the lake and fabricated the kidnaping story to cover her actions. Before Smith was a suspect in the childrens deaths, she told reporters, "My children wanted me. They needed me. And now I cant help them". Normally, relatives will speak of a missing person in the present tense. The fact that Smith used the past tense in this context suggested to trained Federal Bureau of Investigation (FBI) agents that she already viewed them as dead.

length and style.

The third writing style I focus on is the percentage of causation words (such as "because") used in the MD&A section, as these words are used when a person wants to explain something. People may spend more effort explaining what is going on if they try to cover up something. Next, the percentage of positive emotion words (relative to negative emotion words) is used to capture the underlying emotion of the writer. If a manager is trying to hide bad news, then even though the earnings number may be manipulated, the negative emotions reflected in the text may reveal the truth. Finally, the last measure intends to capture the managerial emphasis on future versus past/present. The intuition here is that people are likely to talk more about "future" if they are not doing so well and not so confident about their performance.

It is important to note that the analysis based on these variables is of preliminary nature. A caveat of analyzing these writing style measures of annual reports is that most of the psychology and linguistics research is based on experimental evidence using documents written by individual writers in non-business settings. Annual report is typically written by the management team and attorneys and therefore the external validity of the writing style measures is not established in my setting. As a result, any empirical test is a joint test of the hypotheses and the maintained assumption that the writing style measures capture certain managerial behavior.

I rely on the Linguistic Inquiry and Word Count (LIWC) package to compute the lexical measures. LIWC is a text analysis software program designed by psychologists James W. Pennebaker, Roger J. Booth, and Martha E. Francis and is able to calculate the degree to which people use different categories of words across a wide array of texts.[16]

[16]More details about the software can be found at http://www.liwc.net/ and http://homepage.psy.utexas.edu/homepage/Faculty/Pennebaker/Home2000/Words.html.

## 5.1 Measures

The empirical measures of the five categories of writing styles are as follows. A variable $IvsU$ measuring the degree of self-reference is calculated for each annual report's MD&A section as:

$$IvsU = \ln((1 + Self)/(1 + You + Other)) \tag{3}$$

, where $Self$ is the percentage of first person pronouns, and $You$ and $Other$ are the percentages of second and third person pronouns. The default LIWC dictionary is composed of 2,300 words and word stems with each word or word stem defining one or more word categories or subcategories. There are 20, 14, and 22 words in the "Total first person", "Total second person", and "Total third person" categories respectively in the default dictionary.

Similarly, $EvsI$ captures the frequency of "exclusive" words relative to that of "inclusive" words:

$$EvsI = \ln((1 + Excl)/(1 + Incl)) \tag{4}$$

, where $Excl$ is the percentage of exclusive words (19 words in the LIWC dictionary including "but", "except", and "without") and $Incl$ is the percentage of inclusive words (16 words in the dictionary including "with", "and", and "include").

The next variable $Cause$ is the percentage of causation-related words (49 words in the dictionary including "because", "effect", and "hence"). The forth category is the positive (versus negative) emotion of a document. A variable $PvsN$ is calculated for each annual report's MD&A section as:

$$PvsN = \ln((1 + Posemo)/(1 + Negemo)) \tag{5}$$

, where $Posemo$ is the percentage of positive emotion words (261 words in the LIWC dictionary including "happy", "pretty", and "good") and $Negemo$ is the percentage of negative emotion words (345 in the dictionary including words like "hate", "worthless", and "enemy").

Finally, a variable $FvsP$ is calculated to capture the frequency of future-oriented words versus the past/present-oriented words of annual reports:

$$FvsP = \ln((1 + Future)/(1 + Past + Present)) \tag{6}$$

, where $Future$ is the percentage of future tense verbs (14 words in the LIWC dictionary including "will", "might", and "shall") and $Past$ and $Present$ are the percentages of past and present tense verbs (144 and 256 words in the dictionary respectively).

Table 6 Panel A and Panel B show the summary statistics of the writing style variables and their correlations with $Fog$ and $Length$ and other firm characteristics. More self-reference words in MD&A are associated with more exclusive words, more causation words, fewer positive words, and more future tense verbs. MD&A sections that are more difficult to read (i.e., annual reports with higher MD&A $Fog$) tend to have fewer self-reference words, more exclusive words, more causation words, more positive words, and more discussion about future, as indicated by the negative correlation between Fog and $IvsU$, and positive correlation coefficients between Fog and $EvsI$, $Cause$, $PvsN$, and $FvsP$. In contrast, annual reports with longer MD&A section tend to have more self-reference words, more exclusive words, fewer causation words, more positive words, and more future tense verbs.

## 5.2 Empirical Findings

Table 7 presents the regression results of the writing style measures on current earnings and other control variables. The MD&A section of the annual reports of firms with lower earnings tend to use more self-references, use more exclusive words, and have more discussion about future. The implications of firm performance for $Cause$ and $PvsN$ are statistically insignificant. The results on self-references words exclusive words are not consistent with the joint hypothesis that firms with bad performance hide adverse information strategically and that managers who try to hide adverse information use fewer self-reference and exclusive words.

I next turn to the association of the writing styles with earnings persistence. As discussed in previous sections, the strategic managerial behavior is more likely to be detected in future earnings, rather than current earnings. From Table 8, it can be seen that $IvsU$ and $EvsI$ are not associated with earnings persistence. However, earnings persistence is a function of $Cause$, $PvsN$, and $FvsP$. More specifically, for profitable firms, higher frequency of causa-

tion words (i.e., higher $Cause$) means less persistent earnings, more positive emotion words relative to negative emotion words (i.e., higher $PvsN$) are associated with more persistent earnings, and higher frequency of future tense verbs relative to past/present tense verbs (i.e., higher $FvsP$) indicates lower earnings persistence. For profitable firms, an inter-quartile increase in $Cause$, $PvsN$, and $FvsP$ is associated with an earnings persistence lower by 0.03, higher by 0.04, and lower by 0.04 respectively.

On the other hand, loss firms with more positive emotion words relative to negative emotion words in their MD&A have less persistent earnings. Overall, the evidence suggests that managers who use more causation words, less positive words, and more future tense verbs may be strategically hiding adverse information about future earnings.

# 6 Further Discussions

## 6.1 Unexercised Stock Option Holdings and Incentives to Obfuscate Information

The research design in the paper can't prove a causality between earnings persistence and the annual report readability and other lexical properties. For instance, perhaps bad news is inherently harder to present and requires more complicated language. One way to mitigate this concern is to find a setting where the incentives for managers to obfuscate information is higher and check whether the empirical results are stronger there. This section provides some evidence on this.

Prior research has documented that managers strategically withhold good news before scheduled employee stock options grants (Aboody and Kasznik (2000)). Managers may want to delay the release of bad information if they have lots of unexercised stock options. I link this intuition with the association between readability and earnings persistence. Everything else equal, managers with more unexercised stock options may want to increase the complexity of the annual reports when current good earnings are not persistent.

This is indeed the case. Table 9 Panel A show that the interaction of $UNEX\_OPT$, a

measure of the amount of unexercised employee stock options, and earnings and Fog index loads up negatively, suggesting that our empirical results are stronger for firms with more unexercised executive stock options. Similar results are observed for *Length* (Panel B of Table 9), although the statistical significance is lower.

## 6.2 Future Stock Returns and Annual Report Readability

The evidence in this paper supports the hypothesis that managers try to obfuscate information through more complex disclosures. One possible benefit for this managerial behavior is to delay the incorporation of bad news into stock prices, as prior studies show that the stock market may under-react to the textual information in annual reports (e.g., Li (2006)). This section therefore checks whether the stock prices reflect the implications of annual report readability for future earnings.

I regress the 12-, 24-, 36-, and 48-month stock returns following the 10-K filing date on *Fog* and *Length* and their interactions with current year earnings. The Fama-MacBeth regression results in Table 10 indicate that there is no significant association between annual report readability and length and future stock returns in the next four years.[17] Overall, I conclude that there is no systematic evidence of stock market investors not understanding the implications of annual report readability for future performance. Thus, the benefit for managers to make complex disclosures remains a puzzle. One possible explanation is that the market mis-pricing effect is small and a simple test does not have enough power to detect it in a general setting.

[17]Unreported results based on two sub-samples (small firms, defined as firms with a market value lower than $2 billion, and firms with low institutional ownership defined as firms with institutional ownership lower than 20%) also show no relation between annual report readability and future returns.

# 7 Conclusions

I empirically study the implications of annual report readability and other lexical features of annual report for current performance and earnings persistence. In doing so, this study sheds some light on the relevance of the SEC plain English disclosure regulation. The empirical findings can be summarized as follows. First, annual reports of firms with poor performance are harder to read. The effect is statistically, but not economically, significant. Second, the profits of firms with annual reports that are easier to read are more persistent in the next one to four years. The effect is economically significant: An inter-quartile change in annual readability has about the same impact on profit persistence as accruals. This suggests that managers may be opportunistically choosing the readability of annual reports to hide information from investors. Taken together, the evidence in this paper suggests that, consistent with the motivation behind the plain English disclosure regulation of the Securities and Exchange Commission, managers may be opportunistically structuring the annual reports to hide adverse information from investors. However, there is no apparent correlation between annual report readability and future stock returns, suggesting that the stock market understands this implication.

## Appendix 1: Steps to Calculate the Readability Indices

This Appendix explains the details of calculating the readability indices starting from the raw 10-K filings used in this paper. I first download the 10-K report from Edgar and do the following editing before further analysis. First, the heading information that is contained between <SEC-HEADER> and </SEC-HEADER> is deleted. Second, all the tables that begin with <TABLE> and end with </TABLE> or the paragraphs that contain <S> or <C> are deleted, because <S> and <C> tags used by some firms in presenting tables. Next, all the paragraphs that contain string such as </TEXT>, </DOCUMENT>, <PAGE>, <TYPE> or /PRIVACY-ENHANCED/ are deleted. All the special characters in the format of <...> and <&..>, which are used widely in documents in SEC XML format are replaced with blanks. Finally, to make sure that all tables, tabulated texts, or financial statements are excluded, all the paragraphs with more than 50% of non-alphabetic characters (e.g., white spaces or numbers) are deleted.

The file after the editing is then analyzed using the Fathom package in Perl. The package can calculate the typical text statistics, including number of characters, number of words, percent of complex words (i.e., words with more than three syllables), number of sentences, number of text lines, number of paragraphs, syllables per word and words per sentence. Based on the statistics, the package also produce the summary readability indices used in the paper.

**Appendix 2: Steps to Extract MD&A and Notes to the Financial Statements**

This Appendix explains the details of extracting the MD&A section and Notes to the financial statements from 10-K filings. Starting with the raw 10-K file, I first get rid of the SEC-header information, all the contents between <TABLE> and </TABLE> text, the paragraphs that contain <S> or <C>, all the paragraphs that contain string such as </TEXT>, </DOCUMENT>, <PAGE>, <TYPE> or /PRIVACY-ENHANCED/, and All the special characters in the format of <...> and <&..> using the same process described in Appendix 1.

Within the remaining text, the program identifies a line that satisfies one of the following criteria as the beginning of the MD&A section: (1) The line starts with "management's discussion" or "management's discussion" following some white spaces; (2) The line contains "management's discussion" and ("item"+one or more white space+"7") and does not contain the word "see"; (3) The line starts with "managements discussion" or "managements discussion" following some white spaces; (4) The line contains "managements discussion" and ("item"+one or more white space+"7") and does not contain the word "see". Since many firms refer to the MD&A section in the front-matter of the annual reports, the word "see" serves to identify all such situations. The program identifies a line that satisfy one of the following criteria as the ending of the MD&A section: (1) The line begins with "Financial Statements" or "Financial Statements" following some white spaces; (2) The line contains "item" followed by some one or more white spaces and "8"; (3) The line contains "Supplementary Data"; (4) The line begins with "SUMMARY OF SELECTED FINANCIAL DATA" or "SUMMARY OF SELECTED FINANCIAL DATA" following some white spaces. Most firms have a table of content listing the main sections of the 10-K filing. In some occasions, this table of content is not embedded between <TABLE> and </TABLE> and therefore is not cleaned in the previous steps. As a result, the line in the table of content about MD&A will also be picked up by the program as part of MD&A.

Similarly, the program identifies a line as the beginning of the Notes to the financial statements, if (1) The line starts with "NOTES TO" or some white spaces followed by

"NOTES TO"; and (2) The line does not contain numbers unless there is "for the years ended". The program identifies a line that satisfy one of the following criteria as the ending of the Notes to financial statements: (1) The line contains "Changes in and Disagreements with Accountants" or "DISAGREEMENTS ON ACCOUNTING"; (2) The line constrains "DIRECTORS AND EXECUTIVE OFFICERS"; (3) The line contains "exhibit index".

After the MD&A and the Notes to the financial statements are identified, all the paragraphs with more than 50% of non-alphabetic characters (e.g., white spaces or numbers) are deleted. Finally, the Fathom package is used to calculate the readability measures.

# References


Aboody, David, and Ron Kasznik, 2000, Ceo stock option awards and the timing of corporate voluntary disclosures, *Journal of Accounting and Economics* 29, 73–100.

Baker, H.E., and D.D. Kare, 1992, Relationship between annual report readability and corporate financial performance, *Management Research News* 15.

Barker, Robert, 2002, A three-point plan for SEC reform, *Business Week Online*.

Barnett, A., and K. Leoffler, 1979, Readability of accounting and auditing messages, *Journal of Business Communication* 16.

Bencivenga, Dominic, 1997, Short cut for investors: Why read a prospectus when a profile will do?, *New York Law Journal*.

Bloomfield, Robert J., 2002, The "incomplete revelation hypothesis" and financial reporting, *Accounting Horizons* 16, 233–243.

Clatworthy, Mark, and Michael J. Jones, 2001, The effect of thematic structure on the variability of annual report readability, *Accounting, Auditing & Accountability Journal* 14.

Cohen, Daniel, 2005, Financial reporting quality: Determinants and economic consequences, Working Paper, New York University.

Collins-Thompson, Kevyn, and Jamie Callan, 2005, Predicting reading difficulty with statistical language models, *Journal of the American Society for Information Science and Technology* 56.

Courtis, John K., 1986, An investigation into annual report readability and corporate risk return relationships, *Accounting and Business Research*.

Daines, Robert, 2001, Does delaware law improve firm value?, *Journal of Financial Economics* 62, 525–558.

Davis, Angela K., Jeremy M. Piger, and Lisa M. Sedor, 2005, Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases, Working paper, Washington University.

Dechow, Patricia, and Weili Ge, 2005, The persistence of earnings and cash flows and the role of special items: Implications for the accrual anomaly, *Review of Accounting Studies,* forthcoming.

Dechow, Patricia M., and Catherine M. Schrand, 2004, *Earnings Quality* (Research Foundation of CFA Institute: Charlottesville, Virginia) first edn.

Fields, Thomas D., Thomas Z. Lys, and Linda Vincent, 2001, Empirical research on accounting choice, *Journal of Accounting and Economics* 31.

Firtel, Kenneth B., 1999, Plain English: A repraisal of the intended audience of disclosure under the securities act of 1933, *Southern California Law Review* 72, 851–897.

Francis, Jennifer, Ryan LaFond, Per Olsson, and Katherine Schipper, 2005, The market pricing of accruals quality, *Journal of Accounting and Economics*.

Francis, Jennifer, Dhananjay Nanda, and Per Olsson, 2005, Voluntary disclosure, information quality, and costs of capital, Working Paper, Duke University.

Glassman, Cynthia A., 2005, Remarks at the plain language association international's fifth international conference, http://www.sec.gov/news/speech/spch110405cag.htm.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Healy, Paul, 1977, Can you understand the footnotes to financial statements?, *Accountants Journal*.

Healy, Paul M., and Krishna G. Palepu, 2001, Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature, *Journal of Accounting and Economics* 31.

Jones, Michael J., and Paul A. Shoemaker, 1994, Accounting narratives: a review of empirical studies of content and readability, *Journal of Accounting Literature* 13.

Lang, Mark, and Russell Lundholm, 1993, Cross-sectional determinants of analyst ratings of corporate disclosures, *Journal of Accounting Research* Autumn, 246–271.

Lebar, Mary Ann, 1982, A general semantics analysis of selected sections of the 10-k, the annual report to shareholders, and the financial press release, *The Accounting Review* 57.

Li, Feng, 2006, Do stock market investors understand the risk sentiment of corporate annual reports?, University of Michigan Working Paper.

Miller, Gregory S., 2002, Earnings performance and discretionary disclosure, *Journal of Accounting Research* 40, 173–204.

Muresan, Gheorghe, Michael Cole, Catherine L. Smith, Lu Liu, and Nicholas J. Belkin, 2006, Does familiarity breed content ? taking account of familiarity with a topic in personalizing information retrieval, *Proceedings of the Hawaii International Conference on System Sciences (HICSS-39).*

Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards, 2003, Lying words: predicting deception from linguistic styles, *Personality and Social Psychology Bulletin* 29, 665–675.

Pennebaker, James W., and Laura A. King, 1999, Linguistic styles: Language use as an individual difference, *Journal of Personality and Social Psychology* 77, 1296–1312.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Nierderhoffer, 2003, Psychological aspects of natural language use: Our words, our selves, *Annual Review of Psychology* 54, 547–577.

Riedl, Edward J., and Suraj Srinivasan, 2005, The strategic reporting of special items, Working Paper, Harvard University and University of Chicago.

Schrand, Catherine M., and Beverly R. Walther, 2000, Strategic benchmarks in earnings announcement: The selective disclosure of prior-period earnings components, *Accounting Review* 75, 151–177.

SEC, 1998, *A plain English handbook: how to create clear SEC disclosure documents* (U.S. Securities and Exchange Commission: Washington, DC).

Shapiro, D., 1989, *Psychotherapy of neurotic character* (Basic Books: New York).

Skinner, Douglas J., 2004, What do dividends tell us about earnings quality, Working Paper, University of Chicago.

Sloan, Richard, 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings?, *Accounting Review* 71, 289–315.

Smith, James E., and Nora P. Smith, 1971, Readability: A measure of the performance of the communication function of financial reporting, *The Accounting Review* 46.

Soper, Fred J., and Robert Dolphin, 1964, Readability and corporate annual reports, *The Accounting Review* 39.

Subramanian, R., R.G. Insley, and R.D. Blackwell, 1993, Performance and readability: A comparison of annual reports of profitable and unprofitable corporations, *Journal of Business Communication* 30.

Undeutsch, U., 1967, *Forensic Physchologie [Forensic Psychology]* (Verlag fur Psychologie: Gottingen, Germany).

Watts, Ross L., and Jerry L. Zimmerman, 1986, *Positive Accounting Theory* (Prentice-Hall: Englewood Cliffs, NJ).

**Figure 1 A: Median *Fog* and *Length* of the Whole Annual Report by Calendar Year of the Filing Date**

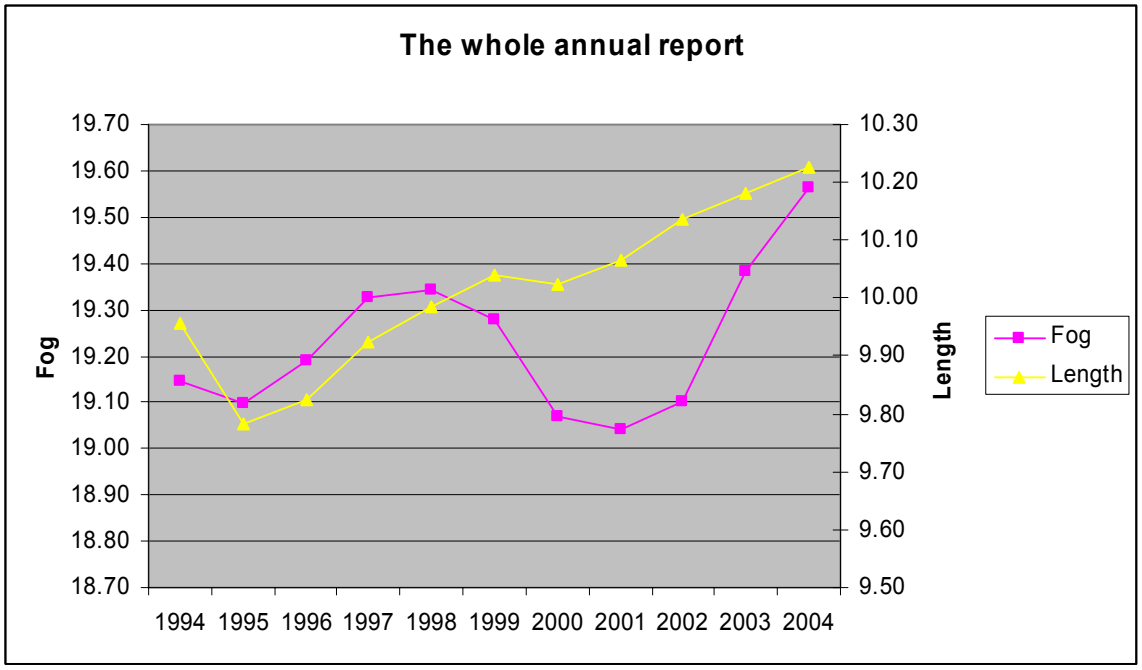



**Figure 1 B: Median *Fog* and *Length* of the MD&A Section by Calendar Year of the Filing Date**

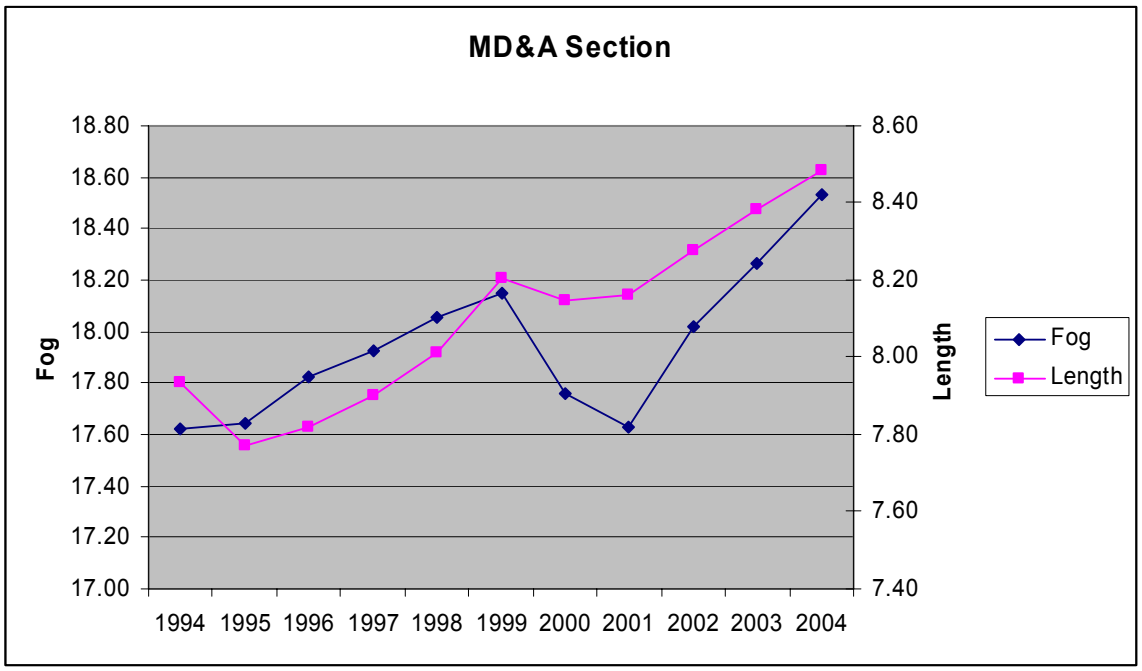
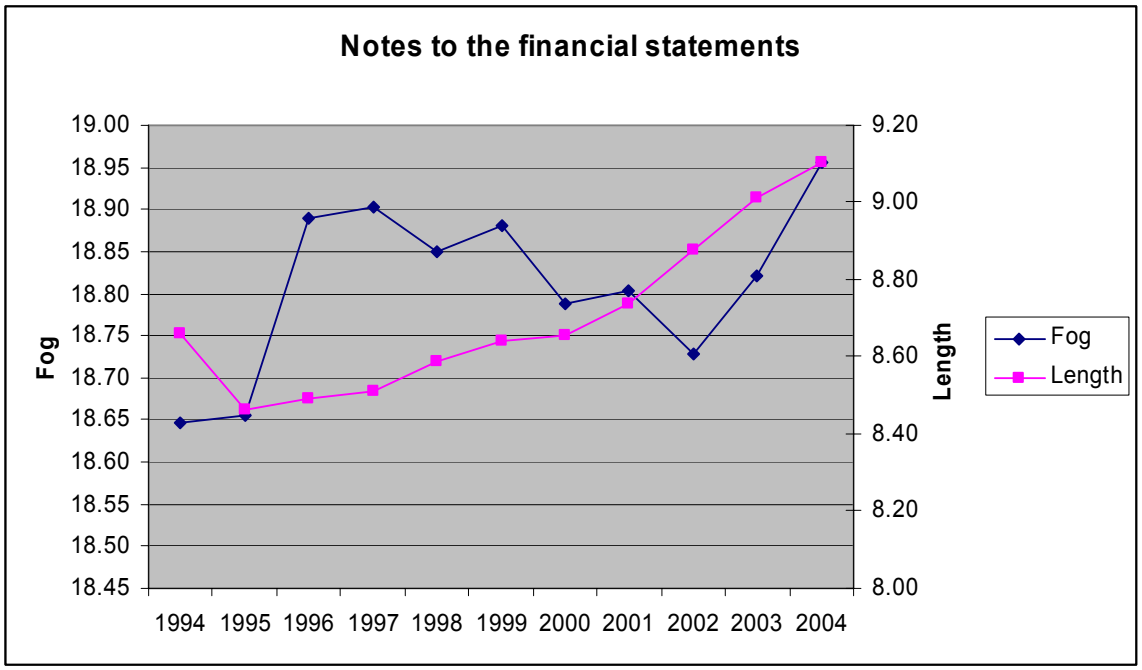


**Figure 1 C: Median *Fog* and *Length* of the Notes to Financial Statements by the Calendar Year of the Filing Date**




Note: *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words in an annual report. Figure 1 A shows the *Fog* and *Length* of the whole annual report. Figure 1 B shows the *Fog* and *Length* of the MD&A section. Figure 1 C shows the *Fog* and *Length* of the Notes to the financial statements.

## Table 1 Panel A: Summary Statistics

| Variable | Mean | Median | Std. Dev. | 1st | 25th | 75th | 99th | Obs |
|---|---|---|---|---|---|---|---|---|
| Year | - | 2000 | - | 1994 | 1997 | 2002 | 2004 | 55,719 |
| Earnings | 0.02 | 0.05 | 0.19 | -0.75 | 0.00 | 0.11 | 0.33 | 55,719 |
| Market-to-book | 2.02 | 1.30 | 2.94 | 0.54 | 1.03 | 2.04 | 11.62 | 51,297 |
| Market value of equity ($MM) | 2022 | 169 | 14,209 | 1 | 44 | 731 | 33,003 | 51,393 |
| Book value of assets ($MM) | 3551 | 271 | 24,875 | 3 | 67 | 1,092 | 57,100 | 55,719 |
| ***Whole annual report*** | | | | | | | | |
| Fog | 19.39 | 19.24 | 1.44 | 16.61 | 18.44 | 20.16 | 23.64 | 55,719 |
| Fog(t)-Fog(t-1) | 0.05 | 0.02 | 1.46 | -3.93 | -0.59 | 0.65 | 4.34 | 44,097 |
| Number of words | 31034 | 23122 | 28057 | 4918 | 15173 | 36926 | 140047 | 55,719 |
| Length | 10.08 | 10.05 | 0.70 | 8.50 | 9.63 | 10.52 | 11.85 | 55,720 |
| Length(t)-length(t-1) | 0.03 | 0.03 | 0.66 | -1.69 | -0.29 | 0.34 | 1.81 | 44,097 |
| ***MD&A section*** | | | | | | | | |
| Fog | 18.23 | 17.98 | 2.55 | 13.66 | 16.66 | 19.44 | 26.12 | 43,335 |
| Fog(t)-Fog(t-1) | 0.06 | 0.02 | 2.33 | -6.52 | -0.70 | 0.76 | 7.00 | 29,989 |
| Number of words | 4665 | 3325 | 5653 | 160 | 1894 | 5782 | 23195 | 43,335 |
| Length | 8.03 | 8.11 | 0.98 | 5.08 | 7.55 | 8.66 | 10.05 | 43,335 |
| Length(t)-length(t-1) | 0.04 | 0.07 | 0.98 | -3.11 | -0.22 | 0.36 | 3.09 | 29,989 |
| ***Notes to the financial statements*** | | | | | | | | |
| Fog | 18.96 | 18.83 | 1.53 | 15.88 | 17.98 | 19.76 | 23.69 | 48,366 |
| Fog(t)-Fog(t-1) | -0.02 | -0.02 | 1.53 | -4.74 | -0.59 | 0.54 | 4.76 | 35,343 |
| Number of words | 12443 | 6135 | 20284 | 1474 | 3855 | 12247 | 95640 | 48,366 |
| Length | 8.90 | 8.72 | 0.92 | 7.30 | 8.26 | 9.41 | 11.47 | 48,366 |
| Length(t)-length(t-1) | 0.06 | 0.06 | 0.84 | -2.41 | -0.16 | 0.30 | 2.49 | 35,343 |

Note: This Panel shows the summary statistics of some the variables in the paper. Year is the calendar year in which an annual report is filed to the SEC Edgar system. Fog is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. Length is the natural logarithm of the number of words in an annual report. Earnings is operating earnings (data178 of Compustat) scaled by assets. Market-to-book is the market value of the firm divided by its book value ((data25*data199+data181)/data6). Market value of equity is calculated as (data25*data199). Size is the logarithm of market value of equity calculated as Log(data25*data199). Book value of assets is data6 from Compustat.

## Table 1 Panel B: Pearson Correlation Matrix

| | Fog (Whole annual report) | Fog (MD&A) | Fog (Notes) | Length (Whole annual report) | Length (MD&A) | Length (Notes) | Market-to-book | Size | Assets |
|---|---|---|---|---|---|---|---|---|---|
| Fog (whole annual report) | | | | | | | | | |
| Fog (MD&A) | **0.368** | | | | | | | | |
| Fog (Notes) | **0.599** | **0.227** | | | | | | | |
| Length (Whole annual report) | **0.377** | **0.112** | **0.250** | | | | | | |
| Length (MD&A) | **0.039** | **-0.189** | **0.014** | **0.264** | | | | | |
| Length (Notes) | **0.241** | **0.096** | **0.383** | **0.656** | **0.194** | | | | |
| Market-to-book | **0.014** | **0.054** | **-0.020** | -0.006 | **-0.023** | **-0.048** | | | |
| Size | **0.007** | **-0.025** | **-0.098** | **0.263** | **0.165** | **0.191** | **0.169** | | |
| Assets | **0.017** | **0.028** | -0.002 | **0.106** | **0.078** | **0.105** | **-0.027** | **0.265** | |

Note: This Panel shows the Pearson correlation coefficient of *Fog* and *Length* of the annual reports with firm characteristics. *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the logarithm of the number of words. Market-to-book is the market value of the firm divided by its book value ((data25*data199+data181)/data6). Market value of equity is calculated as (data25*data199). Size is the logarithm of market value of equity calculated as Log(data25*data199). Book value of assets is data6 from Compustat.

The Pearson correlation coefficient in bold is significant at 0.01 level.

## Table 1 Panel C: Five Industries with the Lowest and Highest *Fog* and *Length*

| *Fog Index* | *Whole annual report* | | *MD&A Section* | | *Notes to financial statements* | |
|---|---|---|---|---|---|---|
| **Lowest** | 32 | Stone, Clay, Glass, and Concrete Products | 31 | Leather and Leather Products | 32 | Stone, Clay, Glass, and Concrete Products |
| | 45 | Transportation by Air | 22 | Textile Mill Products | 26 | Paper And Allied Products |
| | 31 | Leather and Leather Products | 45 | Transportation by Air | 22 | Textile Mill Products |
| | 56 | Apparel and Accessory Stores | 26 | Paper And Allied Products | 57 | Home Furniture, Furnishings, And Equipment Stores |
| | 20 | Food and Kindred Products | 54 | Food Stores | 33 | Primary Metal Industries |
| | 15 | Building construction | 49 | Electric, Gas, And Sanitary Services | 64 | Insurance Agents |
| | 49 | Electric, Gas, and Sanitary Services | 64 | Insurance Agents | 63 | Insurance Carriers |
| | 63 | Insurance Carriers | 80 | Health Services | 80 | Health Services |
| | 80 | Health Services | 63 | Insurance Carriers | 15 | Building Construction General Contractors And Operative Builders |
| **Highest** | 64 | Insurance Agents | 48 | Communications | 67 | Holding And Other Investment Offices |

| *Length* | *Whole annual report* | | *MD&A Section* | | *Notes to financial statements* | |
|---|---|---|---|---|---|---|
| **Lowest** | 65 | Real Estate | 23 | Apparel And Other Finished Fabrics Products | 42 | Motor Freight Transportation And Warehousing |
| | 25 | Furniture And Fixtures | 24 | Lumber And Wood Products (Except Furniture) | 39 | Miscellaneous Manufacturing Industries |
| | 42 | Motor Freight Transportation And Warehousing | 30 | Rubber And Miscellaneous Plastics Products | 38 | Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks |
| | 20 | Food And Kindred Products | 65 | Real Estate | 34 | Fabricated Metal Products, Except Machinery And Transportation Equipment |
| | 30 | Rubber And Miscellaneous Plastics Products | 54 | Food Stores | 65 | Real Estate |
| | 48 | Communications | 62 | Security And Commodity Brokers, Dealers, Exchanges, And Services | 48 | Communications |
| | 63 | Insurance Carriers | 49 | Electric, Gas, And Sanitary Services | 79 | Amusement And Recreation Services |
| | 70 | Hotels, Rooming Houses, Camps, And Other Lodging Places | 63 | Insurance Carriers | 49 | Electric, Gas, And Sanitary Services |
| | 29 | Petroleum Refining And Related Industries | 29 | Petroleum Refining And Related Industries | 29 | Petroleum Refining And Related Industries |
| **Highest** | 49 | Electric, Gas, And Sanitary Services | 60 | Depository Institutions | 63 | Insurance Carriers |

Note: This Panel shows the five 2-digit SIC industries that have the highest *Fog* and *Length* and the five industries that have the lowest *Fog* and *Length*. Firms with fewer than 8 years of data and industries with fewer than 100 firm-years in the sample are not included. *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words.

## Table 1 Panel D: Persistence of *Fog* and *Length*

| ***Fog*** | Quintile 1 | Quintile 2 | Quintile 3 | Quintile 4 | Quintile 5 | N |
|---|---|---|---|---|---|---|
| Year 0 | | | | | 100 | 11094 |
| Year 1 | 6.56 | 10.04 | 14.13 | 24.50 | 44.77 | 8564 |
| Year 2 | 7.42 | 11.35 | 16.56 | 24.82 | 39.85 | 6901 |
| Year 3 | 8.58 | 12.44 | 17.33 | 24.79 | 36.86 | 5418 |
| | | | | | | |
| Year 0 | 100 | | | | | 11091 |
| Year 1 | 56.83 | 20.83 | 9.35 | 6.14 | 6.86 | 8849 |
| Year 2 | 50.80 | 22.16 | 11.28 | 7.79 | 7.97 | 7252 |
| Year 3 | 46.99 | 23.07 | 11.45 | 9.36 | 9.12 | 5790 |

| ***Length*** | Quintile 1 | Quintile 2 | Quintile 3 | Quintile 4 | Quintile 5 | N |
|---|---|---|---|---|---|---|
| Year 0 | | | | | 100 | 11095 |
| Year 1 | 5.00 | 10.41 | 16.19 | 25.59 | 42.81 | 8692 |
| Year 2 | 5.94 | 11.42 | 17.13 | 25.67 | 39.83 | 6964 |
| Year 3 | 6.17 | 12.29 | 17.53 | 25.05 | 38.96 | 5460 |
| | | | | | | |
| Year 0 | 100 | | | | | 11093 |
| Year 1 | 62.31 | 17.27 | 7.91 | 6.73 | 5.78 | 8684 |
| Year 2 | 57.11 | 18.98 | 9.69 | 7.60 | 6.61 | 7107 |
| Year 3 | 54.05 | 19.11 | 10.98 | 8.72 | 7.15 | 5621 |

Note: This Panel shows the transition matrix of *Fog* and *Length* of the whole annual report across quintiles for firms in the 1st and 5th quintiles. *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words. Each year (year 0), firms are sorted into quintiles based on *Fog* or *Length*. In the next three years (year 1 to year 3), the percentages by quintiles for firms that are in the 1st and 5th quintiles in year 0 are calculated and tabulated.

## Table 2 Panel A: Determinants of *Fog*

| | | Dependent variables | | |
|---|---|---|---|---|
| | | [1] | [2] | [3] |
| | **Predicted sign** | ***Fog of the whole annual report*** | ***Fog of the MD&A section*** | ***Fog of the notes to the financial statements*** |
| SIZE | + | 0.019 | -0.015 | -0.064 |
| | | [1.89]* | [-0.95] | [-5.98]*** |
| MTB | + | 0.001 | 0.032 | 0.012 |
| | | [0.13] | [2.52]** | [1.86]* |
| AGE | - | -0.004 | 0.003 | -0.005 |
| | | [-2.47]** | [0.96] | [-2.63]** |
| SI | - | -0.193 | -0.447 | -0.066 |
| | | [-2.01]** | [-2.60]** | [-0.68] |
| RET_VOL | + | 0.438 | 1.326 | 0.532 |
| | | [3.07]*** | [5.00]*** | [4.72]*** |
| EARN_VOL | + | 0.182 | 0.822 | 0.056 |
| | | [2.20]** | [5.68]*** | [0.65] |
| NBSEG | + | -0.002 | 0.029 | 0.033 |
| | | [-0.09] | [0.82] | [1.30] |
| NGSEG | + | -0.062 | -0.074 | -0.081 |
| | | [3.75]*** | [-2.08]** | [-3.71]*** |
| NITEMS | + | -0.471 | -0.821 | -0.684 |
| | | [-1.50] | [-1.25] | [-2.31]** |
| SEO | + | -0.066 | -0.173 | 0.026 |
| | | [-1.69]* | [-3.84]*** | [0.44] |
| MA | + | 0.074 | 0.055 | 0.059 |
| | | [2.76]*** | [0.91] | [2.44]** |
| DLW | +/- | 0.157 | 0.128 | 0.085 |
| | | [4.10]*** | [1.82]* | [1.65] |
| Constant | | 21.712 | 21.854 | 22.877 |
| | | [12.88]*** | [6.14]*** | [14.39]*** |
| Year dummies | | Yes | Yes | Yes |
| Industry dummies | | Yes | Yes | Yes |
| Observations | | 36375 | 28279 | 31331 |
| R-squared | | 0.08 | 0.09 | 0.06 |

## Table 2 Panel B: Determinants of *Length*

| | | Dependent variables | | |
|---|---|---|---|---|
| | | [1] | [2] | [3] |
| | **Predicted sign** | ***Length of the whole annual report*** | ***Length of the MD&A section*** | ***Length of Notes to the financial statements*** |
| SIZE | + | 0.103 | 0.079 | 0.098 |
| | | [18.85]*** | [11.70]*** | [14.60]*** |
| MTB | + | -0.026 | -0.032 | -0.034 |
| | | [-6.01]*** | [-6.76]*** | [-5.05]*** |
| AGE | - | -0.008 | -0.005 | -0.002 |
| | | [-9.56]*** | [-4.70]*** | [-1.79]* |
| SI | - | -0.423 | -0.209 | -0.485 |
| | | [-6.77]*** | [-2.63]** | [-6.17]*** |
| RET_VOL | + | 0.726 | 0.368 | 0.918 |
| | | [9.09]*** | [3.75]*** | [9.11]*** |
| EARN_VOL | + | 0.184 | 0.083 | 0.186 |
| | | [6.44]*** | [1.44] | [5.03]*** |
| NBSEG | + | 0.007 | 0.025 | 0.019 |
| | | [0.75] | [1.91]* | [1.02] |
| NGSEG | + | -0.007 | 0.002 | -0.016 |
| | | [-1.00] | [0.19] | [-1.23] |
| NITEMS | + | -0.261 | 0.103 | -0.242 |
| | | [-1.73]* | [0.64] | [-1.42] |
| SEO | + | 0.03 | 0.032 | 0.006 |
| | | [1.91]* | [1.10] | [0.27] |
| MA | + | 0.074 | 0.012 | 0.099 |
| | | [9.92]*** | [0.71] | [6.83]*** |
| DLW | +/- | 0.089 | 0.076 | 0.097 |
| | | [5.48]*** | [3.15]*** | [2.25]** |
| Constant | Constant | 10.793 | 6.894 | 9.405 |
| | | [13.48]*** | [7.58]*** | [10.34]*** |
| Year dummies | | Yes | Yes | Yes |
| Industry dummies | | Yes | Yes | Yes |
| Observations | | 36375 | 28279 | 31331 |
| R-squared | | 0.18 | 0.09 | 0.13 |

Note:

This Table shows the regression results of *Fog* (Panel A) and *Length* (Panel B) on the following variables and year fixed effects and 2-digit SIC industry fixed effects: SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, SEO, MA, and DLW. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise.

*Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words.

T-statistics are based on standard errors clustered at two-digit SIC industry code level.

**Table 3 Panel A: Firm Performance and Annual Report Fog and Length (level specification)**

| | Dependent variables | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] | [11] | [12] |
| | ***Whole annual report*** | | | | ***MD&A section*** | | | | ***Notes to the financial statements*** | | | |
| | Fog | Fog | Length | Length | Fog | Fog | Length | Length | Fog | Fog | Length | Length |
| Earnings | -0.458 | | -0.508 | | -1.659 | | -0.284 | | -0.185 | | -0.551 | |
| | [-4.44]*** | | [-12.93]*** | | [-8.38]*** | | [-4.93]*** | | [-2.53]** | | [-5.80]*** | |
| Profit/Loss dummy | | -0.163 | | -0.184 | | -0.625 | | -0.095 | | -0.037 | | -0.179 |
| | | [-3.95]*** | | [-17.61]*** | | [-6.28]*** | | [-5.53]*** | | [-1.32] | | [-10.87]*** |
| Constant | 21.399 | 21.606 | 21.606 | 11.076 | 20.820 | 21.656 | 7.618 | 7.75 | 22.420 | 22.475 | 9.605 | 9.835 |
| | [15.50]*** | [15.71]*** | [15.71]*** | [18.39]*** | [6.30]*** | [6.59]*** | [6.48]*** | [6.60]*** | [15.57]*** | [15.47]*** | [13.06]*** | [13.00]*** |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 41100 | 41100 | 41100 | 41100 | 32099 | 32099 | 32099 | 32099 | 35533 | 35533 | 35533 | 35533 |
| R-squared | 0.08 | 0.08 | 0.19 | 0.18 | 0.10 | 0.10 | 0.09 | 0.09 | 0.06 | 0.06 | 0.13 | 0.13 |

**Table 3 Panel B: Firm Performance and Annual Report Fog and Length (change specification)**

| | Dependent variables | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] | [11] | [12] |
| | ***Whole annual report*** | | | | ***MD&A section*** | | | | ***Notes to the financial statements*** | | | |
| | Fog | Fog | Length | Length | Fog | Fog | Length | Length | Fog | Fog | Length | Length |
| Change in Earnings | -0.238 | | -0.194 | | -0.399 | | -0.012 | | -0.317 | | -0.238 | |
| | [-2.79]*** | | [-5.37]*** | | [-4.87]*** | | [-0.23] | | [-3.32]*** | | [-5.47]*** | |
| Earnings +/- dummy | | -0.094 | | -0.053 | | -0.117 | | 0.016 | | -0.066 | | -0.061 |
| | | [-4.85]*** | | [-5.56]*** | | [-4.31]*** | | [1.24] | | [-3.37]*** | | [-5.89]*** |
| Constant | -0.073 | -0.064 | -0.183 | -0.183 | 0.926 | 0.899 | -0.293 | -0.295 | 0.043 | 0.040 | 0.021 | 0.02 |
| | [-0.18] | [-0.16] | [-1.10] | [-1.10] | [0.97] | [0.93] | [-0.58] | [-0.59] | [0.08] | [0.07] | [0.05] | [0.05] |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 34481 | 34481 | 34481 | 34481 | 23606 | 23606 | 23606 | 23606 | 27526 | 27526 | 27526 | 27526 |
| R-squared | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 |

Note: This Table shows the regression results of the annual report readability on firm performance using the level specification (Panel A) and change specification (Panel B). The dependent variables are *Fog* and *Length* of the whole annual report and the MD&A or Note to financial statements. *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4.*Length* is the natural logarithm of the annual report. MD&A Fog and Notes Fog are the Fog index of MD&A section and the Notes to the financial statements. MD&A Length and Notes Length are the length of MD&A section and the Notes to the financial statements. When MD&A Fog or MD&A Length is used in the regression, the MD&A section needs to contain at least 100 words. When Notes Fog or Notes Length is used in the regression, the Notes to the financial statements need to contain at least 1000 words. Earnings is operating earnings (data178 of Compustat) scaled by assets. Profit/Loss dummy is a dummy variable that equals 1 if a company reports a profit and 0 otherwise. Earnings +/- dummy is a dummy variable that equals 1 if a company reports an increase in operating earnings and 0 otherwise.

The control variables include SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, SEO, MA, and DLW. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise.

T-statistics in parentheses are based on standard errors clustered at two-digit SIC industry code level.

**Table 4 Panel A: Earnings Persistence and Annual Report Fog Index (Profit Firm-years)**

| | Dependent variables | | | | | |
|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] |
| | ***The whole annual report*** | | ***MD&A section*** | | ***Notes to the financial statements*** | |
| | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) |
| Earnings(t) | 0.026 | -0.057 | 0.300 | 0.864 | 0.221 | 0.126 |
| | [0.03] | [-0.06] | [0.29] | [0.71] | [0.24] | [0.13] |
| Fog | 0.003 | 0.004 | 0.001 | 0.003 | 0.002 | 0.002 |
| | [4.01]*** | [3.04]*** | [2.16]** | [2.63]** | [2.61]** | [2.85]*** |
| Earnings(t)*Fog | -0.028 | -0.041 | -0.016 | -0.036 | -0.022 | -0.023 |
| | [-3.74]*** | [-2.95]*** | [-3.13]*** | [-3.00]*** | [-2.71]*** | [-3.02]*** |
| Constant | 0.038 | 0.002 | 0.054 | -0.043 | 0.019 | -0.024 |
| | [0.39] | [0.02] | [0.56] | [-0.40] | [0.19] | [-0.26] |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 22798 | 19089 | 18533 | 15744 | 20569 | 17546 |
| R-squared | 0.41 | 0.26 | 0.42 | 0.26 | 0.41 | 0.26 |

**Table 4 Panel B: Earnings Persistence and Annual Report Length (Profit Firm-years)**

| | Dependent variables | | | | | |
|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] |
| | ***The whole annual report*** | | ***MD&A section*** | | ***Notes to the financial statements*** | |
| | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) |
| Earnings(t) | -0.026 | -0.267 | -0.162 | -0.434 | -0.183 | -0.14 |
| | [-0.02] | [-0.22] | [-0.15] | [-0.35] | [-0.20] | [-0.13] |
| Length | 0.005 | 0.006 | 0.002 | 0.001 | 0.002 | 0.004 |
| | [2.56]** | [2.93]*** | [0.91] | [0.56] | [1.90]* | [2.22]** |
| Earnings(t)*Length | -0.060 | -0.075 | -0.019 | -0.021 | -0.025 | -0.036 |
| | [-2.97]*** | [-3.48]*** | [-1.23] | [-0.79] | [-2.87]*** | [-2.31]** |
| Constant | 0.053 | 0.037 | 0.085 | 0.068 | 0.055 | -0.008 |
| | [0.51] | [0.34] | [0.81] | [0.63] | [0.56] | [-0.07] |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 22798 | 19089 | 18533 | 15744 | 20569 | 17546 |
| R-squared | 0.41 | 0.26 | 0.42 | 0.26 | 0.41 | 0.26 |

**Table 4 Panel C: Earnings Persistence and Annual Report Readability (Profit Firm-years)**

| | Dependent variables | | | | |
|---|---|---|---|---|---|
| | [1] Earn(t+1) | [2] Earn(t+2) | | [3] Earn(t+1) | [4] Earn(t+2) |
| Earnings(t) | 0.993 | 1.629 | Earnings(t) | 0.527 | 0.706 |
| | [1.00] | [1.39] | | [0.48] | [0.53] |
| Fog | 0.001 | 0.002 | Length | 0.003 | 0.003 |
| | [1.89]* | [1.02] | | [1.19] | [0.81] |
| Earnings(t)*Fog | -0.012 | -0.024 | Earnings(t)*Length | -0.048 | -0.072 |
| | [-1.53] | [-1.53] | | [-1.96]* | [-2.47]** |
| MD&A Fog | 0.001 | 0.003 | MD&A Length | 0.001 | 0.001 |
| | [1.09] | [2.39]** | | [0.42] | [0.21] |
| Earnings(t)*MD&A Fog | -0.010 | -0.029 | Earnings(t)*MD&A Length | -0.009 | -0.009 |
| | [-1.78]* | [-2.61]** | | [-0.53] | [-0.30] |
| Notes Fog | 0.001 | 0.001 | Notes Length | 0.001 | 0.003 |
| | [1.41] | [0.39] | | [0.58] | [0.77] |
| Earnings(t)*Notes Fog | -0.015 | -0.004 | Earnings(t)*Notes Length | -0.006 | -0.003 |
| | [-1.66] | [-0.33] | | [-0.48] | [-0.10] |
| Constant | -0.015 | -0.12 | Constant | 0.032 | -0.014 |
| | [-0.16] | [-1.30] | | [0.30] | [-0.13] |
| Year dummies | Yes | Yes | Year dummies | Yes | Yes |
| Industry dummies | Yes | Yes | Industry dummies | Yes | Yes |
| Control variables | Yes | Yes | Control variables | Yes | Yes |
| Observations | 17233 | 14813 | Observations | 17233 | 14813 |
| R-squared | 0.42 | 0.27 | R-squared | 0.42 | 0.26 |

Note: This table shows the effect of annual report readability on earnings persistence by regressing future earnings on current earnings, readability index, and their interactions using profit firm-years. The sample is all firm-years that report profits. The dependent variables are earnings of year t+1 to year t+4. *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words. Earnings is operating earnings (data178 of Compustat) scaled by assets. MD&A Fog and Notes Fog are the Fog index of MD&A section and the Notes to the financial statements. MD&A Length and Notes Length are the *Length* of the MD&A section and the Notes to the financial statements. When MD&A Fog or MD&A Length is used in the regression, the MD&A section needs to contain at least 100 words. When Notes Fog or Notes Length is used in the regression, the Notes to the financial statements needs to contain at least 1000 words.

The control variables include ACC, DIV, SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, SEO, MA, DLW *and* their interactions with earnings. Accruals is calculated as (data178-data308)/data6. DIV is a dummy that equals one if a firm has dividend (i.e., data21>0) this year and zero otherwise. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise.

T-statistics in parentheses are based on standard errors clustered at two-digit SIC industry code level.

**Table 5 Panel A: Earnings Persistence and Annual Report Fog Index (Loss Firm-years)**

| | Dependent variables | | | | | |
|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] |
| | ***MD&A section*** | | | | | |
| | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) |
| Earnings(t) | -0.390 | 1.258 | 1.058 | 1.845 | -0.408 | 0.223 |
| | [-0.20] | [0.73] | [0.44] | [0.79] | [-0.25] | [0.11] |
| Fog | -0.004 | -0.005 | -0.001 | -0.003 | -0.001 | -0.004 |
| | [-1.37] | [-1.55] | [-1.04] | [-1.96]* | [-0.23] | [-1.20] |
| Earnings(t)*Fog | -0.014 | -0.011 | 0.006 | 0.000 | -0.004 | -0.015 |
| | [-0.80] | [-0.77] | [0.88] | [0.05] | [-0.43] | [-1.09] |
| Constant | -0.388 | -0.308 | -0.153 | -0.201 | -0.373 | -0.325 |
| | [-1.39] | [-1.06] | [-0.43] | [-0.58] | [-1.42] | [-1.14] |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 6961 | 5205 | 5420 | 4140 | 5898 | 4565 |
| R-squared | 0.41 | 0.29 | 0.41 | 0.29 | 0.41 | 0.29 |

**Table 5 Panel B: Earnings Persistence and Annual Report Length (Loss Firm-years)**

| | Dependent variables | | | | | |
|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] |
| | ***The whole annual report*** | | ***MD&A section*** | | ***Notes to the financial statements*** | |
| | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) | Earn(t+1) | Earn(t+2) |
| Earnings(t) | -0.707 | 2.006 | 1.202 | 1.788 | -0.77 | 0.24 |
| | [-0.40] | [1.10] | [0.55] | [0.74] | [-0.52] | [0.12] |
| Length | -0.003 | -0.006 | 0.004 | 0.006 | 0.005 | -0.003 |
| | [-0.64] | [-1.06] | [1.25] | [1.55] | [1.89]* | [-0.64] |
| Earnings(t)*Length | 0.001 | -0.053 | 0.005 | 0.012 | 0.016 | -0.029 |
| | [0.04] | [-2.35]** | [0.32] | [0.81] | [0.98] | [-1.61] |
| Constant | -0.449 | -0.307 | -0.209 | -0.301 | -0.462 | -0.378 |
| | [-1.90]* | [-1.14] | [-0.66] | [-0.81] | [-1.98]* | [-1.29] |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 6961 | 5205 | 5420 | 4140 | 5898 | 4565 |
| R-squared | 0.41 | 0.29 | 0.41 | 0.29 | 0.41 | 0.29 |

Note: This Panel shows the effect of annual report readability on earnings persistence by regressing future earnings on current earnings, readability index, and their interactions using loss firm-years. The sample is all firm-years that report losses. The dependent variables are earnings of year t+1 to year t+4. *Fog* is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words. Earnings is operating earnings (data178 of Compustat) scaled by assets.

The control variables include ACC, DIV, SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, SEO, MA, DLW *and* their interactions with earnings. Accruals is calculated as (data178-data308)/data6. DIV is a dummy that equals one if a firm has dividend (i.e., data21>0) this year and zero otherwise. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise.

T-statistics in parentheses are based on standard errors clustered at two-digit SIC industry code level.

**Table 6 Panel A: Summary Statistics of Annual Report Writing Styles**

| Variable | Mean | Median | S.D. | 1st | 25th | 75th | 99th | Obs |
|---|---|---|---|---|---|---|---|---|
| IvsU | 0.26 | 0.00 | 0.57 | -0.36 | -0.07 | 0.27 | 1.74 | 43332 |
| EvsI | -1.32 | -1.31 | 0.31 | -2.15 | -1.49 | -1.13 | -0.55 | 43332 |
| Cause | 1.35 | 1.28 | 0.52 | 0.41 | 1.04 | 1.58 | 3.02 | 43332 |
| PvsN | 0.64 | 0.64 | 0.32 | -0.13 | 0.44 | 0.85 | 1.40 | 43332 |
| FvsP | -1.15 | -1.17 | 0.29 | -1.81 | -1.33 | -0.99 | -0.30 | 43332 |

**Table 6 Panel B: Pearson Correlation Coefficient of Annual Report Writing Styles with Fog and Length**

| | Length | MD&A Length | Notes Length | IvsU | EvsI | Cause | PvsN | FvsP | Fog | MD&A Fog | Note Fog |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Length | | | | | | | | | | | |
| MD&A Length | **0.270** | | | | | | | | | | |
| Notes Length | **0.670** | **0.191** | | | | | | | | | |
| IvsU | **0.157** | **0.074** | **0.093** | | | | | | | | |
| EvsI | **0.081** | **0.336** | **0.050** | **0.079** | | | | | | | |
| Cause | -0.010 | **-0.202** | **-0.045** | **0.113** | **0.076** | | | | | | |
| PvsN | **0.097** | **0.147** | **0.097** | **-0.060** | **0.044** | **-0.202** | | | | | |
| FvsP | **0.138** | **0.193** | **0.068** | **0.105** | **0.371** | **0.125** | **-0.078** | | | | |
| Fog | **0.398** | **0.039** | **0.250** | **-0.053** | **0.069** | **0.030** | **0.026** | **0.133** | | | |
| MD&A Fog | **0.114** | **-0.189** | **0.094** | **-0.032** | **0.091** | **0.217** | **0.075** | **0.204** | **0.368** | | |
| Notes Fog | **0.059** | **-0.045** | **0.401** | **-0.155** | **-0.056** | **-0.022** | 0.012 | **-0.025** | **0.074** | **0.047** | |

Note: Panel A shows the summary statistics of annual report length (LENGTH) and five categories of writing styles (IvsU, EvsI, Cause, PvsN, and FvsP). Panel B shows the Pearson correlation coefficient of the variables. The correlation coefficients in bold are significant at 0.01 level. Length is the logarithm of the number of words in an annual report. MD&A Length and Notes Length are the logarithms of the number of words in the MD&A section and the Notes to the financial statements. IvsU is log((1+Self)/(1+You+Other)), where Self is the percentage of first person pronouns in the MD&A section. You and Other are the percentage of second and third person pronouns in the MD&A section. EvsI is log((1+Excl)/(1+Incl)), where Excl is the percentage of exclusive words and Incl is the percentage of inclusive words in the MD&A section. Cause is the percentage of causation words in the MD&A section. PvsN is log((1+Posemo)/(1+Negemo)), where Posemo is the percentage of positive emotion words and Negemo is the percentage of negative emotion words in the MD&A section. FvsP is log((1+Future)/(1+Past+Present)), where Future is the percentage of future tense verbs and Past and Present are the percentages of past and present tense verbs in the MD&A section. Fog is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. Length is the natural logarithm of the number of words. MD&A Fog (Length) and Notes Fog (Length) are the Fog (Length) of the MD&A section and the Notes to the financial statements.

**Table 7: Firm Performance and Writing Styles**

| | Dependent variables | | | | |
|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] |
| | IvsU | EvsI | Cause | PvsN | FvsP |
| Earnings | -0.334 | -0.037 | -0.012 | 0.015 | -0.188 |
| | [-8.49]*** | [-2.26]** | [-0.41] | [0.42] | [-5.60]*** |
| SIZE | 0.036 | -0.011 | 0.003 | 0.023 | 0.012 |
| | [9.63]*** | [-4.53]*** | [1.34] | [7.12]*** | [6.85]*** |
| MTB | 0.004 | 0.002 | 0.004 | 0.001 | 0.005 |
| | [1.58] | [2.45]** | [2.52]** | [0.45] | [7.37]*** |
| AGE | -0.007 | -0.002 | -0.003 | 0.001 | -0.003 |
| | [-7.90]*** | [-7.45]*** | [-5.87]*** | [2.90]*** | [-9.50]*** |
| SI | 0.001 | 0.016 | 0.035 | 0.009 | 0.012 |
| | [0.03] | [1.16] | [1.08] | [0.52] | [0.58] |
| RET_VOL | 0.579 | 0.076 | 0.045 | -0.083 | 0.138 |
| | [11.74]*** | [3.13]*** | [1.00] | [-2.17]** | [5.39]*** |
| NBSEG | 0.001 | -0.006 | -0.018 | -0.002 | -0.009 |
| | [0.04] | [-1.43] | [-1.90]* | [-0.50] | [-2.09]** |
| NGSEG | 0.019 | -0.001 | 0.030 | -0.011 | -0.005 |
| | [0.85] | [-0.29] | [4.28]*** | [-2.92]*** | [-1.01] |
| NITEMS | -0.084 | -0.016 | 0.099 | -0.181 | -0.021 |
| | [-1.31] | [-0.29] | [1.83]* | [-2.24]** | [-0.40] |
| SEO | 0.143 | 0.011 | -0.030 | 0.006 | 0.015 |
| | [7.65]*** | [1.40] | [-2.55]** | [0.62] | [2.04]** |
| MA | 0.021 | -0.016 | -0.003 | 0.010 | -0.006 |
| | [2.62]** | [-3.80]*** | [-0.49] | [2.27]** | [-1.25] |
| DLW | 0.001 | 0.030 | 0.006 | 0.000 | 0.033 |
| | [0.07] | [2.69]*** | [0.46] | [0.04] | [3.46]*** |
| Constant | 0.304 | -1.268 | 0.821 | 1.515 | -1.142 |
| | [0.91] | [-4.26]*** | [2.85]*** | [3.46]*** | [-4.05]*** |
| Year dummies | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes |
| Observations | 32099 | 32099 | 32099 | 32099 | 32099 |
| R-squared | 0.34 | 0.09 | 0.09 | 0.2 | 0.11 |

Note: This table shows the regression results of the annual report writing style measures on firm performance. The dependent variables are IvsU, EvsI, Cause, PvsN, and FvsP. IvsU is log((1+Self)/(1+You+Other)), where Self is the percentage of first person pronouns in the MD&A section. You and Other are the percentage of second and third person pronouns in the MD&A section. EvsI is log((1+Excl)/(1+Incl)), where Excl is the percentage of exclusive words and Incl is the percentage of inclusive words in the MD&A section. Cause is the percentage of causation words in the MD&A section. PvsN is log((1+Posemo)/(1+Negemo)), where Posemo is the percentage of positive emotion words and Negemo is the percentage of negative emotion words in the MD&A section. FvsP is log((1+Future)/(1+Past+Present)), where Future is the percentage of future tense verbs and Past and Present are the percentages of past and present tense verbs in the MD&A section. Earnings is operating earnings (data178 of Compustat) scaled by assets. The control variables include SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, SEO, MA, and DLW. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise.

T-statistics in parentheses are based on standard errors clustered at two-digit SIC industry code level.

## Table 8: Earnings Persistence and Writing Styles

| | Dependent variables | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sample: Profitable firms | | | | | Sample: Loss firms | | | | |
| | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] |
| Earnings | -0.360 | -0.283 | -0.386 | -0.410 | -0.351 | 0.927 | 0.957 | 0.950 | 1.288 | 1.070 |
| | [-0.34] | [-0.26] | [-0.35] | [-0.36] | [-0.32] | [0.41] | [0.42] | [0.42] | [0.62] | [0.47] |
| IvsU | -0.003 | | | | | -0.015 | | | | |
| | [-0.59] | | | | | [-1.95]* | | | | |
| Earnings*IvsU | -0.025 | | | | | -0.022 | | | | |
| | [-0.53] | | | | | [-0.82] | | | | |
| EvsI | | 0.000 | | | | | -0.006 | | | |
| | | [0.05] | | | | | [-0.63] | | | |
| Earnings*EvsI | | 0.005 | | | | | -0.002 | | | |
| | | [0.10] | | | | | [-0.04] | | | |
| Cause | | | 0.003 | | | | | -0.004 | | |
| | | | [1.39] | | | | | [-0.92] | | |
| Earnings*Cause | | | -0.049 | | | | | -0.006 | | |
| | | | [-2.27]** | | | | | [-0.27] | | |
| PvsN | | | | -0.006 | | | | | -0.018 | |
| | | | | [-0.84] | | | | | [-1.40] | |
| Earnings*PvsN | | | | 0.089 | | | | | -0.111 | |
| | | | | [1.71]* | | | | | [-2.25]** | |
| FvsP | | | | | 0.002 | | | | | -0.006 |
| | | | | | [0.46] | | | | | [-0.60] |
| Earnings*FvsP | | | | | -0.118 | | | | | 0.041 |
| | | | | | [-2.77]*** | | | | | [0.80] |
| Constant | 0.101 | 0.095 | 0.106 | 0.108 | 0.093 | -0.275 | -0.279 | -0.267 | -0.211 | -0.279 |
| | [0.97] | [0.87] | [0.95] | [0.94] | [0.86] | [-0.81] | [-0.81] | [-0.79] | [-0.64] | [-0.79] |
| Year dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Industry dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Control variables | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 18507 | 18507 | 18507 | 18507 | 18507 | 5418 | 5418 | 5418 | 5418 | 5418 |
| R-squared | 0.42 | 0.42 | 0.42 | 0.42 | 0.42 | 0.41 | 0.41 | 0.41 | 0.41 | 0.41 |

Note: This table shows the effect of annual report writing styles on earnings persistence by regressing future earnings on current earnings, the writing style measures, and their interactions. The samples in columns [1] to [5] are firms that report profits, and those in columns [6] to [10] are all firm-years that report losses. The dependent variables are earnings of year t+1, scaled by book value of assets. The five categories of writing styles (IvsU, EvsI, Cause, PvsN, and FvsP) are defined as follows: IvsU is log((1+Self)/(1+You+Other)), where Self is the percentage of first person pronouns in the MD&A section. You and Other are the percentage of second and third person pronouns in the MD&A section. EvsI is log((1+Excl)/(1+Incl)), where Excl is the percentage of exclusive words and Incl is the percentage of inclusive words in the MD&A section. Cause is the percentage of causation words in the MD&A section. PvsN is log((1+Posemo)/(1+Negemo)), where Posemo is the percentage of positive emotion words and Negemo is the percentage of negative emotion words in the MD&A section. FvsP is log((1+Future)/(1+Past+Present)), where Future is the percentage of future tense verbs and Past and Present are the percentages of past and present tense verbs in the MD&A section.

The control variables include ACC, DIV, SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, SEO, MA, DLW *and* their interactions with earnings. Accruals is calculated as (data178-data308)/data6. DIV is a dummy that equals one if a firm has dividend (i.e., data21>0) this year and zero otherwise. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise.

T-statistics in parentheses are based on standard errors clustered at two-digit SIC industry code level.

## Table 9: The Effect of Executive Option Holdings

| **Panel A: *Fog* and earnings persistence** | | | | |
|---|---|---|---|---|
| | Dependent variables | | | |
| | [1] Earnings(t+1) | [2] Earnings(t+2) | [3] Earnings(t+3) | [4] Earnings(t+4) |
| Earnings | 1.869 | 0.978 | 0.602 | -1.964 |
| | [1.87]* | [0.61] | [0.42] | [-1.06] |
| Fog | 0.001 | 0.002 | 0.000 | -0.001 |
| | [1.50] | [1.34] | [0.21] | [-0.75] |
| Earnings*Fog | -0.017 | -0.023 | -0.006 | 0.005 |
| | [-1.82]* | [-2.12]** | [-0.50] | [0.35] |
| UNEX_OPT | -0.028 | -0.024 | -0.037 | -0.071 |
| | [-2.45]** | [1.89]* | [-1.88]* | [-2.31]** |
| Earnings *UNEX_OPT | 0.218 | 0.210 | 0.379 | 0.710 |
| | [2.30]** | [2.06]** | [2.24]** | [2.89]*** |
| Fog*UNEX_OPT | 0.001 | 0.001 | 0.002 | 0.004 |
| | [2.34]** | [1.72]* | [1.80]* | [2.30]** |
| Earnings*Fog*UNEX_OPT | -0.011 | -0.011 | -0.020 | -0.037 |
| | [-2.31]** | [-2.00]** | [-2.22]** | [-2.92]*** |
| **Panel B: *Length* and earnings persistence** | | | | |
| | Dependent variables | | | |
| | [1] Earnings(t+1) | [2] Earnings(t+2) | [3] Earnings(t+3) | [4] Earnings(t+4) |
| Earnings | 1.818 | 0.64 | 0.598 | -1.706 |
| | [1.89]* | [0.40] | [0.42] | [-1.01] |
| Fog | 0.004 | 0.002 | 0.002 | 0.000 |
| | [1.94]* | [0.72] | [0.55] | [0.13] |
| Earnings*Fog | -0.036 | -0.03 | -0.022 | -0.008 |
| | [-2.14]** | [-1.64] | [-0.95] | [-0.30] |
| UNEX_OPT | -0.015 | -0.010 | -0.023 | -0.052 |
| | [-1.30] | [-0.79] | [-1.54] | [-2.02]** |
| Earnings *UNEX_OPT | 0.123 | 0.076 | 0.249 | 0.491 |
| | [1.31] | [0.77] | [1.99]* | [2.41]** |
| Fog*UNEX_OPT | 0.001 | 0.001 | 0.002 | 0.005 |
| | [1.17] | [0.62] | [1.42] | [1.99]* |
| Earnings*Fog*UNEX_OPT | -0.012 | -0.008 | -0.025 | -0.049 |
| | [-1.30] | [-0.73] | [-1.94]* | [-2.39]** |

Note: All the regressions in this table are based on the sub-sample of profit firm-years. The dependent variables are earnings of year t+1 to year t+4. Fog is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. *Length* is the natural logarithm of the number of words in annual reports. Earnings is operating earnings (data178 of Compustat) scaled by assets. UNEX_OPT in Panel A is the logarithm of (# of unexercised stock options owned by the CEO/# of shares owned by the CEO), both of which are from the EXECUCOMP database.

The control variables (unreported) include ACC, DIV, SIZE, MTB, AGE, SI, RET_VOL, EARN_VOL, NBSEG, NGSEG, NITEMS, MKT_RET, HINDEX, HTECH, PLIT, SEO, MA, DLW *and* their interactions with earnings. Accruals is calculated as (data178-data308)/data6. DIV is a dummy that equals one if a firm has

dividend (i.e., data21>0) this year and zero otherwise. SIZE is the logarithm of market value of equity calculated as Log(data25*data199). MTB is the market value of the firm divided by its book value ((data25*data199+data181)/data6). AGE is the number of years since a firm shows up in CRSP monthly stock return files. SI is special items (data17) scaled by book value of assets. RET_VOL is the standard deviation of the monthly stock returns in the last year. EARN_VOL is the standard deviation of the operating earnings in the last five fiscal years. NBSEG is the logarithm of one plus the number of business segments and NGSEG is the logarithm of one plus the number of geographic segments. NITEMS is the number of non-missing items on Compustat. SEO is a dummy that equals one if a firm has seasoned equity offering in this year according to SDC Global New Issues database and zero otherwise. MA is a dummy that equals one if a firm appears as an acquirer in this year in SDC Platinum M&A database and zero otherwise. DLW is a dummy that equals one if a company is incorporated in Delaware and zero otherwise. Year and industry-fixed effects are also included.

T-statistics in parentheses are based on standard errors clustered at two-digit SIC industry code level.

## Table 10 Panel A: Fama-MacBeth Regressions of Future Returns on *Fog*

| **Sample: Profit firm-years** | Dependent variables | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] |
| | Ret (t+1) | Ret (t+2) | Ret (t+3) | Ret (t+4) | Ret (t+1) | Ret (t+2) | Ret (t+3) | Ret (t+4) |
| Fog | -0.001 | 0.004 | 0.008 | 0.012 | -0.002 | 0.001 | 0.010 | 0.012 |
| | [-0.16] | [0.65] | [0.99] | [1.27] | [-0.54] | [0.23] | [1.04] | [0.94] |
| Earnings | | | | | -0.222 | -0.562 | 0.285 | 0.049 |
| | | | | | [-0.48] | [-0.71] | [0.41] | [0.05] |
| Earnings*Fog | | | | | 0.014 | 0.024 | -0.017 | -0.009 |
| | | | | | [0.60] | [0.55] | [-0.46] | [-0.17] |
| Constant | 0.199 | 0.120 | 0.048 | -0.011 | 0.219 | 0.180 | 0.023 | 0.001 |
| | [2.73]*** | [1.01] | [0.29] | [-0.06] | [3.29]*** | [1.40] | [0.12] | [0.01] |
| Number of years | 10 | 9 | 8 | 7 | 10 | 9 | 8 | 7 |
| Average observations | 3024 | 2654 | 2438 | 2174 | 3022 | 2652 | 2436 | 2172 |
| Average R-squared | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |

## Table 10 Panel B: Fama-MacBeth Regressions of Future Returns on *Length*

| **Sample: Profit firm-years** | Dependent variables | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] |
| | Ret (t+1) | Ret (t+2) | Ret (t+3) | Ret (t+4) | Ret (t+1) | Ret (t+2) | Ret (t+3) | Ret (t+4) |
| Length | -0.011 | 0.006 | -0.000 | 0.012 | -0.006 | 0.005 | 0.003 | 0.003 |
| | [-0.84] | [0.51] | [-0.02] | [0.63] | [-0.33] | [0.23] | [0.14] | [0.09] |
| Earnings | | | | | 0.496 | -0.046 | 0.305 | -0.628 |
| | | | | | [0.50] | [-0.03] | [0.26] | [-0.29] |
| Earnings*Length | | | | | -0.047 | -0.006 | -0.035 | 0.062 |
| | | | | | [-0.44] | [-0.04] | [-0.29] | [0.26] |
| Constant | 0.295 | 0.137 | 0.214 | 0.101 | 0.249 | 0.158 | 0.181 | 0.194 |
| | [3.13]*** | [1.34] | [1.39] | [0.53] | [1.61] | [0.77] | [0.74] | (0.59] |
| Number of years | 10 | 9 | 8 | 7 | 10 | 9 | 8 | 7 |
| Average observations | 3024 | 2654 | 2438 | 2174 | 3022 | 2652 | 2436 | 2172 |
| Average R-squared | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |

Note: The dependent variables are annual returns of year t+1 to year t+4, starting from the month after the annual report filing date. Fog is the Fog Index calculated as (words per sentence + percent of complex words) * 0.4. Length is the natural logarithm of the number of words in the annual reports. Earnings is operating earnings (data178 of Compustat) scaled by assets.

T-statistics in parentheses are based on the coefficients from the annual cross-sectional regressions.