J Regul Econ (2014) 45:94–113
DOI 10.1007/s11149-013-9236-5

ORIGINAL ARTICLE

Regulation and financial disclosure: The impact of plain English

Tim Loughran · Bill McDonald

Published online: 12 November 2013
© Springer Science+Business Media New York 2013

Abstract In October 1998, the SEC implemented a rule requiring firms to use plain English in their prospectus filings. In addition to the rule, the SEC encouraged the use of plain English in all filings and communication with shareholders. Did the SEC rule significantly impact managers' disclosure style? And, more interestingly, did the SEC's recommendations lead managers to change their disclosure style in filings not under the plain English mandate? Our textual analysis of Form 424, IPO prospectus, and 10-K filings over 1994–2009 finds that the SEC's implementation of the plain English rule substantively impacted managerial behavior. When we focus on 10-K filings, we find that after the 1998 rule, firms are more likely to improve the stylistic components of their filing before an equity issuance and firms with better corporate governance policies are more likely to comply with the rule.

Keywords Plain English · Market regulation · Disclosure · Textual analysis

JEL Classification G12 · G14 · G18 · M41

1 Introduction

Using a large sample of SEC filings over 1994–2009, we study the direct and indirect effects of regulators' attempts to guide positive market outcomes by explicit policy and, separately, by their encouragement to apply the policy in a broader setting. Specifically, we examine the SEC's plain English rule of October 1998. The impetus of the rule is

T. Loughran (✉) · B. McDonald
Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556-5646, USA
e-mail: Loughran.9@nd.edu

B. McDonald
e-mail: mcdonald.1@nd.edu

that investors will be better able to assess and more likely to invest in companies whose financial disclosures are not buried in legal jargon and obtuse language. Importantly, the rule is restricted only to prospectus filings, however, SEC documents and speeches by SEC leaders clearly encourage firms to adopt these principles in all their filings and communications with shareholders.[1]

The plain English rule mandates that firms' prospectuses "must use plain English principles in the organization, language, and design of the front and back cover pages, the summary, and the risk factors section." The rule becomes somewhat less precise, however, when it requires that the writing in these sections of the prospectus "substantially complies with" a list of plain English principles. How successful was the SEC in getting firms to improve the readability of the prospectus and, by encouraging managers to conform in all public disclosures to the rule, the 10-K filings?[2]

Filings covered by the rule include any prospectus or security registration (Forms 424, S-1, F-1, and their variants). We will consider the full sample of 424 filings, which are dominated by debt offerings and, separately, S-1 filings for initial public offerings (IPOs) as our sample of filings mandated by the plain English rule. We also examine all 10-K filings which serve as an important communication channel for managers and provide a form not specifically mandated by the rule. Additionally, the frequency and consistency of 10-K filings allow us to focus on the actions and characteristics of firms in the context of plain English.

To evaluate disclosure style, we create a standardized statistic that aggregates a series of writing components specifically identified by the SEC. The six components are average sentence length, average word length, passive voice, legalese, personal pronouns, and negative/superfluous phrases. This standardized statistic may be useful to other researchers examining document readability or plain English enforcement.

We find that our measure of plain English does "improve" notably in all of the samples after the regulation is enacted. While the prospectus samples (Form 424s and IPO S-1s) increase their compliance dramatically and immediately, the 10-K sample increases over the 2 years following the rule, stabilizes, and then trends upward for the 3 years following the passage of Sarbanes–Oxley. Throughout this paper we use the term "improve" to denote a rise in the plain English measure. We do not determine if in fact the filings are more effective or more accessible as a result of these stylistic suggestions. Whether the spillover of the rule's directives to 10-Ks is attributable to simple encouragement by the SEC or more directly by coercion through delays in disclosure approval, the SEC's plain English mandate appears to have clearly affected both the mandated and non-mandated disclosure documents.

After providing descriptive results for the three samples we then focus on the 10-K filings. Using the 10-K sample, we link our plain English measure to firms having seasoned equity offerings (SEOs). In the year preceding an SEO, we find that firms are more likely to improve the plain English components of their 10-K filing after

[1] See p. 4 of *A Plain English Handbook* (1998) and p. 68 of SEC Release #34-38164 http://www.sec.gov/rules/proposed/34-38164.txt.

[2] In 2010, the plain English guidelines were extended to Form ADV, a form filed by investment advisers to explain their qualifications, operations, and strategies. This extension, however, has no impact on the focus of our research.

the plain English rule takes effect in October, 1998. This behavioral shift could be attributable to either a firm simply updating the template language of their 10-K in light of the SEO, or to the belief that the SEC's recommendations for making financial disclosures more accessible could in fact make their equity offering more palatable in the marketplace.

When we link the 10-K data to the Gompers et al. (2003) Corporate Governance Index, we find that firms with shareholder-friendly governance structures are more likely to file 10-Ks with high plain English scores under the regulatory regime. Thus, as might be expected, firms with solid corporate governance policies are more likely to comply with the SEC requests.

We show that regulators can impact market outcomes both directly and indirectly. The plain English rule requires only changes in prospectus filings, not in 10-Ks. The improvements we observe in our 10-K sample are a result of encouragement by the SEC to adopt the plain English guidelines originally intended for prospectus filings.

Section 2 of the paper offers a background on the 1998 Plain English Rule. Section 3 describes our Form 424, IPO S-1, and 10-K samples, data sources, and parsing procedure. Descriptive results for the three samples are presented in Sect. 4. In Sect. 5, we focus on the 10-K filings vis-à-vis SEO offerings and corporate governance. Finally, Sect. 6 presents our conclusions.

2 The plain English rule and prior literature

The plain English rule became effective October 1, 1998. SEC Staff Legal Bulletin No. 7 provides a summary of the rule and corresponding amendments:

> "Companies filing registration statements under the Securities Act of 1933 must:
> - write the forepart of these registration statements in plain English;
> - write the remaining portions of these registration statements in a clear, understandable manner; and
> - design these registration statements to be visually inviting and easy to read."

Rule 421(d) specifically requires that issuers must:

> "Substantially comply with these plain English principles:
> - short sentences;
> - definite, concrete everyday language;
> - active voice;
> - tabular presentation of complex information;
> - no legal jargon; and
> - no multiple negatives."

Rule 421(b) was later amended, prescribing stylistic approaches to avoid, such as "legal and highly technical business terminology," or "legalistic or overly complex presentations that make the substance of the disclosure difficult to understand."

Although the plain English rule is mandated only for prospectuses, in documentation surrounding the rule's release the SEC clearly encourages conformance with the rule in all corporate filings. Arthur Levitt, then-Chairman of the SEC, in his foreword to *A Plain English Handbook,* concludes with: “I urge you—in long and short documents, in prospectuses and shareholder reports—to speak to investors in words they can understand” (p. 4). The SEC in its proposed rules document states: “Our ultimate goal is to have all disclosure documents written in plain English” (release #34-38164, p. 24), and later in the document: “We also encourage you to use these techniques for drafting your other disclosure documents.” Clearly the SEC is trying to encourage behavior well beyond the specific mandate of the plain English rule.

To gauge the impact of the SEC’s rule and subsequent encouragement to broaden its impact we use textual analysis to examine firms’ compliance with specific plain English mandates in thousands of Form 424, S-1, and 10-K filings. Prior literature has used textual analysis to study newspaper articles, company press releases, and message board postings. For example, Tetlock (2007) links the content of the popular “Abreast of the Market” column with the following day’s stock returns. He finds that higher pessimism in the newspaper column predicts lower following day stock returns. Higher negative word counts in news stories are also linked to lower firm earnings (Tetlock et al. 2008). Some of the prior literature has focused on creating finance-oriented word lists to measure tone (Loughran and McDonald 2011) while others have created Naïve Bayesian machine learning algorithms to gauge tone and content (Li 2010).

Li (2008) examines the overall readability of the 10-K reports, a notion similar to the intent of the plain English rule. Li, using a comparably sized sample to ours, finds that annual reports with lower earnings are more difficult to read. He uses the Fog Index to gauge changes in readability and reports that the mean and median index increases (i.e., 10-K readability declines) over the 1994–2004 sample period. Unlike our paper, his main focus is on linking 10-K readability with current firm earnings and earnings persistence.

Lehavy et al. (2011) link 10-K readability with analyst earnings forecasts. Using the Fog Index, the three authors report that the less readable is the 10-K, the greater is analyst dispersion and the lower is analyst accuracy. You and Zhang (2009) find that 10-K complexity is related to investor underreaction. They define 10-Ks with more words than the median yearly filing as being complex. Investors are found to exhibit delayed reactions to the information content of complex 10-K documents.

Miller (2010) examines how small and large investor trading behavior is affected by 10-K length and readability. He finds that longer and less readable 10-Ks reduce small investor trading volume. Like Li (2008) and Lehavy et al. (2011), Miller (2010) uses the Fog Index as one of his measures of readability.

Loughran and McDonald (2013b) argue that measuring readability in business documents is an elusive function of how readability is defined and show that the traditional Fog Index is substantially misspecified when applied to financial disclosures. The Fog index is defined as 0.4*(average number of words per sentence + percent of complex words), where complex words are those with more than two syllables. They document that the count of complex words is dominated by relatively common business terms and argue that words weighing heavily in the tabulation, words like *financial, company, management*, and *interest*, are unlikely to confound any investor or analyst.

Our focus is not on the general notion of readability, but on management’s adoption of the specific stylistic recommendations of the plain English rule in documents, especially where such changes are not required. In contrast to these papers, beyond the question of readability, we are interested in the impact of the SEC's regulation. Thus our measure of readability is dictated by the specific recommendations promulgated by the SEC.

3 Data

3.1 Samples and parsing procedures

Although electronic filing was not required by the SEC until May 1996, a significant number of documents are available on the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) beginning in 1994. We examine three separate samples of SEC filings. Our initial Form 424 sample contains 272,334 filings over the 1994–2009 sample period. For many debt offerings included in the 424 sample, the form consists only of a table itemizing the debt characteristics along with a few paragraphs describing the offering. Almost 30 % of the sample forms have <1,000 words. Since our focus is on measuring aspects of writing style, we restrict the sample to the 188,474 Form 424 filings containing 1,000 words or more.

Because of the importance of an initial registration for an IPO, we also examine the sample from Loughran and McDonald (2013a) which includes 1,842 US IPOs from 1996 to 2009. Relative to the 424 filings, all of these prospectuses are verbose, providing a useful gauge of writing style. The briefest filing in our IPO sample contains more than 25,000 words.

The sample formation process for the 10-Ks is more complex because we link the data to accounting and market data. Details for the process are reported in Table 1. The initial 10-K sample (including both 10-K and 10-K405 forms) covering 1994–2009 contains 129,988 firm-year observations. We apply a total of nine different screens. The two screens substantially reducing the sample are requiring the 10-K to have a CRSP permanent ID match and to be an ordinary common equity firm (CRSP share type code of 10 or 11). There were, for example, over 10,000 observations for asset-backed securities in the original 10-K sample, primarily attributable to filings for security offerings such as exchange traded funds. These funds were removed from the sample by applying the CRSP ID match and the ordinary common equity filter.

Following Loughran and McDonald (2011), we also require the 10-K to have a minimum length of at least 2,000 words (reducing the sample by 5,659).[3] To limit the sample to firms likely to issue equity, we eliminate companies with a stock price of <\$3.[4] This screen removes 9,284 firm-year observations that would find it difficult

[3] Some firms, like Dawson Geophysical's filing on December 12, 1996, incorporate much of the required material by reference. For example, Dawson Geophysical's entire Management Discussion and Analysis section is the following: "The information required by this Item 7 is hereby incorporated by reference to the Registrant's 1996 Annual Report (pp. 24–27) referred to in Item 1." Such filings are removed from our sample by the 2,000 word requirement.

[4] The probability of issuing equity is inversely related to the firm's stock price. Of the firms that remain in our sample, 7.2 % of those with a stock price ≥ \$20 issue equity in the following year, compared to 6.3 % for those with stock prices between \$10 and \$20, 4.3 % for those with stock prices between \$5 and \$10, and 2.6 % for those with stock prices between \$3 and \$5. One could infer from this pattern that firms with stock prices <\$3 are even less likely to issue equity.

Table 1 10-K sample creation

Sample source/screen	Sample
EDGAR 1994–2009 non-duplicate 10-K filings	129, 988
Minus	
Firms' filings with another10-K filing within prior 180 days	235
Firms with 10-K number of words <2,000	5, 659
Firms without CRSP PERMNO match	50, 093
Firms whose stock is not ordinary common	5, 165
Firms with missing market capitalization data	775
Firms with stock price <$3	9, 284
Firms with missing prior period returns	1, 805
Firms not trade on NYSE/AMEX/NASDAQ	22
Firms with negative or missing book-to-market	1, 713
Final 10-K firm-year sample	55, 237
Number of unique firms	9, 013

to sell shares to investors. We further require the firm to have a positive value for the book-to-market ratio (removing more than 1,700 observations). After applying these filters, the final 10-K sample totals 55,237 firm-year observations representing 9,013 unique firms.[5] Throughout the paper, "year" is the calendar year of a form's filing. So, Google's December 31, 2004, 10-K which was filed on March 30, 2005, would be classified as a 2005 observation.

Figure 1 presents the distribution of the 10-K sample size and firm market capitalization by the filing month. Approximately 57 % of the 10-Ks are filed in the month of March. Most firms have December 31st fiscal year-ends and will wait to file until the latest possible date. The substantially higher median market capitalization in February is an artifact of a recent SEC rule requiring large public float firms to file within 60 days of their fiscal year end, with smaller firms allowed 70 days. (See SEC Release #33-8644 http://www.sec.gov/rules/final/33-8644.txt.) On average, 67, 81, and 91 % of the 10-Ks are filed by the end of the first, second, and third quarters, respectively.

To parse all of the documents in each sample we download the files from the SEC's EDGAR web site, clean extraneous coding from the document (HTML, embedded jpg's, etc.), and parse the document into words and sentences. Palmer (2000) provides a useful discussion from the natural language processing literature on the challenges of this process and emphasizes a simple but important theme that is common throughout the natural language processing literature—"an algorithm that performs very well on a specific corpus may not be successful on another corpus." The formatting and structure of SEC filings are far more complex than those of a traditional novel, which is why we

[5] We also initially considered separately testing the Management Discussion and Analysis (MD&A) segment of the 10-Ks. Parsing out the MD&A section is challenging because of inconsistencies in how it is identified. Loughran and McDonald (2011) discuss these issues and show that using only the MD&A section does not improve the discriminating power in explaining filing date returns.

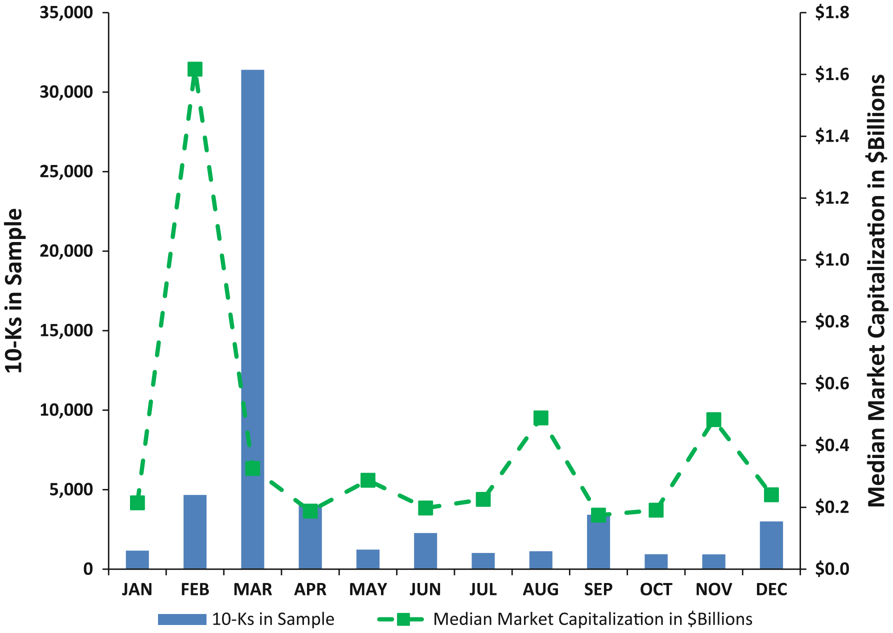

Fig. 1 Number of 10-Ks in sample and median market capitalization in billions of dollars by month

design custom software to parse the documents. A detailed discussion of the parsing process is provided in the "Appendix".

3.2 A measure of plain English

Just as mandating writing style is difficult, so is measuring the degree of compliance. Without deep parsing, which is itself subject to substantial error, at what point does a document meet the threshold of expression in the active voice? What is "clear and understandable"? Unlike other papers in the literature, we create a new readability measure labeled *Plain English* that is anchored in specific examples provided by the SEC documentation surrounding the plain English initiative.

To measure plain English we tabulate the following six components for each document:

1. Average sentence length: The average number of words per sentence in the document. Rule 421(d) emphasizes this characteristic and sentence length is mentioned in specific examples in the Plain English handbook (e.g., pp. 28–29).
2. Average word length: The SEC's documentation emphasizes the use of "short, common words." We count the character length of each word in each filing and average this across all words in the document.
3. Passive: pp. 19–21 of the handbook emphasize the importance of avoiding passive voice. Passive voice can take many forms. We first identify auxiliary verb variants of "to be" including: "to be", "to have", "will be", "has been", "have been", "had

been", "will have been", "being", "am", "are", "is", "was", and "were". Auxiliary verbs followed by a word ending in "ed" or one of 158 irregular verbs are tabulated as passive.

4. Legalese: A count of the words and phrases paralleling those identified in Staff Legal Bulletin No. 7 (http://www.sec.gov/interps/legal/cfslb7a.htm) as inappropriate legal jargon (e.g., "by such forward looking" or "hereinafter so surrendered"). We use a list of 731 legal words from Loughran and McDonald (2011).
5. Personal pronouns: A count of personal pronouns, whose usage the handbook (p. 22) indicates will "dramatically" improve the clarity of writing. The handbook targets first-person plural and second-person singular personal pronouns. Counts are tabulated for "we", "us", "our", "ours", "you", "your", "yours".
6. Other: We combine categories identified in the Plain English handbook whose frequency of occurrence is relatively low. This includes negative phrases, superfluous words and the use of the word "respectively" (see pp. 17–35 of the handbook). Specifically: (1) Negative phrases: A count of 12 negative compound phrases identified on p. 27 of the handbook (e.g., "does not have" or "not certain"); (2) Superfluous—a count of the eight phrases identified as superfluous on page 25 of the handbook (e.g., "because of the fact that" or "in order to"). (3) Respectively—a count of each occurrence of the word "respectively".

We then need to combine the six groups described above into an aggregate measure of plain English. All word/phrase counts are expressed as a proportion relative to the total number of words occurring in the document. Because some of the variables are measured on different scales or their expected proportions might substantially differ, we normalize each of the six components (mean zero, standard deviation of one) and sum. All of the components except personal pronouns are negatively signed in the summation. This process provides the variable we label *Plain English*, where higher values represent documents that better conform to the writing standards promulgated by the SEC.[6]

You and Zhang (2009) use a simple word count to measure 10-K complexity. Loughran and McDonald (2013b) argue that document length provides a useful proxy for readability. As emphasized on p. 11 of *A Plain English Handbook*, however, the goal of the regulation "is clarity, not brevity" and "writing a disclosure in plain English can sometimes increase the length of particular sections …" Therefore we use document length as a control variable (measured as the natural log of the number of words) in our regressions but do not include it in our *Plain English* measure.

3.3 SEO and governance data

We use the Thomson Financial Securities Data (also known as Securities Data Co.) to link 10-K filings with firms issuing seasoned equity during our sample period. Of the 55,237 firm/year observations in the original 10-K sample there are 3,276 SEOs by 2,178 unique firms.

[6] A text file containing our *Plain English* measure for each CIK and filing date combination is available at http://www3.nd.edu/~mcdonald/Word_Lists.html.

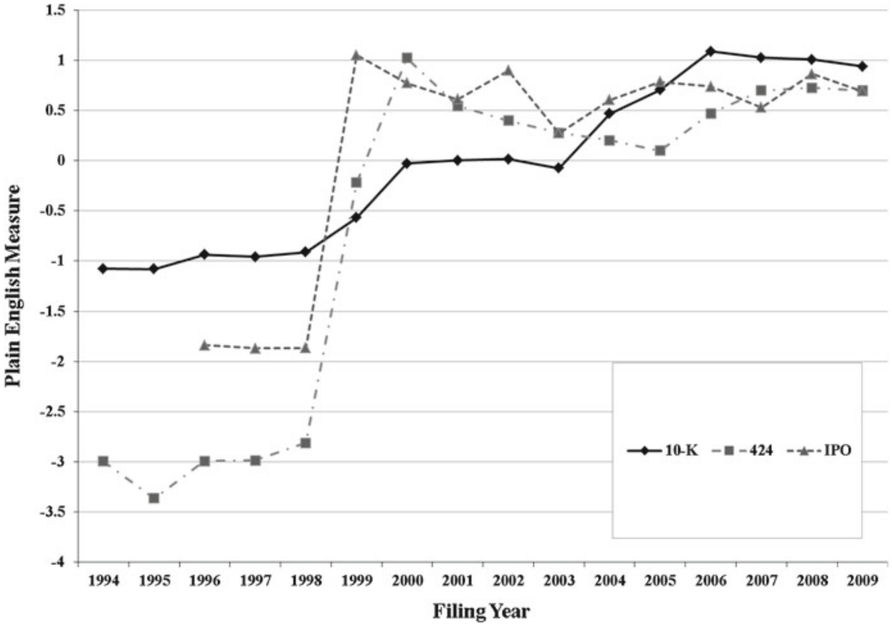

Fig. 2 Mean values of the plain English measure for the sample of 55,237 10-K filings, 188,474 424 filings, and 1,842 IPO prospectuses by year. The plain English rule took effect in October 1998

We obtain the Gompers et al. (2003) Corporate Governance Index from http://finance.wharton.upenn.edu/~metrick/data. The authors use 24 different governance rules to assign scores ranging from 1 to 24—democratic to dictatorship, respectively, using their terminology—to gauge shareholder rights. The higher the governance index, the more dictatorial are the firm's polices (and the weaker shareholder rights). The lower the index, the more democratic the company's policies are. Data are available only for the years 1995, 1998, 2000, 2002, 2004, and 2006. Thus, our 10-K sample size for this measure is 8,927.

4 Descriptive results for the Form 424, S-1, and 10-K sample

Before focusing on the 10-K filings, we begin by reporting in Fig. 2 the means for the plain English measure for each of the three samples (Form 424s, IPO S-1s, and 10-Ks).[7] As the vast majority of the forms filed in 1998 were written before the effective date in October of that year, the rule's impact should become apparent in the 1999 averages.

The results for the mandated disclosures (labeled 424 and IPO) are dramatic. Both increase by over three standard deviations over the 2 years subsequent to the rule's effective date, with the IPOs' entire increase taking place in 1999, while the 424 sample continued to increase in 2000. The level of compliance across the mandated

[7] If instead we plot the median values, the figure remains essentially the same.

forms appears to have stabilized after the initial increase with a slight downward drift in the following years. The sample size for each year's mean for the 424 filings trends strictly upward from 1,614 forms in 1994 to 18,877 in 2009. Note that the distribution of sample size for the IPOs is very different, with the sample peaking and falling with the overall IPO market. Consistent with this point, the sample averages around 130 observations per year, with a minimum of 21 in the 2008, the initial period of a very deep recession, and a maximum of 402 in 1999, the peak of the internet boom.

For the 10-K sample in Fig. 2 the time series is less dramatic and, although the trend is clearly positive, the impact is less concentrated. There are two clear increases in our *Plain English* measure surrounding regulatory events. The first occurs after the plain English rule. Our average measure rises from −1.0 in 1998 to 0.0 by 2000.

The other increase occurs after passage of Sarbanes–Oxley Act (SOX) in 2002. Our average plain English measure is 0.0 in 2002 and 1.0 by 2005. Although SOX did not specifically mandate a 10-K writing style, the law did increase enforcement/penalties for white collar crimes and clearly motivated managers to care more about their document creation and presentation. This result indicates that even in the 10-K sample, whose style mandate was only an artifact of a rule related to prospectuses, the plain English rule had a substantial impact on the textual presentation.

To insure that the upward trend in plain English is not simply due to an influx of newly listed firms (which tend to be smaller firms) or due to firms issuing seasoned equity, we report the average plain English variable following two different screens. First, we examined the subsequent trend in the variable for all firms that had 1998 data. In untabulated results, when IPOs are not allowed to enter our sample, we observe results with the same general upward trend as presented in Fig. 2. Firms with available 1998 data have average plain English values of −0.99 in 1998, −0.41 in 2002, and 0.56 in 2009. Second, when we remove all firms that subsequently have an SEO at any point during our sample period, the same pattern exists. For non-SEO firms, the average plain English is −0.93 in 1998, −0.25 in 2002, and 0.81 in 2009. New listings or firms that issue seasoned equity are not driving the upward trend in readability.

5 The 10-K sample, SEOs, and corporate governance

For the remainder of the paper we will focus on the 10-K filings, which allow us to examine a critical channel of communication with shareholders. Summary statistics for all of the 10-K sample variables are reported in Table 2. The sample is divided into two periods: before the October 1, 1998, plain English rule [column (1)] and after [column (2)]. The last column of the table lists the summary statistics for the entire period. The *Plain English* measure reports substantially higher values during the second period.

For the individual six components of *Plain English*, all report the trend advocated by the SEC with the exception of word length. For example, the average words per sentence declines from 25.273 to 24.490 while the percent of legalese in the average 10-K document drops from 2.038 to 1.638 %. The largest change between the periods is in the increased use of personal pronouns (0.171 vs. 1.291 %). All of the differences

Table 2 Variable means for the 10-K sample

Variable	(1) 1994–September 1998	(2) October1998–2009	(3) 1994–2009
Plain English measure	−0.974	0.363	0.000
Average words per sentence	25.273	24.490	24.702
Average word length	5.444	5.490	5.478
Plain English—passive (%)	1.079	1.058	1.064
Plain English—legalese (%)	2.038	1.638	1.747
Plain English—personal pronouns (%)	0.171	1.291	0.987
Plain English—other (%)	0.189	0.185	0.186
SEO dummy	0.055	0.061	0.059
Governance index	9.074	9.072	9.072
Number of words	34,546.47	44,269.81	41,631.83
Size (in billions)	$1.89	$3.36	$2.96
Age (years)	15.910	16.101	16.049
Pre-event market model alpha (%)	4.455	6.156	5.694
Pre-event market modelroot-mean-square-error	0.029	0.031	0.031
Book-to-market	0.579	0.628	0.615
Nasdaq dummy	0.540	0.585	0.573
Sample size	14,986	40,251	55,237

The number of observations for the Governance Index variable for columns 1–3 is 2,476, 6,451, and 8,927 respectively

in the six *Plain English* components are statistically significant at the 1 % level between the two time periods.

About 6 % of our sample had an SEO in the year after the 10-K filing date. The occurrence of seasoned equity issuance increased from 5.5 to 6.1 % between the two periods. For the governance index, the average level of about 9.07 remains essentially the same across the two time periods.

Table 2 also reports the control variables for subsequent regressions, where the control variables are selected because of their importance in earlier papers on textual analysis and equity issuance. The firm specific control variables are:

(1) *Words*—the 10-K's word count;
(2) *Size*—the market capitalization on the day before the file date (day $t - 1$);
(3) *Book-to-market*—the book-to-market ratio taken from data reported within the prior year and as defined in Fama and French (2001);
(4) *Age(in years)*—number of years a firm is listed on CRSP at the time of the filing;
(5) *Pre-alpha*—the alpha from a market model regression of daily data from the year prior to the filing date using the CRSP value-weighted market index as the market proxy and excluding the 5 days prior to the file date;[8]

[8] Price and number of shares must be available within the prior 22 days and there must be at least 60 observations for the market model regressions to be included in the sample.

(6) *Pre-rmse*—the root mean-square-error from the previously defined market model regression;
(7) *Nasdaq dummy*—a dummy variable set equal to one for firms trading on the Nasdaq stock exchange.

Note that in the subsequent regressions we use the natural log of words, size, and book-to-market, however, we report them in Table 2 in their native format for interpretive purposes.

We find that the average number of words in 10-K filings rises from 34,546 to 44,270 between the pre- and post-1998 periods, consistent with the evidence of Li (2008). The average size of the sample firm is $2.96 billion. Due to our positive book-to-market ratio requirement and the $3 price screen, the sample is tilted towards larger market capitalization firms. Table 2 also reports in the later period a higher percentage of the sample universe listing their shares on the Nasdaq exchange, versus the Amex or the NYSE.

Our results generally contrast the evidence in both Li (2008) and Miller (2010) concerning manager's response to the SEC's plain English rule. In his Figure 1A, Li (2008) reports that the Fog Index slightly increases from 1999 to 2004 (the end of his sample). That is, he finds that readability generally declines in the same time period in which we report improvement in the plain English measure. In another difference to our paper, he reports a steep decrease in median annual report readability (i.e., increases in Fog Index) from 2001 to 2004. In contrast, Miller (2010) finds no major changes in the Fog Index in the years surrounding implementation of the plain English rule.

The difference between our evidence and that of Li (2008) and Miller (2010) is due to the designated readability measure. The other papers use the Fog Index to evaluate changes in manager's 10-K writing style. Instead, we examine whether managers specifically follow the plain English rules. Recall that the Fog Index has only two components: (1) average words per sentence and (2) percentage of words with more than two syllables. As noted in Table 2, average words per sentence declined during our sample period while average word length (a reasonable proxy for percentage of complex words) was up slightly. Thus, there was not much of a change in the Fog Index during 1994–2009, as reported in Miller (2010). Yet, managers clearly changed their writing style to more closely follow the plain English guidelines (i.e., less passive voice and legalese, and more personal pronouns). By examining compliance of the plain English rules, we can document that managers did indeed change their 10-Ks to reflect the new guidelines.

5.1 Industry results

Does *Plain English* vary across industries in the 10-K filings? Fig. 3 documents the wide variability of *Plain English* across the Fama and French (1997) 48 industries. Firms are classified into the 48 categories based on SIC codes (self-reported by the firms) taken from the 10-K filings. The worst industries in terms of *Plain English* are Tobacco Products, Insurance, Defense, Precious Metals, and Aircraft. The tobacco industry's low score on the *Plain English* measure is due to the large amounts of the 10-K devoted to describing the litigation risk of their industry. As an example, Reynolds

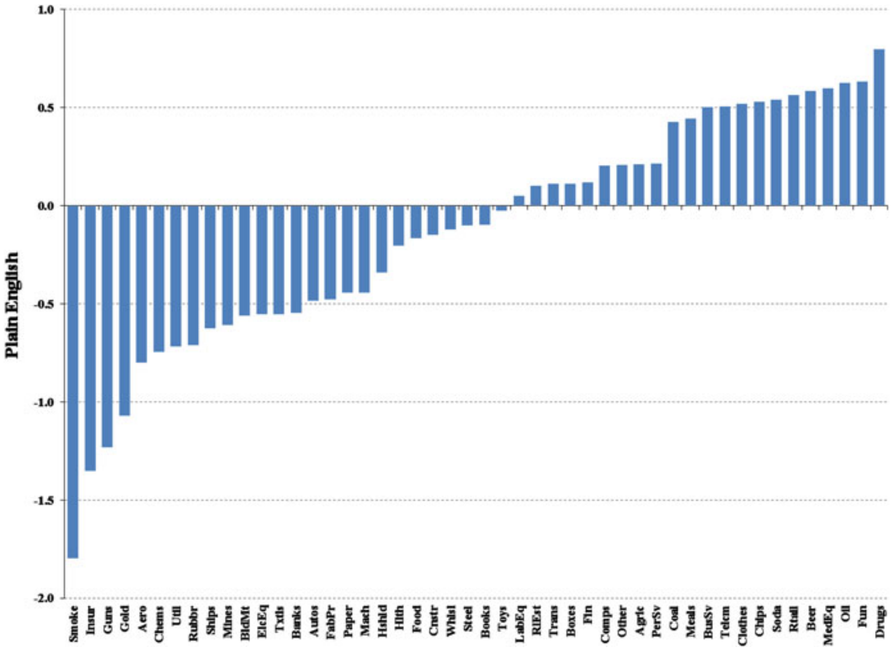

Fig. 3 Plain English measure by Fama–French industry

American (formerly R. J. Reynolds Tobacco) in 2006, accounted for one of the most extreme percentages of legal words in a 10-K (over 5.5 % of all words were legal).

The five industries with the highest values of *Plain English* are Pharmaceutical Products, Entertainment, Petroleum, Medical Equipment, and Alcoholic Beverages. High usage of personal pronouns accounts for the strong average *Plain English* value for the Pharmaceutical Product industry. The petroleum industry scores well partly due to its low average word length (driven by its common usage of the word *oil*). There appears to be a slight pattern of industries that are more consumer oriented (versus traditional manufacturing) having better *Plain English* values. To control for the year-to-year changes in *Plain English* documented in Fig. 2 and the large differences in *Plain English* across industries, our subsequent regressions will include year and Fama–French industry fixed effects.

5.2 Plain English and seasoned equity offerings

Recall that about 6 % of our sample firms had an SEO in the year after the SEO filing date. The regressions reported in Table 3 examine whether companies' proximity to an SEO is associated with an increase in their use of plain English, especially after 1998. The dependent variable, Δ *Plain English*, is the *Plain English* value in year t minus the firm's *Plain English* value in year $t-1$. Firms without a prior *Plain English* value are removed from the regressions.

Table 3 Regressions of the change in plain English (Δ plain English) on SEO classifications and control variables for the 10-K sample

	Full sample (N = 44,298)		
	(1)	(2)	(3)
SEO dummy		0.257	0.024
		(4.69)	(0.55)
SEO dummy * PE dummy			0.578
			(20.42)
SEO dummy * SOX dummy			0.111
			(2.29)
Control variables			
Log(# of words)	−0.417	−0.422	−0.421
	(−5.47)	(−5.51)	(−5.50)
Log(size)	0.056	0.055	0.055
	(5.83)	(5.75)	(5.65)
Log(book-to-market)	0.010	0.013	0.014
	(0.50)	(0.66)	(0.70)
Age (in years)	−0.001	−0.001	−0.001
	(−2.99)	(−2.76)	(−2.93)
Pre-alpha	0.243	0.215	0.202
	(1.73)	(1.57)	(1.51)
Pre-rmse	5.928	5.794	5.790
	(6.60)	(6.33)	(6.37)
Nasdaq dummy	−0.067	−0.066	−0.067
	(−3.56)	(−3.52)	(−3.56)
R-square (%)	3.81	3.95	4.06

The dependent variable, Δ Plain English, is the change in Plain English from the firm's prior year filing. Plain English is a standardized measure that aggregates six writing components specifically suggested by the SEC (sentence length, word length, passive voice, legalese, personal pronouns, and negative/superfluous phrases). SEO dummy has a value of one if the firm issued equity in the year after the 10-K filing, and zero otherwise. PE dummy is set to one for time period after October 1998 and before July of 2002, zero otherwise. SOX dummy is set to one for 10-K filings after July of 2002 (the signing date of the Sarbanes–Oxley Act), else zero. Included in all of the regressions but not tabulated are an intercept, Fama and French (1997) industry dummies, and year dummies. The t statistics are in parentheses with the standard errors clustered by year and industry

The key independent variable, *SEO dummy*, takes the value of one if the firm issued seasoned equity in the year following the 10-K filing; otherwise it takes a value of zero. Because we are interested in the impact of the plain English rule, we also include an interaction term where the dummy variable (*PE dummy*) is set equal to one in the period following the October 1998 enactment of the rule and before the signing of Sarbanes–Oxley and is multiplied by the *SEO dummy*. We also have an interaction term where the *SEO dummy* is multiplied by the *SOX dummy* (set to one for the period following the enactment of the Sarbanes–Oxley legislation in July of 2002, else zero).

The independent variables in the Table 3 regressions are *SEO dummy, SEO dummy* PE dummy*, and *SEO dummy* SOX dummy*, our control variables, and dummies for Fama–French industry and calendar year. In all of the regressions, the t statistics are in parentheses and based on the standard errors clustered by year and industry.

Column (1) reports the coefficients and t statistics for the model containing only the control variables as independent variables. The coefficient on the *Log(words)* variable is negative and statistically significant. The more words in a 10-K, the less likely the firm's plain English value will change. Column (1) also reports that younger firms and larger companies have bigger changes in the plain English measure. Companies with more pre-period volatility are also more likely to better comply with the plain English rules.

The second column of Table 3 includes *SEO dummy* as an additional independent variable. The coefficient on *SEO dummy* is positive (0.257) and statistically significant at conventional levels (t statistic of 4.69). Thus, proximity to an SEO causes firms to more fully comply with the plain English rule.

A more careful partitioning of the rule's impact is provided in column (3), where we can see the changes in writing style associated with SEOs prior to and following the plain English initiative. From column (3), the *SEO dummy* variable is not significant in predicting the change in plain English in the pre-regulation period. Notice that the coefficients on *SEO dummy*PE dummy* and *SEO dummy * SOX dummy* are both positive and statistically significant. The coefficient on *SEO dummy*PE dummy* is larger and has a smaller standard error, suggesting that there was a larger impact immediately following the plain English rule than after the signing date for SOX.

A priori, whether management would attempt to clarify or obfuscate disclosures prior to an equity offering is not clear. One could imagine a scenario where management is less than transparent in an attempt to impart a positive spin on the disclosures. Our results, however, indicate that managers appear to improve the clarity of their writing, as measured by the plain English components, prior to issuing SEOs, and that this change was influenced by the SEC recommendations. The evidence for the *Plain English* measure is consistent with the Healy and Palepu (1993, 1995) hypothesis that managers who expect to issue equity can use voluntary disclosure to influence investors' perceptions of the firm. Following the SEC recommendations in October of 1998, the overall writing quality of the 10-K (as measured by *Plain English*), increases prior to issuing equity, even after controlling for other factors.

5.3 Plain English and corporate governance

The Gompers et al. (2003) Corporate Governance Index is a widely-used proxy for shareholder rights. If our *Plain English* variable does capture management's tendency to respond to SEC encouragement, one might expect that firms with shareholder friendly management are more likely to comply with the SEC recommendations. In Table 4, we report regression results using the level of *Plain English* as the dependent variable. The independent variables are the Gompers et al. (2003) Corporate Governance Index, our control variables, and dummies for the calendar year and Fama–

Table 4 Regressions of plain English readability measure on the Gompers et al. (2003) Corporate Governance Index and other variables for the 10-K sample

	(1)	(2)	(3)
Governance Index		−0.032	0.010
		(−2.89)	(0.64)
PE dummy			0.558
			(3.90)
Governance Index * PE dummy			−0.062
			(−3.02)
Control variables			
Log(# of words)	−0.358	−0.349	−0.347
	(−2.94)	(−2.89)	(−2.85)
Log(size)	−0.026	−0.026	−0.027
	(−0.95)	(−0.95)	(−0.97)
Log(book-to-market)	−0.113	−0.111	−0.110
	(−1.76)	(−1.75)	(−1.74)
Age (in years)	−0.012	−0.011	−0.011
	(−3.36)	(−3.10)	(−3.10)
Pre-alpha	−0.064	−0.063	−0.074
	(−0.43)	(−0.44)	(−0.52)
Pre-rmse	23.630	22.890	22.925
	(4.88)	(4.67)	(4.63)
Nasdaq dummy	0.330	0.316	0.313
	(3.54)	(3.48)	(3.49)
R-square (%)	23.00	23.15	23.29

The dependent variable, Plain English, is a standardized measure that aggregates six writing components specifically suggested by the SEC (sentence length, word length, passive voice, legalese, personal pronouns, and negative/superfluous phrases). PE dummy is zero for time periods prior to October 1998, one otherwise. Note that higher values of the Governance Index implies less shareholder friendly management. Included in each regression but not tabulated are an intercept, year dummies, and industry dummies. The sample size is 8,927 for each regression. The t statistics are in parentheses with the standard errors clustered by year and industry

French industry. In all of the regressions, the t statistics are in parentheses and based on the standard errors clustered by year and industry.

In the Table 4 regressions, the sample is reduced to 8,927 observations due to data availability of the Governance Index. The first regression in Table 4 includes only the control variables. From this regression, we find that firms listed on Nasdaq, younger firms, higher prior volatility, and 10-Ks with fewer words have better *Plain English* values after controlling for size, industry, and calendar year.

When the Governance Index variable is added in the second regression, there is little change in significance levels or coefficient values of the other control variables. The coefficient on the Governance Index variable is negative and statistically significant in the second regression, which implies that the higher the index (the more dictatorial the firm), the lower the use of plain English. In column (3) when we partition the effect into pre- and post-plain English, we once again see that this effect is only significant following the SEC's adoption of the rule. Thus, our results suggest that shareholder-friendly management is also more responsive to the SEC's encouragement in preparing their 10-K disclosures.

6 Conclusion

The style directives of the SEC are relatively amorphous and have been encouraged beyond the scope of the plain English rule. Thus it is interesting to document the measurable impact of these rules in both mandated and non-mandated filings. Our textual analysis of Form 424, S-1, and 10-K filings over 1994–2009 provides clear evidence of managers responding to the SEC's rule and, in the case of 10-Ks, their encouragement and/or intimidation to broadly comply with the plain English initiative. Whether or not the plain English rule makes the disclosures more transparent and useful for investors valuing publicly traded securities are open questions for future research to address.

To gauge manager's response to the rule, we create a standardized measure that aggregates six writing components specifically suggested by the SEC (sentence length, word length, passive voice, legalese, personal pronouns, and negative/superfluous phrases). This plain English inspired metric may be of interest to other researchers examining the rule's effect. The first finding is that management, in creating prospectuses and 10-Ks, clearly responded to the plain English writing guidelines immediately after enactment of the October 1998 Act. For the annual 10-K filing we observe a second increase in our *Plain English* measure occurring after passage of the Sarbanes–Oxley Act of 2002. In the 10-K sample, five of the six writing components show statistically significant improvement after October of 1998, e.g., average sentence length decreased, use of personal pronouns became more common, and the percentage of 10-K text devoted to legal terminology dropped.

Second, higher plain English values relate to issuing seasoned equity to outside investors under the regulatory regime. This improvement in plain English usage is concentrated after October of 1998. Third, companies with more democratic corporate governance policies have much higher plain English values in their 10-Ks than companies with weaker governance policies during the period when the plain English rule was in effect.

In sum, our results indicate that the plain English rule has produced a dramatic impact based on the SEC's style guidelines for writing. These changes are observed in both mandated and non-mandated filings. Managers consider writing style of sufficient importance to improve their prose in anticipation of seeking additional equity funding. And, as might be expected, shareholder-friendly managers produce 10-Ks that are more user-friendly. Collectively these results suggest that both the SEC and management consider writing style important.

Acknowledgments We thank Robert Battalio, Michael Crew (Editor), Andrew Ellul, Margaret Forster, Paul Gao, Kathleen Hanley, Steven Kachelmeier, Feng Li, Ray Pfeiffer, Jennifer Marietta-Westberg, Simona Mola, Richard Sloan, two anonymous referees, and seminar participants at the 2009 American Finance Association annual meeting, University of Amsterdam, and University of Notre Dame for helpful comments. We are grateful to Betsy Laydon for research assistance

Appendix

Downloading the documents

We use the master.idx file from the SEC web site to identify filings from 1994 to 2009. We then programmatically download each Form 424, IPO S-1, 10-K or 10-K405 filing for subsequent parsing. Note that until 2003, a box on the front page of the 10-K form was to be check-marked if a "disclosure of delinquent filers pursuant to Item 405" was not included in the current filing, nor anticipated to be disclosed in statements incorporated by reference or amendments. If this box were checked, the form was filed as a 10-K405. In 2001, almost one-third of 10-K filings were 10-K405 forms. According to the SEC, because there was confusion and inconsistency in making this choice, the 405 provision was eliminated after 2002. As this choice has no impact on the focus of our study, we included both 10-K and 10-K405 forms in our sample and make no distinction between the two throughout the analysis.

Parsing the documents

Parsing is done using a series of programs written by the authors. We use the following sequence to parse each filing:

1. Download text version of each filing and store as string variable.
2. Remove graphics—increasingly through time, the filings have ASCII encoded graphics embedded in the file. ASCII encoding of a graphic increases the size of a file by orders of magnitude. For example, the median file size for the year 2000 was approximately 270 KB and the largest filing without graphics was 5.7 MB. Texas Utilities' year 2000 filing included graphics and was 20.4 MB.
3. Extract SIC code from SEC header.
4. Remove SEC header.
5. Re-encode—convert HTML "&XXX" codes back to text, e.g., &nbsp = space.
6. Remove tables—remove all characters between <TABLE> and </TABLE> whose character count is more than 10 % numbers.
7. Remove HTML—the quantity of HTML contained in the documents increased substantially beginning in 2000. Many documents contain much more HTML than text.
8. Remove abbreviations—counting words per sentence is important for the readability measures. This is typically done by removing abbreviations and then counting the number of sentence terminators and the number of words. For traditional text, this is quite effective after eliminating a few common abbreviations. Parsing finan-

cial disclosures, however, is much more difficult because they contain a variety of abbreviations and use periods to delineate section identifiers or as spacers. Liberman and Church (1992) find that 47 % of the periods occurring in the *Wall Street Journal* are associated with abbreviations. We created a program that is more exhaustive in identifying abbreviations than the routine used in the PERL Fathom package. Because the PERL Fathom package does not deeply parse for abbreviations, it will tend to report more sentences than actually contained in a filing, thus making the average number of words per sentence downward biased.

9. Convert lists to sentences—as in the Fathom package, our sentence count is based on the number of sentence terminators. One challenge in parsing financial disclosures into sentences is that the documents often contain lists separated by semicolons or commas that should not be treated as a single sentence. Redish (2000) notes the problem of measuring readability in texts with extensive lists. Our program attempts to identify such lists based on punctuation and line spacing. Where the program determines that a sequence of text is a list, commas or semicolons delineating the list items are replaced with periods. In addition, to avoid counting the periods in section headers (e.g., Sect. 1.2.), ellipses, or other cases where a period is likely not terminating a sentence, there must be at least 20 characters between two periods for the token to be treated as a sentence.
10. Creating word and phrase counts—the cleaned document is next divided into tokens based on word boundaries using a regular expression. Each token is compared with a master dictionary file to determine if the token is a word. Only tokens of two or more letters are counted as words, thus the words “I” and “a” are not counted. The words for each document are then loaded into a dictionary for that specific filing containing the words and their counts. Word counts are derived from this dictionary. Phrases for the *Plain English* variable are identified by applying regular expressions to the cleaned document.

References

A Plain English Handbook: How to create clear SEC disclosure documents. (1998). Office of Investor Education and Assistance, U.S. Securities and Exchange Commission, http://www.sec.gov/pdf/handbook.pdf

Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, *43*(2), 153–193.

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

Gompers, P., Ishii, J., & Metrick, A. (2003). Corporate governance and equity prices. *The Quarterly Journal of Economics*, *118*(1), 107–156.

Healy, P., & Palepu, K. (1993). The effect of firms’ financial disclosure strategies on stock price. *Accounting Horizons*, *7*(1), 1–11.

Healy, P. M., & Palepu, K. G. (1995). The challenges of investor communication: The case of CUC International, Inc. *Journal of Financial Economics*, *38*(2), 111–140.

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

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

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve Bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049–1102.

Liberman, M., & Church, K. (1992). Speech and language processing. In S. Furui & M. Sondhi (Eds.), *Advances in speech signal processing* (pp. 791–832). New York: Marcel Dekker.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2013a). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*(2), 307–326.

Loughran, T., & McDonald, B. (2013b). Measuring readability in financial disclosures. *Forthcoming in the Journal of Finance*.

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

Palmer, D. D. (2000). Tokenisation and sentence segmentation. In D. Robert, H. Moisl, & H. L. Somers (Eds.), *Handbook of natural language processing* (pp. 11–36). London: Taylor Francis.

Redish, J. (2000). Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation*, *24*(3), 132–137.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437–1467.

You, H., & Zhang, X. J. (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies*, *14*(4), 559–586.