Journal of Accounting and Economics 64 (2017) 221–245

Contents lists available at ScienceDirect

Journal of Accounting and Economics

journal homepage: www.elsevier.com/locate/jacceco

The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation☆

Travis Dyer[a], Mark Lang[a,*], Lorien Stice-Lawrence[b]

[a] Kenan-Flagler Business School, The University of North Carolina, Chapel Hill, NC 27516-3490 United States
[b] Leventhal School of Accounting, University of Southern California, Los Angeles, CA United States

ARTICLE INFO

Article history:
Received 2 March 2016
Revised 28 February 2017
Accepted 31 July 2017
Available online 8 September 2017

Keywords:
Disclosure regulation
Textual analysis
Latent Dirichlet Allocation
Financial reporting

ABSTRACT

We document marked trends in 10-K disclosure over the period 1996–2013, with increases in length, boilerplate, stickiness, and redundancy and decreases in specificity, readability, and the relative amount of hard information. We use Latent Dirichlet Allocation (LDA) to examine specific topics and find that new FASB and SEC requirements explain most of the increase in length and that 3 of the 150 topics—fair value, internal controls, and risk factor disclosures—account for virtually all of the increase. These three disclosures also play a major role in explaining the trends in the remaining textual characteristics.

1. Introduction

Investors, preparers, regulators, and standard setters have expressed concern that corporate disclosure has become longer, more redundant, less readable, less specific, and more boilerplate over time (Li, 2008; KPMG, 2011; SEC, 2013). In 2013, the SEC began a comprehensive review of regulation with the intent of identifying excessive, unduly complex, and redundant disclosure (SEC, 2013). Similarly, the FASB has an ongoing agenda project, the Disclosure Framework, evaluating the effectiveness of textual disclosure (FASB, 2012). A variety of explanations have been offered for why disclosure might be changing over time including increases in litigation concerns, business complexity, globalization, regulation, and new mandatory disclosures (KPMG, 2011; SEC, 2013; Monga and Chasan, 2015). In this paper we quantify a variety of 10-K disclosure attributes and provide initial descriptive evidence on trends in these attributes over time.[1]

While there is a substantial academic literature on trends in the characteristics of quantitative accounting data (particularly earnings and book value) over time,[2] the magnitude, economic determinants, specific content, and attributes of trends in textual disclosure have received less attention. At least in part, this likely reflects the challenge in assessing the content of 10-K textual disclosure and in categorizing and quantifying disclosure for a large number of lengthy, complex documents,

☆ We thank Beth Blankespoor for sharing code to measure the numeric content of disclosure. We thank Joanna Wu (the editor), Greg Miller (the discussant and referee), and workshop participants at University of Bristol, Cornell University, University of Exeter, Georgetown University, Indiana University, University of North Carolina, Stanford University, University of Utah, and the 2016 *Journal of Accounting and Economics* Conference for helpful comments and suggestions. An earlier version of this paper was entitled "The Ever-Expanding 10-K: Why Are 10-Ks Getting So Much Longer (and Does It Matter)?" The Internet Appendix can be accessed at http://tinyurl.com/hrep57s.

* Corresponding author.
E-mail address: Mark_Lang@unc.edu (M. Lang).

[1] Because of the descriptive nature of our analysis, we take as given the disclosure attributes highlighted by regulators and prior research and do not take a position on whether they are, in fact, desirable.

[2] See, for example, Collins et al. (1997), Francis and Schipper (1999), Lev and Zarowin (1999), Givoly and Hayn (2000), Dichev and Tang (2008), Demerjian (2011), Srivastava (2014), and Bushman et al. (2016).

http://dx.doi.org/10.1016/j.jacceco.2017.07.002
0165-4101/

especially given that disclosure of a given topic often appears in multiple sections of the 10-K and any given passage often combines multiple topics.

In many ways, the issues in assessing 10-K content are similar to those faced in other literatures. For example, researchers in journalism have been interested in trends in coverage of the *New York Times* (Blei, 2012), those in literature in understanding topical trends in poetry (Rhody, 2012), in politics understanding trends in Senate discourse (Grimmer, 2010), in history understanding historical trends using the content of State Department cables (Chaney et al., 2015), and in science understanding topical trends in journals such as *Science* (Blei and Lafferty, 2007). In all of these domains, the challenge is in analyzing trends in corpuses far too large for humans to manually review and to summarize them in a way that is easily interpretable.

Following that literature, we use a natural language processing technique, Latent Dirichlet Allocation (LDA), to understand the changing content of 10-Ks.[3] LDA is a Bayesian computational linguistic technique that identifies the latent topics in a corpus of documents.[4] It is well suited to understanding the text of the 10-K because it permits analysis of the topical content of a large group of lengthy documents over time in an objective and replicable matter and relies on a very limited set of assumptions that are likely to be met in 10-K disclosure. Further, it is specifically designed to infer proportions of content for documents which contain multiple topics, even if the topics are entangled, which is important given that 10-Ks comprise a large number of interspersed topics.[5] It permits the proportion of the 10-K related to each topic to vary across documents so it is well-suited to examining topical trends in textual disclosure. As a result, we can deconstruct the 10-K by topic irrespective of whether topics appear in, for example, the footnotes, risk factors, or Management's Discussion and Analysis (MD&A). We can then assess trends in the discussion of topics over time and relate them to changes in specific disclosure requirements (e.g., new FASB standards, SEC requirements, and regulatory events such as SOX) and other events (e.g., changes in litigation risk, mergers and acquisitions, etc.).

Additionally, once the topic model is trained, it permits us to identify paragraphs by topic, so that we can track where specific topics occur within the 10-K. This allows us to identify the extent to which, for example, FASB requirements (e.g., footnotes) create redundancy with SEC requirements (e.g., risk factors and MD&A) by highlighting which topics tend to be redundant within the 10-K as well as across time and across firms. Similarly, by accumulating text within a topic, we can identify the topical sources of textual attributes that prior literature suggests may be important such as boilerplate, redundancy, stickiness, and lack of specificity.

In our empirical analysis, we examine the text of 10-Ks for 10,452 firms and 75,991 firm-years over the period 1996 to 2013. We begin by documenting trends in textual characteristics that have been identified by prior research as potentially affecting the informativeness of disclosure, including length (Loughran and McDonald, 2014), readability (Miller, 2010), boilerplate (Lang and Stice-Lawrence, 2015), redundancy (Cazier and Pfeiffer, 2015b), specificity (Hope et al., 2016), stickiness (Brown and Tucker, 2011), and the relative prevalence of informative numbers in the text or "hard" information (Blankespoor, 2016).

We document clear and consistent trends across all measures. Median text length doubled from 23,000 words in 1996 to nearly 50,000 in 2013, while redundancy, boilerplate, and stickiness increased nearly monotonically and readability, specificity, and the relative mix of hard information showed clear decreases. Given these trends, we next investigate their topical sources.

Prior literature (e.g., Cazier and Pfeiffer, 2015b) suggests that variables such as size, industry-composition, complexity, one-time events, litigation, and SEC oversight affect textual attributes such as length and readability in the cross section. Consistent with assertions by commentators such as Monga and Chasan (2015), it could be the case that those factors also change in the time series in ways that explain trends in textual attributes. We examine textual attributes after controlling for a wide variety of company-level variables suggested by the prior literature and for a constant sample of firms, but similar patterns persist. While those variables are significant cross-sectional determinants of textual attributes, including them in a regression framework does not explain the trend in disclosure characteristics over time.

Given that readily observable firm-level attributes do not explain the trends we observe, we use LDA to examine the topical content and characteristics of the additional disclosure. Our analysis suggests that the corpus of 10-Ks comprises 150 topics, which we aggregate into 13 broader categories for ease of discussion. The four categories which account for the bulk of 10-K length are Performance; Compliance with specific accounting and disclosure standards; Industry-Specific disclosure; and Employee-Related disclosure. However, only disclosure related to Compliance with specific accounting and disclosure standards increased substantially over time. Within this category, three topics explain the vast majority of the increase: fair value and impairment disclosure, discussion of internal controls, and risk factor disclosure.

To ensure that we have accurately identified the content of these three topics, we demonstrate that they are associated as expected with underlying economic attributes (special items, internal control weaknesses, and return variability). Then, we examine patterns in disclosure around the events that should have increased disclosure of these topics (implementation

[3] LDA is also used commercially. For example, the *New York Times* uses LDA to recommend articles to subscribers by inferring topics from articles they have read and identifying articles with similar content (Spangher, 2015).

[4] "Latent" refers to the fact that LDA is designed to infer the underlying topics in a document, "Dirichlet" refers to the family of probability distributions used in the estimation, and "Allocation" refers to the fact that the estimation allocates words to topics.

[5] For example, the discussion of pensions might include a discussion of foreign currencies and appear in MD&A as well as in risk factors and the footnotes.

of SFAS 157, Sarbanes-Oxley (SOX), and Item 1A). We document sharp increases in the length of these three topics in the years in which their associated standards were implemented, consistent with the LDA topics effectively capturing disclosure in response to standards. Disclosure associated with these topics is not limited to a single section of the annual report but extends across all of the major sections.[6] Similarly, the pattern in disclosure length for these three topics largely explains the increase in disclosure length for the 10-K as a whole.

In our third set of analyses, we link topical disclosure at the paragraph level, in particular that relating to the three major increasing topics, to other textual attributes of the 10-K. We demonstrate that fair value/impairment, risk factor, and internal control disclosure tend to have relatively high levels of redundancy, stickiness, and boilerplate, and low levels of readability, specificity, and hard information. Further analyses indicate that the increasing prevalence of these three topics contributed significantly to the overall increases in redundancy, stickiness, and boilerplate, and the decreases in readability, specificity, and the mix of hard information.

Finally, we examine cross-sectional variation in fair value, internal control, and risk factor disclosures. We document consistent patterns of increased length associated with these topics for disparate subsamples of firms suggesting that firms significantly increased disclosure length even when the additional disclosure may not have been as relevant. Further, we find that firms for which the requirements were potentially less relevant often responded by providing disclosure that was particularly high in boilerplate, redundancy, complexity, and stickiness, and lacking in hard information, although less so for fair values where firms appear to have exercised more flexibility to tailor disclosure based on materiality.

Overall, our evidence identifies clear trends in textual attributes and suggests that a substantial portion can be explained by disclosure in response to recent regulatory changes. While the fact that disclosure associated with new requirements increased over time is not in and of itself surprising, our results provide several potential contributions. First, we believe it is important to quantify the extent to which attributes of 10-K textual disclosure have, in fact, changed over time and attempt to distinguish among various explanations. While the three primary disclosure topics that we identify are logical candidates to explain the increase in 10-K disclosure length over time, it is noteworthy that they explain such a large proportion of the overall increase in length, as well as in other attributes such as complexity, redundancy, boilerplate, stickiness and lack of specificity. In contrast, economic factors from the prior literature (e.g., litigation risk, business complexity, and globalization) and the wide variety of other new requirements that were enacted during the sample period have limited ability to explain the disclosure trends that we document.

Second, we develop and demonstrate the value of natural language processing techniques such as LDA in understanding trends in the underlying content of textual disclosure. To our knowledge, ours is the first research to focus directly on trends in 10-K content over time, and we believe that LDA has the potential to be a powerful tool for understanding trends in the content of financial text because it provides an approach for evaluating topical coverage for large samples of lengthy documents on a consistent and objective basis over time. While summary quantitative measures such as length, redundancy, and readability are useful in providing aggregate characterizations of the accessibility and informativeness of documents, it is important to develop techniques that permit insight into the underlying content of disclosure in order to make these attributes interpretable. LDA permits the researcher to identify specific disclosure topics, highlight trends, isolate causes, and evaluate potential economic outcomes. Beyond 10-Ks, LDA has the potential to provide insight into trends in the content of other disclosures such as press releases, SEC speeches, conference calls, and articles in the business press.

Third, we use LDA to link specific topics to textual characteristics of annual reports that have been studied in prior research, providing a mechanism to assess their topical content, and provide systematic evidence across a number of dimensions on the trends in these characteristics. This is especially important given that prior literature focuses on textual outcomes at an aggregate level and generally does not incorporate the fact that discussions of different topics will have different textual attributes. LDA provides the opportunity to reinterpret the existing literature on outcomes of these attributes factoring in the actual content of the discussion to which they relate.

This research is subject to important caveats. First, topics from LDA (much like factors in factor analysis) require interpretation by the researcher. As discussed in the research design section, we follow the prior computational linguistics literature in identifying the appropriate number of topics. In addition, we review the word lists and read representative paragraphs for each topic to ensure the content matches the label and investigate the timing of changes in major topics around regulatory changes to ensure they behave as expected. As a result, we are confident that our interpretations are reasonable and consistent with the behavior of the topics in our corpus.

Second, our results are descriptive and do not allow us to draw normative conclusions.[7] We focus on a set of textual attributes that academic research, regulators, and investors suggest may be important attributes of disclosure. However, we

[6] Because the three major topics that we identify are linked to individual standards, it might be tempting to assume that analysis of specific sub-sections within the 10-K would be sufficient to draw conclusions. However, not only is it difficult to reliably identify specific disclosures within each of the sections of the 10-K for a large sample, but our results indicate that disclosure topics, and our three main topics of interest, tend to be spread throughout the sections of the 10-K. For example, the median topic has material (greater than 100 words) discussion in three separate sections for many of the documents in our sample, while our three main topics have material discussions spread in five major sections across firms. This wide dispersion of topics across the sections of the 10-K makes LDA a particularly appropriate tool in this setting because it is able to identify text relating to a single topic that is interspersed throughout a lengthy document.

[7] To some extent, the challenge here is analogous to the literature on trends in characteristics of net income over time in that the researcher cannot make normative statements given the variety of stakeholders and other sources of information. Rather, the literature relies on characteristics such as the ability of earnings to predict future cash flows or the association with returns which prior research or standard setters have suggested are likely to be

acknowledge the limitations of these textual measures to capture meaningful aspects of disclosure in the specific context of 10-Ks, especially for sophisticated financial statement users. Closely related, while the textual attributes we consider have been linked to various aspects of informativeness in prior research using aggregate 10-K text, it is possible that those results do not apply to specific topics we consider.[8]

2. Background and related research

As noted above, financial reporting research has traditionally focused on quantitative data, particularly summary statistics such as net income and shareholders' equity, reflecting in large part the relative ease of assessing associations between quantitative data, coupled with an inherent assumption of unlimited information-processing capacity on the part of investors (see, for example, the papers cited in Footnote 1). More recently, researchers have begun to explore determinants of textual attributes of the 10-K.[9] While prior studies focus on cross-sectional determinants of textual characteristics, our results suggest that those factors have limited ability to explain trends in reporting over time.

In our analysis we focus on a broad set of textual attributes: length, readability, redundancy, boilerplate, specificity, stickiness, and the number of numbers in the text relative to the number of words (which we refer to as the relative mix of hard and soft information). We examine multiple characteristics because no single attribute can conceptually or empirically capture all aspects of disclosure that are relevant to financial statement users. We are guided by attributes that have been studied in the prior academic literature. We acknowledge that the interpretation of these textual attributes is limited by the fact that we do not directly measure the usefulness of the actual disclosure content, that different users (e.g., sophisticated and unsophisticated investors) may be affected by the same attributes differently, and that different types of information may lend themselves to disclosure with different attributes (e.g., some topics may lend themselves to disclosure which is more redundant, boilerplate, or expressed in longer sentences). To our knowledge, ours is the first academic paper to focus on identifying the magnitude, content, and causes of time trends in textual disclosure, although these trends have received substantial attention by practitioners.

First, prior literature in academia and practice has discussed effects of less readable and lengthy disclosure, sometimes referred to as disclosure "overload" (KPMG, 2011). Similar to the incomplete revelation hypothesis in Bloomfield (2002) in which statistics that are costly to extract are not fully incorporated into price, these attributes have been shown to decrease information impounded at the time of the filing and increase subsequent price drift (Lee, 2012; You and Zhang, 2009).[10]

Similarly, redundancy of disclosure within a document, re-use of the same firm's disclosure from a prior period (disclosure "stickiness"), and generic and standardized disclosure (often referred to as "boilerplate") are all discussed by the FASB in its invitation to comment on the disclosure framework project (FASB, 2012), and the SEC has urged firms to evaluate boilerplate disclosure and indicated that redundancies between FASB and SEC disclosure requirements will be a focus going forward (Higgins, 2014; SEC, 1998, 2013). Cazier and Pfeiffer (2015b) show that redundant disclosure leads to less efficient price discovery, while Brown and Tucker (2011) find that MD&As that are updated less over time ("sticky" disclosures) have muted stock price responses. Lang and Stice-Lawrence (2015) empirically link the use of boilerplate to decreased liquidity, analyst following, and institutional ownership for an international sample.

Lastly, we examine the specificity of disclosure and the relative amount of hard information. Regulators have expressed concern that textual disclosure has become increasingly vague and less likely to be supported by quantitative data (SEC, 1998). To capture this, we calculate specificity as how often the text refers to specific people, places, organizations, times, or numbers. Hope et al. (2016) show that more specific risk disclosures lead to greater market reactions and better risk assessments by analysts. To measure the extent to which narrative disclosure is supported by quantitative data (the relative mix of hard and soft information), we measure the number of informative numbers in the 10-K (i.e., excluding dates and section numbers) relative to the total number of words. This gives a sense of the quantitative density of disclosure, because text that contains numbers is more verifiable and precise than general descriptions of topics. Blankespoor (2016) documents an increase in quantitative disclosure after the introduction of XBRL, consistent with firms providing more quantitative data when users' processing costs decrease.

We use LDA to identify notable changes in disclosure content over our period and the extent to which changes in topical content influence trends in each of these disclosure attributes. As noted earlier, LDA has been used in many other literatures; however, it has only recently been used in accounting and finance. For example, Huang et al. (2016) employ LDA to examine differences between the topics discussed in conference calls and analyst reports, Hoberg and Lewis (2017) use LDA to examine the content of a firm's MD&A in years surrounding fraud, and Ball et al. (2015) use LDA to identify topics within

desirable attributes. However, the results are innately descriptive and it is up to the standard setter to decide how to apply them in practice and up to future research to link the earnings characteristics to usefulness in particular settings.

[8] For example, while Cazier and Pfeiffer (2015b) provide evidence that redundant disclosure is associated with less efficient price formation and Loughran and McDonald (2014) suggest that complex 10-K disclosure is associated with a muted stock price response, it is possible that redundant and complex disclosure on specific topics increases information content for subsets of investors. There are clearly opportunities for future research investigating the effects of disclosure attributes like complexity and redundancy in the context of specific topics and investor groups.

[9] For example, Li (2008) links Fog to poor performance, Cazier and Pfeiffer (2015a) link length to complexity, and Cazier and Pfeiffer (2015b) link redundancy to obfuscation.

[10] Additionally, disclosure length and Fog have been shown to lead to greater market uncertainty (Loughran and McDonald, 2014) and less investment and trading by small investors (Lawrence, 2013; Miller, 2010).

Table 1
Descriptive statistics.

	Mean	Std Dev.	Q1	Median	Q3	N
Words	45,349	31,454	24,678	37,370	55,852	75,991
Redundant Words	3902	5476	1112	2276	4284	75,991
Boiler Words	13,271	10,073	6584	10,882	16,521	75,538
Sticky Words	25,735	15,612	14,303	22,500	33,167	69,526
Specificity	51.89	13.45	42.26	50.75	60.32	75,991
HardInfoMix	18.65	5.25	14.80	17.93	21.76	75,991
Fog	21.34	1.25	20.54	21.21	21.95	75,991
BigN	0.77	0.42	1.00	1.00	1.00	75,991
Assets ($)	3551.30	11,733.65	91.85	391.85	1747.65	75,991
Intangibles	0.12	0.17	0.00	0.03	0.18	75,991
MTB	3.14	4.21	1.17	1.90	3.33	75,991
Leverage	0.21	0.20	0.02	0.16	0.34	75,991
BusSeg	1.95	1.47	1.00	1.00	3.00	75,991
ForSeg	1.14	1.64	0.00	1.00	2.00	75,991
NYSE	0.43	0.50	0.00	0.00	1.00	75,991
Loss	0.31	0.46	0.00	0.00	1.00	75,991

Assets are measured in millions and scaled to be in constant year 1996 dollars.

MD&A. While the prior literature confirms that LDA has the potential to organize textual disclosure for numerical analysis, it has not, to our knowledge, been applied to understanding trends in 10-K disclosure or to identifying the topical sources of constructs such as length, readability, redundancy, specificity, boilerplate, stickiness, or the mix of hard information.

3. Data

We generate a database of text using SEC 10-K filings spanning the years 1996 to 2013.[11] Control variables in our reported analyses are obtained from CRSP and Compustat. Following Loughran and McDonald (2014) we remove firms with negative market-to-book ratios. The intersection of our data constraints results in a sample of 10,452 firms and 75,991 firm-years. Definitions for all of our variables are included in the Appendix.

Table 1 provides descriptive sample statistics. The median firm included 37,370 words in their annual report. Based on the Fog index, reading and comprehending the median annual report requires approximately 21.21 years of formal education. The median annual report has 2276 words (6% of the 10-K for the median firm) in sentences that are repeated verbatim throughout the annual report, 10,882 words (29%) in sentences containing boilerplate phrases, and 22,500 words (67%) in sentences containing "sticky" phrases.[12] Median levels of Specificity and HardInfoMix of 50.75 and 17.93 indicate that the median 10-K includes about 51 and 18 specific terms and informative numbers, respectively, for every 1000 words of text. Lastly, approximately 77% of the sample firms are audited by a Big "N" auditor, and 31% report a loss.

Fig. 1 provides initial evidence on trends in reporting over our sample period, including length, readability, redundancy, boilerplate, stickiness, specificity, and the mix of hard information. Perhaps most prominent and relevant for our purposes is the increase in length depicted in Fig. 1 Panel A and the near monotonicity of this increase. While there is some evidence of larger increases around SOX in the early 2000s and the financial crisis, especially for firms in the 75th percentile, the increase for the median firm has been remarkably continuous. The number of words for the median firm increased from about 23,000 in 1996 to almost 50,000 in 2013.

In terms of other attributes, the pattern for redundant words in Fig. 1 Panel B is similarly conspicuous, with the median firm increasing from 800 words in redundant sentences in 1996 to almost 3300 in 2013.[13] Fig. 1 Panels C and D suggest similar upward trends in the amount of boilerplate and stickiness, with both tripling over time, indicating an increasing tendency for firms to repeat disclosure from year to year and to use generic disclosure. On the other hand, readability for the median firm decreased monotonically over the twelve years since 2000, following an increase between 1998 and 2000 that was likely a result of the SEC's plain English requirements in 1998.[14] Results in Panels E and F also suggest clear decreases in other attributes of informative disclosure, with specificity and the relative mix of hard information exhibiting nearly monotonic decreases over the period.

The preceding analysis is descriptive, but it provides strong initial evidence that trends in textual disclosure have been systematic and that attributes which prior research suggests have the potential to reduce the informativeness of disclosure have increased, while those which prior research suggests may enhance informativeness have decreased. Further, the consistency in trends among the attributes suggests the possibility that the same underlying factors may be driving the series.

[11] We include only those filings in 1996 that were issued after electronic filing on EDGAR became mandatory.

[12] Similar to Lang and Stice-Lawrence (2015), we identify boilerplate as 4-word phrases that are extremely common across all firms in a given fiscal year. "Sticky" phrases are 8-word phrases that are repeated from the same firm's prior year report. See the Appendix for further details.

[13] Our measure of redundancy almost certainly understates the true level because we err on the side of being conservative by requiring verbatim repetition of sentences. Conclusions are consistent if we relax our criteria by not requiring that all words in a sentence be repeated verbatim.

[14] Fog is defined as the average number of words per sentence plus the percent of words containing more than two syllables, multiplied by 0.4, and can be interpreted as the number of years of formal education an individual would need to read and understand a given document.

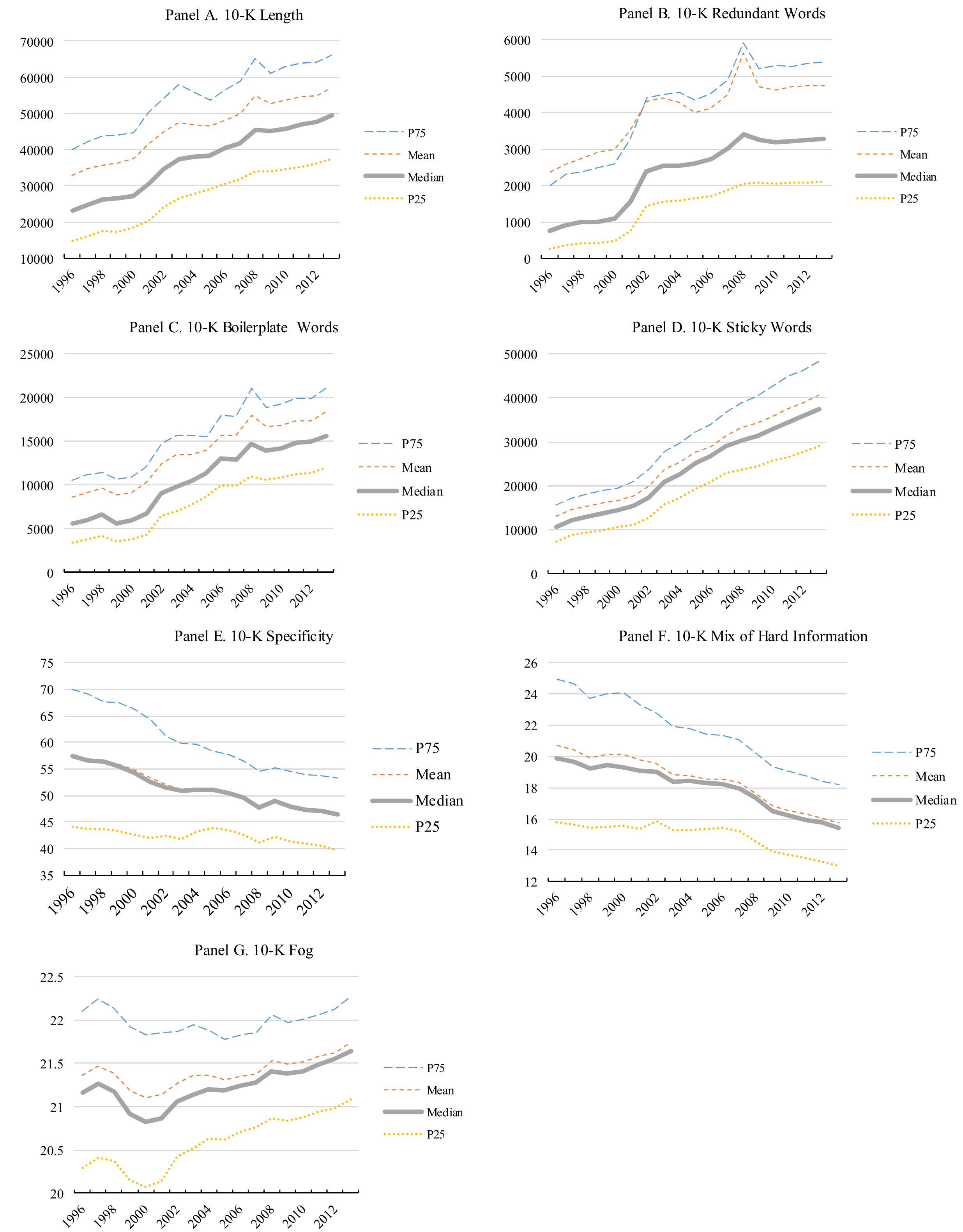


Fig. 1. Trends in textual attributes over time.

Table 2
Determinants of textual attributes.

	LnWords (1)		LnRedundantWords (2)		LnBoilerWords (3)		LnStickyWords (4)	
	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Trend	0.038***	64.000	0.091***	70.997	0.060***	92.101	0.070***	129.175
BigN	0.047***	5.741	0.026	1.598	0.018**	2.137	0.039***	5.281
LnAssets	0.142***	49.101	0.173***	31.040	0.128***	41.352	0.125***	46.359
Intangibles	0.104***	4.730	0.249***	6.156	0.215***	9.314	0.024	1.201
MTB	0.005***	7.573	0.004***	2.610	0.004***	5.019	0.005***	7.345
Leverage	0.164***	7.836	0.254***	6.324	0.182***	8.135	0.059***	3.111
Age	−0.006***	−18.239	−0.005***	−7.715	−0.005***	−14.611	−0.007***	−20.011
BusSeg	0.019***	7.391	0.034***	6.952	0.009***	3.331	0.014***	5.932
ForSeg	0.008***	3.913	0.009**	2.249	0.009***	4.041	0.012***	6.233
NYSE	−0.008	−0.847	0.024	1.349	0.013	1.333	−0.011	−1.232
Loss	0.231***	38.710	0.353***	30.710	0.215***	33.747	0.153***	28.996
Observations	75,991		75,991		75,538		69,526	
R-squared	0.367		0.297		0.374		0.613	
Industry Fixed Effects	Y		Y		Y		Y	

	Specificity (5)		Hard_Info_Mix (6)		Fog (7)	
	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Trend	−0.738***	−49.150	−0.320***	−54.286	0.022***	15.629
BigN	−1.801***	−8.599	−0.719***	−8.268	0.089***	4.650
LnAssets	−0.503***	−7.373	−0.395***	−13.754	0.117***	18.265
Intangibles	0.378	0.717	0.598***	2.865	0.343***	6.927
MTB	−0.070***	−3.844	−0.067***	−8.204	0.000	0.059
Leverage	3.784***	7.346	3.099***	14.745	0.405***	8.626
Age	0.133***	16.512	0.052***	15.585	−0.006***	−7.206
BusSeg	0.408***	6.343	0.139***	5.268	−0.009	−1.505
ForSeg	0.353***	6.477	−0.016	−0.740	−0.006	−1.129
NYSE	1.054***	4.800	0.337***	3.706	0.096***	4.713
Loss	−2.032***	−14.078	−1.086***	−18.906	0.156***	11.610
Observations	75,991		75,991		75,991	
R-squared	0.143		0.186		0.110	
Industry Fixed Effects	Y		Y		Y	

Determinants of textual attributes, including cross-sectional determinants and a time-series trend (*Trend*). Assets are inflation-adjusted to be in constant year 1996 dollars. All continuous variables are winsorized at the 1st and 99th percentile by year. Standard errors are adjusted for heteroskedasticity and clustered by firm. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In the next section, we examine the extent to which determinants previously used to explain cross-sectional variation in textual attributes explain the trends that we observe.

4. Why have textual attributes changed over time?

There are several potential explanations for the changes in 10-K characteristics over time. First, they might reflect a change in sample composition if, for example, more firms with intangible assets (and potentially lengthier and more complex corresponding disclosure) began trading publicly during the sample period.[15] However, untabulated analysis indicates that all seven of our attributes continue to trend very similarly for a constant sample.

Alternatively, some practitioners have suggested that changes in 10-K disclosure over time may be the result of changes in the economic fundamentals of firms (Monga and Chasan, 2015; FASB, 2012; SEC, 2013). For example, factors such as business complexity, leverage, size, auditor, and profitability have been shown to be important cross-sectional determinants of textual attributes (Cazier and Pfeiffer, 2015a,b; Li, 2008). In Table 2 we report results for regressions where we explain our textual outcomes using variables such as size, auditor, NYSE membership, complexity in terms of numbers of business segments or operating segments, market-to-book ratio, leverage, intangibles, and losses. Although we do not discuss all of the coefficients in this table for parsimony, results are generally consistent with expectations. Length increases with size, complexity, Big-N auditor, market-to-book, leverage, and losses. Firms reporting losses tend to have vague and "foggy" discussions (lower readability, specificity, and hard information mix), consistent with the obfuscation hypothesis in Li (2008). However, the *Trend* variable remains strongly significant for all of the textual attributes.[16]

[15] For example, Srivastava (2014) suggests that trends in value relevance of accounting data can be explained by changes in the sample composition of publicly-traded firms.

[16] In untabulated results, we include additional control variables related to litigation risk, R&D, ownership, analyst following, number of comment letters filed for the firm and its peers, the number of Accounting, Auditing and Enforcement Releases for peer firms in the industry, unexpected earnings, mergers,

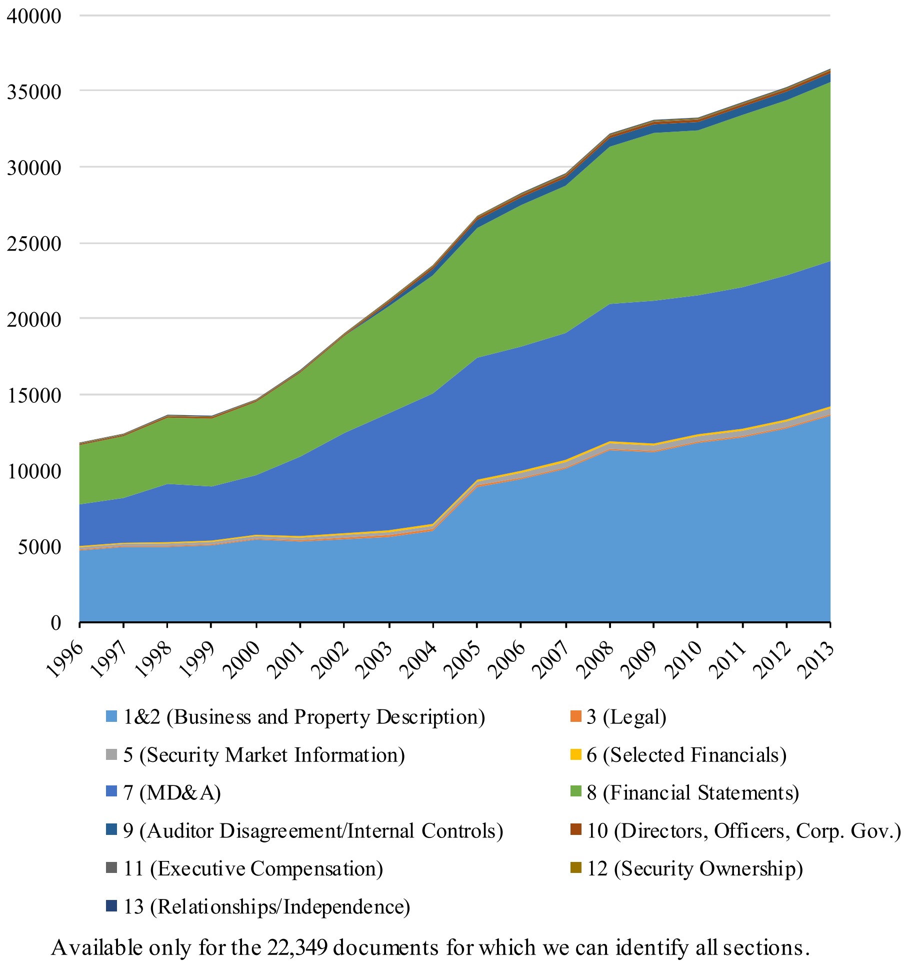


Fig. 2. Median 10-K section length by year.

The preceding analysis suggests that disclosure attributes are influenced by firm-specific variables in predictable ways. However, the trends remain significant after controlling for these variables. Another possibility is that the change in overall length is driven by increases in specific sections of the 10-K due to changes in disclosure requirements by either the SEC or the FASB (KPMG, 2011; White, 2013).

Fig. 2 plots the median length for sections of the 10-K. Of the eleven sections, three make up about 90% of the total text in the most recent year: Sections 1 and 2 (Business and Property Descriptions), Section 7 (Management's Discussion and Analysis), and Section 8 (Financial Statements and Footnotes). Sections 1, 2 and 7 reflect SEC requirements and Section 8 reflects FASB standards. Most noticeable from Fig. 2 is the fact that the length of each of the major sections has increased substantially and at roughly the same rate over time. As a consequence, the proportion of the 10-K in each section has remained similar over time, with Sections 1 and 2 comprising 36% of the total in 1996 as compared with 35% in 2013, Section 7 comprising 21% in 1996 vs. 25% in 2013, and Section 8 comprising 30% in both 1996 and 2013. We find similar results when we examine the rest of our textual attributes at the section-level (untabulated for parsimony), with the textual attributes trending within all of the major sections and the relative contribution of the major sections to overall attributes remaining relatively constant over time. Thus while changes to disclosure standards may be important determinants of overall changes in 10-K disclosure, these results suggest that overall changes in disclosure do not reflect requirements that are unique to a specific section but rather reflect content that spans multiple sections, including sections under the purviews of both the FASB and SEC.

The preceding analyses suggest that firm-level determinants from prior research and specific sections of the 10-K do not fully explain the trends in textual disclosure. In the next section we use LDA to identify the topical content of the disclosure and, most importantly, to quantify the topics that account for the bulk of the changes in overall length that we observe.

and market-wide returns. In all cases, we find similarly significant trend coefficients, but inclusion of these additional controls decreases sample size by nearly half due to data requirements.

This more nuanced analysis at the topic level allows us to study drivers of trends in a more detailed way than is possible from analyzing text at the document- or even the section-level.

5. Using LDA to explain the change in 10-K length

LDA is an unsupervised Bayesian machine-learning approach developed by Blei et al. (2003) to identify the topics contained in a large corpus of text. LDA uses the probability of words co-occurring within documents to identify sets of topics and their associated words and is conceptually similar to factor analysis, where the model produces topics instead of factors. As in factor analysis, the computer identifies the words associated with a topic and the researchers assign a label to the topic based on their assessment of the likely content given the set of words and their probabilities. LDA is particularly useful in our setting because it allows us to identify the mix of topics in the overall 10-K and within each section, even though multiple topics may be interwoven in any given section and any individual topic may occur in multiple sections. This allows us to identify the topics contained in each 10-K and trends in their proportions over time.

As noted earlier, LDA has been used in a variety of contexts to investigate trends in the content of textual disclosure over time and is designed to analyze large numbers of documents, each of which potentially contains multiple latent topics (e.g., the *New York Times*, French poetry, or State Department cables). It makes a minimal number of assumptions that are likely to be at least approximately met in 10-K disclosure. First, it assumes that the overall corpus of documents contains a finite number of topics, implying that every document consists of a mix of those topics. With input from the researcher, LDA helps to estimate the number of topics in the overall corpus (in our case 150 topics) as well as the proportion of each topic in each document (the proportion can vary across documents or over time, and not every document need contain every topic).[17] Second, LDA assumes that specific words appear with different frequencies across topics. LDA estimates the frequency of each word within a topic (a given word may appear across multiple topics with different frequencies and not every word need appear in every topic). As a result, the output from applying LDA to the population of 10-Ks is the proportion of each of the 150 topics that appears in each 10-K (e.g., a given 10-K might be 1.5% about Pensions), and the relative weights of words in each topic (e.g., the word "actuarial" might be 10 times more likely to occur in the Pension topic than the word "derivative"). While the researcher helps to determine the number of topics that are generated by the model (in our case 150), that choice is guided by a specific methodology discussed below. We use the MALLET software developed by McCallum (2002) to apply LDA to our sample and generate topics for our document collection using collapsed Gibbs sampling.[18]

Because LDA is an unsupervised method, it is replicable and free of researcher bias. However, because the topics can sometimes be difficult to interpret, it is important that the researchers help to select the number of topics generated by the model. Following prior literature, we use a variety of criteria to ensure that we identify the appropriate number of interpretable topics. First, as proposed in Blei et al. (2003), we measure the "perplexity" of the topic model (defined more formally in the Appendix) for topic models with between 10 and 400 topics and observe that perplexity begins leveling off at 150 topics. Because lower perplexity indicates that the model is a better fit for the observed data, this indicates that the model performance gains relatively little from increasing the number of topics after that point.

Although perplexity is a good general guide, and lower perplexity will always lead to models with at least marginally better fit relative to held-out data, the increase in fit is sometimes at the expense of interpretability due to overfitting. Chang et al. (2009) discuss how increasing the number of topics to produce ever finer partitions can make the model less useful because it becomes almost impossible for humans to differentiate between many of the topics. They propose a task in which the overall interpretability of a particular LDA model is measured by how often a human coder agrees with the topics chosen using the model. We perform this "word intrusion" task by providing research assistants with sets of six words, five of which the computer suggests belong in the same topic and a sixth which appears commonly in 10-Ks but which the model did not assign to that topic (an "intruder" word). The extent to which the human coder agrees with the computer on the assignment of words to a topic is a measure of the effectiveness of the technique in capturing meaningful topics. We perform this word intrusion task for topic models of 150, 200, and 250 topics (more details in the Appendix) and find that the 150-topic model has the best interpretability (i.e., the fewest disagreements between the LDA model and human coders). As a consequence, we use LDA topics from the 150-topic model.

Because it is difficult to present details on 150 topics concisely, in our initial analyses we manually group each of the LDA topics into thirteen broad categories. To form these categories, two individuals with financial backgrounds (one MBA student with work experience in banking and one of the authors) independently evaluated each of the 150 topic word lists and determined the best fit of each topic into broader category groupings.[19] Category labels are for parsimony and ease of interpretation and do not affect the statistical analysis. The Internet Appendix includes the full list of all 150 topics, the top 20 words most frequently associated with each topic, a topic label created by the researchers, and a "representative

[17] LDA generates "topic loadings" which can be interpreted as the proportion of the document comprising each topic and which, for a single document, sum to one. Our model allows the prominence of topics (the alpha hyperparameter) to vary across the entire corpus of all 10-Ks so that topics that appear in relatively few documents (e.g., industry-specific topics such as healthcare) are given less prominence while topics that are used in more documents (e.g., accounting policies) are given more prominence. This essentially means that common topics are allowed to be "bigger" than others so that they have a consistently higher topic loading on average.

[18] More details are available at http://people.cs.umass.edu/~mccallum/code.html.

[19] In most cases, the two coders agreed on categorization, but in cases in which the coders disagreed one of the authors judged the best fit.

Table 3
Description of topic categories.

Category label	Description	
Business operations & strategy	Topics relating to day-to-day company operations or strategy (discussion of customers is in the topic relating to performance). EX: Products, advertising, accounts receivable, contractors, software, and systems.	
Business structure & M&A	Discussions of the current business structure and organization, or changes to these. EX: Subsidiaries, partnerships, acquisitions, bankruptcy and reorganization, trusts, joint ventures.	
Compliance with SEC & accounting standards	Discussions of SEC requirements and accounting standards, or disclosures to comply with these requirements. EX: Issuance of new accounting standards, discussion of regulatory documents for the annual report or prospectus, (management's certification of) internal controls, fair value disclosure, required risk factor disclosures.	
Contracts & legal	Disclosure about legal agreements or proceedings. EX: Provisions of contracts, litigation.	
Employees & executives	Disclosure about employees and executives. EX: Salaries and benefits, retirement, unions, executive backgrounds, indemnification agreements, code of conduct.	
Geographic location	Discussions about various specific geographic regions (mostly in relation to regional operations). EX: Southwestern United States, China, Midwest, Latin America.	
Intellectual property & R&D	Intellectual property and research and development. EX: Patents, laboratory research, licensing rights.	
Investments, securities, derivatives	Discussion of the firm's investments. EX: General investment activity and risk, securities investment and trading revenue, REITs, derivatives.	
Loans, debt, banking	All discussions relating to loans and debt. EX: Loan obligations, payments, rates, and collateral; mortgages; debentures; and default.	
Performance, revenues, and customers	Discussion of performance, revenue, and customers. EX: Performance summary, clients and revenue, customer accounts, distribution to customers, growth, special items.	
Property and leasing	Topics relating to properties. EX: Leases, tenant-landlord issues, and transactions.	
Stock and options	Discussions relating to the company's own stock, including options, warrants, and dividends.	
Industry-specific disclosure categories	Healthcare & medical	Insurance
	Energy, resources, and utilities	Transportation
	Media, communications, and leisure	Technology industry
	Consumer products	Other industry-specific

For more detailed information on the specific topics that are included in each category, see the Internet Appendix.

paragraph" identified using a procedure similar to Hoberg and Lewis (2017). For all other details relating to the specifics of our LDA procedure and the generation of representative paragraphs, please see the Appendix.

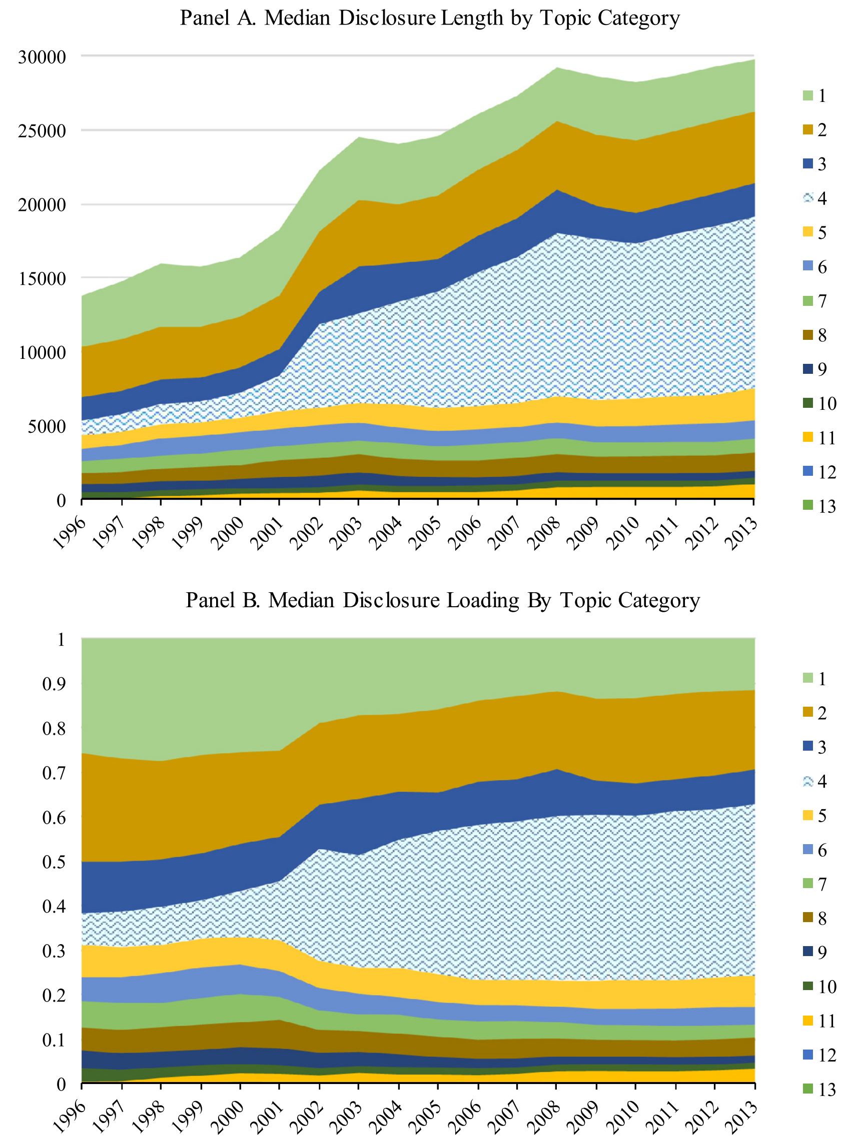
Table 3 lists the broad categories into which we group the topics in our analysis, along with brief descriptions.[20] For example, "Business Operations and Strategy" refers to discussion of day-to-day business operations such as products, advertising, and information systems; "Business Structure and M&A" refers to discussion of subsidiaries and partnerships, as well as mergers, acquisitions, and other corporate transactions; and "Loans, Debt and Banking" refers to discussion of the firm's financing. Of the categories, the five that constitute the largest portion of 10-K text, especially in the early part of the sample period, are: "Performance, Revenues, and Customers," which is primarily discussion of the performance and revenue generation of the firm; "Industry Specific Disclosure," which includes topics that are unique to specific industries (e.g., healthcare or transportation); "Employees and Executives," which includes descriptions of executives and executive compensation plans; "Compliance with SEC and Accounting Standards," which is text associated with specific reporting requirements; and "Investments, Securities, and Derivatives," which includes descriptions of financial instruments. The objective identification of topics by the LDA procedure and our more subjective grouping into categories allows us to disaggregate the overall trend in length into the portions attributable to individual types of disclosure. We construct a pseudo topic length by multiplying the topic loadings by the length of the total document to estimate the number of words used to discuss each topic.

Fig. 3 Panel A plots the median number of words in each of the broader categories over time. In general, the pattern is clear. Most topics have remained relatively constant over the sample period and therefore do not explain the overall increase in 10-K length. The notable exception is "Compliance with SEC and Accounting Standards" which increased markedly during the sample period.[21] Essentially all of the increase in the length of disclosure for the median firm over the sample period appears to be associated with the Compliance with SEC and Accounting Standards category. Fig. 3 Panel B provides a similar trend analysis but expressed as the median proportion of total disclosure (i.e., scaling the proportion of disclosure on each topic so that the total adds up to 1).[22] Again, we see that the proportion of disclosure related to the Compliance category has

[20] Our categorization is admittedly subjective and is only for expositional purposes (later analyses split out the 150 subtopics which were determined by LDA). The number of categories and their contents were selected based on the perception of similarity of content. To validate the categories, we created vectors for each topic using the rank of the top 20 words and calculated the average cosine similarity of topics within each category (Aletras and Stevenson, 2014). We then compared these average similarities to the similarity of topics within 1,000 benchmark categorizations based on randomly assigning topics to categories. The results suggest that our grouping exhibits more word similarity than would be expected at random at the 0.001 level.

[21] The title "Compliance" is not intended to be pejorative. The category reflects topics for which the categorizers could clearly identify the disclosure as a response to a specific SEC or FASB requirement and the text did not fit naturally into any other topic.

[22] Although the proportions of all topics and topic categories within individual documents sum to one, the sum of median proportions within a given year may not. For expositional clarity, we scale the sum of all median proportions by year to sum to one; the inferences from the unscaled graph are identical.



1. *Performance, Revenues, and Customers*
2. *Industry-Specific Disclosure*
3. *Empoyees & Executives*
4. *Compliance with SEC & Accounting Standards*
5. *Loans, Debt, Banking*
6. *Business Operations & Strategy*
7. *Stock and Options*
8. *Business Structure & M&A*
9. *Contracts & Legal*
10. *Geographic Location*
11. *Investments, Securities, Derivatives*
12. *Intellectual Property & R&D*
13. *Property and Leasing*

Although topic loadings within documents sum to 1, median loadings across firms do not; for presentation purposes, median yearly loadings in Panel B are scaled to sum to 1.

Fig. 3. Disclosure over time by LDA topic category.

increased markedly as a proportion of the total length over the sample period, while the proportions of the other categories (by construction) have decreased.

The preceding analysis provides preliminary, although circumstantial, evidence on the source of the increase in 10-K length over time. Although a general increase in the length of the Compliance category may not be surprising given the introduction of new requirements over our sample period, the fact that the increases are limited to disclosure in the "Compliance" category is potentially more surprising because one might also have anticipated increases in categories such as "Business Operations and Strategy," "Business Structure and M&A," or "Performance, Revenues and Customers," with, for ex-

Table 4
Top 15 increasing LDA topics.

Change in median length	Topic title	Topic category
4317.41	Fair value/impairment	*Compliance with SEC and accounting standards*
2216.18	Internal control disclosure	*Compliance with SEC and accounting standards*
2092.66	Risk factor disclosures	*Compliance with SEC and accounting standards*
590.19	Customer accounts	*Performance, revenues, and customers*
356.85	Financing (Facilities)	*Loans, debt, banking*
343.78	Accounting Standards Codification (ASC)	*Compliance with SEC and accounting standards*
328.74	Derivatives	*Investments, securities, derivatives*
270.25	Acquisitions	*Business structure & M&A*
216.31	Exhibits incorporated by reference	*Compliance with SEC and accounting standards*
212.47	Growth	*Performance, revenues, and customers*
143.79	Foreign currency exchange	*Business operations & strategy*
116.33	Special items	*Performance, revenues, and customers*
46.70	Litigation	*Contracts & legal*
37.89	CEO/CFO certification of internal controls	*Compliance with SEC and accounting standards*
19.93	Pension and retirement plans	*Employee & executives*

The change in median length for each topic is calculated as the average median length for 2013 and 2012 (in words) minus the average of median length for 1996 and 1997.

ample, general trends in business complexity, firm size, or globalization over time. Further, the magnitude of the increase is substantial, with textual disclosure in the Compliance category increasing approximately ten-fold over the sample period.

Although the components of the broader categories are somewhat subjective, the analysis of the subcomponents is objective because LDA determines the 150 individual topics and assigns specific text to them. Table 4 reports the top increasing topics by length. The top three increasing topics are all part of the Compliance with SEC and Accounting Standards category, namely fair value/impairment, internal control, and risk factor disclosures. Notably, these three topics alone make up the bulk of the increase in overall length with increases of 4300, 2200, and 2100 words, respectively, compared to an increase of less than 600 words for the next most increasing topic, customer accounts. While we would expect these topics to have increased in length given the implementation of new standards (KPMG, 2011; White, 2013), it is noteworthy that they make up such a large proportion of the total increase in disclosure length, especially given the substantial number of other new FASB and SEC requirements during the sample period.[23] Because of the large magnitude of the increases in the lengths of these three topics compared to all other topics, we focus on them in our remaining analyses. Examining them individually by year allows us to establish when (and, indirectly, why) these topics increased so substantially.

One potential concern with the preceding analysis is that LDA may be substituting disclosure that had previously appeared in other topics into our Top 3 topics as a mechanical effect of more standardized language following disclosure guidance. To ensure that is not the case, we investigated whether there are potentially offsetting decreases in any other topics during our sample period. Consistent with the notion that disclosure is added but seldom eliminated, none of the other topics decreased in length over our sample period enough to account for the increase in our Top 3 topics. To examine the issue more formally, we identified the 3, 5 and 10 topics most closely related to each of our Top 3 topics based on cosine similarity (Brown and Tucker, 2011). Netting changes in disclosure length of each of our Top 3 topics with changes in related topics yields similar increases in net disclosure to those reported for our Top 3 topics alone, suggesting that substitution across topics does not explain the increases that we observe.

The first of these three topics relates to fair value and impairment disclosure. Its top words according to the LDA procedure include: “fair,” “reporting,” “consolidated,” “impairment,” “control,” “future,” “recognized,” “estimated,” “expected,” and “asset.” [24] Because SFAS 157 is the most important standard to affect fair value accounting, we expect that much of the disclosure categorized under this topic will be related to that standard. In Table 5 we list the representative paragraphs for each of our Top 3 increasing topics, including the Fair Value/Impairment topic. The representative paragraph for this topic relates to the effect of fair values for evaluating goodwill impairment (in addition to establishing a framework for measuring fair value accounting, SFAS 157 specifically amended SFAS 142 relating to goodwill impairment). Examination of paragraphs with a high loading of the fair value topic indicates that the grouping reflects fair value discussion on a range of topics including derivatives, investment securities, and other investments. Because the representative paragraph technique is inherently biased toward more standardized paragraphs (e.g., those that cite the relevant standards), we provide an additional sample paragraph to provide a more comprehensive view of the range of discussions that fall within this topic (additional sample paragraphs for all three topics are in the Internet Appendix). This paragraph discusses the use of fair values in yearly evaluations of debt and equity securities, also related to SFAS 157.

The next topic relates to internal control disclosure, with the top five words being “control,” “internal,” “reporting,” “registrant,” and “material.” The representative paragraph is the auditor’s opinion on management’s assessment of internal control, required under SOX Section 404, with an additional sample paragraph that is from management’s discussion of the effectiveness of their internal control system. Of all of our Top 3 topics, internal control disclosures tend to follow the word-

[23] These three subtopics taken together account for about 9,000 of the 10,000 total median word increase in the Compliance topic over the sample period.
[24] Note that “value” was excluded from the LDA procedure because it is extremely common; therefore, it cannot appear as a keyword for any topic.

Table 5
Representative paragraphs for top 3 increasing topics.

Topic	Representative paragraph
Fair Value/Impairment	In accordance with GAAP, the Company has the option to first assess qualitative factors to determine whether it is necessary to perform a more detailed quantitative impairment test. If the Company is able to determine through the qualitative assessment that it is more likely than not that the fair value of a reporting unit exceeds its carrying value, no further evaluation is necessary. However, if the Company concludes otherwise, then the Company is required to perform the first step of the two-step impairment test by calculating the reporting unit's fair value and comparing the fair value to the reporting unit's carrying amount, including goodwill. If a reporting unit's fair value exceeds its carrying value, the second step of the impairment test is not required and no impairment loss is recognized. If a reporting unit's carrying value exceeds its fair value, the second step of the impairment test is performed to measure the amount of the impairment loss and an impairment charge is recorded equal to the difference between the carrying value of the reporting unit's goodwill and the implied fair value of the reporting unit's goodwill. The implied fair value of goodwill is determined in the same manner as the amount of goodwill recognized in a business combination where the excess of the fair value of the reporting unit over the fair value of the identifiable net assets of the reporting unit is the implied fair value of goodwill. See Note 5 Goodwill and Intangible Assets, Net.
Fair value/impairment (additional example paragraph)	Each reporting period we review all of our investments in equity and debt securities, except for those classified as trading, to determine whether a significant event or change in circumstances has occurred that may have an adverse effect on the fair value of each investment. When such events or changes occur, we evaluate the fair value compared to our cost basis in the investment. We also perform this evaluation every reporting period for each investment for which our cost basis exceeded the fair value in the prior period. The fair values of most of our investments in publicly traded companies are often readily available based on quoted market prices. For investments in nonpublicly traded companies, management's assessment of fair value is based on valuation methodologies including discounted cash flows, estimates of sales proceeds and appraisals, as appropriate. We consider the assumptions that we believe hypothetical marketplace participants would use in evaluating estimated future cash flows when employing the discounted cash flow or estimates of sales proceeds valuation methodologies.
Internal control disclosure	Also, in our opinion, management's assessment, included in Management's Report on Internal Control over Financial Reporting appearing under Item 8, that the Company maintained effective internal control over financial reporting as of December 31, 2004 based on criteria established in Internal Control - Integrated Framework issued by the Committee of Sponsoring Organizations of the Treadway Commission (COSO), is fairly stated, in all material respects, based on those criteria. Furthermore, in our opinion, the Company maintained, in all material respects, effective internal control over financial reporting as of December 31, 2004, based on criteria established in Internal Control Integrated Framework issued by the COSO. The Company's management is responsible for maintaining effective internal control over financial reporting and for its assessment of the effectiveness of internal control over financial reporting. Our responsibility is to express opinions on management's assessment and on the effectiveness of the Company's internal control over financial reporting based on our audit. We conducted our audit of internal control over financial reporting in accordance with the standards of the Public Company Accounting Oversight Board (United States). Those standards require that we plan and perform the audit to obtain reasonable assurance about whether effective internal control over financial reporting was maintained in all material respects. An audit of internal control over financial reporting includes obtaining an understanding of internal control over financial reporting, evaluating management's assessment, testing and evaluating the design and operating effectiveness of internal control, and performing such other procedures as we consider necessary in the circumstances. We believe that our audit provides a reasonable basis for our opinions.
Internal control disclosure (additional example paragraph)	Management, including our CEO and CFO, does not expect that our internal controls will prevent or detect all errors and all fraud. A control system, no matter how well designed and operated, can provide only reasonable, not absolute, assurance that the objectives of the control system are met. Further, the design of a control system must reflect the fact that there are resource constraints, and the benefits of controls must be considered relative to their costs. In addition, any evaluation of the effectiveness of controls is subject to risks that those internal controls may become inadequate in future periods because of changes in business conditions, or that the degree of compliance with the policies or procedures deteriorates.
Risk factor disclosures	The Company's future performance depends to a significant degree upon the continued contributions of its officers and key management, sales and technical personnel, many of whom would be difficult to replace. The loss of any of these individuals could have a material adverse effect on the Company's business, financial condition, results of operations and business prospects. In addition, the Company's future success and ability to manage growth will be dependent upon its ability to hire additional highly skilled employees for a variety of management, engineering, technical and sales and marketing positions. The competition for such personnel is intense, however, and there can be no assurance that the Company will be able to attract, assimilate or retain sufficient qualified personnel to achieve its future business objectives. The failure to do so could have a material adverse effect on the Company's business, financial condition, results of operations and business prospects. See "Risk Factors – Dependence on Key Personnel."
Risk factor disclosures (Additional example paragraph)	In addition, such events could materially adversely affect our reputation with our customers, associates, and vendors, as well as our operations, results of operations, financial condition and liquidity, and could result in litigation against us or the imposition of penalties or liabilities, which may not be covered by our insurance policies. Moreover, a security breach could require us to devote significant management resources to address the problems created by the security breach and to expend significant additional resources to upgrade further the security measures that we employ to guard such important personal information against cyberattacks and other attempts to access such information and could result in a disruption of our operations, particularly our online sales operations.

To see Representative Paragraphs for the remaining topics, as well as additional example paragraphs for these topics, see the Internet Appendix.

Table 6
Numeric counterparts of top 3 increasing topics.

	LnFair value/Impairment (1)		LnInternal control (2)		LnRisk factors (3)	
	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Special items	−1.573***	−11.063				
Material weaknesses			0.634***	31.807		
Risk					11.748***	17.519
Trend	0.475***	192.961	0.429***	220.859	0.252***	94.042
BigN	0.418***	11.238	−0.190***	−7.765	0.381***	9.299
LnAssets	0.214***	20.816	0.052***	6.835	0.155***	10.653
Intangibles	1.408***	16.848	0.019	0.294	−0.106	−0.990
MTB	−0.010***	−3.649	0.015***	7.087	0.006	1.602
Leverage	−0.123	−1.547	−0.706***	−11.958	−0.299***	−3.052
Age	0.001	0.941	−0.000	−0.147	−0.047***	−28.576
BusSeg	−0.010	−1.020	−0.050***	−6.375	−0.077***	−5.850
ForSeg	0.113***	12.572	−0.028***	−4.049	−0.011	−0.901
NYSE	−0.054	−1.594	−0.046**	−1.970	−0.251***	−5.495
Loss	0.035	1.349	−0.361***	−17.886	0.365***	13.235
Observations	75,991		75,991		75,987	
R-squared	0.596		0.579		0.315	
Industry Fixed Effects	Y		Y		Y	

Determinants of logged length for the top three increasing topics, fair value/impairment, internal control, and risk factor disclosures. For each topic of interest, we identify a numeric counterpart that we think captures similar information to the textual disclosure. *Material Weaknesses* is coded 0 before the implementation of SOX in order to preserve the entire sample period; this is appropriate because there was no requirement to identify or disclose internal control weaknesses prior to this period. Assets are inflation-adjusted to be in constant year 1996 dollars. All continuous variables are winsorized at the 1st and 99th percentile by year. Standard errors are adjusted for heteroskedasticity and clustered by firm. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

ing of the associated standard most closely. As discussed later, this standardization is also reflected in the associated textual characteristics.

Our last main topic is risk factor disclosure. The top five words in this topic are "results," "future," "ability," "result," and "adversely." This type of language is consistent with risk factor disclosures that are intended to provide information on future events that might adversely affect firm performance. Disclosure under this topic is relatively broad as reflected in the fact that the representative paragraph describes the loss of key talent and personnel as a risk factor for the firm, while the additional paragraph discusses risks associated with possible security breaches.[25] Although some firms disclosed risk factors voluntarily throughout our sample period, the SEC mandated this disclosure in Item 1A of the 10-K in 2005.

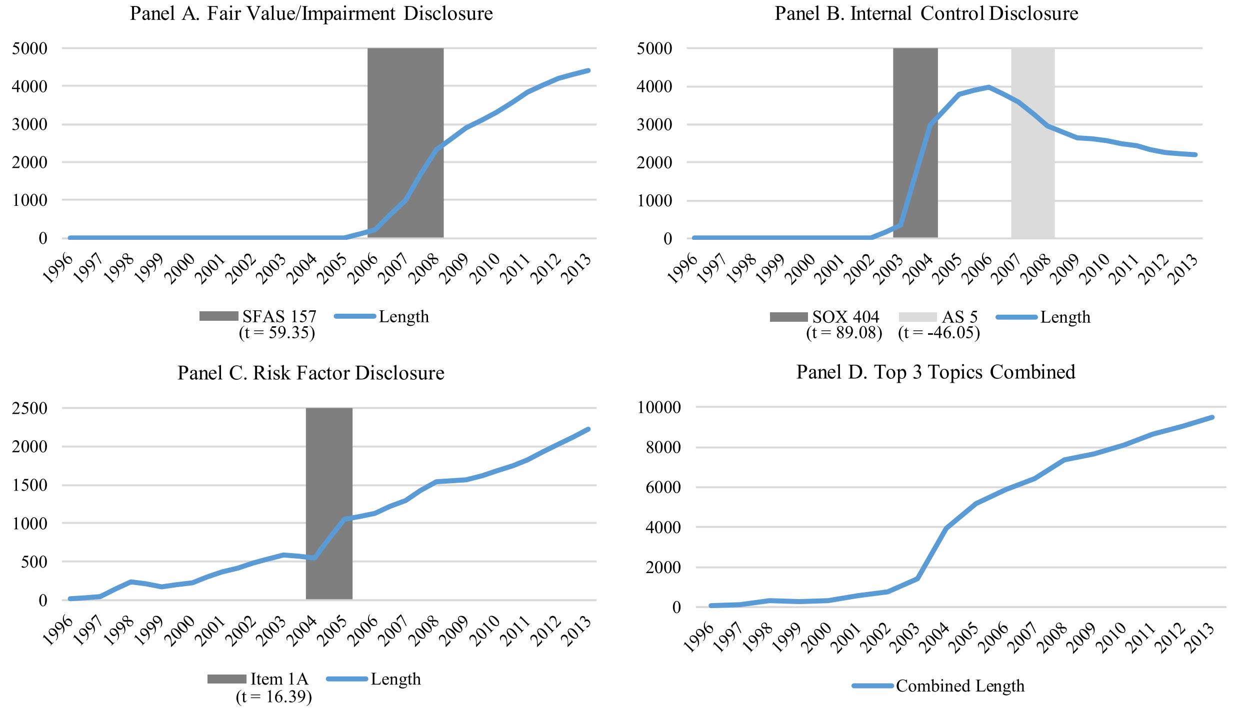
As further support that these three top increasing topics capture the type of disclosure that we have attributed to them, we identify specific firm attributes that should be associated with each of the three topics and link them with the length of these topics. In the case of Internal Controls, we expect significant additional text for firms with internal control weaknesses; for Fair Value/Impairments we expect additional text for firms with substantial one-time items, in particular impairments; and for Risk Factors we expect additional text for firms with substantial market risk.[26] Results in Table 6 indicate that special items (including impairments), internal control weaknesses, and risk all have significant and predictable associations with their relevant disclosure topics.[27] The *Trend* variable remains strong and positive for each topic, suggesting that changes in economic circumstance, as we measure them, do not explain the time trends.

We investigate more closely the timing of these trends in fair value, risk factor, and internal control disclosure to assess patterns around the associated regulatory events and the probability that increases in these topics could explain increases in overall 10-K length. Fig. 4 plots the trends for these topics over time and provides evidence consistent with expectations. Panel A, which plots the length of the Fair Value topic, is interesting for several reasons. First, recall that SFAS 157, "Fair Value Measurements," was passed in 2006 and required for fiscal years beginning after November 15, 2007, with early adoption encouraged. That timeline is very consistent with the path of disclosure around 2006–2008, with virtually no disclosure for that topic pre-2007, an initial substantial increase during 2007 likely reflecting early adopters, and the bulk of the increase during 2008. The fact that the pattern is consistent with expectations is reassuring because it suggests that, while LDA is a naïve Bayesian approach, it can identify discussion associated with specific topics quite crisply irrespective of where it

[25] It is striking that LDA is able to identify risk factor disclosure as relating to the same topic irrespective of the specific nature of the risk—loss of key personnel, cyber-hacking, litigation, sales disruption, water quality, etc.

[26] We use overall return volatility as a measure of risk. Results are consistent if risk is measured based on beta (Campbell et al., 2014) or firm-level litigation risk.

[27] The coefficient on special items is negative, consistent with the notion that special items are generally negative (e.g., losses) and that larger negative special items are associated with lengthier text. Results are consistent using impairment (a subset of special items) or if we replace signed special items with the absolute value (with a significantly positive coefficient).



SFAS 157 was passed in September 2006 and became effective for fiscal reporting periods commencing after November 15, 2007 with early adoption encouraged. SOX 404 became effective for accelerated filers for fiscal years ending on or after June 15, 2004, and for all other firms by 2007. AS 5 was effective for fiscal years ending on or after November 15th, 2007. Item 1A became mandatory for firms filing with the SEC for fiscal periods ending on or after December 1, 2005. Panel D plots the median of the sum of these three topics over time. Heteroskedasticity robust t-statistics clustered by firm are obtained from regressions which regress length on an indicator which takes on a value of one for years after the relevant disclosure requirement was introduced for each topic, controlling for the overall time trend.

Fig. 4. Median length of top 3 increasing topics over time.

appears within a document. This is important because, although LDA has been applied in other contexts, it has not been used to identify text associated with specific accounting rules.

Second, and more importantly, the figure indicates that disclosure around SFAS 157 was a major source of additional length in the typical 10-K. Recall that the median length of the Compliance category in Fig. 3 increased by about 10,000 words; in comparison, the increase in disclosure pertaining to SFAS 157 alone was nearly 4300 words. To assess the statistical significance of the increase, we regress the length of the Fair Value topic on an indicator that takes on a value of one for years after the introduction of SFAS 157, controlling for the overall trend. Fig. 4 reports the heteroscedasticity robust t-statistic clustered by firm in the legend below Panel A (analogous t-statistics are reported for the regulatory changes in the other panels). The increase in length associated with SFAS 157 is highly significant (t-statistic = 59.35). It is also interesting to note that this increase does not appear to have been temporary. The text associated with this topic leveled out to some extent after the 2008 mandatory adoption date but continued to rise, albeit more gradually, through 2013, suggesting that additional disclosure was necessitated with application of the standard (and related guidance) over time.[28]

The second largest increase is disclosure related to internal controls. Recall that SOX internal control certifications were required for fiscal years starting in 2004 and 2005. Panel B shows a distinct increase for the LDA topic we label Internal Controls between 2004 and 2005, leveling off in 2006, suggesting that LDA correctly identified internal control disclosure. More importantly, Fig. 4 suggests that internal control discussion is an important determinant of the increase in 10-K length, especially between 2004 and 2006. Unlike fair value disclosure which continued to increase, text associated with internal controls dropped somewhat between 2007 and 2008 before leveling off at about 2100 words, down from a high of 3900 words in 2006. This drop coincides with the introduction of Auditing Standard 5 (AS5) by the PCAOB for fiscal years ending on or after November 15, 2007. Among other changes, AS 5 allows auditors to issue a combined report on both the financial statements and the internal controls over financial reporting whereas previously auditors were required to issue two

[28] While the sharp increase in text around 2007 seems clearly related to the implementation of the new standard, the continued increase could reflect implementation guidance or evolving economic circumstances. The fact that the trends are robust to controls for underlying economics lends credence to the implementation guidance explanation.

separate reports.[29] The increase in length around adoption of SFAS 157 is highly significant (t-statistic = 89.08) as is the decrease around AS 5 (t-statistic = −46.05).

The third major source of the increased length is forward-looking disclosure associated with risk factors, depicted in Panel C. While not specifically required in the 10-K prior to 2005 (although required in prospectuses for debt and equity offerings), firms often provided risk factor disclosures voluntarily when they made forward-looking statements (Campbell et al., 2014; Nelson and Pritchard, 2016). Beginning in 2005, the SEC emphasized the importance of adequate risk factor disclosures and required that they be discussed in a separate section of the 10-K (Item 1A). As a result, we expect an increase in the discussion of risk factors throughout our sample period as SEC interest increased, but with a substantial increase around 2005 when the new rules became effective. The graph for the risk factor topic displays the predicted pattern, with a gradual increase through 2004 followed by a substantial jump in 2005 and a more gradual increase subsequent to 2005. The increase in length around adoption of Item 1A is highly significant (t-statistic = 16.39). Similar to fair values, the increase in disclosure around the effective date does not appear to have been temporary, with an increasing subsequent trend likely reflecting the SEC's continuing focus on implementation along with evolving economic circumstances. By 2013, median risk factor disclosure had increased by almost 2300 words.

Fig. 4 Panel D displays the sum of the three topics over time, which combine to explain an increase of almost 10,000 words. Further, there is a close similarity between the increase in Compliance disclosure from Fig. 3 and the sum of the three components in Fig. 4 Panel D suggesting that those three factors explain the bulk of the increase in Compliance disclosure (which, in turn, explains most of the increase in total 10-K length).

Overall, the additional detail that our LDA analysis provides allows us to dig more deeply into the causes and content of the additional length in 10-K disclosure than would be possible with an analysis at the document- or section-level. Perhaps most notable is the extent to which, despite the number of additional SEC and FASB requirements during our sample period, the bulk of the increase in textual disclosure relates to three topics, two under the purview of the SEC and one under the purview of the FASB.[30]

6. Do changes in topical disclosure length reflect disclosure requirements?

A potential issue with the preceding results is that it is not possible to observe what firms would have disclosed in the absence of the new requirements. For example, it is possible that disclosure of risk factors, fair values, and internal controls would have increased irrespective of the requirements and that the new standards simply reflect changing demands for information. The patterns in Fig. 4 provide some evidence on that point because the timing of the changes in disclosure coincides quite tightly with the new requirements. In addition, the regressions in Table 2 suggest that the trends are robust to a wide variety of economic, regulatory, and litigation-related controls suggesting that general economic trends do not explain the increased length.[31] However, it is possible that the relation between our textual attributes and economic determinants is not stable over time or that our analysis excludes important variables.

An alternate approach is to consider a comparison sample of non-U.S. firms that were not subject to the same regulatory changes as our primary sample. Although non-U.S. firms experienced some mandatory changes in fair value disclosure because the IASB issued IAS 39 on fair values around the same time that the FASB issued SFAS 157, they were not subject to the SEC risk factor and internal control requirements. As a result, we would not expect to see the same pattern in risk factor or internal control disclosure for a non-U.S. sample if new SEC reporting requirements, and not changes to fundamentals, drive our results.

To investigate that possibility, we analyze annual reports for a sample of 16,038 non-U.S. firm-years from Lang and Stice-Lawrence (2015) and examine trends in disclosure using our trained LDA model. Untabulated results suggest there is virtually no disclosure on the internal control or risk factor topics either prior to or subsequent to the change in U.S. regulations and no evidence of an increase over time. Consistent with the timing of IAS 39, disclosure on the fair value topic increases for the non-U.S. firms, especially during the 2006–2008 period, although the increase is not as large as for U.S. firms. Subject to the caveat that the non-U.S. sample may not be entirely comparable with the U.S. sample, these results suggest that the disclosure changes for the U.S. sample, particularly those relating to risk factors and internal controls, primarily reflect changes in regulatory requirements.

7. Does the additional text contribute to trends in the other textual attributes?

Having documented that much of the increase in 10-K length appears to be a result of increases in specific topics associated with accounting and other regulatory action during the mid-2000 s, we next investigate the textual characteristics

[29] We observe a similar decrease when examining only the subset of firms that never reported an internal control weakness, suggesting that a higher incidence of firms with internal control weaknesses around initial implementation of SOX is not the sole driver of this peak and that firms without internal control weaknesses also experienced an initial increase, and subsequent decrease, in their discussion of the topic.

[30] In the 1996–2009 period (up until the codification), for example, the FASB issued forty-four pronouncements (not including amendments), yet only SFAS 157 appears to explain the bulk of the increase in text length.

[31] While Table 2 is estimated across all topics, Table 4 replicates the analysis at the specific topic level and yields very similar conclusions, suggesting that economic trends are less likely to explain the trends in disclosure.

Table 7
Topic-level textual characteristics.

	Redundancy Median	Boiler % Median	Stickiness Median	Specificity Median	HardInfoMix Median	Fog Median
Top 3 Increasing Topics						
Risk factors	0.6%	29.4%	85.9%	13.75	1.27	21.88
Internal control	12.9%	86.9%	90.0%	26.42	4.30	26.56
Fair Value/impairment	2.7%	36.1%	84.0%	18.70	6.15	21.45
Categories						
Compliance	9.3%	48.2%	74.1%	39.61	15.12	22.86
Other disclosure	4.9%	23.4%	64.8%	54.14	18.92	20.84
Contracts & legal	7.5%	34.4%	58.1%	38.70	7.97	21.04
Business Op. & strategy	2.0%	14.1%	75.7%	47.05	11.45	21.38
Business Struct. & M&A	2.9%	26.6%	63.6%	65.17	22.92	21.09
Empoyees & executives	1.1%	27.2%	56.1%	58.20	14.03	21.25
Geographic location	0.0%	9.8%	62.4%	134.33	24.10	19.92
Intellectual Prop. & R&D	0.0%	15.0%	81.5%	34.72	9.35	21.02
Investments, Sec. & Deriv.	0.9%	21.9%	79.5%	41.13	16.29	20.23
Loans, debt, banking	4.1%	19.4%	64.9%	68.00	30.86	20.67
Performance, Rev. & Cust.	3.6%	27.6%	69.6%	66.35	26.78	20.33
Property & leasing	0.0%	16.6%	71.7%	54.64	22.86	19.57
Stock & options	2.0%	43.5%	68.7%	61.78	31.87	18.67
Industry-specific	2.7%	10.6%	75.4%	50.53	12.94	20.70

Descriptive statistics for textual characteristics measured at the topic- and category-level. *Other disclosure* is the combination of all disclosure categories other than *Compliance*. Variable definitions given in the Appendix.

of this additional disclosure. We focus on characteristics identified as potentially important by prior research and investigate the extent to which our Top 3 topics explain the trends documented in Fig. 1. An advantage of LDA is that we can apply it at the paragraph level to evaluate textual characteristics within subsets of text. In particular, we use our trained LDA model to re-analyze each paragraph in the 10-K and estimate paragraph-level topic loadings (essentially the probability that the paragraph belongs to a specific topic) in a process called "inferencing." We then assign the paragraph to the topic which has the highest loading.[32] This allows us to assign all paragraphs to individual topics and thus measure the textual characteristics of disclosure relating to that topic.

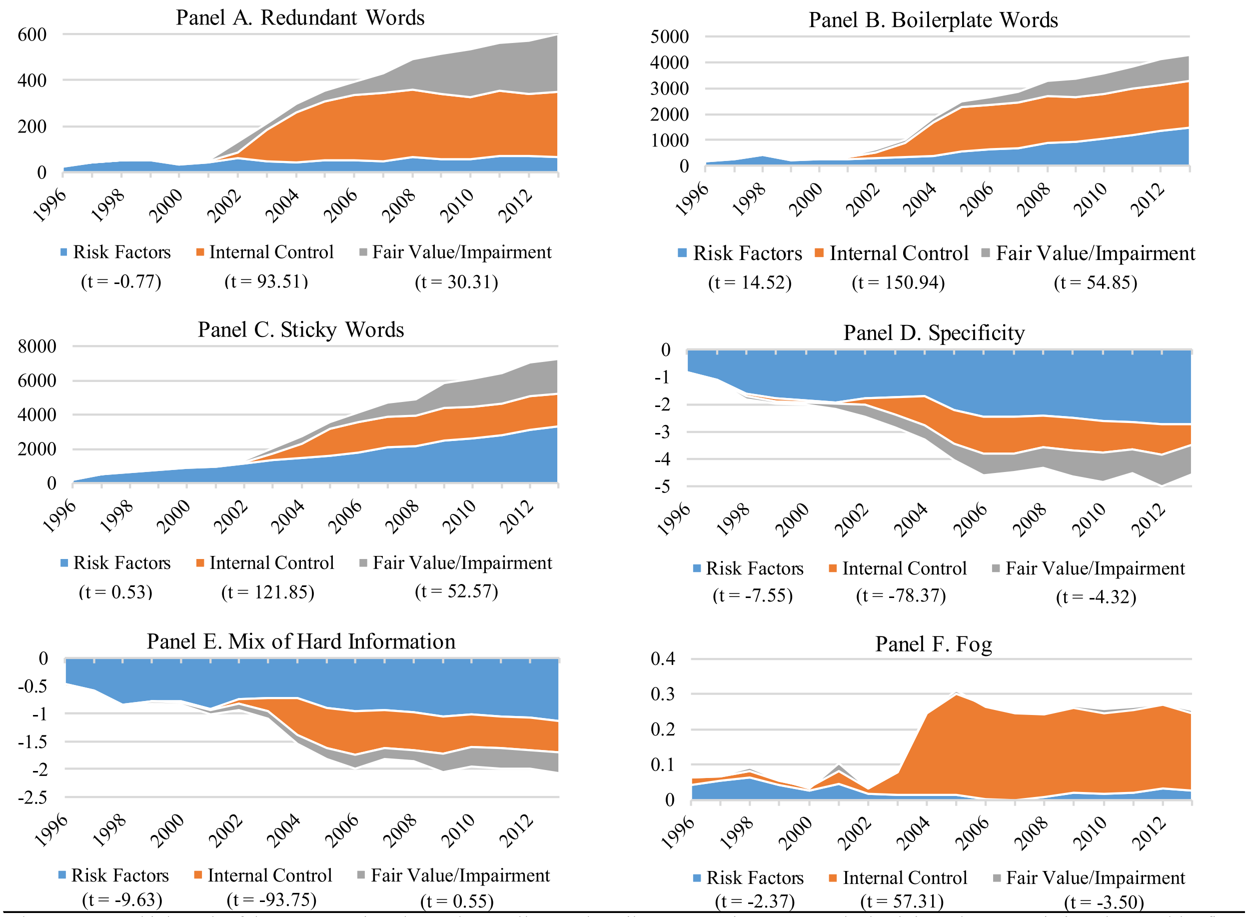
Table 7 provides descriptive statistics for each of the categories (as well as for our Top 3 increasing topics) aggregated across all paragraphs within each category (topic). Redundancy, boilerplate, and stickiness are expressed in percentage terms so that the relative redundancy, boilerplate, and stickiness of disclosure can be directly compared across categories (topics) of different lengths.[33] The broad category statistics in Table 7 are generally consistent with expectations.[34] Comparing the Compliance category to all other categories ("Other Disclosure Categories" in Table 7), Compliance has substantially higher levels of boilerplate, Fog, stickiness, and redundancy and lower levels of hard information and specificity. This fact, coupled with the earlier finding that the proportion of the 10-K devoted to Compliance has increased substantially over time, suggests that the overall increase in these attributes could be the result of increases in the proportion of the 10-K representing Compliance disclosure.

In terms of the Top 3 increasing topics, the descriptive statistics suggest that internal control disclosures tend to have high levels of redundancy, Fog, boilerplate, and stickiness, in most cases higher than for any category, including other Compliance disclosure. This is not altogether surprising because SOX requirements are fairly specific in terms of disclosure, leading to high levels of boilerplate. Risk factor disclosures also tend to have high Fog and stickiness, and fair value disclosure has high stickiness. All three topics tend to have relatively low levels of specificity and the mix of hard information. The attributes of risk factor and fair value disclosures are noteworthy because firms have more flexibility and these disclosures are intended to convey timely firm-specific information. Given the textual attributes of the Compliance category, and the Top 3 topics in particular, their increasing prevalence in the 10-K could help to explain the overall trends in disclosure characteristics documented earlier.

[32] This procedure introduces noise because a given paragraph may discuss multiple topics, which would cause the textual characteristics for paragraphs assigned to a given topic to revert to the mean. We do not exclude paragraphs that include multiple topics so that we can aggregate our statistics to the document level, but find similar (if not stronger) results when we instead impose a cutoff loading (probability) of 0.5 to assign paragraphs to specific topics.

[33] To reduce noise, descriptive statistics for each topic (category) are only calculated for documents which have at least 100 words in paragraphs assigned to that topic (category). We calculated, but for parsimony do not tabulate, the proportion of negative, uncertain, and litigious words in paragraphs relating to each topic and category based on the Loughran and McDonald business words dictionary. These measures also behaved according to expectations for our categories and Top 3 topics (e.g., risk factor disclosure contains a relatively high amount of negative and uncertainty words, the Contracts & Legal category tends to use litigious words, etc.), providing additional assurance that our paragraph-level approach is effective at identifying paragraphs relating to particular topics.

[34] For example, the Contracts & Legal category has a high mean level of redundancy and low levels of quantitative data, consistent with redundant legal language and few numbers, and the Intellectual Property & R&D category is the least specific, consistent with firms choosing to give vague descriptions of R&D to decrease proprietary costs.



The extent to which each of the Top 3 topics changed overall textual attributes over time. Heteroskedasticity robust t-statistics clustered by firm are obtained from regressions which regress the overall attribute on an indicator which takes on a value of one for years after the relevant disclosure requirement was introduced for each topic, controlling for the overall time trend.

Fig. 5. Contribution of the top 3 topics to textual attributes over time.

In Fig. 5, we investigate the extent to which increases in fair value, risk factor, and internal control disclosure contribute to trends in other textual attributes. Recall from Fig. 1 that Fog, boilerplate, redundancy, and stickiness have increased over time while specificity and the relative mix of hard information have decreased. Because we can identify specific portions of the text relating to each topic, we can compare the levels of each of our textual attributes calculated using the entire text to the attributes of the remaining text after paragraphs relating to the Top 3 topics are removed. Essentially this allows us to compare the actual trends to trends in the "counterfactual" text if those three topics had not been included. We plot the difference between attributes calculated with and without our Top 3 topics to show the contribution of each of the topics to trends in our set of textual attributes. To assess statistical significance, we report under each panel and textual attribute label in Fig. 5 the t-statistic associated with the change in that attribute around the implementation of the associated regulation.[35]

Each of our Top 3 topics helps to explain the trends in disclosure attributes over our sample period. Internal control and fair value disclosures are more important drivers of redundancy, boilerplate and stickiness. The increase in Fog, on the other hand, is largely explained by increases in internal control disclosure, consistent with the very high level of Fog for that disclosure in Table 7, particularly after the introduction of SOX in 2004.[36] Similarly, all three topics are associated with the decline in specificity, while risk factors and internal controls help explain the decrease in the mix of hard information documented in Fig. 1.

Overall, the results from this analysis provide strong evidence that the increase in length in the Top 3 topics also helps explain the patterns in redundancy, boilerplate, stickiness, Fog, specificity and the mix of hard information. Further, the

[35] In particular, we regress the attribute on a trend variable and an indicator that takes on a value of one for years after the requirement was introduced, clustering by firm. We report the t-statistic from the indicator variable.

[36] As an example, the representative paragraph relating to internal control disclosure in Table 5 has a Fog score of 29.2. The results in Fig. 5 suggest that risk factor and fair value disclosures decreased overall Fog slightly, consistent with those topics having low levels of Fog relative to other Compliance-related disclosure.

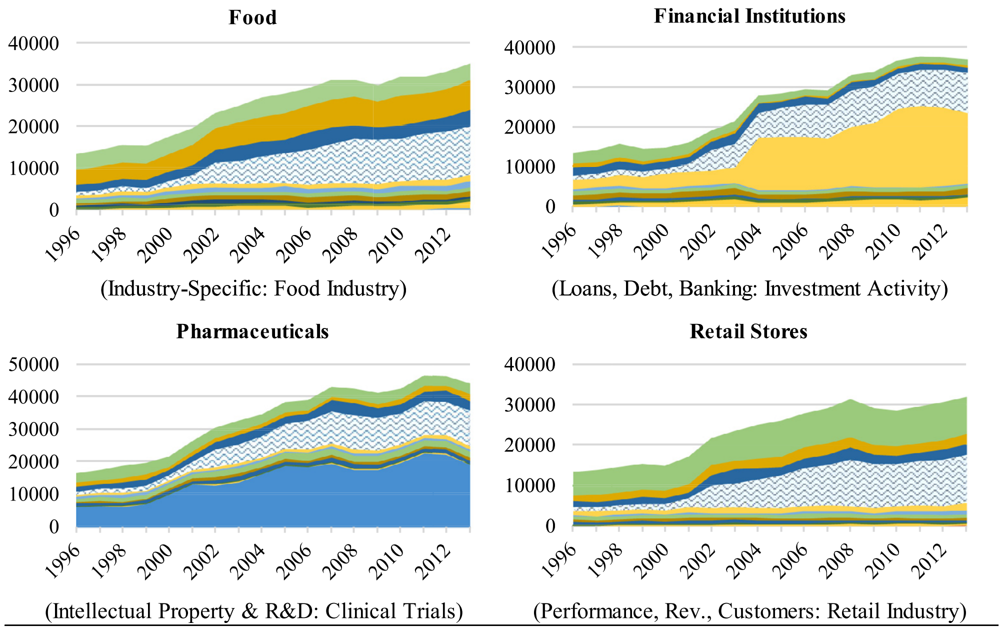


Figure 3, Panel A plotted separately for firms in 4 of the Fama-French 17 Industry Grouping industries. The most prominent category (other than Compliance) for each industry and the most prominent topic within that category are listed beneath each figure.

Fig. 6. LDA topic categories over time for 4 example industries.

effects are consistent with the topic-level descriptive statistics from Table 7, and with the trends by topic in Fig. 4. The importance of each of the Top 3 topics varies across textual attributes, highlighting the importance of evaluating the effects of a given topic in terms of specific attributes.

8. Cross-Sectional variation

The preceding analysis allows us to identify the sources of aggregate changes in disclosure length and link these changes to changes in other textual attributes. However, if there is heterogeneity in the extent to which new disclosure requirements affect firms, then focusing on changes for the median firm masks the potential spectrum of outcomes across firms. For example, if risk plays a large role in determining how firms respond to risk factor disclosure requirements (e.g., high-risk firms provide substantially more risk factor disclosure), then we might find that the median change in disclosure that we document is not representative of disclosure changes for high- and/or low-risk firms. Similarly, cross-industry variation in fundamentals could affect the extent to which disclosure regulations are relevant.

In Fig. 6 we plot trends in median disclosure length over time at the category level (similar to Fig. 3) for firms in 4 of the Fama-French 17-industry groupings (Food, Financial Institutions, Pharmaceuticals, and Retail Stores), chosen to illustrate the range of disclosure across industries with very different underlying economics (graphs for the remaining 13 industries are in the Internet Appendix). Two primary points are worth noting.

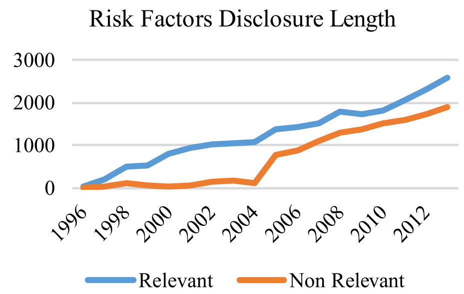
First, as expected, there are substantial differences in disclosure across firms with different underlying industry fundamentals. To illustrate the variety of disclosure by industry, below each panel in Fig. 6 we note the largest category, other than Compliance, and corresponding topic for the industry. For example, Food is a very representative industry, with proportions of disclosure in the major categories similar to the sample as a whole from Fig. 3. The major source of non-Compliance disclosure is in the Industry-Specific category, specifically in the Food Industry topic. Retail Trade is generally quite similar to the overall sample, but with greater discussion in the Performance, Revenues and Customers category relating specifically to the Retail Industry topic. Pharmaceuticals are notable for the extensive discussion in the Intellectual Property and R&D category, particularly related to Clinical Trials. Financial Institutions have a greater proportion of disclosure in the Loans, Debt and Banking category, particularly related to Investment Activity and, not surprisingly, disclosure in that category increased substantially around the financial crisis. While these results are not surprising, they illustrate that there is substantial heterogeneity in disclosure driven by underlying economics and that LDA is effective in capturing that heterogeneity.

Second, and more importantly, each of the industries in Fig. 6 (as well as those in the Internet Appendix) experience substantial increases in 10-K length over the sample period. Further, for all industries, a substantial portion of the increase in length is due to increases in Compliance-related disclosure and the magnitude of the increase in Compliance-related disclosure is consistently about 9000–10,000 words, similar to the sample as a whole. Overall, the increase in Compliance-related length appears to have been pervasive across firms with very different underlying economics. Results at the topic-level for the Top 3 Topics provide similar inferences and are provided in the Internet Appendix.

The fact that we observe similar trends across a range of industries suggests that our results are not limited to a sub-sample of firms, which is interesting for several reasons. First, it further confirms that the patterns we document are likely the result of changes in regulatory requirements and do not simply reflect the underlying economics of firms (i.e., it seems unlikely that firms across industry groupings would all simultaneously experience similar economic shocks). Second, the similarity in trends across a broad spectrum of firms suggests that, although the disclosure requirements were more likely to be relevant for some types of firms than others, they resulted in substantial increases in disclosure length across a wide cross-section of firms including those for which the additional disclosure may have been less relevant.

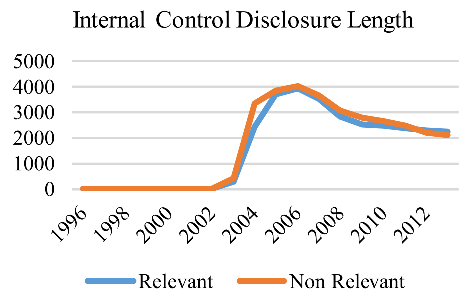
To further explore that possibility, we compare disclosure attributes of our Top 3 Topics for subsets of firms for which each of these disclosures was ex ante likely to be more and less relevant. In particular, we compare risk factor disclosure for firms in the top and bottom total risk quintiles, internal control disclosure for the top and bottom internal control

Panel A. 10-K Attributes of Firms for which Risk Factor Disclosures are Most/Least Relevant



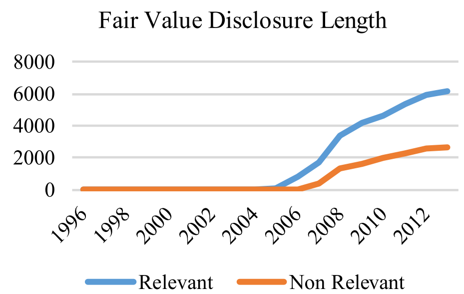
Risk Factor Disclosure Attributes	**t-value**
Difference (low relevance - high relevance)	
Redundancy	2.65
Boilerplate	23.48
Stickiness	12.69
Fog	20.97
Specificity	-11.18
Number of Numbers	-6.76

Panel B. 10-K Attributes of Firms for which Internal Control Disclosures are Most/Least Relevant



Internal Control Disclosure Attributes	**t-value**
Difference (low relevance - high relevance)	
Redundancy	41.05
Boilerplate	77.45
Stickiness	95.38
Fog	94.48
Specificity	33.12
Number of Numbers	-43.10

Panel C. 10-K Attributes of Firms for which Fair Value Disclosures are Most/Least Relevant



Fair Value Disclosure Attributes	**t-value**
Difference (low relevance - high relevance)	
Redundancy	-18.50
Boilerplate	-15.57
Stickiness	8.62
Fog	12.33
Specificity	-1.50
Number of Numbers	-11.09

Attributes of risk factor, internal control, and fair value disclosure of firms for which these disclosures are most and least relevant based on total risk (stock volatility), internal control risk (earnings announcement lag), and intangibles, respectively. Each panel displays disclosure length for that topic for low and high relevance firms, and also presents the difference between the other disclosure attributes of low and high relevance firms.

Fig. 7. Disclosure attributes of firms for which the top 3 topics are most/least relevant.

risk quintiles,[37] and fair value disclosure for the top and bottom intangible asset quintiles. In Fig. 7, we plot the trends in disclosure length for the top and bottom quintiles and report t-statistics comparing redundancy, boilerplate, stickiness, fog, specificity and mix of hard information across the two groups. Overall, the results suggest that (particularly for risk factors and internal controls) in both extreme quintiles firms tended to substantially increase disclosure in response to the new regulations, but for those for whom the new standards were potentially less relevant (e.g., risk factor disclosure for low-risk firms), disclosure tended to be more redundant, boilerplate, sticky, and foggy, with lower levels of hard information.

In particular, results in Fig. 7 Panel A indicate that firms in the top risk quintile had substantial risk disclosure before the introduction of Item 1A in 2005 (consistent with Nelson and Pritchard, 2016) and increased their risk factor disclosure relatively less in response to the new requirements. Low-risk firms, on the other hand, increased their disclosure substantially in response to the new requirements. It appears that, when risk disclosure was voluntary, firms for which it was more relevant were more likely to disclose, while the primary increase associated with Item 1A was for firms for which it was less likely to be relevant. In terms of the six textual attributes, Fig. 7 suggest that firms for which the risk factors were less of an issue responded by providing disclosure with high levels of boilerplate, stickiness, and redundancy, and low levels of specificity, readability, and hard information compared with the high-risk firms.

Fig. 7 Panel B suggests that firms with high and low internal control risk experienced almost identical time series trends in internal control disclosure. As with risk factor disclosure, the increase for low internal control risk firms was characterized by high levels of boilerplate, redundancy, stickiness and Fog and less hard information (although specificity increased). In contrast to internal control and risk factor disclosure, Panel C suggests that the results for fair value disclosure were more nuanced. Fair value disclosure for high relevance (high intangibles) firms increased by more than double that of low relevance firms and other textual attributes do not appear to differ consistently between the two sets of firms, suggesting that increases in fair value disclosure tended to be more customized to the specific circumstances of the firm.

Overall, these results suggest that firms for which risk factor and internal control issues were less salient responded to the new requirements with disclosure that was high in Fog, boilerplate, redundancy, and stickiness, and lacking in hard information. Fair value disclosure, on the other hand, tended to be more customized to the specific circumstances of the firm. While we cannot make normative statements, the results suggest implementation variation across disclosure requirements in terms of customization. To the extent that standardization affects disclosure informativeness, it may be worth investigating ways to encourage more customized disclosure (or potentially establishing robust materiality cutoffs to limit disclosure for firms for which it may be less relevant).

9. Conclusions and implications for future research

We provide initial empirical evidence quantifying trends in disclosure topics and attributes. First, we find that the length of 10-K disclosure has increased dramatically and that this trend cannot be explained by changes in observable firm-level characteristics from the prior literature. Similarly, boilerplate, Fog, redundancy, and stickiness of disclosure have increased substantially while specificity and the mix of hard information have decreased. Second, LDA-based analysis indicates that the majority of the increase in length is the result of disclosure associated with three new SEC and FASB requirements: fair value disclosure associated with SFAS 157, internal control disclosure under SOX, and risk factor disclosure mandated by the SEC in Item 1A. Increases in these three disclosures also help explain increases in boilerplate, redundancy, and stickiness, as well as decreases in readability, specificity, and the relative mix of hard information. Finally, firms for which the new risk factor and internal control rules were likely to be less relevant responded with disclosure that is characterized by higher levels of Fog, boilerplate, stickiness, and redundancy, and lower levels of hard information.

While our analysis is subject to caveats, particularly because we do not test the informativeness of the textual attributes we measure, we believe our findings have the potential to contribute to future research on topics such as disclosure attributes and the effects of regulation. First, there is scope for more focused research on the causes and, in particular, the consequences of the dramatic increase in overall length of the 10-K, and associated increases in redundancy, Fog, boilerplate, and stickiness, and decreases in specificity and hard information we document. Second, our results suggest a more nuanced approach to help isolate sources of disclosure attributes. For example, while all three of the main topics we identify are important in explaining the overall trend in disclosure attributes, fair value disclosure is an important driver of increases in redundancy and stickiness; risk factor disclosure is an important driver of decreases in specificity and hard information; and internal control disclosure is a primary driver of increases in Fog and boilerplate.

Finally, our results for risk factors and internal controls suggest scope for research examining potential approaches for targeting disclosure requirements that might be differentially relevant across firms depending on their underlying economic circumstances. Could, for example, requirements be tailored to individual settings through better application of materiality standards? Similarly, there may be a natural role for technology in dealing with disclosure that tends to be particularly redundant, sticky, or boilerplate through use of hyperlinks between sections in the 10-K (e.g., material that is repeated in MD&A and the footnotes), between years (e.g., risk factors that remain constant over time), or across firms (e.g., boilerplate that is shared across an industry). This could allow investors to more easily identify key changes in disclosure, either relative

[37] Following research on internal information quality such as Gallemore and Labro (2015), we measure internal control risk using the length of the lag between the fiscal year end and earnings announcement, consistent with firms with better internal control quality (lower risk) having a shorter lag.

to other firms in the industry or relative to the same firm over time, while ensuring ready access to the redundant text for investors who wish to access it.

Given the difficulty in addressing some of these issues in a purely archival setting, there is a natural role for laboratory or survey research on the informativeness of alternative disclosure approaches once topics and specific passages of text have been identified that are, for example, redundant, complex, and lacking in specificity, to better understand how these attributes affect users of financial information in specific contexts. For instance, it is ultimately an empirical question whether standardized disclosure in a given context is preferable (e.g., because it makes it easier to identify anomalous disclosure) or whether customized disclosure is more informative. Similarly, an attribute such as boilerplate in internal control disclosure might be beneficial to unsophisticated investors but not to sophisticated investors. More broadly, our results suggest that techniques such as LDA have promise in allowing researchers to further open the "black box" of textual disclosure and objectively categorize the underlying content in a manner that can be efficiently applied to large numbers of lengthy documents.

Appendix

Variable definitions

Textual Variables	
Variable	Description
Words	The number of words used in the 10-K.
Boiler words	The number of words in sentences that include at least one 4-word phrase that is shared by at least 75% of all firms in a given fiscal year (similar to Lang and Stice-Lawrence, 2015).
Boiler%	The percent of boilerplate words in a given portion of text.
Fog	The Gunning (1952) Fog index, where Fog = 0.4*(average number of words per sentence + percent of complex words), where complex words are the words in excess of two syllables.
HardInfoMix	The number of informative numbers (e.g., omitting dates, section numbers, etc.) in disclosure text identified using the method in Blankespoor (2016), scaled by the total number of words. This ratio is then multiplied by 1000.
Redundant words	The number of words in sentences that are repeated verbatim in other portions of the 10-K.
Redundancy	The percent of redundant words in a given portion of text.
Specificity	The number of entities (locations, people, organizations, dollar amounts, percentages, dates, or times) identified by the Stanford Named Entity Recognizer (NER) tool, scaled by the total number of words (see Hope et al., 2016, for more details). This ratio is then multiplied by 1000.
Sticky Words	The number of words in sentences that include at least one 8-word phrase that is identical to a phrase used in the prior year's 10-K.
Stickiness	The percent of sticky words in a given portion of text.
Fair Value/Impairment Loading (Length)	The loading (length) of the fair value/impairment topic identified by the LDA model. (The length is calculated by multiplying the loading by the total length of the text.)
Internal Control Loadings (Length)	The loading (length) of the internal control topic identified by the LDA model. (The length is calculated by multiplying the loading by the total length of the text.)
Risk factor disclosures loading (Length)	The loading (length) of the risk factor disclosures topic identified by the LDA model. (The length is calculated by multiplying the loading by the total length of the text.)

Other Variables	
Variable	Description
Age	Age is calculated as the number of years since the first year a firm was listed in Compustat.
BigN	Takes the value of 1 for the following five audit firms: Arthur Andersen, Ernst and Young, Pricewaterhousecoopers, and KPMG. Missing values are set to 0.
BusSeg	The number of business segments [BUSSEG]. If missing, 0.
ForSeg	The number of foreign segments [GEOSEG]. If missing, 0.
Intangibles	The percentage of assets classified as intangible [INTAN]/[*AT*]. If missing, 0.
Leverage	The long-term and current period debt scaled by total assets ([*DLTT*] + [*DLC*])/[*AT*].
LnAssets	The natural logarithm of total assets [AT].
Loss	An indicator variable that takes on the value of 1 when net income [NI] is below 0, and 1 otherwise.
Material Weaknesses	The number of material weaknesses in internal controls that a firm reports for that period.
MTB	The market value of equity [CSHO]*[PRCC_F] divided by the book value of equity [CEQ].
NYSE	Takes on a value of 1 if the firm is listed on either NYSE (EXCHCD = 1) or AMEX (EXCHCD = 2), and 0 otherwise.
Risk	The standard deviation of daily returns during the fiscal year.
Special Items	Special items [SPI], scaled by total assets [AT]. If missing, 0.
Trend	A variable that increments by 1 each year, starting at 0 in the fiscal year 1996.

Sample restrictions

Before analyzing any 10-K filings, we exclude all amended and small business (Form 10-KSB) filings from our sample, as well as those filed before June 1, 1996 (when electronic filing was still voluntary) and Form 10-K405s. We also exclude documents containing fewer than 3000 words and those that are missing basic data from Compustat (assets, net income, shares outstanding, price, and book value of equity).

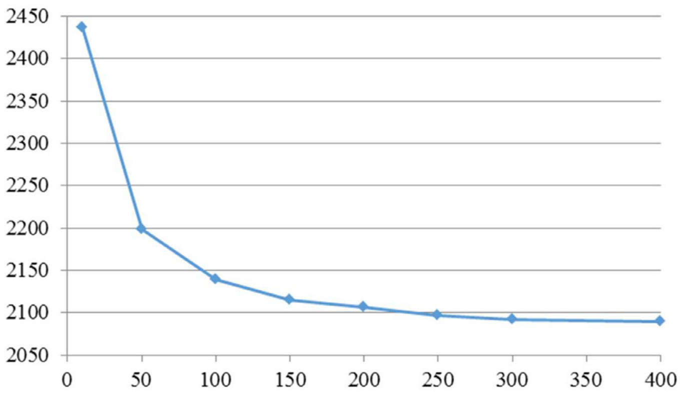


Fig. A.1. Perplexity by number of topics.

10-K Cleaning Procedures

We use Perl to remove all HTML and non-relevant text from the 10-K filings in our sample using procedures similar to those used in Li (2008). First, we remove all header and appendix information, including the SEC header section at the start of all 10-K documents, as well as any graphics, zip files, xml files, excel files, 101 exhibits, 100 exhibits, pdf files, and XBRL. Second, we remove all HTML text from the file using the HTML:: Parser Perl module. Any remaining tags such as ⟨TEXT⟩, ⟨PAGE⟩, ⟨DOCUMENT⟩, and ⟨TYPE⟩ are removed following Miller (2010). We also delete lines with ⟨S⟩ and ⟨C⟩ following Miller (2010). Third, we implement character restrictions to the document. We delete lines with fewer than 20 characters or 15 alphanumeric characters, which removes lines of just numbers as well as section headings. Following Li (2008) we further delete paragraphs with more than 50 percent non-alphabetic characters. Additionally, we remove paragraphs with fewer than 80 characters (Blankespoore, 2016).

Identifying 10-K item sections

In order to uniquely identify each of the item sections within the 10-K, we implement the first two steps in the 10-K cleaning procedure described above and then generate unique identifiers for all instances where a reference to an item section was used in the 10-K document.[38] This identifier tracks the sequence in which the references to the section were used. In order to identify which reference is the true starting location of the item section, we iteratively remove section references that are inconsistent with logical ordering of the section numbers. In this iterative process, we take the last full sequence, if multiple sequences exist, which removes tables of content. If two references are referenced only once and are neighboring sections, we remove references between. If an identified reference does not have the necessary sections that follow or precede it (e.g. Section 7 is not followed by Section 8, or preceded by Section 6), then it is removed. For those documents where this iterative process is reduced to a unique sequence of all of the required sections, we break apart the 10-K at the locations of each section reference, and then perform the final step of the 10-K cleaning procedure on each section separately. Finally, we impose minimum word limits for some sections, to ensure that we are capturing the actual section and not just a reference to the section. For Sections 1, 7 and 8 this threshold level is 50 words. For sections 10, 11, and 12 the level is 20 words. The length of the remaining sections was subject to too much natural variability for us to determine a reasonable cutoff. This process allows us to identify all of the sections in 22,349 10-K filings.

Perplexity

The formula for perplexity from Blei et al. (2003) is:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

It is a function of the per-word likelihood, $p(w_d)$, and the number of words in each document, N_d. Perplexity decreases as the likelihood of the model increases, or in other words when the statistical fit is better.

In order to calculate the perplexity plotted below, we trained the model on 90% of our data and then calculated the perplexity using a random hold-out sample of the remaining 10% of the observations. (Fig. A.1)

[38] We do not separately identify Section 2 or any section beyond 14 as Section 2 was often combined with Section 1 and those beyond 14 were not consistent throughout our sample period.

Word intrusion task

In order to identify the topic model with the best fit paired with high interpretability, we perform a word intrusion task for LDA models estimated using 150, 200, and 250 topics. We choose these three models because the incremental decrease in perplexity after 150 topics is relatively small whereas there are obvious gains to model fit for less than 150 topics.

The word intrusion task is structured as follows. A human coder is presented with a set of 6 words in a random order. Five of the words are the words with the highest probability of appearing in Topic X according to the model, whereas the sixth word has a low probability of occurring in Topic X.[39] Participants in the task are asked to choose the word which does not match the other five words, or in other words the "intruder" word. For example, the set of 6 words could be: debt, loan, facility, term, inventory, revolving. In this case, the intruder word is "inventory," as the rest of the topic is about debt. These groups of six words were generated for each potential topic in each of the three potential models and presented (unlabeled) to the coders in a random order. Our two human coders reviewed each group of words and chose an intruder word.[40] The relevant statistic is the percent of the time the human coders agreed with the model, where high agreement indicates high interpretability of the topics. For both coders, the model with the highest interpretability was the 150-topic model so we use this in all of our subsequent tests.[41]

Word constraints in the LDA procedure

We place a few constraints on the documents that we use when estimating our LDA model. We first remove all common stopwords such as "is," "the," and "and" as these are not useful in classifying topics and decrease performance, and all words that do not occur in at least 100 documents. Additionally, certain words that are extremely common in firm annual reports (such as "company" and "value") are so common that they prevent the model from estimating. All words that occur in every document or are in the top 0.1% most common words are excluded. These words are listed below:

company will value information years upon company's fiscal rate based report sales management services form costs related tax ended certain market credit products amount period net including operations securities cash time statements income section common assets shares business plan year date interest december agreement stock may financial million shall

Paragraph-level analysis

Measuring textual attributes at the topic level

We measure textual attributes at the topic level by first breaking each document into paragraphs.[42] We then use our trained topic model to "infer" topics at the paragraph level; essentially this takes the probabilities per word that we calculated at the document level for the entire corpus, and then uses the observed words in the given paragraph to estimate the topic loadings of all of the topics for that paragraph. This can be done out of sample (for documents not in the training sample), but we use it solely on paragraphs from the original sample (see Blei et al., 2003 for more information on inferencing). We then identify the topic for which the paragraph has the highest loading (i.e. the topic that is discussed the most) and we assign that paragraph to that topic.

After performing the above process for all paragraphs in the corpus, we end up with 150 groups of paragraphs, where each group consists of paragraphs that share the same dominant topic. We can then calculate textual attributes, such as Fog, redundancy, and specificity for each paragraph and estimate the average statistics for these measures at the topic level (i.e. for all paragraphs sharing the same topic).

Identifying representative paragraphs

We follow an approach very similar to Hoberg and Lewis (2017) to identify the representative paragraphs for our LDA topic model, except that we use inferencing at the paragraph level instead of cosine similarity to identify the most prominent topic in each paragraph. That is, for each topic we identify the 1000 paragraphs with the highest topic loading for that topic and then retain only the middle tercile of documents by length (number of words). Among the remaining 333 paragraphs, we compare each paragraph with all of the other paragraphs using cosine similarity and select the paragraph that has the highest average similarity with the other paragraphs.

One thing to keep in mind with this procedure is that it favors picking paragraphs with more "standardized" content because this will be shared across many firms. For example, in a simple example where all documents contain two paragraphs about Topic A, where one of the paragraphs quotes verbatim the accounting standard that applies to that topic and the

[39] The intruder words that we select are among the 15% least probable words for the given topic. Following Chang et al. (2009) we further constrain these words to be relatively common in at least one other topic to prevent coders from identifying them as the intruder words by virtue of their rare usage in any financial topic. Thus our intruder words must be in the top 20 most common words in at least one other topic.

[40] Both coders have a background in accounting and business and are familiar with financial terminology.

[41] In particular, the first coder agreed with the model 86%, 84%, and 79% of the time for 150, 200, and 250 topics, respectively. The second coder agreed with the 150-topic model 89% of the time, with the 200-topic model 83% of the time, and with the 250-topic model 81% of the time.

[42] "Paragraphs" is a loose description. Specifically, what we refer to as "paragraphs" are portions of text separated from all other text by two end of line markers (e.g., carriage returns).

other paragraph describes the application of that standard to each particular firm, then this process would select the standard paragraph as the representative paragraph. In other words, this process may ignore important variability in discussions of a topic that can arise across firms in favor of picking standardized and potentially "boilerplate" paragraphs. Inferences from these paragraphs must therefore be drawn with that caveat in mind.

References

Aletras, N., Stevenson, M., 2014. Measuring the similarity between automatically generated topics. In: EACL, pp. 22–27.
Ball, C., Hoberg, G., Maksimovic, V., 2015. Disclosure, business change and earnings quality. Unpublished Working Paper.
Blankespoor, E., 2016. The impact of information processing costs on firm disclosure choice: evidence from the XBRL mandate.
Blei, D., 2012. Probabilistic topic models. Commun. ACM 55, 77–84.
Blei, D., Lafferty, J., 2007. A correlated topic model of science. Ann. Appl. Stat. 1, 17–35.
Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Bloomfield, R., 2002. The "incomplete revelation hypothesis" and financial reporting. Account. Horiz. 16, 233–243.
Brown, S., Tucker, J., 2011. Large-sample evidence on firms' year-over-year MD&A modifications. J. Account. Res. 49, 309–346.
Bushman, R., Lerman, A., Zhang, X., 2016. The changing landscape of accrual accounting. J. Account. Res. 54, 41–78.
Campbell, J.L, Chen, H., Dhaliwal, D.S., Lu, H., Steele, L.B., 2014. The information content of mandatory risk factor disclosures in corporate filings. Rev. Account. Stud. 19, 396–455.
Cazier, R.A., Pfeiffer, R.J., 2015a. Why are 10-K filings so long? Account. Horiz. 30, 1–21.
Cazier, R.A., Pfeiffer, R.J., 2015b. Say again? Assessing redundancy in 10-K disclosures. Unpublished Working Paper.
Chaney, A., Wallach, H., Blei, D., 2015. Who, what, when, where, and why? A computational approach to understanding historical events using state department cables. Unpublished Working Paper.
Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M., 2009. Reading tea leaves: How humans interpret topic models. In: Bengio, Y, Schuurmans, D, Lafferty, JD, Williams, CKI, Culotta, A (Eds.), Advanced Neural Information Processing Systems. Curran Associates, New York, pp. 288–296.
Collins, D., Maydew, E., Weiss, I., 1997. Changes in the value-relevance of earnings and book values over the past forty years. J. Account. Econ. 24, 39–67.
Demerjian, P., 2011. Accounting standards and debt covenants: has the "balance sheet approach" led to a decline in the use of balance sheet covenants. J. Account. Econ. 52, 178–202.
Dichev, I., Tang, V., 2008. Matching and the changing properties of accounting earnings over the last 40 years. Account. Rev. 83, 1425–1460.
Financial Accounting Standards Board (FASB), 2012. Disclosure Framework: Invitation to Comment. Norwalk, CT.
Francis, J., Schipper, K., 1999. Have financial statements lost their relevance? J. Account. Res. 37, 319–352.
Gallemore, J., Labro, E., 2015. The importance of the internal information environment for tax avoidance. J. Account. Econ. 60, 149–167.
Givoly, D., Hayn, C., 2000. The changing time-series properties of earnings, cash flows and accruals: has financial reporting become more conservative? J. Account. Econ. 29, 287–320.
Grimmer, J., 2010. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. Polit. Anal. 18, 1–35.
Gunning, R., 1952. The Technique of Clear Writing. McGraw-Hill, New York.
Higgins, K., 2014. Disclosure Effectiveness: Remarks before the American Bar Association Business Law Section Spring Meeting. Available at: https://www.sec.gov/News/Speech/Detail/Speech/1370541479332.
Hoberg, G., Lewis, C., 2017. Do fraudulent firms produce abnormal disclosure? J. Corp. Finance 43, 58–85.
Hope, O., Hu, D., Lu, H., 2016. The benefits of specific risk-factor disclosures. Rev. Account. Stud. 21, 1005–1045.
Huang, A., Lehavy, R., Zang, A., Zheng, R., 2016. Analyst information discovery and interpretation roles: a topic modeling approach. Management Science, Forthcoming.
KPMG, 2011. Disclosure overload and complexity: hidden in plain sight. Available at: http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/disclosure-overload-complexity.pdf.
Lang, M., Stice-Lawrence, L., 2015. Textual analysis and international financial reporting: large sample evidence. J. Account. Econ. 60, 110–135.
Lawrence, A., 2013. Individual investors and financial disclosure. J. Account. Econ. 56, 130–147.
Lee, Y., 2012. The effect of quarterly report readability on information efficiency of stock prices. Contemp. Account. Res. 29, 1137–1170.
Lev, B., Zarowin, P., 1999. The boundaries of financial reporting and how to extend them. J. Account. Res. 37, 353–385.
Li, F., 2008. Annual report readability, current earnings, and earnings persistence. J. Account. Econ. 43, 221–247.
Loughran, T., Mcdonald, B., 2014. Measuring readability in financial disclosures. J. Finance 69, 1643–1671.
McCallum, A, 2002. MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu.
Miller, B.P., 2010. The effects of reporting complexity on small and large investor trading. Account. Rev. 85, 2107–2143.
Monga, V., Chasan, E., 2015. The 109,894-word annual report; as regulators require more disclosures, 10-Ks reach epic lengths; how much is too much? Wall Street J. (Online), Jun 2, 2015.
Nelson, K., Pritchard, A., 2016. Carrot or stick? the shift from voluntary to mandatory disclosure of risk factors. J. Empirical Legal Stud. 13, 266–297.
Rhody, L., 2012. Topic modeling and figurative language. J. Digit. Humanit. 2, 19–35.
Securities and Exchange Commission (SEC), 1998. A Plain English Handbook: How to Create Clear SEC Disclosure Documents. Available at: https://www.sec.gov/pdf/handbook.pdf. SEC Offices, Washington D.C.
Securities and Exchange Commission (SEC), 2013. Report on Review of Disclosure Requirements in Regulation S-K. Available at: http://www.sec.gov/news/studies/2013/reg-sk-disclosure-requirements-review.pdf. SEC Offices, Washington D.C.
Spangher, A., 2015. Building the Next New York Times Recommendation Engine. The New York Times.
Srivastava, A., 2014. Why have measures of earnings quality changed over time? J. Account. Econ. 57, 196–217.
White, M.J., 2013. The Path Forward On Disclosure. National Harbor, MD.
You, H., Zhang, X., 2009. Financial reporting complexity and investor underreaction to 10-K information. Rev. Account. Stud. 14, 559–586.