Available online at www.sciencedirect.com
ScienceDirect
Journal of Accounting and Economics 45 (2008) 221–247
JOURNAL OF Accounting & Economics
www.elsevier.com/locate/jae

Annual report readability, current earnings, and earnings persistence ☆

Feng Li*

Ross School of Business, University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109, USA

Received 28 February 2006; received in revised form 12 February 2008; accepted 13 February 2008
Available online 4 March 2008

Abstract

This paper examines the relation between annual report readability and firm performance and earnings persistence. I measure the readability of public company annual reports using the Fog index from the computational linguistics literature and the length of the document. I find that: (1) the annual reports of firms with lower earnings are harder to read (i.e., they have a higher Fog index and are longer); and (2) firms with annual reports that are easier to read have more persistent positive earnings.

JEL classification: D82; G18; M41; M45; G14

Keywords: Disclosure; Annual report readability; Profitability; Earnings persistence

1. Introduction

If markets react less completely to information that is less easily extracted from public disclosures, then managers have more incentive to obfuscate information when firm performance is bad (Bloomfield, 2002). Consistent with this "management obfuscation hypothesis," prior research finds that management is willing to be more forthcoming in the disclosure of information when their respective firms are performing well (e.g., Lang and Lundholm, 1993; Schrand and Walther, 2000). I test this hypothesis by examining whether annual reports of firms that are performing poorly are more difficult and complicated, and whether the positive earnings of firms with more complex annual reports are less persistent and the negative earnings of the firms with more complex annual reports are more persistent in the immediately following years.

☆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 Daniel Cohen, John Core, Ilia Dichev, Richard Frankel, Scott Richardson, Doug Skinner, Suraj Srinivasan, Franco Wong, and the participants of the Hosmer Lunch Workshop at the University of Michigan and the 2006 JAE conference participants for comments. The comments and suggestions of Bob Holthausen (the editor) and Rob Bloomfield (the discussant and referee) greatly improved the paper. All errors remain my own.
*Tel.: +1 734 936 2771; fax: +1 734 936 0274.
E-mail address: Feng@umich.edu

0165-4101/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jacceco.2008.02.003

I measure annual report readability using two variables. The first variable is the *Fog* index from the computational linguistics literature, which statistic combines the number of words per sentence and the number of syllables per word to create a measure of readability. The *Fog* index proposes that, assuming everything else to be equal, more syllables per word or more words per sentence make a document harder to read. The second measure of readability is the length of the annual report. The reasoning behind this choice is the presumption that longer documents are more deterring and require higher costs of information-processing. 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 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 lower earnings persistence when they are profitable. The effect is significant both economically and statistically. An inter-quartile change in readability has a similar impact on earnings persistence when compared to the effect of an inter-quartile change in the absolute amount of total accruals scaled by book value of assets.

Other lexical features of the annual reports also have systematic associations with earnings persistence—confirming the readability-based findings. For profitable firms, a higher frequency of causation words (such as "because") in the Management Discussion and Analysis (MD&A) section is associated with less persistent earnings; a higher occurrence of positive emotion words (relative to negative emotion words) is associated with more persistent earnings; and a higher frequency of future tense verbs (relative to past/present tense verbs) is indicative of a lower earnings persistence. On the other hand, loss (i.e., non-profitable) firms with a higher occurrence of positive emotion words (relative to negative emotion words) in their MD&A have less persistent earnings.

Viewed collectively, the evidence in this paper suggests a clear correlation between the linguistic features of annual reports and firm performance. This paper contributes to the literature in three ways. First, this is the first large-sample study to examine the cross-sectional variation in annual report readability and other lexical properties and their subsequent 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 finding lends further support to the "incomplete revelation hypothesis" in Bloomfield (2002).

Second, much of the literature on empirical disclosure has focused on the determinants and consequences of the amount of disclosure (e.g., Miller, 2002) and most papers have small sample sizes—mainly due to the manually coded measure of disclosure quality.[1] 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 possess a lower quality of disclosure, 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). However, the prior research generally does not examine the association between firm disclosure quality and earnings quality.[2] 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 et al., 2005a), few of these examine the implications of disclosure quality on future earnings. This paper extends the previous literature by illustrating that the quality of disclosure is correlated with earnings persistence and contains information regarding earnings quality.

Several caveats are in order. As discussed in Bloomfield (2008), there are several alternative explanations to my empirical findings. For instance, bad news may be inherently more complicated to articulate and investors could demand more information from managers when there is bad news. Future research is required to distinguish between these explanations. Second, an annual report is just one of the many ways that managers communicate with investors. Managers face more constraints communicating to investors through annual

[1]A few exceptions include Rogers (2004) and Berger et al. (2006).

[2]One exception is Francis et al. (2005b), who examine the relation between voluntary disclosure and accrual quality.

reports than through other channels (e.g., conference calls). Therefore, alternate communication channels may provide better models in which to explore the relation between readability and firm performance.

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 several additional empirical tests in Section 6. Section 7 concludes.

2. Literature and hypotheses

2.1. Literature

Existing research on corporate disclosure has focused mostly on the amount of disclosure made by firms (Healy and Palepu, 2001). An important dimension of disclosure—the lexical properties—has not been studied systematically, even though regulators and investors pay a significant amount of attention to these properties. For instance, the SEC has continually attempted to make public company prospectuses more readable and easier to comprehend. In several Securities Act Releases after the 1933 Securities Act, a higher level of clarity in the disclosure documents was encouraged—with an emphasis on not compromising full and fair disclosure (Firtel, 1999). In 1967, the SEC constituted an internal study group in order to examine and make recommendations for improving its disclosure regime. This study resulted in the 1969 "*Wheat Report*." Among other findings, *the Wheat Report* noted that the average investor could not readily understand the complicated prospectuses; the report therefore recommended that companies avoid unnecessarily complex, lengthy or verbose writing.

In October 1998, the SEC issued new plain English disclosure guidelines that encouraged the use of plain English in the drafting and formatting of all prospectuses in registered public offerings by domestic and foreign issuers. The SEC's Investor Ed Office published and posted the following on its website: "A plain English handbook: how to create clear SEC disclosure documents" in order to provide practical tips for drafting disclosure documents. For instance, when drafting the front and back cover pages, the summary and the risk factors sections, an issuer must comply with the following six basic principles: short sentences; definitive, concrete, everyday language; the use of the 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).

Surprisingly—given the importance of the corporate disclosures to regulators and investors—there is little large-sample empirical evidence on these documents' linguistic features. Jones and Shoemaker (1994) provide a review of 32 studies in the fields of accounting, business communication and management on the readability of annual report narratives (26 studies), tax law (3 studies) and accounting textbook (3 studies).

Most of these studies attempt to assess the readability 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 within the financial statements of 50 New Zealand firms. Lebar (1982) studies the Forms 10-Ks, annual reports and press releases 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 quite difficult to read and may be classified as technical literature, which classification risks "being inaccessible to a large proportion of private lay shareholders" (Jones and Shoemaker, 1994). Some studies 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 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; Barnett and Leoffler, 1979) and corporate profitability (Courtis, 1986; Baker and Kare, 1992; Subramanian et al., 1993). The evidence found therein is also mixed and inconclusive. For instance, Courtis (1986) does not find a strong correlation between readability and net profits and return on capital. However, Subramanian et al. (1993) find that the annual reports of profitable firms are significantly easier to read than those of poor performers.

It is important to bear in mind, however, that the sample sizes of the previous studies are very small. Only two of the 32 studies reviewed by Jones and Shoemaker (1994) have a sample size larger than 100. Among the 16 papers examined in Table I of Clatworthy and Jones (2001), 14 have a sample size of 50 or smaller and the largest sample size is 120. This fact may explain the mixed findings of the prior studies. What's more important is that none of the prior studies examine the implications of disclosure features for earnings persistence, which implications are likely to be more important than current profitability when examining management obfuscation.

In this paper, I extend this literature by 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

The management obfuscation hypothesis argues that managers have incentives to obfuscate information when firm performance is poor because the market may react with a delayed incorporation of the information contained in complicated disclosures (Bloomfield, 2002). The maintained assumption behind this argument 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; Bloomfield, 2002), managers may want to strategically hide adverse information through less transparent disclosures. In particular, Bloomfield (2002) argues that managers make many decisions motivated, at least partly, by a desire to make it more difficult for investors to uncover information that the managers do not want uncovered—as it would 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 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 line items on income statements are also less persistent than those disclosed solely in footnotes (Riedl and Srinivasan, 2005). The managerial obfuscation hypothesis thus predicts a negative relation between a firm's current performance and its annual report's level of complexity.

However, this hypothesized relation between disclosure readability and a firm's current performance may not be significant. First, annual reports contain a large amount of financial information regarding current and historical performance. Hence, the benefit of writing more complicated annual reports in order to hide adverse information regarding current performance seems slight. Second, if good current reported earnings are due (partially) to strategic manipulation, then managers may not necessarily want to make the annual reports easier to read.

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 be an attempt to hide or delay the future discovery of adverse information. Therefore, I further examine the implication of annual report readability for future performance—focusing particularly on earnings persistence. Earnings persistence captures information regarding future earnings and provides a better setting in which to examine potential management obfuscation behavior.

2.2.2. Future performance

The aforementioned presupposition regarding the relation between disclosure quality and a firm's current performance can be extended to its future performance. Opportunistic managers may have incentives to make the annual report more difficult to read if good earnings of the current year are transitory or if poor earnings are persistent. On the other hand, firms with better future performance may want to disclose information more transparently in order to lower the information-processing costs and to distinguish themselves from the "lemons." In other words, to the extent that complicated annual reports can hide the transitory nature of good news or the permanent nature of bad news by increasing investors' information-processing costs, the management obfuscation hypothesis predicts that the profits (losses) of firms with more complex annual reports are less (more) persistent.

Most prior studies on the subject of 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 et al., 2005a; Cohen, 2005). A few papers study the relation between disclosure quality and earnings quality: Francis et al. (2005b) find a positive relationship between voluntary disclosure quality and the accruals quality; Riedl and Srinivasan (2005) examine the implications for earnings persistence with regard to whether special items are recognized as line items on the income statements 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 the 10-K filings from Edgar for every remaining firm-year. Those firm-years that do not have electronic 10-K filings on Edgar are then excluded.[3] (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 3,000 words or 100 lines of remaining text are dropped. The calculation of the annual report readability is based on the remaining text. Details of these steps are presented in Appendix A. It is important to delete the tables and financial statements in this step—since the readability indices are designed for text rather than for numbers or tables. (5) Finally, firm-years that have operating earnings (scaled by book value of 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 dates between 1994 and 2004. Since most of the firms have a December fiscal year end, my sample mainly covers the fiscal years 1993–2003.

3.2. The readability measures

I use two statistics to measure the annual report readability. The first is the *Fog* index from the computational linguistics literature. 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.[4] 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 as follows:

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

where complex words are defined as words with three syllables or more. The relation between the *Fog* and reading ease is as follows: $FOG \geqslant 18$ 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 presumed to be higher, assuming everything else to be equal, longer documents seem to be more deterring and more difficult to read. Therefore, the length of an annual report could be used strategically by managers in order to make an annual report less transparent and to hide adverse information from investors. The SEC has consistently suggested that companies avoid lengthy

[3]SEC has electronic Edgar filings available online from 1994.

[4]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 $(11.8 * \text{syllables\_per\_word}) + (0.39 * \text{words\_per\_sentence}) - 15.59$, rates text by a U.S. grade school level. Therefore, a score of 8.0 means that the document could be understood by an average eighth grader. The Flesch Reading Ease rates text on a 100 point scale and is calculated as $206.835 - (1.015 * \text{words per sentence}) - (84.6 * \text{syllables per word})$. The higher the Flesch Reading Ease index, the easier the text will be. The empirical results based on the Kincaid Index and the Flesch Index are similar to those based on the *Fog* index and are therefore unreported. For more information about these readability measures, see http://www.plainlanguage.com/Resources/readability.html.

sentences and documents (SEC, 1998). Practitioners also use lengthy documents as examples of bad and complex disclosure (e.g., Barker, 2002). There are pros and cons to 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 disclosure. I define the length of annual reports 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 some extreme values.

I use the Lingua::EN:Fathom package of the Perl language to analyze the raw 10-K files and calculate *Fog* and *Length*.[5] This program has been used in various fields, including information science and business communication (e.g., Collins-Thompson and Callan, 2005; Muresan et al., 2006).

To check the validity of the Perl program in calculating the *Fog*, I first compare the calculation with those from other studies. Smith and Smith (1971) manually calculate the Flesch Reading Index of some randomly selected footnotes of the 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 the Perl program are 24.44 and 24.63, respectively, which figures are similar to their manually calculated numbers.

A second way of checking the validity of the calculation is to compare it with the results based on manual calculations or other computer programs using the same text.[6] I randomly select three paragraphs from 10 annual reports and manually count the number of words per sentence and syllables per word. The difference between the results from the manual calculations and the Perl programs is smaller than 5% in most cases, which results confirm the validity of the program.

One concern regarding the use of syntactical features such as the *Fog* index in order to measure readability is that the results may not reflect actual comprehension difficulty (Jones and Shoemaker, 1994). However, the fact that I focus on the relative readability of the annual reports in a cross-section mitigates this concern.

I calculate the *Fog* and *Length* for both the entire annual report and sub-sections of the document. In particular, I focus on two sub-sections: the MD&A and the Notes to the financial statements (hereafter referred to as Notes). The MD&A section contains managers discussion of past performance and future outlook; Notes have detailed assumptions regarding the reported financial numbers. Details of electronically extracting the sub-sections are presented in Appendix B. Companies use different formats in their annual reports and the electronic extraction of MD&A and Notes is by no means 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 1,000 words in order to be included in the analysis.

3.3. Summary statistics

Panel A of Table 1 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 entire annual report are 19.4 and 19.2, respectively, which statistics are ''unreadable'' according to the standard 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 editorials 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 significantly 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*

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

[6]For unexplained reasons, the Kincaid Index calculated by Microsoft WORD does not score above grade 12, although the original formula scores up to a graduate school level. As a result, it is not appropriate to check the validity using WORD.

Table 1
(A) Summary statistics; (B) Pearson correlation matrix; (C) Persistence of *Fog* and *Length* for the first and fifth quintile percentages

Variable	Mean	Median	Std. Dev.	1st	25th	75th	99th	*N*
(A)								
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 ($MM)	2,022	169	14,209	1	44	731	33,003	51,393
Book value of assets ($MM)	3,551	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	31,034	23,122	28,057	4918	15,173	36,926	140,047	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	4,665	3,325	5,653	160	1,894	5,782	23,195	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	12,443	6,135	20,284	1,474	3,855	12,247	95,640	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

	Fog (whole annual report)	Fog (MD&A)	Fog (notes)	Length (whole annual report)	Length (MD&A)	Length (notes)	Market-to-book	Size	Assets
(B)									
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**	

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	*N*
(C)						
Fog						
Year 0					100	11,094
Year 1	6.56	10.04	14.13	24.50	44.77	8,564
Year 2	7.42	11.35	16.56	24.82	39.85	6,901
Year 3	8.58	12.44	17.33	24.79	36.86	5,418
Year 0	100					11,091
Year 1	56.83	20.83	9.35	6.14	6.86	8,849
Year 2	50.80	22.16	11.28	7.79	7.97	7,252
Year 3	46.99	23.07	11.45	9.36	9.12	5,790

Table 1 (*continued*)

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	*N*
Length						
Year 0					100	11,095
Year 1	5.00	10.41	16.19	25.59	42.81	8,692
Year 2	5.94	11.42	17.13	25.67	39.83	6,964
Year 3	6.17	12.29	17.53	25.05	38.96	5,460
Year 0	100					11,093
Year 1	62.31	17.27	7.91	6.73	5.78	8,684
Year 2	57.11	18.98	9.69	7.60	6.61	7,107
Year 3	54.05	19.11	10.98	8.72	7.15	5,621

(A) 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 book value of 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. All data item numbers refer to the Compustat item numbers.
(B) This panel shows the Pearson correlation coefficients 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 natural 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. All data item numbers refer to the Compustat item numbers.
The Pearson correlation coefficient in bold is significant at 0.01 level.
(C) 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.

index between Reader's Digest and TIME magazine is about 2.[7] Likewise, the variation in the year-to-year change in the *Fog* and *Length* is not small. The standard deviation of the change in the *Fog* index is 1.46 and that of *Length* is 0.66. The 25th and the 75th percentile of the year-to-year change in the *Fog* are −0.59 and 0.65, respectively.

Panel A also presents the readability of the MD&A and the Notes. The MD&A section of the annual report is much easier to read than the document as a whole, with the mean (median) *Fog* index being 18.23 (17.98). Moreover, the variation in the MD&A readability is much larger than that for the entire annual report—with the standard deviation of the *Fog* being 2.55 and the inter-quartile range at about 2.8. The Notes 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 23,122, 3,325, and 6,135, respectively.

Fig. 1A plots the median level of the *Fog* and *Length* of the annual reports for the sample firms over time.[8] Interestingly, there is an obvious drop in the *Fog* in the years immediately following 1999, suggesting that the plain English disclosure guidelines issued in 1998 might have forced companies 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. The SEC Critical Accounting Policies proposal and the Sarbanes-Oxley Act regulation may have contributed to this change. In contrast, the *Length* of the annual reports experience a steady increase over time.

Figs. 1B and C plot the median level of the *Fog* and *Length* of the MD&A and the Notes sections. The drop in the year 2000 of the readability of the whole annual report observed in Fig. 1A comes primarily from the

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

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

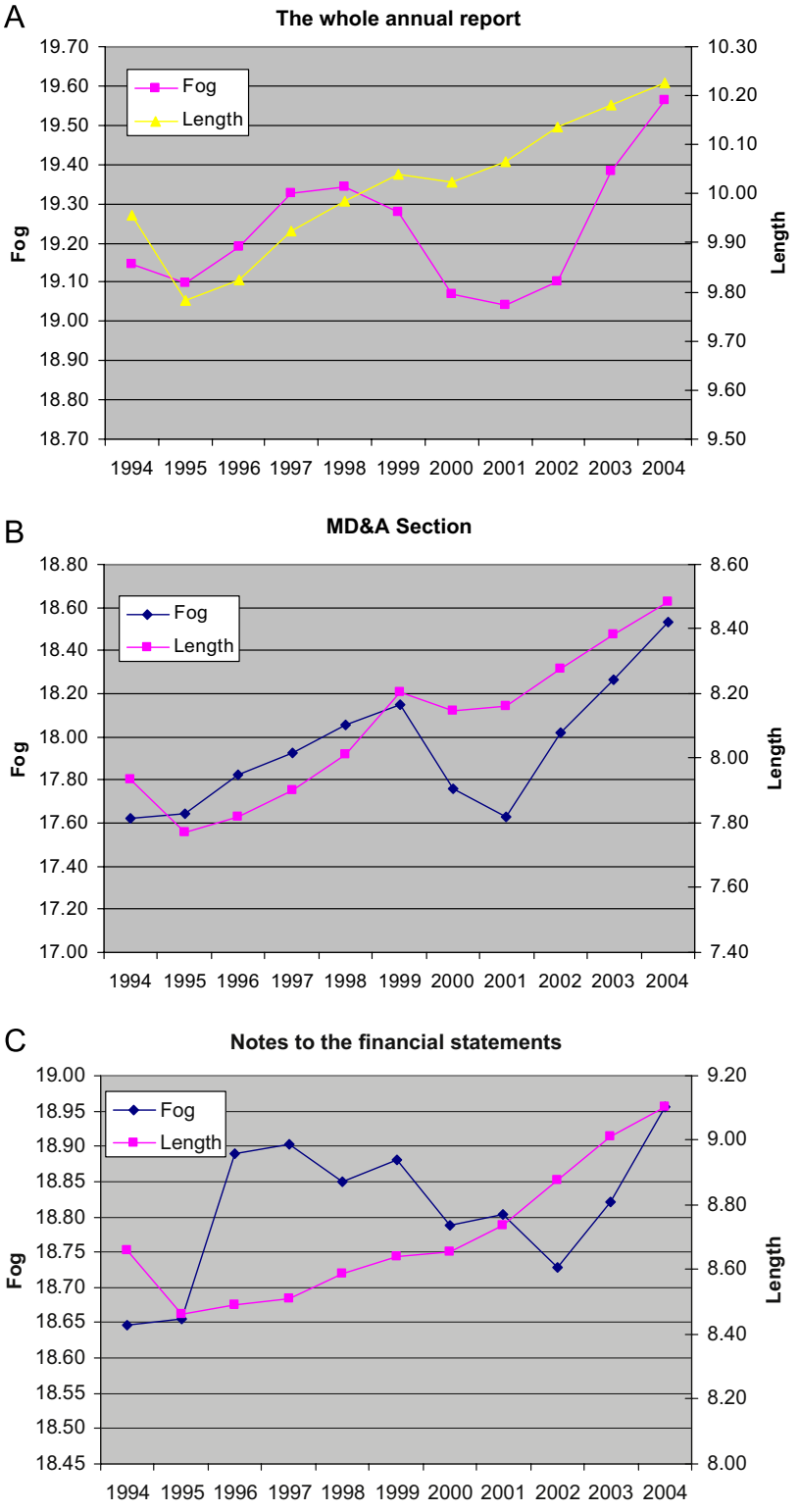

Fig. 1. (A) Median *Fog* and *Length* of the whole annual report by calendar year of the filing date. (B) Median *Fog* and *Length* of the MD&A section by calendar year of the filing date. (C) Median *Fog* and *Length* of the notes to financial statements by the calendar year of the filing date. *Note*: *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. (A) shows the *Fog* and *Length* of the whole annual report. (B) shows the *Fog* and *Length* of the MD&A section. (C) shows the *Fog* and *Length* of the Notes to the financial statements.

MD&A section (not from the Notes). Both the MD&A and the Notes sections experience dramatic increases in the *Fog* in 2003 and 2004.

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

There is a strong correlation between the readability of MD&A section, the Notes, and the annual report as a whole. The Pearson correlation coefficient between the *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 longer annual reports, as evidenced by the correlation coefficient of 0.263 between *Length* and firm size. Growth firms (firms with higher market-to-book ratio) do not seem to file different annual reports, with the Pearson correlation coefficients between market-to-book and the *Fog* being 0.014 and the Pearson coefficient between *Length* and market-to-book being −0.006, both of which results carry little economic magnitude.

Panel C of Table 1 presents the persistence of annual report readability for firms in the first and fifth quintiles of the *Fog* and *Length*. Every year, firms are sorted into five quintiles based on the *Fog* or *Length*. For firms in the first and fifth quintiles, I track their readability level over the following three years. For instance, there are 11,094 (100%) firm-years in the fifth quintile of the *Fog* in Year 0. In the next year, 44.77% of these firms still remain in the fifth quintile, 24.50% switch to Quintile 4, 14.13% are in Quintile 3, 10.04% are in Quintile 2, and 6.56% join Quintile 1. Overall, there seems to be some time-series variation in annual report readability. Of the firms in the fifth quintile of the *Fog* in Year 0, only about 61% stay in Quintiles 4 and 5; the remainder 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.[9] Ex ante, there are many factors that might non-strategically affect annual report readability. It is important to empirically document the determinants and control for them in my later empirical tests when I look for strategic interactions between firm performance and annual report readability. The factors examined here include the following variables:

- Size: Size captures many aspects of a firm's operational and business environment. For instance, the accounting literature has used firm size as a proxy for a firm's political cost (e.g., Watts and Zimmerman, 1986). Hence, I include *SIZE*—defined as the 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 larger 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—is included as a potential determinant of annual report readability. Growth firms may have more complex and uncertain business models—and thus more complex annual reports.
- Firm age: Older firms may exhibit different annual report readability because there is less information asymmetry and less 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's first appearance in the CRSP monthly stock return files (*AGE*).

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

- Special items: Firms with a significant amount of special items are more 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. Assuming everything else to be 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: Communication to investors by firms with more volatile business environments is presumably more complicated. I use firm-specific stock return volatility (*RET_VOL*, measured as the standard deviation of the monthly stock returns in the prior year) and earnings volatility (*EARN_VOL*, measured as the standard deviation of the operating earnings during the prior five fiscal years) in order to capture the volatility of business.
- Complexity of operations: Firms with more complex operations are more likely to have 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 the Compustat segment files at the end of a fiscal year.
- Financial complexity: Firms that have more complex financial situations are also more likely to have 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, then the situation is more financially complex.
- Firm events: Unusual firm events may require extra and more detailed disclosures. I create two dummies, *MA* and *SEO*, in order to capture firm-year specific merger-and-acquisition and seasoned equity offering events. *MA* is set to 1 for a year in which a company appears in the SDC Platinum M&A database as an acquirer and 0 otherwise; *SEO* is set to 1 for a year in which 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, have 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 annual reports have a different level of readability.

In addition, I include year and industry fixed effects as potential determinants of the readability.[10]

Table 2 presents the results of regressing the *Fog* and *Length* on their potential determinants. Since the readability of annual reports is likely to be correlated within industries, the standard errors are clustered at the two-digit SIC industry level. In column [1] of Table 2 Panel B, the *Fog* index of the entire annual report is regressed on the variables with year and industry fixed effects. Larger firms, firms with more volatile business, firms with merger and acquisition activities, and firms incorporated in Delaware have more complex annual reports, as evidenced by the positive and significant coefficients on *SIZE*, *RET_VOL*, *EARN_VOL*, *MA*, and *DLW*.[11] On the other hand, *AGE*, *SI*, *NGSEG* and *SEO* are negatively associated with the 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 counterintuitive result is the negative coefficient on *NGSEG*, suggesting that firms with more geographic segments tend to have less complicated annual reports. The explanatory power of all the variables examined collectively, however, seems relatively low, since the adjusted *R*-squared of the regression is only 8% and half of this explanatory power derives from industry dummies.

[10]As an alternative specification, I drop the year dummies and include the accumulated CRSP value-weighted stock market returns in the last 12 months in the regression in order 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 higher risks of litigation may wish to write their annual reports more rigorously, therefore ending up with annual reports that are harder to read (Bencivenga, 1997). Hence, I construct an industry-specific litigation risk using the Securities Class Action Clearinghouse Database from the Stanford Law School. When the industry-specific variables (the Herfindahl Index, the high-tech industry dummy and the litigation risk) are included in the regression, the industry fixed effects are omitted from the regressions. 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.

[11]*NBSEG*, is positively related to the *Fog* if industry fixed effects are not included and becomes insignificant if industry dummies are controlled.

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. Perhaps this is because growth opportunities are more difficult to describe than are assets-in-place in the management discussion and analysis sections. Another interesting difference is that the associations between *SI*, *RET_VOL*, and *EARN_VOL* and readability are much stronger for MD&A than they are the whole document. For instance, the coefficient on *EARN_VOL* is 0.822 (with a *t*-statistic of 5.68) in column [2], while the coefficient is 0.182 (with a *t*-statistic of 2.20) in column [1]. This result suggests that more negative special items and more volatile business environments prove more difficult to explain in the MD&A section.

Table 2
(A) Summary statistics of the determinants of *Fog* and *Length*; (B) Determinants of *Fog*; (C) Determinants of *Length*

Variable	Mean	Median	Std. Dev.	1st	25th	75th	99th	*N*
(A)								
AGE	10.99	8.00	9.86	0.00	3.00	16.00	38.00	48,893
SI	−18.69	0.00	365.96	−481.00	−2.20	0.00	68.63	54,804
RET_VOL	0.15	0.12	0.11	0.03	0.08	0.19	0.57	44,045
EARN_VOL	0.08	0.04	0.37	0.00	0.02	0.08	0.69	45,238
NBSEG	3.75	3.00	3.47	1.00	1.00	4.00	16.00	33,432
NGSEG	4.15	3.00	4.23	1.00	1.00	5.00	20.00	30,117
NITEMS	211.37	223.00	37.66	95.00	207.00	232.00	256.00	55,720
SEO	0.06	–	–	–	–	–	–	55,720
MA	0.27	–	–	–	–	–	–	55,720
DLW	0.21	–	–	–	–	–	–	57,114

Dependent variable		[1] *Fog* of the whole annual report	[2] *Fog* of the MD&A section	[3] *Fog* of the Notes to the financial statements
Independent variable	Predicted sign			
(B)				
SIZE	+	0.019[1.89]*	−0.015[−0.95]	−0.064[−5.98]***
MTB	+	0.001[0.13]	0.032[2.52]**	0.012[1.86]*
AGE	−	−0.004[2.47]**	0.003[0.96]	−0.005[−2.63]**
SI	−	−0.193[2.01]**	−0.447[−2.60]**	−0.066[−0.68]
RET_VOL	+	0.438[3.07]***	1.326[5.00]***	0.532[4.72]***
EARN_VOL	+	0.182[2.20]**	0.822[5.68]***	0.056[0.65]
NBSEG	+	−0.002[0.09]	0.029[0.82]	0.033[1.30]
NGSEG	+	−0.062[3.75]***	−0.074[−2.08]**	−0.081[−3.71]***
NITEMS	+	−0.471[1.50]	−0.821[−1.25]	−0.684[−2.31]**
SEO	+	−0.066[1.69]*	−0.173[−3.84]***	0.026[0.44]
MA	+	0.074[2.76]***	0.055[0.91]	0.059[2.44]**
DLW	+/−	0.157[4.10]***	0.128[1.82]*	0.085[1.65]
Year dummies		Yes	Yes	Yes
Industry dummies		Yes	Yes	Yes
Observations		36,375	28,279	31,331
Adj. *R*-squared		0.08	0.09	0.06

Dependent variable		[1] *Length* of the whole annual report	[2] *Length* of the MD&A section	[3] *Length* of the Notes to the financial statements
Independent variable	Predicted sign			
(C)				
SIZE	+	0.103[18.85]***	0.079[11.70]***	0.098[14.60]***
MTB	+	−0.026[−6.01]***	−0.032[−6.76]***	−0.034[−5.05]***
AGE	−	−0.008[−9.56]***	−0.005[−4.70]***	−0.002[−1.79]*
SI	−	−0.423[−6.77]***	−0.209[−2.63]**	−0.485[−6.17]***
RET_VOL	+	0.726[9.09]***	0.368[3.75]***	0.918[9.11]***

Table 2 (*continued*)

Dependent variable		[1] *Length* of the whole annual report	[2] *Length* of the MD&A section	[3] *Length* of the Notes to the financial statements
Independent variable	Predicted sign			
EARN_VOL	+	0.184[6.44]***	0.083[1.44]	0.186[5.03]***
NBSEG	+	0.007[0.75]	0.025[1.91]*	0.019[1.02]
NGSEG	+	−0.007[−1.00]	0.002[0.19]	−0.016[−1.23]
NITEMS	+	−0.261[−1.73]*	0.103[0.64]	−0.242[−1.42]
SEO	+	0.03[1.91]*	0.032[1.10]	0.006[0.27]
MA	+	0.074[9.92]***	0.012[0.71]	0.099[6.83]***
DLW	+/−	0.089[5.48]***	0.076[3.15]***	0.097[2.25]**
Year dummies		Yes	Yes	Yes
Industry dummies		Yes	Yes	Yes
Observations		36,375	28,279	31,331
Adj. *R*-squared		0.18	0.09	0.13

This table shows the summary statistics of the potential determinants of *Fog* and *Length* (Panel A) and the regression results of *Fog* (Panel B) and *Length* (Panel C) on the determinants and year fixed effects and 2-digit SIC industry fixed effects. The determinants are *SIZE*, *MTB*, *AGE*, *SI*, *RET_VOL*, *EARN_VOL*, *NBSEG*, *NGSEG*, *NITEMS*, *SEO*, *MA*, and *DLW*. *SIZE* is the natural 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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to *SDC* Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in SDC Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. All the regressions are estimated with an intercept included but the intercept is not reported. All data item numbers refer to the Compustat item numbers.
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 shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

In column [3], the dependent variable is the *Fog* of the Notes. 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 readability of the Notes. When a firm is involved in M&A transactions, the Notes 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 firms' financial complexity very well.

Panel C of Table 2 presents the results of regressing annual report length on potential determinants. The determinants of the length of the whole annual report, the MD&A section and the Notes are quite similar. Larger 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., the *Fog* and *Length*). Table 3 presents the results of regressing the *Fog* and *Length* on earnings (scaled by book value of assets) using both level specification (Panel A) and change specification (Panel B). In all the regressions, the variables used

Table 3
(A) Firm performance and annual report *Fog* and *Length* (level specification); (B) Firm performance and annual report *Fog* and *Length* (change specification)

Dependent variable	Whole annual report				MD&A section				Notes to the financial statements			
	[1] *Fog*	[2] *Fog*	[3] *Length*	[4] *Length*	[5] *Fog*	[6] *Fog*	[7] *Length*	[8] *Length*	[9] *Fog*	[10] *Fog*	[11] *Length*	[12] *Length*
(A)												
Independent variable												
Earnings	−0.458[−4.44]***		−0.508[−12.93]***		−1.659[−8.38]***		−0.284[−4.93]***		−0.185[−2.53]**		−0.551[−5.80]***	
Profit/loss dummy		−0.163[−3.95]***		−0.184[−17.61]***		−0.625[−6.28]***		−0.095[−5.53]***		−0.037[−1.32]		−0.179[−10.87]***
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	41,100	41,100	41,100	41,100	32,099	32,099	32,099	32,099	35,533	35,533	35,533	35,533
Adj. *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

	Whole annual report				MD&A section				Notes to the financial statements			
	[1] Δ*Fog*	[2] Δ*Fog*	[3] Δ*Length*	[4] Δ*Length*	[5] Δ*Fog*	[6] Δ*Fog*	[7] Δ*Length*	[8] Δ*Length*	[9] Δ*Fog*	[10] Δ*Fog*	[11] Δ*Length*	[12] Δ*Length*
(B)												
Independent variable												
Change in earnings	−0.238[−2.79]***		−0.194[−5.37]***		−0.399[−4.87]***		−0.012[−0.23]		−0.317[−3.32]***		−0.238[−5.47]***	
Earnings ± dummy		−0.094[−4.85]***		−0.053[−5.56]***		−0.117[−4.31]***		0.016[1.24]		−0.066[−3.37]***		−0.061[−5.89]***
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	34,481	34,481	34,481	34,481	23,606	23,606	23,606	23,606	27,526	27,526	27,526	27,526
Adj. *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

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, the MD&A section or Notes to financial statements in Panel A and year-to-year change in *Fog* and *Length* in Panel B. *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 the 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 need to contain at least 1,000 words. Earnings is operating earnings (data178 of Compustat) scaled by book value of 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 (coefficients unreported) include *SIZE*, *MTB*, *AGE*, *SI*, *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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to *SDC* Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in *SDC* Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. All data item numbers refer to the Compustat item numbers. All the regressions are estimated with an intercept included but the intercept is not reported.

t-Statistics shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

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 economic 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 in order to control for within-industry correlation of the annual report readability.

The negative coefficients on earnings indicate that firms with higher earnings have annual reports that are easier to read (i.e., they have a lower *Fog* and are shorter). In columns [1] and [3] of Panel A, the coefficients on earnings are −0.458 (with a *t*-statistic of −4.44) and −0.508 (with a *t*-statistic of −12.93) when used to explain the *Fog* and length of the whole annual report. Replacing the earnings level with a profit/loss dummy, which equals 1 if a company reports a profit and 0 otherwise, yields similar results: The coefficients on the dummies are −0.163 (with a *t*-statistic of −3.95) and −0.184 (with a *t*-statistic of −17.61) in columns [2] and [4]. These results indicate that the annual reports of loss firms are more difficult to read than those of profit firms.

Table 4
(A) Earnings persistence and annual report *Fog* index (profit firm-years); (B) earnings persistence and annual report *Length* (profit firm-years); (C) earnings persistence and annual report readability (profit firm-years)

Dependent variable	The whole annual report		MD&A section		Notes to the financial statements	
	[1] $Earn_{(t+1)}$	[2] $Earn_{(t+2)}$	[3] $Earn_{(t+1)}$	[4] $Earn_{(t+2)}$	[5] $Earn_{(t+1)}$	[6] $Earn_{(t+2)}$
(A)						
Independent variable						
$Earnings_{(t)}$	0.026[0.03]	−0.057[−0.06]	0.300[0.29]	0.864[0.71]	0.221[0.24]	0.126[0.13]
Fog	0.003[4.01]***	0.004[3.04]***	0.001[2.16]**	0.003[2.63]**	0.002[2.61]**	0.002[2.85]***
$Earnings_{(t)} * Fog$	−0.028[−3.74]***	−0.041[−2.95]***	−0.016[−3.13]***	−0.036[−3.00]***	−0.022[−2.71]***	−0.023[−3.02]***
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	22,798	19,089	18,533	15,744	20,569	17,546
Adj. *R*-squared	0.41	0.26	0.42	0.26	0.41	0.26
	The whole annual report		MD&A section		Notes to the financial statements	
	[1] $Earn_{(t+1)}$	[2] $Earn_{(t+2)}$	[3] $Earn_{(t+1)}$	[4] $Earn_{(t+2)}$	[5] $Earn_{(t+1)}$	[6]$Earn_{(t+2)}$
(B)						
Independent variable						
$Earnings_{(t)}$	−0.026[−0.02]	−0.267[−0.22]	−0.162[−0.15]	−0.434[−0.35]	−0.183[−0.20]	−0.14[−0.13]
Length	0.005[2.56]**	0.006[2.93]***	0.002[0.91]	0.001[0.56]	0.002[1.90]*	0.004[2.22]**
$Earnings_{(t)} * Length$	−0.060[−2.97]***	−0.075[−3.48]***	−0.019[−1.23]	−0.021[−0.79]	−0.025[−2.87]***	−0.036[−2.31]**
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	22,798	19,089	18,533	15,744	20,569	17,546
Adj. *R*-squared	0.41	0.26	0.42	0.26	0.41	0.26

	[1] $Earn_{(t+1)}$	[2] $Earn_{(t+2)}$		[3] $Earn_{(t+1)}$	[4] $Earn_{(t+2)}$
(C)					
Independent variable					
$Earnings_{(t)}$	0.993[1.00]	1.629[1.39]	$Earnings_{(t)}$	0.527[0.48]	0.706[0.53]
Fog	0.001[1.89]*	0.002[1.02]	*Length*	0.003[1.19]	0.003[0.81]
$Earnings_{(t)} * Fog$	−0.012[1.53]	−0.024[1.53]	$Earnings_{(t)} * Length$	−0.048[−1.96]*	−0.072[−2.47]**
MD&A Fog	0.001[1.09]	0.003[2.39]**	MD&A Length	0.001[0.42]	0.001[0.21]
$Earnings_{(t)} *$ MD&A*Fog*	−0.01[−1.78]*	−0.029[−2.61]**	$Earnings_{(t)} *$ MD&A Length	−0.009[−0.53]	−0.009[−0.30]
Notes *Fog*	0.001[1.41]	0.001[0.39]	Notes Length	0.001[0.58]	0.003[0.77]
$Earnings_{(t)}$*Notes *Fog*	−0.015[−1.66]	−0.004[−0.33]	$Earnings_{(t)}$*Notes Length	−0.006[−0.48]	−0.003[−0.10]

Table 4 (*continued*)

	[1] $Earn_{(t+1)}$	[2] $Earn_{(t+2)}$		[3] $Earn_{(t+1)}$	[4] $Earn_{(t+2)}$
(C)					
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	17,233	14,813	Observations	17,233	14,813
Adj. *R*-squared	0.42	0.27	Adj. *R*-squared	0.42	0.26

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 a positive earnings. The dependent variables are earnings of year $t+1$ ($earn_{(t+1)}$) and year $t+2$ ($earn_{(t+2)}$). *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 book value of assets. MD&A *Fog* and Notes *Fog* are the *Fog* index of the 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 1,000 words.

The control variables (coefficients unreported) 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 1 if a firm has dividend (i.e., data21 > 0) this year and 0 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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to SDC Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in SDC Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. All data item numbers refer to the Compustat item numbers. All the regressions are estimated with an intercept included but the intercept is not reported.

t-Statistics shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

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 reports in a more readable way than in the immediately preceding 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 the *Fog* and *Length* (columns [1] and [3]). Columns [2] and [4] illustrate that, on average, the change in the *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 the *Fog* stems mainly from the MD&A section. In column [5] of Table 3 Panel A, the coefficient on earnings is −1.659 (with a *t*-statistic of −8.38), more than three times the coefficient in column [1] when the *Fog* of the entire annual report is used. On the other hand, while the *Fog* of the Notes is negatively associated with earnings, the coefficient on earnings is much smaller: −0.185 (with a *t*-statistic of −2.53) in column [9] of Table 3 Panel A and −0.037 (with a *t*-statistic of −1.32) in column [10], which results are less than half of the coefficients in columns [1]–[2].

However, the relation between earnings and *Length* derives more from the Notes section than from the MD&A section. Splitting annual reports into MD&A and Notes shows that the Notes (coefficient on earnings is −0.551 with a *t*-statistic of −5.80 in column [11] of Panel A of Table 3) is more negatively correlated with earnings than MD&A (column [7] coefficient −0.284 and a *t*-statistic of −4.93). This finding suggests that length of the Notes is more likely to be used as a strategic deterrence to investors. The change specification further confirms that the negative relation between firm performance and annual report length is stronger in the Notes than it is in the MD&A section.

However, the incremental *R*-squared of earnings as a means of explaining the *Fog* and *Length* is trivial. Comparing column [1] of Table 3 Panel A with column [1] of Table 2 Panel B reveals that adding current

earnings increases the *R*-squared by 0.00. This finding suggests that economic performance is not a first-order determinant of annual report readability. To gauge the economic size of the effects, I perform the following calculation. On average, increasing a firm's earnings from 0.00 (25th percentile of the sample) to 0.11 (75th percentile) will lead to a decrease in the *Fog* index of about 0.05. This is minimal compared to the variation of the *Fog* in the sample (Table 1). Put in different terms, the annual reports of firms at the 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 0.163 (0.184), which finding is also minimal.

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 result is consistent with the hypothesis that managers do not appear to make annual reports much more complex in order to mask poor current performance.

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 "foggier" or longer annual reports are less persistent. Panel A of Table 4 presents the regression results of one- and two-year ahead earnings on the current year's earnings, the Fog, and their interaction using a sample of *all firm-years with positive earnings*.[12] The interaction term captures the change in earnings persistence as annual report readability changes. In all of 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 accruals (*ABSACC*) and a dividend dummy (*DIV*, which equals 1 if a company pays dividend and 0 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 are not reported.

In all cases, the interaction term is negative. For instance, in columns [1] and [2] of Table 4 Panel A, in which the *Fog* of the entire annual report is used to explain year $t+1$ and $t+2$ earnings persistence, the interaction term coefficients are -0.028 (with a *t*-statistic of -3.74 with the standard errors clustered at industry-level) and -0.041 (with a *t*-statistic of -2.95). This means that, as the *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. Assuming everything else to be equal, for an inter-quartile increase in the *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 a *Fog* index of greater than 18 have lower earnings persistence than do those with a *Fog* index of 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 (columns [3]–[6]) 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 greater (Table 1) and the overall economic effect of MD&A readability on earnings persistence is comparable to the readability of the whole annual report.

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

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 the positive earnings sample, the coefficient on the interaction term has a coefficient of −0.060 (with a *t*-statistic of −2.97). The effect of annual report length on earnings persistence is economically large: an increase of length from 9.63 (the 25th percentile from Table 1 Panel A) to 10.52 (the 75th percentile) implies an earnings persistence lowered by 0.05. The length of the Notes is more associated with earnings persistence than is the MD&A length: In column [5] of Table 4 Panel B, the coefficient on the interaction of earnings with the length of the Notes is −0.025 (with a *t*-statistic of −2.87); in column [3], the coefficient on the interaction of earnings with MD&A length is −0.019 (with a *t*-statistic of −1.23). Overall, it seems that the length of the annual report is negatively related to performance and earnings persistence, and this effect is stronger in the Notes than in the MD&A section.

Table 5
(A) Earnings persistence and annual report *Fog* index (loss firm-years); (B) earnings persistence and annual report Length (loss firm-years)

Dependent variable	The whole annual report		MD&A section		Notes to the financial statements	
	[1] $Earn_{(t+1)}$	[2] $Earn_{(t+2)}$	[3] $Earn_{(t+1)}$	[4] $Earn_{(t+2)}$	[5] $Earn_{(t+1)}$	[6] $Earn_{(t+2)}$
(A)						
Independent variable						
$Earnings_{(t)}$	−0.390[−0.20]	1.258[0.73]	1.058[0.44]	1.845[0.79]	−0.408[−0.25]	0.223[0.11]
Fog	−0.004[−1.37]	−0.005[−1.55]	−0.001[−1.04]	−0.003[−1.96]*	−0.001[−0.23]	−0.004[−1.20]
$Earnings_{(t)}$*Fog*	−0.014[−0.80]	−0.011[−0.77]	0.006[0.88]	0.000[0.05]	−0.004[−0.43]	−0.015[−1.09]
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	6,961	5,205	5,420	4,140	5,898	4,565
Adj. *R*-squared	0.41	0.29	0.41	0.29	0.41	0.29
(B)						
$Earnings_{(t)}$	−0.707[−0.40]	2.006[1.10]	1.202[0.55]	1.788[0.74]	−0.770[−0.52]	0.240[0.12]
Length	−0.003[−0.64]	−0.006[−1.06]	0.004[1.25]	0.006[1.55]	0.005[1.89]*	−0.003[−0.64]
$Earnings_{(t)}$*Length*	0.001[0.04]	−0.053[−2.35]**	0.005[0.32]	0.012[0.81]	0.016[0.98]	−0.029[−1.61]
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	6,961	5,205	5,420	4,140	5,898	4,565
Adj. *R*-squared	0.41	0.29	0.41	0.29	0.41	0.29

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 loss firm-years. The sample is all firm-years that report losses. The dependent variables are earnings of year $t+1$ ($earn_{(t+1)}$) and year $t+2$ ($earn_{(t+2)}$). *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 book value of assets.

The control variables (coefficients unreported) 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 1 if a firm has dividend (i.e., data21 > 0) this year and 0 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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to SDC Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in SDC Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. All data item numbers refer to the Compustat item numbers. All the regressions are estimated with an intercept included but the intercept is not reported.

t-Statistics shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

Panel C of Table 4 includes the readability of the whole annual report, the MD&A section and the Notes in one regression in order to examine which part of the annual report has the largest impact on earnings persistence. The results indicate that the *Fog* of the entire 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 be more heavily concentrated in 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 the MD&A *Fog* and earnings is −0.010 (with a *t*-statistic of −1.78) and that of the Notes *Fog* is −0.015 (with a *t*-statistic of −1.66). Hence, although marginally insignificant, the effect of the Notes *Fog* on earnings persistence is comparable to that of the MD&A *Fog* in economic magnitude. Overall, it seems that the readability of both the MD&A section and the Notes section contain information regarding 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 the *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 in order to present less persistent good news. On the other hand, I do not find significant evidence that firms make their annual reports more difficult to read in order to hide more persistent bad news.[13]

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.[14] Managerial strategic disclosure is just one of the possible explanations for my findings. One approach used to mitigate this concern is to go beyond readability and to examine other features of the annual reports.[15]

In particular, I focus on five categories of writing styles of the MD&A section: the relative frequencies of self-referential words, exclusive words, causation words, positive emotion words, and future tense verbs. Research in psychology shows that words, that tell how people are expressing themselves can often be more informative than what the people are expressing (Undeutsch, 1967; Pennebaker and King, 1999; Pennebaker et al., 2003; Shapiro, 1989) and that liars and truth-tellers communicate in qualitatively different ways.[16]

[13]The results are empirically robust. 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 retaining firms with at least 10 years worth of data. I then run the tests again, adding firm dummies in the regressions, and the results still hold. Second, I include earnings-squared as an additional explanatory variable in order to control for possible non-linearity in earnings persistence; the results (unreported) yielded 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, using a sub-sample of firm-years which have no special items, I further examine whether unusual events related to special items are driving the empirical findings. 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 in order 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 following year and more than 80% of them remain profitable each year in the following one to four years. It seems unlikely that litigation is a first-order concern for these firms.

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

[15]Davis et al. (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. Nelson and Pritchard (2007) examine the use of cautionary language in annual reports. Hutton et al. (2003) consider the impact of supplementary statements on the informativeness of management earnings forecasts and Baginski et al. (2004) study managerial attributions in management earnings forecasts. Kothari and Short (2006) analyze the content of more than 100,000 disclosure reports by management, analysts, and news reporters and find that the tone of the disclosures is related to cost of capital.

[16]A caveat of analyzing these measures of writing style with regard to 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. An annual report is typically written by the management team and attorneys and, therefore, the external validity of measures of the writing style is not

More specifically, Newman et al. (2003) find that when people tell the truth, they are more likely to use first-person singular pronouns and more exclusive words such as "except," "but," "without," and "excluding." Therefore, the first two measures that I examine are the percentages of self-referential and exclusive words in the MD&A section of the annual report[17]:

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

where *Self* is the percentage of first-person pronouns (20 words in the LIWC dictionary), and *You* and *Other* are the percentages of second-person pronouns (14 words in the dictionary) and third-person pronouns (22 words in the dictionary); and

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

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 third writing style on which I focus 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 are more apt to spend more effort explaining what is going on if they are attempting to cover something up. *Cause* is the percentage of causation-related words (49 words in the dictionary including "because," "effect," and "hence"). The fourth writing style captures 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 "hate," "worthless," and "enemy").

Finally, the last measure intends to capture the managerial emphasis on the future versus the past/present. The assumption here is that people are likely to talk more about "future" if they are not doing well and are not confident about their performance.

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

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 presents the results of regressing the writing-style measures on current earnings and other control variables.[18] The MD&A sections of the annual reports of firms with lower earnings tend to use more self-referential words, more exclusive words, and more discussions about the future. The implications of firm performance for *Cause* and *PvsN* are statistically insignificant. The results on self-referential words and 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-referential 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 in

(*footnote continued*)
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 behaviors.

[17]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; the program is able to calculate the degree to which people use different categories of words across a wide array of texts. 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. 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.

[18]The psychology and linguistics literature provides very little guidance on the determinants of the writing-style variables, especially for this paper's setting. I report the empirical tests based on the same set of control variables as in the previous tests. The results are robust to including a sub-set or different combinations of the control variables or firm fixed effects.

Table 6
Firm performance and writing styles

Dependent variable	[1] *IvsU*	[2] *EvsI*	[3] *Cause*	[4] *PvsN*	[5] *FvsP*
Independent variable					
Earnings	−0.334[−8.49]***	−0.037[−2.26]**	−0.012[−0.41]	0.015[0.42]	−0.188[−5.60]***
SIZE	0.036[9.63]***	−0.011[−4.53]***	0.003[1.34]	0.023[7.12]***	0.012[6.85]***
MTB	0.004[1.58]	0.002[2.45]**	0.004[2.52]**	0.001[0.45]	0.005[7.37]***
AGE	−0.007[−7.90]***	−0.002[−7.45]***	−0.003[−5.87]***	0.001[2.90]***	−0.003[−9.50]***
SI	0.001[0.03]	0.016[1.16]	0.035[1.08]	0.009[0.52]	0.012[0.58]
RET_VOL	0.579[11.74]***	0.076[3.13]***	0.045[1.00]	−0.083[−2.17]**	0.138[5.39]***
NBSEG	0.001[0.04]	−0.006[−1.43]	−0.018[−1.90]*	−0.002[−0.50]	−0.009[−2.09]**
NGSEG	0.019[0.85]	−0.001[−0.29]	0.03[4.28]***	−0.011[−2.92]***	−0.005[−1.01]
NITEMS	−0.084[−1.31]	−0.016[−0.29]	0.099[1.83]*	−0.181[−2.24]**	−0.021[−0.40]
SEO	0.143[7.65]***	0.011[1.40]	−0.03[−2.55]**	0.006[0.62]	0.015[2.04]**
MA	0.021[2.62]**	−0.016[−3.80]***	−0.003[−0.49]	0.01[2.27]**	−0.006[−1.25]
DLW	0.001[0.07]	0.030[2.69]***	0.006[0.46]	0.000[0.04]	0.033[3.46]***
Year dummies	Yes	Yes	Yes	Yes	Yes
Industry dummies	Yes	Yes	Yes	Yes	Yes
Observations	32,099	32,099	32,099	32,099	32,099
Adj. *R*-squared	0.34	0.09	0.09	0.20	0.11

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 percentages of second and third person pronouns in the MD&A section. *EvsI* is $\log((1+\text{Excl})/(1+\text{Incl}))$, 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 book value of 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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to SDC Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in SDC Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. All data item numbers refer to the Compustat item numbers. All the regressions are estimated with an intercept included but the intercept is not reported.
t-Statistics shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

current earnings. From Table 7, 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, a higher frequency of causation 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 a 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.[19] 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.

[19]The summary statistics of the writing-style results are not tabulated.

Table 7
Earnings persistence and writing styles

Dependent variable	Sample: profitable firms					Sample: loss firms				
Independent variable										
Earnings	−0.360[−0.34]	−0.283[−0.26]	−0.386[−0.35]	−0.410[−0.36]	−0.351[−0.32]	0.927[0.41]	0.957[0.42]	0.950[0.42]	1.288[0.62]	1.070[0.47]
IvsU	−0.003[−0.59]					−0.015[−1.95]*				
Earnings∗*IvsU*	−0.025[−0.53]					−0.022[−0.82]				
EvsI		0.000[0.05]					−0.006[−0.63]			
Earnings∗*EvsI*		0.005[0.10]					−0.002[−0.04]			
Cause			0.003[1.39]					−0.004[−0.92]		
Earnings∗*Cause*			−0.049[−2.27]**					−0.006[−0.27]		
PvsN				−0.006[−0.84]					−0.018[−1.40]	
Earnings∗*PvsN*				0.089[1.71]*					−0.111[−2.25]**	
FvsP					0.002[0.46]					−0.006[−0.60]
Earnings∗*FvsP*					−0.118[−2.77]***					0.041[0.80]
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	18,507	18,507	18,507	18,507	18,507	5,418	5,418	5,418	5,418	5,418
Adj. *R*-squared	0.42	0.42	0.42	0.42	0.42	0.41	0.41	0.41	0.41	0.41

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 [1–5] are firms that report a positive earnings, and those in columns [6–10] 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 percentages of second and third person pronouns in the MD&A section. *EvsI* is $\log((1+\text{Excl})/(1+\text{Incl}))$, 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 (coefficients unreported) 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 1 if a firm has dividend (i.e., data21 > 0) this year and 0 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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to SDC Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in SDC Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. All data item numbers refer to the Compustat item numbers. All the regressions are estimated with an intercept included but the intercept is not reported.

t-Statistics shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

The associations between the writing styles of annual reports and earnings could potentially help sort out the alternative explanations to the findings (Bloomfield (2008)). For instance, as discussed in Bloomfield (2008), the attribution theory in psychology predicts that when people try to explain a bad outcome, they are more likely to attribute the outcome to other parties and hence refer more often to other people than to themselves. The negative relation between earnings and *IvsU* suggests that the empirical results are perhaps not driven by the attribution bias of managers.

6. Further analyses

6.1. Unexercised stock option holdings and incentives to obfuscate information

The research design in the paper cannot establish causality between earnings persistence and the annual report readability. For instance, one alternative explanation is that perhaps bad news is inherently more difficult to present and requires more complicated language. One way to mitigate this concern is to find a setting in which the incentive for managers to obfuscate information is stronger and to check whether the empirical results are stronger there. This section provides some evidence on this matter.

Prior research has documented that managers strategically withhold good news before scheduled employee stock options grants (Aboody and Kasznik, 2000) and managers' disclosure behavior is associated with their trading incentives (Rogers, 2004). Managers may want to delay the release of bad information if they have an abundance of unexercised stock options. I link this intuition with the association between readability and earnings persistence. Assuming everything else to be 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. Panel A of Table 8 shows that the interaction of *UNEX_OPT*, a measure of the amount of unexercised (but exercisable) employee stock options, and earnings and the *Fog* index loads up negatively, suggesting that our empirical results are stronger for firms with more unexercised executive stock options. Untabulated descriptives statistics on *UNEX_OPT* show that for firms at the 25th percentile of *UNEX_OPT*, an increase in *Fog* by 1 reduces one-year ahead earnings persistence by 0.014; on the other hand, for firms at the 75th percentile, an increase in *Fog* by 1 implies a one-year ahead earnings persistence lower by 0.039. Similar results are observed for *Length* (Panel B of Table 8), although the statistical significance is lower.

6.2. Future stock returns and annual report readability

Managers may benefit from writing more complicated annual reports by delaying the incorporation of bad news into stock prices, as prior studies show that the stock market may under-react to the textual information found 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-month stock returns following the 10-K filing date on the *Fog* and *Length*. The Fama–MacBeth regression results in Table 9 indicate that there is no significant association between annual report readability and length and future stock returns.[20] The change in the *Fog* has no predictive power for the following year's stock returns either. However, the change in *Length* ($Length_t - Length_{t-1}$) is negatively associated with the following year's returns (column [6], with a t-statistic of -3.72), suggesting that the stock market does not fully understand the implications of annual report length for future performance and that managers could benefit from obfuscating information through lengthy disclosures. Overall, there is mixed evidence on whether managers successfully delay the incorporation of bad news into stock prices by writing more complicated annual reports.

[20]Unreported results based on two sub-samples (small firms, defined as firms with a market value of less 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.

Table 8
The effect of executive option holdings. (A) *Fog* and earnings persistence; (B) length and earnings persistence

	Dependent variables			
	[1] $Earn_{(t+1)}$	[2] $Earn_{(t+2)}$	[3] $Earn_{(t+3)}$	[4 $Earn_{(t+4)}$]
(A)				
Earnings	1.869[1.87]*	0.978[0.61]	0.602[0.42]	−1.964[−1.06]
Fog	0.001[1.50]	0.002[1.34]	0.000[0.21]	−0.001[−0.75]
Earnings * *Fog*	−0.017[−1.82]*	−0.023[−2.12]**	−0.006[−0.50]	0.005[0.35]
UNEX_OPT	−0.028[−2.45]**	−0.024[−1.89]*	−0.037[−1.88]*	−0.071[−2.31]**
Earnings * *UNEX_OPT*	0.218[2.30]**	0.210[2.06]**	0.379[2.24]**	0.710[2.89]***
*Fog***UNEX_OPT*	0.001[2.34]**	0.001[1.72]*	0.002[1.80]*	0.004[2.30]**
Earnings * *Fog* * *UNEX_OPT*	−0.011[−2.31]**	−0.011[−2.00]**	−0.020[−2.22]**	−0.037[−2.92]***
Observations	7,407	6,235	5,122	4,051
Adj. *R*-squared	0.56	0.39	0.31	0.28
(B)				
Earnings	1.818[1.89]*	0.640[0.40]	0.598[0.42]	−1.706[−1.01]
Length	0.004[1.94]*	0.002[0.72]	0.002[0.55]	0.000[0.13]
Earnings * *Length*	−0.036[−2.14]**	−0.030[−1.64]	−0.022[−0.95]	−0.008[−0.30]
UNEX_OPT	−0.015[−1.30]	−0.010[−0.79]	−0.023[−1.54]	−0.052[−2.02]**
Earnings * *UNEX_OPT*	0.123[1.31]	0.076[0.77]	0.249[1.99]*	0.491[2.41]**
Length * *UNEX_OPT*	0.001[1.17]	0.001[0.62]	0.002[1.42]	0.005[1.99]*
Earnings * *Length* * *UNEX_OPT*	−0.012[−1.30]	−0.008[−0.73]	−0.025[−1.94]*	−0.049[−2.39]**
Observations	7,407	6,235	5,122	4,051
Adj. *R*-squared	0.56	0.39	0.31	0.28

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 book value of assets. *UNEX_OPT* is the logarithm of (number of exercisable but unexercised stock options owned by the CEO/number of shares owned by the CEO), both of which numbers are from the EXECUCOMP database.

The control variables (coefficients 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 1 if a firm has dividend (i.e., data21 > 0) this year and 0 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 1 plus the number of business segments and *NGSEG* is the logarithm of 1 plus the number of geographic segments. *NITEMS* is the number of non-missing items on Compustat. *SEO* is a dummy that equals 1 if a firm has seasoned equity offering in this year according to SDC Global New Issues database and 0 otherwise. *MA* is a dummy that equals 1 if a firm appears as an acquirer in this year in SDC Platinum M&A database and 0 otherwise. *DLW* is a dummy that equals 1 if a company is incorporated in Delaware and 0 otherwise. Year and industry fixed effects are also included. All data item numbers refer to the Compustat item numbers. All the regressions are estimated with an intercept included but the intercept is not reported.

t-Statistics shown in brackets are based on standard errors clustered at two-digit SIC industry level. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

7. Conclusions

This paper provides the first large-sample evidence on the determinants and implications of the lexical properties of corporate disclosures. In particular, I study the implications of annual report readability and other lexical features of annual report for current performance and earnings persistence. The empirical findings can be summarized as follows. First, annual reports of firms with poor performance are more difficult 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. The effect is economically significant: an inter-quartile change in annual readability has about the same impact on profit persistence as do accruals. Viewed

Table 9
Fama–MacBeth regressions of future returns on *Fog* and *Length*

Dependent variable	[1] $Ret_{(t+1)}$	[2] $Ret_{(t+1)}$	[3] $Ret_{(t+1)}$	Dependent variable	[4] $Ret_{(t+1)}$	[5] $Ret_{(t+1)}$	[6] $Ret_{(t+1)}$
Independent variable							
$Fog_{(t)}$	−0.001[−0.16]	−0.002[−0.54]			−0.011[−0.84]	−0.006[−0.33]	
$Earnings_{(t)}$		−0.222[−0.48]				0.496[0.50]	
$Earnings_{(t)} * Fog_{(t)}$		0.014[0.60]				−0.047[−0.44]	
$Fog_{(t)} - Fog_{(t-1)}$			−0.020[−0.28]				−0.022[−3.72]***
Constant	0.199[2.73]***	0.219[3.29]***	0.233[2.73]***		0.295[3.13]***	0.249[1.61]	0.233[2.73]***
Number of years	10	10	9	Number of years	10	10	9
Average observations	3024	3022	3373	Average observations	3024	3022	3373
Average adj. *R*-squared	0.00	0.00	0.00	Average adj. *R*-squared	0.00	0.00	0.00

The dependent variables are annual returns of year $t+1$ (the 12-month returns 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 book value of assets.
t-Statistics shown in brackets are based on the coefficients from the annual cross-sectional regressions. ***/**/* means significance at 0.01, 0.05, and 0.10 level, respectively.

collectively, the evidence in this paper suggests that managers may be opportunistically structuring the annual reports to hide adverse information from investors.

Appendix A. 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 perform 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 are used by some firms to present tables. Next, all the tags in the format of ⟨...⟩ and ⟨&...⟩, which are used widely in documents in SEC HTML or XML format documents, are replaced with blanks. Finally, to make sure that all the tables, tabulated text, 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 the 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 produces the summary readability indices used in the paper.

Appendix B. Steps to extract MD&A and Notes to the financial statements

This appendix explains the details of extracting the MD&A section and Notes from 10-K filings. Starting with the raw 10-K file, I first delete the SEC-header information, all the contents between ⟨TABLE⟩ and ⟨/TABLE⟩ text, the paragraphs that contain ⟨S⟩ or ⟨C⟩, all the tags in the format of ⟨...⟩ and ⟨&...⟩ are removed using the same process described in Appendix A.

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 some white spaces followed by "managements discussion" or "managements discussion"; or (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 satisfies one of the following criteria as the *ending* of the MD&A section: (1) the line begins with some white spaces followed by "Financial Statements" or "Financial Statements"; (2) the line contains "item" followed by one or more white spaces and the number "8"; (3) the line contains "Supplementary Data"; or (4) the line begins with some white spaces followed by "SUMMARY OF SELECTED FINANCIAL DATA" or "SUMMARY OF SELECTED FINANCIAL DATA." Most firms have a table of contents listing the main sections of the 10-K filing. In some instances, this table of contents 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 contents about MD&A will also be picked up by the program as part of the MD&A.

Similarly, the program identifies a line as the *beginning* of the Notes, if: (1) the line starts with "NOTES TO" or some white spaces followed by "NOTES TO"; and (2) the line does not contain any number except when it follows "for the years ended." The program identifies a line that satisfies one of the following criteria as the *ending* of the Notes: (1) the line contains "Changes in and Disagreements with Accountants" or "DISAGREEMENTS ON ACCOUNTING"; (2) the line contains "DIRECTORS AND EXECUTIVE OFFICERS"; or (3) the line contains "exhibit index."

After the MD&A and the Notes 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, D., Kasznik, R., 2000. CEO stock option awards and the timing of corporate voluntary disclosures. Journal of Accounting and Economics 29, 73–100.

Baginski, S.P., Hassell, J.M., Kimbrough, M.D., 2004. Why do managers explain their earnings forecasts? Journal of Accounting Research 42, 1–29.

Baker, H.E., Kare, D.D., 1992. Relationship between annual report readability and corporate financial performance. Management Research News 15, 1–4.

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

Barnett, A., Leoffler, K., 1979. Readability of accounting and auditing messages. Journal of Business Communication 16, 49–59.

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

Berger, P., Chen, J., Li, F., 2006. Firm-specific information and cost of equity capital, Working Paper, University of Chicago.

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

Bloomfield, R., 2008. Discussion of annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, this issue.

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

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

Collins-Thompson, K., Callan, J., 2005. Predicting reading difficulty with statistical language models. Journal of the American Society for Information Science and Technology 56, 1448–1462.

Courtis, J.K., 1986. An investigation into annual report readability and corporate risk return relationships. Accounting and Business Research, 285–294.

Daines, R., 2001. Does Delaware law improve firm value? Journal of Financial Economics 62, 525–558.

Davis, A.K., Piger, J.M., Sedor, L.M., 2005. Beyond the numbers: an analysis of optimistic and pessimistic language in earnings press releases. Working Paper, Washington University.

Dechow, P.M., Ge, W., 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, P.M., Schrand, C.M., 2004. Earnings Quality, first ed. Research Foundation of CFA Institute, Charlottesville, Virginia.

Firtel, K.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, J., LaFond, R., Olsson, P., Schipper, K., 2005a. The market pricing of accruals quality. Journal of Accounting and Economics 2, 295–327.

Francis, J., Nanda, D., Olsson, P., 2005b. Voluntary disclosure, information quality, and costs of capital. Working Paper, Duke University.
Glassman, C.A., 2005. Remarks at the Plain Language Association International's Fifth International Conference. ⟨http://www.sec.gov/news/speech/spch110405cag.htm⟩.
Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. American Economic Review 70, 393–408.
Healy, P., 1977. Can you understand the footnotes to financial statements? Accountants Journal, 219–222.
Healy, P.M., Palepu, K.G., 2001. Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. Journal of Accounting and Economics 31, 405–440.
Hutton, A., Miller, G., Skinner, D., 2003. The role of supplementary statements with management earnings forecasts. Journal of Accounting Research 41, 867–890.
Jones, M.J., Shoemaker, P.A., 1994. Accounting narratives: a review of empirical studies of content and readability. Journal of Accounting Literature 13, 142.
Kothari, S.P., Short, J.E., 2006. The effect of disclosures by management, analysts, and financial press on the equity cost of capital: a study using content analysis. Working Paper MIT.
Lang, M., Lundholm, R., 1993. Cross-sectional determinants of analyst ratings of corporate disclosures. Journal of Accounting Research Autumn, 246–271.
Lebar, M.A., 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, 176–189.
Li, F., 2006. Do stock market investors understand the risk sentiment of corporate annual reports? University of Michigan Working Paper.
Miller, G.S., 2002. Earnings performance and discretionary disclosure. Journal of Accounting Research 40, 173–204.
Muresan, G., Cole, M., Smith, C.L., Liu, L., Belkin, N.J., 2006. Does familiarity breed content? taking account of familiarity with a topic in personalizing information retrieval. In: Proceedings of the Hawaii International Conference on System Sciences.
Nelson, K.K., Pritchard, A.C., 2007. Litigation risk and voluntary disclosure: the use of meaningful cautionary language. Working Paper, University of Michigan.
Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M., 2003. Lying words: predicting deception from linguistic styles. Personality and Social Psychology Bulletin 29, 665–675.
Pennebaker, J.W., King, L.A., 1999. Linguistic styles: language use as an individual difference. Journal of Personality and Social Psychology 77, 1296–1312.
Pennebaker, J.W., Mehl, M.R., Nierderhoffer, K.G., 2003. Psychological aspects of natural language use: our words, our selves. Annual Review of Psychology 54, 547–577.
Riedl, E.J., Srinivasan, S., 2005. The strategic reporting of special items. Working Paper, Harvard University and University of Chicago.
Rogers, J.L., 2004. Disclosure quality and management trading incentives. Working paper, University of Chicago.
Schrand, C.M., Walther, B.R., 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, D.J., 2004. What do dividends tell us about earnings quality. Working Paper, University of Chicago.
Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Accounting Review 71, 289–315.
Smith, J.E., Smith, N.P., 1971. Readability: a measure of the performance of the communication function of financial reporting. The Accounting Review 46, 552–561.
Soper, F.J., Dolphin, R., 1964. Readability and corporate annual reports. The Accounting Review 39, 358–362.
Subramanian, R., Insley, R.G., Blackwell, R.D., 1993. Performance and readability: a comparison of annual reports of profitable and unprofitable corporations. Journal of Business Communication 30, 49–61.
Undeutsch, U., 1967. Forensic Physchologie [Forensic Psychology]. Verlag fur Psychologie, Gottingen, Germany.
Watts, R.L., Zimmerman, J.L., 1986. Positive Accounting Theory. Prentice-Hall, Englewood Cliffs, NJ.