The Primacy of Numbers in Financial and Accounting Disclosures: Implications for Textual Analysis Research

Federico Siano
Boston University - Questrom School of Business
fsiano@bu.edu

Peter Wysocki[*]
Boston University - Questrom School of Business
wysockip@bu.edu

September 2019

Abstract:

Numbers are central to financial and accounting disclosures, yet current textual analysis research generally ignores and deletes numbers within disclosures. We hypothesize and show that the prevalence of numbers within a corporate disclosure text is highly correlated with proxies for the readability of the disclosure. More importantly, we show that prior findings on the links between proxies for disclosure readability and various economic outcomes are largely subsumed by the prevalence of numbers within a disclosure. We discuss implications for past and future research that attempts to analyze the determinants, attributes and outcomes of financial and accounting disclosures.

Key Words: *Accounting Quality, Analyst Following, Disclosure; Linguistics; Quantitative Information; Readability; Textual Analysis.*

Data Availability: From publicly-available data sets.

[*]We wish to thank George Papadakis, Taylor Wiesen and workshop participants at the 2018 Lisbon Accounting Conference, ESMT Berlin and Cass Business School for helpful discussions and comments.

"It's clearly a budget. It's got a lot of numbers in it."
---- George W. Bush, May 5, 2000

1. Introduction

The goal of this study is to provide evidence on the links between the often-ignored numbers *within* the text of corporate disclosures, the textual attributes of these disclosures, and associated economic outcomes. While prior textual analysis research in accounting and finance almost universally ignores or deletes numbers *within* the text of corporate disclosures[1], we present new evidence of a fundamental association between the prevalence of numbers within a business text and the readability of the text. First, we establish a foundational link between the prevalence of numbers and the readability of business articles published in the *Wall Street Journal*. We hypothesize and find that the prevalence of numbers within an article is associated with the use of less complex language (down to the sentence level), and thus more "readable" text. Second, we document a similarly strong association between the prevalence of numbers and the readability of the Management Discussion and Analysis (MD&A) section of firms' 10-K and 10-Q financial reports. Finally, we hypothesize and show that the prevalence of numbers within MD&A disclosures largely subsumes and explains two key findings in the textual analysis literature that disclosure readability is directly associated with firm profitability (see Li, 2008) and analyst following (see Lehavy et al., 2011).

This study provides new insights into one of the more active research areas in accounting and finance that focuses on the determinants, attributes and outcomes of corporate disclosures (see,

[1] See recent surveys of the textual analysis literature in accounting and finance by Li (2011), Das (2014), Kearney and Liu (2014), Loughran and McDonald (2016), and Dyer et al. (2017b).

for example, Healy and Palepu, 2001, and Leuz and Wysocki, 2016). More recently, empirical researchers have applied new textual and linguistic analysis tools to characterize the textual attributes of corporate disclosures and then show that these attributes are associated with key outcomes such as reported profitability, trading behavior, analyst following, retail investor choices, cost of capital, earnings management, and firm valuation.[2] Notwithstanding these advances, the textual analysis literature almost universally ignores or deletes numbers *within* the text of accounting and financial documents.[3] However, these numbers directly capture and summarize performance and financial position and are arguably the disclosures of primary interest for many stakeholders, while the surrounding disclosure text often plays a secondary role of describing or providing context for the disclosed numbers and quantitative information. Therefore, the current practice of ignoring numbers within disclosure texts leads to a correlated omitted variable problem for researchers that can affect inferences about the direct determinants and outcomes of the textual attributes of corporate disclosures.

Our empirical analysis has three main parts related to four primary hypotheses. In the first part, we examine the links between the numbers and word complexity and readability of generic business texts. Specifically, we analyze a large set of *Wall Street Journal* articles that are not primarily focused on companies' earnings reports. This dataset is used to establish the existence of structural association (for business/economic documents) between the prevalence of numbers in a document and the "readability" of the document in a setting that is not confounded by

[2] Key outcomes that have be correlated with the textual attributes of corporate disclosures include profitability (Li, 2008), trading behavior (Miller, 2010), analyst following (Lehavy et al., 2011), retail investor choices (Lawrence, 2013), cost of capital (Bonsall and Miller, 2017), earnings management (Lo et al., 2017), and firm valuation (Hwang and Kim, 2017).

[3] The notable exception is found in Lundholm et al. (2014) who examine the relative readability and "number of numbers" in annual reports of foreign firms compared to domestic U.S. firms. While Lundholm et al. (2014) do not correlate disclosure readability with the "number of numbers" within a disclosure, they do find that foreign firms' disclosures have both higher average readability and greater average "number of numbers" compared to U.S. firms.

managerial disclosure incentives. In the second part, we analyze the textual disclosures within the Management Discussion and Analysis (MD&A) section of firms' 10-K SEC EDGAR filings from 1994-2017. Similar to recent textual analysis studies in accounting, we calculate the readability (as captured by both the *Fog* index [Gunning, 1952] and the *Bog* index [Bonsall et al., 2017]) of the MD&A section of each report, but we also tabulate the prevalence of numbers within the MD&A section. We then examine whether controlling for the prevalence of numbers within the text of a financial report affects previously-documented findings of a positive link between MD&A disclosure readability and reported profitability and a negative relation between disclosure readability and analyst following. In the final section, we replicate the prior analyses using a previously-unexplored set of corporate disclosures based on a novel U.S. SEC dataset from 1987-1993 of 10-Q financial reports from the pre-EDGAR era.[4]

Overall, our findings show a strong association between the prevalence of numbers within business documents (*Wall Street Journal* articles and the MD&A sections of 10-K and 10-Q reports) and the complexity of words within and the readability of the documents. More importantly, we show that the prevalence of numbers within the text of corporate disclosures largely subsumes two key findings from the early textual analysis literature that disclosure readability is directly associated with firm profitability (see Li, 2008) and analyst following (see Lehavy et al., 2011). Our findings suggest that ignoring numbers within corporate disclosure texts is likely to impact researchers' inferences about the links between specific textual attributes of the disclosures and numerous accounting, financial and economic outcomes. Like prior textual analysis research, our

[4] This novel sample of pre-EDGAR 10-Q disclosures provides a number of contributions and robustness tests including: (i) it introduces new and potentially useful data to other researchers, (ii) it allows for "out of sample" tests of prior empirical findings, and (iii) it is potentially less subject to recent corporate disclosure trends that can introduce noise into the textual analysis of disclosures including the use of boilerplate disclosures, the profusion of embedded tables and images in recent EDGAR filings, and corporate disclosure "bloat" to comply with new reporting regulations (see, for example, Cazier and Pfeiffer, 2016; and Dyer et al., 2017a).

current findings primarily document associations, so there remain open questions about the mechanisms underlying the associations. For example, does the disclosure of numbers fundamentally *cause* other textual attributes, or are both the prevalence of numbers and other textual attributes the joint outcomes of managers' unobserved latent disclosure objectives?[5]

Our findings on the prevalence of numbers within the text of corporate disclosures also contribute to the literature on accounting quality which proposes a number of competing measures of disclosure quality, complexity and comparability based on the *quantity* of numbers presented in accounting reports (see, for example, Chen et al., 2015, Hoitash et al., 2017, and Hoitash and Hoitash, 2018). Our evidence suggests that more quantitative information disclosed *within* the text of corporate disclosures is associated with higher quality disclosures. This aligns with the empirical evidence for a financial statement disaggregation measure proposed by Chen et al. (2016), but contrasts with the "number of recognized numbers" evidence for XBRL-coded financial statement data presented in Hoitash and Hoitash (2018).

The structure of the remainder of the paper is as follows: In Section 2, we discuss the related literature. In Section 3, we present our hypotheses. Section 4 presents the data samples and describes the key variables used in the empirical tests. Section 5 discusses the main empirical tests as well as the robustness analyses. Finally, in Section 6, we summarize our results, outline our conclusions and discuss future work.

2. Related literature

Our research is related to and has implications for two main streams of existing accounting and finance research: (1) research on the textual attributes of disclosures and their connections

[5] It should be noted that our findings for the sample of *WSJ* news articles (not focused on corporate earnings reports) suggests that the existence of sentence-level quantitative information structurally *leads to* the use of less complex language in the sentence.

with various financial, accounting, disclosure and economic outcomes, and (2) empirical research that attempts to characterize the quality, complexity and comparability of firms' reported numbers and disclosures.

First, our paper contributes to the emerging empirical literature in accounting and finance that develops and uses textual analysis tools to better understand the content and characteristics of the text of accounting and financial disclosures (see, for example, Li, 2011; Das, 2014; Kearney and Liu, 2014; Loughran and McDonald, 2016; and, Dyer et al., 2017b). A large part of this literature focuses on quantifying the readability (or language complexity) of accounting and financial disclosures and then correlating disclosure readability with various outcomes.

Many disclosure readability studies focus on a company's annual report readability (often highlighting the MD&A section of the report) and correlate this text attribute with a wide array of issues and outcomes. For example, Ertugrul et al. (2017) investigate the links between annual report readability and corporate borrowing costs, while Lee (2012) examines the impact of readability on equity market efficiency. Ginesti et al. (2018) examine the link between annual report readability and corporate board of director characteristics. Lo et al. (2017) explore the connections between annual report readability and earnings management. Lim et al. (2018) correlate corporate strategy with annual report readability. Other studies focus on investor issues related to annual report readability such as investors' processing fluency (Rennekamp, 2012), small versus large shareholders' trading activity (Miller, 2010), retail investors' trading decisions (Lawrence, 2013), and investor demand for information from foreign firms (Lundholm et al., 2014). Other studies attempt to differentiate between possible competing determinants of annual report readability such as managerial obfuscation versus a firm's underlying operational complexity (see, for example, Guay et al. 2016, and Bushee et al., 2018). There are also a series of

recent studies that examine the readability of other business texts and attempt to make connections with closely-related outcomes. For example, De Franco et al. (2015) examine the possible determinants and implications of analyst report readability. Laksmana et al. (2012) and Hooghiemstra et al. (2017) examine compensation discussion and analysis (CD&A) readability and managerial obfuscation incentives. Inger et al. (2018) examine the association between tax footnote readability and firms' tax avoidance strategies.

While all of these disclosure readability studies acknowledge and control for a wide array of financial, corporate, board, managerial, and investor characteristics in the empirical tests, almost all of these studies delete or ignore the quantitative information contained within the text of the disclosures. However, this almost-universal methodological choice is at odds with almost 50 years of capital markets research that recognizes the prominent role of *quantitative* financial and accounting information and how this information is linked with accounting, financial and economic decisions. Clearly, one should expect that the quantitative information *within* the text of corporate disclosures should also be associated with various outcomes and decisions. Thus, the current paradigm in the literature that ignores numbers *within* disclosure texts leads to a correlated omitted variables problem for researchers that is likely to affect inferences about the direct determinants and outcomes of the textual attributes of corporate disclosures.

Second, this study relates to the growing literature that attempts to better define, measure and understand the implications of reporting quality, complexity and comparability. Recent related innovations in this literature have focused on the amount of quantitative information presented in accounting reports. For example, Chen et al. (2015) suggest and implement a new measure of accounting quality based on the amount of disaggregation of reported numbers in companies'

financial statements.[6] Their evidence suggests that the greater quantity of disaggregated numbers reported in the income statement, balance sheet and statement of cash flows is associated with better capital market outcomes (i.e., more numbers and higher accounting quality are related). Similarly, Hoitash et al. (2017) and Hoitash and Hoitash (2018) focus on the quantity of numbers reported in firms' financial statements. For example, Hoitash and Hoitash (2018) use XBRL tags to tabulate the number of unique quantitative items recognized and reported in firms' financial statements. They argue and present evidence consistent with the notion that more (XBRL-tagged) items reported in a firm's financial statements reflect greater accounting complexity (i.e., lower quality) and thus greater difficulty for stakeholders to process the accounting information.

Our study complements and extends this line of research by exploring the amount of quantitative information presented *within* the text of corporate disclosures. In addition, our analyses attempt to connect this quantitative information with other textual attributes and "soft" information contained in the text of firms' disclosures. Siano (2019) also examines the connection between numbers disclosed within the text of 10-K filings and a company's financing and capital market horizons.

3. Hypotheses

The financial accounting and capital markets literature of the past 50 years has overwhelmingly focused on quantitative information and the numbers *recognized* in firms' financial statements (see, for example, Kothari, 2000; and Lee, 2000). For both economic and pragmatic reasons, this literature has generally avoided dealing with and characterizing the "soft" information contained in the text of corporate disclosures. But, as noted in Section 2 above, the

[6] Drake et al. (2016) also count the number of unique, non-missing *Compustat* items in the financial statements and use it as a control variable in their examination of the use of historical EGDAR filings by investors.

emerging textual analysis literature in accounting and finance has made significant recent advances in analyzing the content and attributes of textual disclosures. However, almost all textual analysis studies use a similar methodology that ignores or removes numbers from the text of accounting and financial documents. We argue that this methodological choice to remove or ignore numbers is problematic because, consistent with the accumulated theory and evidence from the mainstream accounting literature, these numbers are likely to be of primary interest to the users of the financial disclosure.[7] While this certainly does not rule an important role for the surrounding text, this text is arguably built on the scaffolding of the disclosed accounting numbers and the surrounding text characterizes and describes the numbers and quantitative information. Thus, we argue that the presence and prevalence of numbers *within* disclosures should, at the very least, be related to the language used in the text of corporate disclosures. Furthermore, while difficult to unequivocally demonstrate, it is likely that the presence of numbers within a disclosure *causally* influences the chosen language and textual attributes of the disclosure.

Our reading of the academic linguistics literature reveals a paucity of discussion, let alone theory and evidence, about the interplay between language and numbers. Similar to the historical aversion of accounting researchers to deal with "soft" language in disclosures, it appears that academic linguists in the humanities have generally ignored hard numbers within corpora.[8] Thus, the extant linguistics and textual analysis literatures provide little guidance on the possible links between numbers and words within documents. However, we argue that one should expect a strong

[7] Given the widely-known contracting and valuation roles of accounting numbers, it should be uncontroversial to expect that stakeholders would seek out quantitative information not only in the financial statements, but also *within* the text of accounting reports. Therefore, analyzing the text of an accounting report without considering the quantitative information (i.e., numbers) would be an incomplete approach at best.

[8] The apparent "avoidance" of numbers by academic linguistics is more of an observation about research themes rather than a commentary on the research methods used in the field. It is certainly the case that contemporary linguistics research relies heavily on advanced quantitative, computational and statistical methods.

link between the presence of numbers in a text and the type of words and the structure of language used to describe the numbers. Specifically, when a text presents and describes measurable and factual quantitative information (i.e., numbers), it seems natural that the associated words are more likely to be objective, concise, precise, free of rhetoric, unambiguous, verifiable, and to apply commonly-agreed-upon concepts[9]. Thus, we would expect that documents that have a greater prevalence of numbers are less likely to use complex and longer words. This leads to our first hypothesis (stated in the null form):

H1: The use of complex words, at both the sentence level and overall document level, is unrelated to the frequency of numbers reported in the document.

The linguistics and textual analysis literatures have also highlighted the readability of a document as an important document attribute (see, for example, Gunning, 1958, and Li, 2011). One of the more widely-used empirical proxies for document readability is the Gunning (1958) *Fog* measure which captures document "readability" as a combination of the average numbers of words per sentence and the percent complex words (words with more than two syllables):

$$Fog = 0.4*(Average\ number\ of\ words\ per\ sentence + \%\ Complex\ words) \quad (1)$$

Given the motivating arguments for Hypothesis H1, we would also expect that documents with more numbers within the text to be more "readable" as measured by the *Fog* index (given that the *Fog* index is, in part, mechanically derived from the number of complex words in a

[9] Lundholm et al. (2014) argue that the "number of numbers" in the text of an annual report captures the amount of factual information in the disclosure.

document). Thus, our second hypothesis (stated in the null form) is:

H2: The readability of a document, as captured by the *Fog* index, is unrelated to the frequency of numbers reported in a business document.

As an extension of Hypothesis H2, we also expect that other empirical measures of document readability are likely to be correlated with the frequency of numbers reported in a business document (i.e., the *Bog* index as investigated in Bonsall et al., 2017).

Next, we turn to the possible implications of ignoring numbers within the text of a corporate disclosure. As discussed in section 2, there is a growing body of evidence that disclosure readability (primarily operationalized in research studies using the *Fog* index) is associated with a host of other accounting, financial and economic outcomes. Two of the more prominent early findings related to disclosure readability are: (i) firms with higher reported profits have higher disclosure readability, as captured by lower *Fog* (Li, 2008), and (ii) firms with higher disclosure readability have lower analyst following (Lehavy et al., 2011). Given the expectation of a strong connection between the prevalence of numbers within a disclosure and the readability of a disclosure, empirical tests that ignore the prevalence of numbers may suffer from a correlated omitted variable problem which can bias the estimated association between disclosure readability and other outcomes. The direction of the bias depends on the covariance between the regressors and the omitted variables. Given that we do not have strong priors on the covariance structure of the regressors, we do not form a directional prediction about how controlling for the prevalence of numbers will *directionally* affect the previous unconditional association between disclosure readability and reported profitability or analyst following. However, given the existence of this correlated omitted variable, we do expect that the prevalence of numbers will impact previously-

estimated associations, and thus we state our third and fourth hypotheses in null form as:

H3: The prevalence of numbers with the text of a disclosure does not affect the empirical association between reported profitability and disclosure readability.

H4: The prevalence of numbers with the text of a disclosure does not affect the empirical association between disclosure readability and analyst following.

4. Description of data samples and key variables

In the following subsections, we (a) summarize the data samples used in our empirical tests, (b) describe the main variables used in our analyses, and (c) summarize the results of the regression analyses used to test our hypotheses.

4.1. Data samples

We use three complementary data samples to provide insights on the association between numbers and the readability of business texts. The three samples are described below.

4.1.1. Sample of Wall Street Journal articles

Using the *Dow Jones Factiva* database, we collect a sample of *Wall Street Journal (WSJ)* news articles published in a time span of two decades, between 1996 and 2016. These articles are primarily text, but also contain accounting, financial or economic numbers. We therefore use the WSJ news as a benchmark to determine if there is a foundational association between the prevalence of numbers within a news article and the textual attributes of the news article (at both the sentence and document level). We organize our data collection into two main categories of articles: "Macro-Economy News" and "Corporate News". "Macro-Economy News" includes articles about "Monetary Policy", "International Trade" and "Economic Commentaries"[10]. "Corporate News" discuss "Corporate Earnings Reports" and "Management Moves". We decide not to limit our analysis to companies' news because they are likely to have content that is influenced by corporate managers' disclosure incentives, especially when those articles (re)report on a firm's financial accounting performance ("Corporate Earnings Reports")[11]. Among the collected documents, we analyze those that are characterized by at least 100 words in both the quantitative (i.e. includes numbers) and the non-quantitative (i.e. does not include numbers) portions of text (i.e. each document has a minimum of 200 total words). We also exclude articles with less than 5 sentences that contain numbers and less than 5 sentences that do not contain

[10] The "International Trade" articles include news describing "Physical Trade", "Trades and External Payments" and "Trade Figures". "Economics Commentaries" encompass "Economic News", "Economic Indicators" and "Editorials".

[11] Our survey of the contents of *WSJ Corporate Earnings Reports* shows that many articles use very similar language and content as the original 10-K or 10-Q filed by a company. As a result, the content of *Corporate Earnings Reports* likely captures managers' underlying disclosure incentives for the original 10-K or 10-Q filings and these incentives could influence the observed association between the numbers within and readability of a *Corporate Earnings Report* news article.

numbers[12]. Consistent with prior literature, we apply these filters to limit spurious associations[13].

4.1.2. 10-K filings from SEC EDGAR database

Our second sample is based on the universe of 10-K filings for U.S. issuers from the SEC EDGAR database between 1994 and 2017. Within 10-K filings, we study the Management Discussion and Analysis (MD&A) section to limit the processing of generic, cautionary, forward-looking or other types of statements driven by regulatory compliance (see, for example, Dyer et al., 2017a) that are not directly related to firms' economic transactions.

4.1.2.1.Data gathering and pre-processing of EDGAR 10-K filings

We download the plain text version of 10-K reports from the Bill McDonald's "Stage One 10-X" dataset available on the Notre Dame *Software Repository for Accounting and Finance*. We process the 10-K filings in two steps. First, we extract the type (e.g., 10-K or 10-K Amendment), the CIK, the filing date, the report date and the MD&A section from each filing. Next, we analyze the MD&A sections and elaborate the relevant textual variables.

We begin by excluding 10-K Amendments, 10-KSB and 10-KSB Amendments from the sample. We subsequently use the Python *Glob* and *Regular Expression* modules to parse the MD&A section of 10-K reports. We code starting signals (e.g., "ITEM 7", "Item 7") and ending signals (e.g., "ITEM 7A", "ITEM 8") to delimit the Management Discussion and Analysis section and develop tailored conditional statements to handle cases of multiple starting and/or ending signals. Whenever the MD&A section of a 10-K cannot be identified with sufficient reliability, we

[12] We keep in the sample documents that do not report any numbers but contain at least 100 words and 5 sentences (10% of the total number of articles). For these documents, in the absence of metrics about quantitative sentences, we use averages from the rest of the news articles.

[13] In robustness tests we replicate our analysis using the unfiltered set of WSJ articles and find qualitatively similar results.

exclude the entire document from the sample. The number of excluded documents varies from year to year and is, on average, in the range 10%-20% of the total filings. We manually check 50 MD&A sections of 50 different companies in different years to confirm the reliability of the outlined coding strategy.

4.1.2.2. Textual analysis methodology for analyzing 10-K MD&A text

We utilize the *Natural Language Toolkit Library* (NLTK) in Python to analyze text and extract relevant information. We start by sentence-tokenizing (i.e. dividing into sentences) each MD&A section and then word-tokenize (i.e. divide into words) each sentence. Sentences are identified through punctuation delimiters, but the NLTK *Library* functions also allow one to control for common textual features within 10-K reports that could lead to an improper sentence identification such as: (i) the possibility of abbreviations (e.g. U.S.A.) or (ii) the presence of decimal numbers which digits are separated by a period (e.g. "increased by 21.5%."). Tables are excluded from the relevant text of MD&A sections. We mark a sentence, or a set of sentences, as a table whenever (i) the number of white spaces is at least 200 and the ratio of numbers to words is greater than 0.25 or (ii) the ratio of numbers to words is greater than 0.50[14]. For each document we count (i) the total number of relevant sentences; (ii) the total number of sentences containing numbers; (iii) the total number of words; (iv) the total number of complex words (i.e. words with more than two syllables); (v) the total number of numbers; and (vi) the total number of dates. To count words and complex words we use the *lexicon_count*[15] and *difficult_words*[16] functions in the

[14] These thresholds balance the trade-off between including too many tables (more likely to happen for high numbers of white spaces and ratios of numbers to words) and excluding too many relevant numbers (more likely to happen for low numbers of white spaces and ratios of numbers to words).
[15] The default *lexicon_count* function excludes a list of "easy words" from the count. We modify the function so that all words are counted.
[16] The *difficult_words* function uses the *Pyphen* library for word hyphenation. We manually test the function on 30 MD&A sections and find an accuracy of 85%.

Textstat package. To count numbers, we first identify what a relevant number is. The MD&A section of corporate filings includes a wide variety of numbers. They can take the form of monetary amounts, percentage changes, ratios, dates or even numbers expressed in words. For the purpose of this analysis, we select the numbers most likely to convey quantitative information and only track and tabulate the frequencies of these types of numbers. Specifically, our parsing algorithm identifies and counts a "number" in the following cases: (i) the number is preceded by a dollar sign ("$"); (ii) the number is followed by the words million/billion/trillion; (iii) the number is followed by a percentage sign ("%") or by the words "percent" / "pct". In addition, the software identifies numbers in parentheses (negative sign) and/or for which the previous markers (i) are preceded by one or two white spaces; (ii) are not preceded by any white spaces; (iii) are capitalized, fully or in part (applies to words). With regards to dates, we first identify years, months, days and then count them as one or multiple dates depending on their structure and whether or not they are located in proximity one to the other.[17]

4.1.3. 10-Q filings from the pre-EDGAR era

Our third data sample is used for assessing robustness and for out-of-sample testing and is based on a novel set of 10-Q filings for U.S. issuers in the pre-EDGAR (pre-1994) era. We chose this data sample for three reasons: (1) to introduce and establish the properties of a new data set for textual analysis researchers in accounting and finance, (2) to apply tests that may be less subject to some recent trends in corporate disclosures in the post-1994 era including growing use of boilerplate and bloat in disclosures driven by regulatory compliance (see, for example, Dyer et al., 2017a), and (3) to analyze a sample of machine-readable filings that include fewer tables, graphs

[17] We include an extensive set of conditional statements to properly identify the elements of a date and avoid double counting. We manually check 100 dates in different formats in 50 documents and find an accuracy of 99%.

and binary files compared to more recent EDGAR filings.

4.1.3.1. History of SEC electronic filings in the pre-EDGAR era

In 1983, U.S. Securities and Exchange Commission (SEC) commenced the construction of an electronic disclosure system with the goal to significantly reduce the use of paper filings and to increase transparency and availability of company data. Starting in 1984, companies could file their statements electronically on a voluntary basis. In 1987, Congress requested that the SEC run tests on a significant group of registrants for a period of at least six months before any electronic filing could be mandated for all regulated firms. Between January and June of 1994, the SEC evaluated the filings submitted electronically by firms belonging to the voluntary pilot group and certified the success of the project to the U.S. Congress. We refer to this sample of electronic filings from 1987-1993 as "pre-EDGAR filings". In December 1994, the SEC made final its rules mandating electronic filing, effective from January 30, 1995 (Release No. 33-7122). The new EDGAR system began to operate in 1995, although electronic filing became mandatory for all companies at the end of 1996, after various phase-in periods.

4.1.3.2. Data gathering and processing of pre-EDGAR 10-Q filings

The documents investigated are retrieved from the *SEC Online Database* available through *LexisNexis Academic*. Our sample includes filings between 1987, the first year in which data are available in the *SEC Online Database*, and 1993. Represented firms are public companies traded on the New York Stock Exchange, American Stock Exchange, or the NASDAQ National Market System. *SEC Online* provides the full text of filings together with categorical information such as the type of document (e.g. 10-Q, 10-K), the filing date, the document date, the company name, the CUSIP number associated to the company's security, the TICKER symbol, the stock exchange in which securities are traded, the SIC code, the fiscal year-end and information on the auditor. Each

regulatory filing begins with a marker and has a table of contents which titles divide the documents into sections. Given this convenient and repetitive structure, we are able to download the *SEC Online* filings in bulk and to parse them through text analysis tools.

We start the data gathering process by downloading all the available *SEC Online* filings, in ".txt" format, between January 1987 and December 1994. The marker *[*Summary]*, found in the most part of documents, is used to separate one form from the other. In all cases where this marker is absent, we add it manually to the filings. For our sample, we select only 10-Q filings and only parse the MD&A section of these filings. We do exclude 10-Q amendments from our sample. Our parsing algorithm collects the company's CUSIP number and document date that are found at the beginning of each 10-Q filing. The parsing algorithm then identifies the start and end of the MD&A and extracts and parses all text from this section of each 10-Q filing.

4.2. Description of textual analysis variables

4.2.1. Main textual analysis variables

In order to provide descriptive evidence on numbers, words, and their possible connections within a document, we create the following variables:

Numbers/Words: the ratio of the "number of numbers" divided by the total words count in the document.

Numbers/WordsQuant: the ratio of the "number of numbers" divided by the words count in the quantitative section of a document.

Numbers/Sentences: the ratio of the "number of numbers" to the total sentences count in a document.

Words/Sentences: the ratio of the total "number of words" to the total number of sentences in a

document. It is a measure of one dimension of the 'readability' of a document and is used as an input into the *Fog* index (Gunning, 1952).

Complex Words/Words: the ratio of "complex words" to total words in a document (complex words are those with more than 2 syllables). This measure is another dimension of the 'readability' of a document and also is used as an input into the *Fog* index (Gunning, 1952).

1/Words: the ratio of 1 to the total number of words in a document. This measure represents the scaling factor we use for numbers and is a proxy for the document's length.

4.2.2. Document readability (Fog index)

Similar to numerous recent papers that examine disclosure readability and linguistic complexity, we use the Gunning (1952) *Fog* index to measure the readability of a document. The empirically-derived *Fog* index is derived from the average numbers of words per sentence and the percent complex words (words with more than two syllables):

$$Fog = 0.4*(Average\ number\ of\ words\ per\ sentence + \%\ Complex\ words) \quad (1)$$

We calculate the components of the *Fog* index following the methodology previously outlined. In a set of robustness tests, we use a number of pre-packaged functions to compute *Fog*.[18] We find that our algorithm is the most conservative and that the results are generally comparable. We also use, as an alternative readability measure, the *Stylewriter Bog* index that was first introduced in the accounting literature by Bonsall et al. (2017).[19]

[18] We use the *gunning_fog* function included in the Python *Textstat* package, the *gunningfog_score* function included in the Python *Textatistic* package and the *fog* function within the Perl Lingua::EN::Phatom package.

[19] Details on the Bog Index algorithm can be found on the Stylewriter website: http://www.stylewriter-usa.com/stylewriter-editing-readability.php.

4.2.3. Other variables

To construct the variables included in the regression analyses using the 10-K Edgar filings, we download and match data from the following repositories over the period 1994-2017: (i) the Annual Fundamental table and the Segments table of Compustat-Capital IQ containing companies' fundamentals, (ii) the CRSP Monthly and Daily Stock File that include securities' prices, (iii) the *I/B/E/S* Summary and Surprise tables with information about analysts following, (iv) *Reuters 13F* with information about institutional ownership, (v) the Bog Index data available on Brian P. Miller's website; (vi) the Accounting Reporting Complexity (Arc) measure available on Udi and Rani Hoitash's website. The definitions of the key outcome and control variables are reported thereafter.

Operating earnings: the contemporaneous annual Compustat operating earnings scaled by total assets.

Operating earnings volatility: the standard deviation of scaled annual operating earnings for the last 5 fiscal years.

Size: the natural logarithm of beginning of period market value of equity from Compustat.

MTB: the beginning of period market value of equity divided by its book value from Compustat.

Returns Volatility: the standard deviation of CRSP monthly stock returns in the last fiscal year.

Age: number of years since a firm shows up within the CRSP Monthly Stock File.

Special Items: special items scaled by total assets from Compustat.

Business Segments: natural logarithm of one plus the number of business segments from Compustat Segments.

Geographic Segments: natural logarithm of one plus the number of geographic segments from Compustat Segments.

Delaware: a binary variable equal to 1 if the company is incorporated in Delaware and 0

otherwise.

Analyst following: the number of I//B/E/S analysts in the first consensus annual earnings forecast date following the 10-K filing[20].

Growth: compounded average growth rate of Compustat sales over the prior 5 years.

10-K News: absolute value of the cumulative market-adjusted return for the 10-K filing window [0,1] from CRSP.

Adv: advertising expenses as a percentage of operating expenses from the prior fiscal year from Compustat.

R&D: research and development expenses as a percentage of operating expenses from the prior fiscal year from Compustat.

InstInv: percentage of institutional ownership for the quarter prior to the 10-K filing from Reuters 13F.

Arc: the natural logarithm of the number of XBRL tags of a firm's financial statements from Hoitash and Hoitash (2018).

Industry membership: binary variables identifying a firm's Fama-French industry membership based on a 48-industry categorization.

For the robustness tests based on pre-EDGAR 10-Q filings (Section 5.5), the definitions of the key *quarterly* outcome and control variables are:

Operating earnings: the contemporaneous quarterly (q) Compustat operating earnings scaled by total assets.

Operating earnings volatility: the standard deviation of scaled quarterly operating earnings for the last 12 quarters.

[20] The log-transformed value of one plus Analyst Following is also used in robustness checks and qualitatively similar results are obtained. Missing values of Analyst Following are replaced with zero.

5. Empirical results

Our empirical analysis has three main parts related to our four hypotheses. We first examine the links among numbers, word complexity and the readability of generic business texts. We then extend this analysis to corporate filings using a comprehensive sample of SEC EDGAR SEC 10-K reports filings from 1994-2017. Next, we examine whether controlling for the prevalence of numbers within the text of a 10-K report affects previously-documented findings of a positive link between disclosure readability and reported profitability and a negative relation between disclosure readability and analyst following. Finally, we assess the robustness of these findings using a novel out-of-sample dataset of 10-Q filings from the pre-EDGAR era.

The presentation of our empirical results follows the order of our four hypotheses: the connection between numbers and word complexity in generic business texts; the association between the prevalence of numbers in a corporate disclosure and the readability of the corporate disclosure; and how the prevalence of numbers may affect inferences related to disclosure readability and firm profitability and analyst following.

5.1. Relation between numbers, word complexity and readability – WSJ articles

Table 1 presents the descriptive statistics for the partitioned subsamples of *Wall Street Journal* news articles for 1992. We consider 1,095 news articles not directly related to corporate earnings reports (hereafter referred to as the sample of *Main News Articles*) comprising the 4 columns labeled "Economic News and Indicators", "International Trade", "Monetary Policy", and "Tracking the Economy". We also separately consider a sample of 923 *WSJ* news articles that cover "Corporate Earnings Reports" (presented as a holdout sample in the last column). On average, the ratio of numbers to words for the *Main News Articles* is 9.6% (with a high of 13.1% for "Tracking the Economy" articles).

Using this *Main News Articles* sample, we compare the textual properties of sentences that contain numbers (quantitative sentences) to those that do not contain numbers (non-quantitative sentences). We first examine average sentence length as a dimension of text readability. As indicated in the row labeled *Mean # words per sentence,* there are some differences in the average sentence length between quantitative (i.e., contains at least one number) and non-quantitative sentences. On average, there is just over one more word per sentence for quantitative sentences compared to non-quantitative sentences for the full sample of 1,095 *Main News Articles.* Given that the average sentence length of the non-quantitative sentences is 10.8 words, this means that, on average, quantitative sentences are 9.3% longer than non-quantitative sentences as measured by words. This suggests that quantitative sentences are "less readable" along the sentence length dimension.

We next turn to the use of complex words (words that are more than two syllables). As indicated in the next row labeled *Mean # complex words per sentence,* there are also differences between quantitative and non-quantitative sentences. On average, there are 0.35 more complex words per sentence for non-quantitative sentences compared to quantitative sentences for the full sample of 1,095 *Main News Articles.* This is an economically meaningful difference because the average number of complex words per sentence for non-quantitative sentences is 2.96. This means that quantitative sentences use, on average, 11.8% fewer complex words than non-quantitative sentences. This finding suggests that quantitative sentences are "more readable" based on this second dimension of fewer complex words. These findings are consistent with the arguments behind *Hypothesis H1*. Specifically, we predicted that quantitative sentences would use fewer long complex words because quantitative information is more compatible with language that is concise, precise, unambiguous, and free of rhetoric; and applies or uses commonly-agreed-upon

verifiable concepts.

The most-commonly used measure of sentence and document readability is the *Fog* index (Gunning, 1952). Therefore, we also compare the *Fog* index for quantitative and non-quantitative sentences for the full sample of 1,095 *Main News Articles.* Given that the *Fog* index is essentially a linear combination of sentence length and complex words, one might conclude that the opposing effects of sentence length and word complexity across quantitative and non-quantitative sentences would "cancel each other out" for this sample. However, the empirical *Fog* index (equation (1)) places an order of magnitude more weight on a 1% difference in word complexity compared to a 1% difference in sentence length to rate the overall composite "readability" of a sentence. Thus, as summarized in the next row in Table 2 labeled *Mean Fog of sentences,* there is a very large difference between the overall *Fog "*readability" of quantitative versus non-quantitative sentences. On average, the *Fog* index is almost 4 points higher (or 24.6%) higher for nonquantitative sentences compared to quantitative sentences (i.e., sentences that include numbers) for the sample of *Main News Articles*. This difference is both economically and statistically significant (p-value <0.01). In other words, nonquantitative sentences are far more "*Foggy"* and thus less readable than quantitative sentences. These findings are consistent with the arguments behind *Hypothesis H2* and the results suggest that the predicted *Fog* differences between quantitative and nonquantitative sentences are driven by differences in word complexity.

Importantly, we confirm in untabulated analyses that the presented results do not depend upon the choice of 1992 as our reference year. In fact, we sample at random five additional years in the period 1996-2016 and find consistent average results.

As discussed earlier, the above findings for the sample of 1,095 *Main News Articles* provide insights on the association between numbers and readability for articles that are unlikely

to be affected by corporate reporting incentives. However, we also perform the readability comparisons for a separate sample of 923 *WSJ* news articles focused on *Corporate Earnings Reports.* As shown in the last column of Table 1, we find even stronger readability differences between quantitative and non-quantitative sentences for this sample. On average, the *Fog* index is 53% (7.89 points) higher for nonquantitative sentences compared to quantitative sentences within *WSJ Corporate Earnings Reports.*

5.2. Relation between prevalence of numbers and disclosure readability - 10-K evidence

We next turn to the sample of *SEC EDGAR 10-K* filings for the years 1994-2017. The main sample consists of 77,144 annual observations from EDGAR filings from 1994-2017 with available data to calculate the *Fog* index from the text of the MD&A section of a firm's 10-K filing and matching *Compustat* data to calculate the key control variables identified in Section 4.

Table 2 presents the correlations among the main variables. The key correlation of interest captures the possible association between the commonly-used *Fog* index (used to capture disclosure readability) and the prevalence of numbers *within* the text of the MD&A disclosure (captured by the ratio *#'s/Words*). Consistent with the arguments motivating *Hypothesis H2*, we find that the Pearson correlation is -0.446. This is economically significant and it is larger than any of the other correlations presented in Table 2 (or even in other studies examining the properties of disclosure readability). Given the possible concern that *#'s/Words* may just capture the overall length of a disclosure (the denominator of this variable), we also present the correlation between *Fog* and the inverse of the length of a disclosure in words (*1/Words)*. As shown in Table 2, this correlation is -0.1 and much weaker than the *Fog* to *#'s/Words* correlation. Also, the correlation between *#'s/Words* and *1/Words* is only +0.15. This evidence supports our hypothesis that the prevalence of numbers within a disclosure is a unique and potentially very important (both

economically and statistically) correlated omitted variable that has the potential to affect inferences about previously-documented associations between disclosure readability and other outcomes.

5.3. Impact of numbers on the profitability-readability relation - 10-K evidence

Table 3 presents a replication of the profitability-readability regression originally estimated in Li (2008). We use a sample of 63,119 annual firm-year observations from EDGAR filings from 1994-2017. In column (1) of Table 3, we estimate a regression that is very similar to Table 3 in Li (2008) using the same *Compustat* explanatory variables. The dependent variable is the *Fog* index for the MD&A section of a firm's 10-K filing. Similar to Li (2008), we control for *Size, MTB, Earnings Volatility, Industry Fixed Effects, and Period (year-firm) Fixed Effects.* Consistent with the findings of Li (2008), we find in regression column (1) that the MD&A *Fog* is strongly negatively related to contemporaneous reported firm profitability (i.e., a very statistically-significant negative coefficient on *Operating Earnings* of -0.77). Thus, this finding is consistent with the original findings in Li (2008) that firms with lower profitability tend to have less readable (higher *Fog*) disclosures.

However, the regression in column (1) of Table 3 does not control for the prevalence of numbers in the MD&A section of the 10-K filing. Therefore, in column (2) of Table 3 we include the ratio of *#'s/Words* in the MD&A as an additional explanatory variable. Not surprisingly, the explanatory power of the regression increases from 22% to 33%. More importantly, the association between firm profitability and *Fog* is much weaker and the statistical significance drops dramatically. In the remaining columns of Table 3, we report additional regression specifications which also include firm fixed effects. These regressions show that the inclusion of the ratio of *#'s/Words* dramatically affects the documented associations presented in Li (2008). In some cases, the association between firm profitability and *Fog* is no longer significant. Overall, the claimed

association between disclosure readability and firm profitability does not appear to be as robust as previous evidence might suggest. Clearly, the prevalence of quantitative disclosures within the MD&A text is an important correlated (and previously-omitted) disclosure characteristic.

In Table 4, we replicate the original findings of Li (2008) using an updated index of text readability based on the *Bog* index (see the StyleWriter *Bog* index used in Bonsall et al., 2017). Using this alternate measure of disclosure readability, the regressions presented in Table 4 are very consistent with the findings in Table 3. Specifically, the association between firm profitability and *Bog* is much weaker and the statistical significance drops dramatically when one includes of the ratio of *#'s/Words* as a key correlate in the regressions. Again, this suggests that the apparent association between disclosure readability (as captured by the alternate *Bog* index) and firm profitability does not appear to be as robust.

5.4. Relation between analyst following and disclosure readability - 10-K evidence

Table 5 outlines a replication of the analyst following regressions originally presented in Lehavy et al. (2011). We use a sample of 44,370 firm-year observations derived from *EDGAR* 10-K filings between 1994 and 2017 and match the firms with analyst forecast data from I/B/E/S. In column 1 of Table 5, we estimate a regression that is very similar to Lehavy et al. (2011) using similar *Compustat* explanatory variables. The dependent variable is the *Number of Analysts* who follow a firm during the period. Similar to Lehavy et al. (2011), we control for *Size, MTB, Earnings Volatility, Industry Fixed Effects, and Period (year) Fixed Effects.* Consistent with the findings of Lehavy et al. (2011), we find in column (1) of Table 5 that analyst following is positively related to the MD&A *Fog* of the contemporaneous 10-K filing (i.e., a statistically- significant positive coefficient on *Fog* of 0.06). Thus, this finding is consistent with the original finding in Lehavy et al. (2011) that firms with less readable (higher *Fog*) disclosures tend to attract more analysts and

this finding is consistent with an information intermediary/processing role for analysts.

However, the regression in column (1) of Table 5 does not control for the prevalence of numbers in the MD&A section of the 10-K filing. Thus, we cannot be certain that readability is the disclosure attribute that is directly associated with analyst following. Alternately, analyst following could be influenced by the (lack of) quantitative information in a firm's disclosures. Therefore, in column (2) of Table 5 we replace *Fog* with the ratio of *Numbers/Words* in the MD&A as an explanatory variable for analyst following. In this regression specification, we find that analyst following is negatively related to the prevalence of numbers in the MD&A *Fog* of the contemporaneous 10-K filing (i.e., a statistically-significant negative coefficient on *Fog* the ratio of *Numbers/Words* of -26.6). Thus, this finding supports the notion that firms with fewer disclosed numbers in the MD&A tend to attract more analysts and is also consistent with an information intermediary/processing role for analysts. Furthermore, consistent with *Hypothesis H4* on the possible confounding effects of the prevalence of numbers on the explanatory role of document readability, we find that the previously-significant relation between *Analyst Following* and *Fog* is no longer significant after controlling for *Number/Words* in the regression. Interestingly, the coefficient on the ratio of *Numbers/Words* is almost unchanged and remains statistically significant. These findings are also confirmed in the regression specifications presented in columns (3) and (4) of Table 5 that include the *Bog* (Bonsall et al., 2017) and *Arc* (Hoitash and Hoitash, 2018) variables. Thus, the claimed association between analyst following and disclosure readability documented in Lehavy et al. (2011) does not appear to robust. On the other hand, the prevalence of numbers in the MD&A is more robust and it appears to subsume and explain the Lehavy et al. (2011) *Fog* effect.

5.5. Robustness tests using 10-Q data from pre-EDGAR era

We next turn to the sample of *10-Q* filings for the years 1987-1993 in the pre-EDGAR era. The main sample consists of 20,154 firm-quarter observations with available data to calculate the *Fog* index from the text of the MD&A section of a firm's 10-Q filing and matching *Compustat* data to calculate the key control variables. Table 6 presents the correlations among the main variables. The key correlation of interest captures the possible association between the commonly-used *Fog* index (used to capture disclosure readability) and the prevalence of numbers *within* the text of the MD&A disclosure (captured by the ratio *Num/Words*). Again, consistent with the arguments motivating *Hypothesis H2*, we find that the Pearson correlation is -0.46 and is quite similar to the findings for the *EDGAR 10-K* data. The other correlations are also quite similar to the annual 10-K data presented in Table 2.

5.6. 10-Q evidence on the impact of numbers on the profitability-readability relation

Table 7 presents a replication of the profitability-readability regression originally estimated in Li (2008). We use a sample of 20,254 firm-quarter observations derived from 10-Q filings between 1987 and 1993. In column (1) of Table 7, we re-estimate quarterly regressions similar to the annual regressions presented in column (1) of Table 3. Again, the dependent variable is the *Fog* index for the MD&A section of a firm's 10-Q filing. Consistent with the findings of Li (2008), we find in regression column (1) that the MD&A *Fog* is strongly negatively related to contemporaneous reported firm profitability (i.e., a statistically-significant negative coefficient on *Operating Earnings* of -1.97). In column (2) of Table 7 we include the ratio of *Numbers/Words* in the MD&A as an additional explanatory variable. Again, the explanatory power of the regression dramatically increases (from 13.9% to 32.8%) and the association between firm profitability and *Fog* is no longer significant. This re-affirms our 10-K results that the claimed association between

disclosure readability and firm profitability does not appear to robust.

To help provide a more complete picture of the association between the prevalence of numbers, readability, and profitability, we also estimate another set of regressions in Table 8. The regressions use the same *Compustat* explanatory variables as Table 7, but the dependent variable in Table 8 is the prevalence of numbers within the MD&A (ratio of *Numbers/Words*). Column (1) of Table 8 shows the regression results without including *Fog* as an explanatory variable. We find that the ratio of *Numbers/Words* shows a strong positive association with contemporaneous reported firm profitability (i.e., a statistically-significant positive coefficient on *Operating Earnings* of 0.06). This finding suggests an important link between the level of profitability and the propensity of managers to include quantitative disclosures *within* the MD&A. However, *Fog* is clearly a correlated omitted variable. Therefore, in column (2) of Table 8, we include the MD&A *Fog* as an additional explanatory variable. Again, not surprisingly, the explanatory power of the regression increases from 4.5% to 25.5% and the coefficient on *Fog* is negative and strongly significant. However, the more interesting finding is that the coefficient on *Operating Earnings* remains essentially unchanged and remains strongly significant. Overall, these findings suggest a fundamental and robust link between profitability and the disclosure of quantitative information in the MD&A text. Moreover, this link appears to mediate the previously-claimed relation between profitability and MD&A readability. Overall, these findings are consistent with the issues raised in *Hypothesis H3* and suggest that the benchmark findings in Li (2008) are not as robust as previously thought and the claimed links between firm performance and disclosure readability are more nuanced than indicated by prior research.

5.7. 10-Q evidence on the relation between analyst following and disclosure readability

Table 9 outlines a replication of the analyst following regressions originally presented in

Lehavy et al. (2011). We use a sample of 15,383 firm-quarter observations derived from 10-Q filings between 1987 and 1993 and match the firms with analyst forecast data from I/B/E/S. In column 1 of Table 9, we again estimate a regression that is very similar to Lehavy et al. (2011) using similar *Compustat* explanatory variables. The dependent variable is the *Number of Analysts* who follow a firm during the period. We find in column (1) of Table 9 that analyst following is positively related to the MD&A *Fog* of the contemporaneous 10-Q filing (i.e., a statistically-significant positive coefficient on *Fog* of 0.04). Thus, this finding is consistent with the original finding in Lehavy et al. (2011) that firms with less readable (higher *Fog*) disclosures tend to attract more analysts. In column (2) of Table 9 we replace *Fog* with the ratio of *Numbers/Words* in the MD&A as an explanatory variable for analyst following. In this regression specification, we find that analyst following is negatively related to the prevalence of numbers in the MD&A *Fog* of the contemporaneous 10-Q filing (i.e., a statistically-significant negative coefficient on *Fog* the ratio of *Numbers/Words* of -5.36). This supports the notion that firms with fewer disclosed numbers in the MD&A tend to attract more analysts and also consistent with an information intermediary/processing role for analysts. Finally, in column (3) of Table 9 we include both *Fog* and the ratio of *Numbers/Words* in the MD&A as explanatory variables for analyst following. This specification is motivated by *Hypothesis H4* on the possible confounding effects of the prevalence of numbers in 10-Q filings. We find that the previously-significant relation between *Analyst Following* and *Fog* is no longer significant after controlling for *Number/Words* in the regression. Similar to our main annual 10-K results, the prevalence of numbers in 10-Q filings is more robustly associated with analyst following and it appears to subsume and explain the Lehavy et al. (2011) *Fog* effect.

6. Conclusions and Future Work

The majority of accounting and finance research over the past 50 years has focused on the determinants and use of *quantitative* information for investment, contracting and business decisions. Only more recently has the literature started to better understand and characterize the non-quantitative and textual information in accounting and financial documents and communications. Researchers have made significant advances using textual analysis tools to document the associations between the textual attributes of accounting and financial documents and various economic outcomes. However, existing textual analysis techniques almost universally remove or ignore *quantitative* information from the text of accounting and financial disclosures.

Consistent with the quantitative focus of the traditional literature, we argue that numbers *within* the text of accounting and financial disclosures should be of primary interest to stakeholders and that the surrounding text is likely to play a secondary role of describing the disclosed numbers. Thus, we argue that the prevalence of numbers within disclosures should be related to, if not a primary determinant of, the language used in the text of corporate disclosures. We present empirical evidence that strongly supports this view.

We utilize document datasets from *Wall Street Journal* articles and firms' 10-K and 10-Q filings to document a strong association between the prevalence of numbers in a business document and the complexity and readability of the document. These associations are found even at the sentence level of business documents and disclosures. More importantly, we present empirical evidence that two key findings from the textual analysis literature are affected by the presence of numbers within firms' disclosures. Specifically, the associations between disclosure readability and reported profitability and analyst following (see, Li, 2008, and Lehavy et al., 2011) become largely insignificant after one acknowledges and controls for the prevalence of numbers

within the text of the disclosures. These results are consistent with the view that numbers disclosed within the body of textual disclosures are key correlates that are linked to a firm's disclosure strategies and the outcomes of these strategies. This reinforces the historical view of the central role of quantitative information in accounting and financial reports.

Overall, our findings suggest that ignoring numbers within the text of corporate disclosures can impact researchers' inferences about the links between textual attributes and various accounting, finance and economic outcomes. Our findings can help researchers reinterpret past findings about the possible determinants and outcomes related to the textual attributes of corporate disclosures. Furthermore, our evidence suggests that future research on disclosure readability (and possibly other textual attributes) should explicitly model or control for the prevalence of numbers within the text of disclosures. In summary, we suggest that future textual analysis research should embrace, rather than avoid, numbers.

Our empirical findings also suggest that greater amounts of quantitative information *within* the text of disclosures are associated with a higher quality disclosure. This finding aligns with the financial statement measure of Chen et al. (2015) which is based on greater disaggregation of financial numbers presented in the income statement, balance sheet and statement of cash flows. On the other hand, our findings and those of Chen et al. (2015) contrast with the Hoitash and Hoitash (2018) finding that the presence of more numbers (based on XBRL coding) reported in the financial statements is associated with greater accounting complexity (i.e., lower accounting quality).

References

Allee, K., and M. DeAngelis, 2015. The structure of voluntary disclosure narratives: evidence from tone dispersion. *Journal of Accounting Research* 53, 241–74.

Asay, S., B. Elliott, and K. Rennekamp, 2017. Disclosure readability and the sensitivity of investors' valuation judgments to outside information. The *Accounting Review*.

Bonsall, S., A. Leone, B. Miller, and K. Rennekamp, 2017. A plain English measure of financial reporting readability. *Journal of Accounting and Economics* 63, 329–57.

Bonsall, S., and B. Miller, 2017. The impact of narrative disclosure readability on bond ratings and the cost of debt capital. *Review of Accounting Studies*.

Bushee, B., Gow, I., Taylor, D., 2018. Linguistic complexity in firm disclosures: obfuscation or information? *Journal of Accounting Research*.

Cazier, R., and R. Pfeiffer, 2016. Why are 10-K filings so long? *Accounting Horizons* 30, 1-21.

Chen, S., B. Miao, and T. Shevlin, 2015. A new measure of disclosure quality: The level of disaggregation of accounting data in annual reports. *Journal of Accounting Research* 53, 1017-54.

Das, S., 2014. Text and context: language analytics in finance. *Foundations and Trends in Finance* 8, 145–261.

De Franco, G., O. Hope, D. Vyas, and Y. Zhou, 2015. Analyst report readability. *Contemporary Accounting Research* 32, 76–104.

Drake, M., D. Roulstone, and J. Thornock, 2016. The usefulness of historical accounting reports. *Journal of Accounting and Economics* 61, 448-64.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017a. The evolution of 10-K textual disclosure: evidence from latent dirichlet allocation. *Journal of Accounting and Economics*.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017b. What have we learned and where do we go with textual research? A discussion of Cazier and Pfeiffer. *Journal of Financial Reporting*.

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

Friberg, R., and T. Seiler, 2017. Risk and ambiguity in 10-Ks: An examination of cash holding and derivatives use. *Journal of Corporate Finance* 45, 608-631.

Ginesti, G., C. Drago, R. Macchioni, and G. Sannino, 2018. Female board participation and annual report readability in firms with boardroom connections. *Gender in Management*.

Guay, W., D. Samuels, and D. Taylor, 2016. Guiding through the fog: financial statement complexity and voluntary disclosure." *Journal of Accounting and Economics* 62, 234–69.

Gunning, R., 1952. The technique of clear writing. New York, NY: *McGraw-Hill Intl Book Co*.

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

Hoitash, R., U. Hoitash, A. Kurt, and R. Verdi, 2017. An XBRL-based measure of financial statement comparability. *MIT Sloan School of Management Working Paper.*

Hoitash, R., and U. Hoitash, 2018. Measuring accounting reporting complexity with XBRL. *The Accounting Review*.

Hooghiemstra, R., Y. Kuang, and B. Qin, 2017. Does obfuscating excessive CEO pay work? The influence of remuneration report readability on say-on-pay votes. *Accounting and Business Research*.

Hwang, B., and H. Kim, 2017. It pays to write well. *Journal of Financial Economics*.

Inger, K., M. Meckfessel, M., Zhou, and W. Fan, 2018. An examination of the impact of tax avoidance on the readability of tax footnotes. *Journal of the American Taxation Association*.

Kearney, C. and S. Liu, 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33, 171–85.

Kothari, S., X. Li, and J. Short, 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review* 84, 163970.

Laksmana, I., W. Tietz, and Y. Yang, 2012. Compensation discussion and analysis (CD&A): readability and management obfuscation. *Journal of Accounting and Public Policy* 31, 185-203.

Lang, M., and L. Stice-Lawrence, 2015. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics* 60, 110–35.

Lawrence, A., 2013. Individual investors and financial disclosure. *Journal of Accounting and Economics* 56, 130-147.

Lee, Y., 2012. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research*.

Lehavy, R., F. Li, K. Merkley, 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review* 86, 1087–115.

Leuz, C. and Wysocki, 2016. The economics of disclosure and financial reporting regulation: evidence and suggestions for future research. *Journal of Accounting Research* 54, 525-622.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45, 221–47.

Li, F., 2011. Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature* 29.

Lim, E., K. Chalmers, and D. Hanlon, 2018. The influence of business strategy on annual report readability. *Journal of Accounting and Public Policy*.

Lo, K., F. Ramos, and R. Rogo, 2017. Earnings management and annual report readability. *Journal of Accounting and Economics* 63.

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

Loughran, T., and B. McDonald, 2014. Measuring readability in financial disclosures. *Journal of Finance* 69, 1643–71.

Loughran, T., and B. McDonald, 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54, 187–230.

Lundholm, R., R. Rogo, and, J. Zhang, 2014. Restoring the tower of Babel: how foreign firms communicate with US investors. *The Accounting Review* 89, 1453-1485.

Miller, B., 2010. The effects of reporting complexity on small and large investor trading. *The Accounting Review* 85, 2107–43.

Rennekamp, K., 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research* 50, 1319–54.

Siano, F., 2019. Finding the Narrative in the Numbers: Long-Term Investors' Demand for Accounting Information. *Boston University – Questrom School of Business.*

You, H., and X. Zhang, 2009. Financial disclosure complexity and investor underreaction to 10-K information. *Review of Accounting Studies* 14, 559–86.

Table 1: Differences in Textual Attributes Across Quantitative and Non- Quantitative Sentences in *WSJ* News Articles

		Category of Wall Street Journal Article (Year = 1992)				
		Main News Articles (1,095 news articles)				*Holdout Sample*
		Economic News & Indicators	International Trade	Monetary Policy	Tracking the Economy	Corporate Earnings Reports
Number of articles		776	238	30	51	923
Mean ratio #'s/words	Full Article	9.0%	11.6%	3.4%	13.1%	14.2%
Mean # words per sentence	Sentence includes #'s	11.26*	11.73	13.35*	19.75*	11.89
	Sentence without #'s	10.68	11.11	11.79	10.46	11.94
Mean # complex words per sentence	Sentence includes #'s	2.75*	2.09*	3.05*	2.64*	2.08*
	Sentence without #'s	2.90	2.87	3.27	4.17	3.37
Mean *Fog* of Sentences	Sentence includes #'s	16.96*	16.48*	16.69*	14.68*	14.77*
	Sentence without #'s	20.57	21.38	20.51	20.27	22.66

This table presents across sub-sample comparisons of sentence-level textual attributes for a sample of *Wall Street Journal* news articles from Lexis-Nexis for the calendar year 1992. The main sample (*Main News Articles)* consists of 1,095 news articles that contain both text and financial information. Sentences within each article are divided into quantitative sentences (includes numbers) and nonquantitative sentences (without numbers). The *mean ratio #'s/Words* captures the average ratio of "number of numbers" to "number of words" within each category (quantitative vs. nonquantitative sentences). *Complex words* are defined as words with more than 2 syllables. The *mean ratio # complex words per sentence* captures the average number of *complex words per sentence* within each sentence category (quantitative vs. nonquantitative sentences). *Fog* is the Gunning (1952) *Fog* index calculated as 0.4*(words per sentence + percent of complex words) for each sentence in a document for each sentence category (quantitative vs. nonquantitative sentences). * indicates significant differences in mean of a variable across quantitative (sentences with #'s) and non-quantitative (sentences with #'s) subsamples at <0.01 level.

Table 2: Correlations Among Key Variables for 10-K MD&A Sample

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Fog	1.000													
(2) Bog	0.475*	1.000												
(3) Arc	0.168*	0.095*	1.000											
(4) #'s/Words	-0.446*	-0.377*	0.015	1.000										
(5) #'s/WordsQuant	-0.492*	-0.379*	-0.075*	0.605*	1.000									
(6) #'s/Sentences	-0.220*	-0.265*	0.048*	0.951*	0.475*	1.000								
(7) 1/Words	-0.099*	-0.230*	-0.126*	0.146*	0.219*	0.071*	1.000							
(8) 1/Sentences	0.004	-0.184*	-0.120*	0.086*	0.146*	0.042*	0.984*	1.000						
(9) Earnings	-0.102*	-0.217*	0.149*	0.164*	0.128*	0.148*	0.011*	-0.004	1.000					
(10) Earnings Volatility	0.055*	0.163*	-0.212*	-0.133*	-0.093*	-0.126*	0.009	0.018*	-0.516*	1.000				
(11) Size	0.205*	0.169*	0.379*	-0.176*	-0.202*	-0.132*	-0.190*	-0.170*	0.334*	-0.206*	1.000			
(12) MTB	0.034*	0.129*	-0.177*	-0.100*	-0.014*	-0.098*	0.020*	0.025*	-0.225*	0.341*	0.193*	1.000		
(13) Returns Volatility	-0.040*	0.081*	-0.259*	-0.044*	-0.009	-0.051*	0.021*	0.021*	-0.390*	0.365*	-0.357*	0.167*	1.000	
(14) Analysts	0.146*	0.109*	0.207*	-0.158*	-0.145*	-0.129*	-0.123*	-0.111*	0.212*	-0.134*	0.684*	0.104*	-0.221*	1.000

This table shows pairwise Pearson correlations between the key variables used in the 10-K disclosure analyses. Sample of 77,144 annual observations from EDGAR filings from 1994-2017. *Fog* is the Gunning (1952) Fog index calculated as 0.4*(avg. words per sentence + percent of complex words) of the MD&A section of a firm's 10-K filing. *Bog* is the StyleWriter Bog index used in Bonsall et al. (2017). *Arc* is the natural logarithm of the accounting reporting complexity measure developed in Hoitash and Hoitash (2018). The *#'s/Words* ratio is calculated as the number of numbers over the number of words (excluding numbers) within the MD&A text of the 10-K filing. The *#'s/WordsQuant* ratio is computed using only the words found in the quantitative portion of text. The *#'s/Sentences* ratio is calculated as the average number of numbers per sentence in the MD&A text of the 10-K filing. *1/Words* is the inverse of the number of words (excluding numbers) contained in the MD&A section of a firm's 10-K filing. *1/Sentences* is the inverse of the number of sentences contained in the MD&A section of a firm's 10-K filing. *Earnings* are the contemporaneous annual Compustat operating earnings scaled by total assets. *Earnings Volatility* is the standard deviation of scaled annual operating earnings for the last 4 years. *Size* is the natural logarithm of beginning of period market value of equity. *MTB* is the beginning of period market value of equity divided by its book value. *Returns Volatility* is the standard deviation of buy-and-hold CRSP stock returns over the last 12 months. *Analysts* is the natural logarithm of the number of analysts following a firm in the first consensus date following the 10-K filing date. The aforementioned explanatory variables are winsorized at the 1% level. * shows significance at the .01 level.

Table 3:
Replication of Li (2008) - The Association between 10-K MD&A Readability (*Fog*) and Annually Reported Profitability

Fiscal Years: 1994-2017 (unit of analysis: firm-year)

Explanatory Variable	Sign Li (2008)	Dependent Variable: 10-K MD&A *Fog*				
Operating Earnings (a)	**(-)**	**-0.77*** [15.6]**	**-0.49*** [9.0]**	**-0.11** [2.6]**	**-0.14* [1.8]**	**-0.15** [2.0]**
Earnings Volatility	(+)	0.14** [2.4]	0.01 [0.2]	-0.05 [-0.8]	0.08 [0.7]	0.08 [0.7]
Size	(+)	0.16*** [19.2]	0.12*** [14.4]	0.00 [0.1]	0.02 [0.9]	0.02 [0.8]
MTB	(+)	-0.06*** [9.4]	-0.06*** [8.4]	-0.03*** [4.6]	-0.02** [2.0]	-0.02* [1.8]
Returns Volatility	(+)	0.62*** [6.4]	0.25*** [2.8]	0.03 [0.4]	0.35*** [3.1]	0.36*** [3.2]
Age	(-)	-0.00* [1.9]	-0.00 [1.2]	-0.00 [0.6]	-0.00 [0.2]	-0.00 [0.1]
Special Items	(-)	-0.54*** [6.5]	-0.35*** [4.5]	-0.22*** [3.5]	-0.25* [1.9]	-0.26** [2.0]
Business Segments	(?)	0.05 [1.6]	0.08*** [3.0]	0.02 [0.8]	-0.06 [1.5]	-0.06 [1.4]
Geographic Segments	(?)	-0.01 [0.5]	-0.02 [0.9]	0.01 [0.5]	0.02 [0.3]	0.02 [0.4]
Delaware	(+)	0.20*** [6.6]	0.16*** [5.7]			
Numbers/Words	**Our HP (-)**		**-43.92*** [35.4]**	**-39.99*** [22.8]**	**-33.75*** [12.4]**	**-33.69*** [12.3]**
1/Words	(?)		59.02 [1.5]	119.12*** [2.9]	-105.50 [0.8]	-106.14 [0.8]
Arc	(?)				0.06*** [2.8]	
Arc Scaled	(?)					-0.04 [0.5]
Observations		63,119	63,119	63,693	16,716	16,716
Adjusted R-squared		0.22	0.33	0.70	0.75	0.75
SE Cluster		Firm	Firm	Firm	Firm	Firm
Industry Fixed Effects		YES	YES	YES	YES	YES
Year Fixed Effects		YES	YES	YES	YES	YES
Firm Fixed Effects		NO	NO	YES	YES	YES

T-statistics reported in square parentheses. The superscript stars (*, **, ***) indicate significance at 10%, 5%, and 1% levels respectively. Variable definitions for annual data outlined in Table 2 and Section 4.2. *Industry Fixed Effects* are based on Fama-French 17-industry definitions.

Table 4: The Association between 10-K Filing Readability (*Bog*) and Annually Reported Profitability [Fiscal Years: 1994-2017 (unit of analysis: firm-year)]

Explanatory Variable	Pred. Sign	Dependent Variable: 10-K Filing *Bog*				
Operating Earnings (a)	**(-)**	**-3.81*** **	**-3.00*** **	**-0.92*** **	**-1.01*** **	**-1.03*** **
		[-14.4]	**[-9.2]**	**[-4.4]**	**[-3.2]**	**[-3.2]**
Earnings Volatility	(+)	0.02	-0.32	0.24	0.43	0.42
		[0.1]	[-1.2]	[0.8]	[0.8]	[0.8]
Size	(+)	0.70***	0.56***	0.20***	0.33***	0.35***
		[17.2]	[14.2]	[3.8]	[3.8]	[4.0]
MTB	(+)	-0.29***	-0.26***	-0.15***	-0.23***	-0.23***
		[-8.8]	[-8.0]	[-6.1]	[-4.9]	[-5.0]
Returns Volatility	(+)	6.20***	4.79***	1.49***	1.73***	1.78***
		[13.8]	[11.1]	[4.9]	[3.6]	[3.7]
Age	(-)	-0.03***	-0.03***	-0.05***	0.02	0.02
		[-6.3]	[-5.4]	[-2.6]	[0.9]	[1.0]
Special Items	(-)	-2.80***	-2.01***	-1.93***	-1.36**	-1.41**
		[-6.8]	[-4.9]	[-6.7]	[-2.3]	[-2.4]
Business Segments	(?)	1.32***	1.37***	0.81***	0.37**	0.38**
		[10.3]	[11.1]	[7.3]	[2.0]	[2.0]
Geographic Segments	(?)	0.20*	0.13	0.17*	0.41**	0.42**
		[1.8]	[1.2]	[1.6]	[2.4]	[2.5]
Delaware	(+)	0.98***	0.86***			
		[6.8]	[6.1]			
Numbers/Words	**Our HP (-)**		**-115.26*** **	**-82.69*** **	**-43.73*** **	**-43.24*** **
			[-22.5]	**[-17.0]**	**[-6.8]**	**[-6.7]**
1/Words	(?)		-1,576.05***	-616.03***	-13.21	34.89
			[-13.2]	[-6.2]	[-0.1]	[0.2]
Arc	(?)				0.23**	
					[2.2]	
Arc Scaled	(?)					-0.43
						[-0.5]
Observations		61,580	61,580	62,128	16,214	16,214
Adjusted R-squared		0.40	0.45	0.79	0.84	0.84
SE Cluster		Firm	Firm	Firm	Firm	Firm
Industry Fixed Effects		YES	YES	YES	YES	YES
Year Fixed Effects		YES	YES	YES	YES	YES
Firm Fixed Effects		NO	NO	YES	YES	YES

T-statistics reported in square parentheses. The superscript stars (*, **, ***) indicate significance at 10%, 5%, and 1% levels respectively. Variable definitions for annual data outlined in Table 2 and Section 4.2. *Industry Fixed Effects* are based on Fama-French 17-industry definitions.

Table 5:
Replication of Lehavy et al. (2011) - Association between Analyst Following & 10-K MD&A Readability (*Fog* & *Bog indices*). Sample 1994-2017 (unit of analysis: firm-year)

Explanatory Variable	Sign Lehavy et al. (2011)	Dependent Variable: I/B/E/S Analyst Following			
Fog	**(+)**	**0.06****	**-0.01**	**0.01**	**0.00**
		[2.4]	**[0.3]**	**[0.1]**	**[0.0]**
Bog	(?)			-0.03*	-0.04*
				[1.8]	[-0.9]
Size	(+)	2.96***	2.95***	3.56***	3.54***
		[68.1]	[68.4]	[51.0]	[52.1]
Growth	(+)	0.49***	0.48***	0.75***	0.76***
		[7.6]	[7.5]	[4.8]	[4.9]
Business Segments	(-)	-0.67***	-0.67***	-0.59***	-0.60***
		[5.8]	[5.8]	[3.5]	[3.6]
10-K News	(+)	2.56***	2.42***	2.51*	2.46*
		[4.2]	[3.9]	[1.9]	[1.9]
Adv	(+)	5.95***	5.73***	8.23***	8.29***
		[5.2]	[5.0]	[4.2]	[4.3]
R&D	(+)	1.28***	1.11***	1.69***	1.77***
		[5.6]	[4.8]	[3.5]	[3.7]
Returns Volatility	(+)	2.53***	2.20***	7.92***	7.86***
		[5.9]	[5.1]	[6.8]	[6.8]
InstInv	(+)	0.61***	0.61***	-0.16	-0.16
		[3.0]	[3.0]	[-0.5]	[-0.5]
Num/Words	**Our HP (-)**		**-26.60*****	**-45.69*****	**-45.28*****
			[6.1]	**[4.9]**	**[4.9]**
1/Words	(?)		-32.97	749.60**	1,143.34**
			[-0.2]	[2.2]	[2.4]
Arc	(?)			-0.31*	
				[-1.7]	
Arc Scaled	(?)				-1.60
					[-1.0]
Observations		44,370	44,370	14,125	14,125
Adjusted R- squared		0.62	0.63	0.66	0.66
SE Cluster		Firm	Firm	Firm	Firm
Industry Fixed Effects		YES	YES	YES	YES
Year Fixed Effects		YES	YES	YES	YES

T-statistics reported in square parentheses. The superscript stars (*, **, ***) indicate significance at 10%, 5%, and 1% levels respectively. Variable definitions for annual data outlined in Table 2 and Section 4.2. *Industry Fixed Effects* are based on Fama-French 17-industry definitions.

Table 6: Correlations Among Key Variables for 10-Q MD&A Sample

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Fog	Num/Words	1/Words	Num/Sent	1/Sent	Operating Earnings scaled	Operating Earnings Volatility	Log Size	M/B
(1)	Fog	1.00								
(2)	Num/Words	(0.46)	1.00							
(3)	1/Words	0.39	0.13	1.00						
(4)	Num/Sent	(0.15)	0.19	(0.09)	1.00					
(5)	1/Sent	0.44	0.04	0.65	(0.08)	1.00				
(6)	Operating Earnings scaled	(0.06)	0.06	0.09	0.01	0.05	1.00			
(7)	Operating Earnings Volatility	0.12	(0.06)	0.04	(0.02)	0.02	(0.17)	1.00		
(8)	Log Size	(0.22)	0.01	(0.20)	0.02	(0.12)	0.33	(0.31)	1.00	
(9)	M/B	0.03	(0.01)	0.07	(0.01)	0.03	0.31	0.20	0.25	1.00

This table shows Pearson correlations between the key variables used in the 10-Q disclosure analyses. Sample of 20,154 quarterly observations from pre-EDGAR filings from 1987-1993. *Fog* is the Gunning (1952) *Fog* index calculated as 0.4*(avg. words per sentence + percent of complex words) of the MD&A section of a firm's 10-Q filing. The *Num/Words* ratio is calculated as the number of numbers over the number of words (excluding numbers and stop words) within the MD&A text of the 10-Q filing. *1/Words* is the inverse of the number of words (excluding numbers and stop words) contained in the MD&A section of a firm's 10-Q filing. The *Num/Sent* ratio is calculated as the average number of numbers per sentence in the MD&A text of the 10-Q filing. *1/Sent* is the inverse of the number of sentences contained in the MD&A section of a firm's 10-Q filing. *Operating earnings* are the contemporaneous quarterly Compustat operating earnings scaled by total assets. *Operating earnings volatility* is the standard deviation of scaled quarterly operating earnings for the last 12 quarters. *Size* is the natural logarithm of beginning of period market value of equity. *M/B* is the beginning of period market value of equity divided by its book value. section of firms' 10-Q filings. The aforementioned explanatory variables are winsorized at the 1% level. Correlations with absolute magnitude greater than 0.03 are statistically significant with p-value <0.01.

Table 7: Replication of Li (2008) of the Association between 10-Q MD&A Readability and Quarterly Reported Profitability

Explanatory Variable	Pred. Sign from Li (2008)	Dependent Variable: 10-Q MD&A *Fog*	
Operating Earnings (q)	(-)	-1.97*** [-2.9]	0.52 [0.9]
Operating Earnings Variability	(+)	3.72*** [3.7]	0.91 [1.0]
Size	(-)	-0.29*** [-27.7]	-0.29*** [-31.5]
MTB	(+)	0.24*** [11.2]	0.22*** [11.4]
Numbers/Words	(-) Our Hypotheses H1 and H2		-39.31*** [-75.3]
FF-17 Industry Fixed Effects		Included	Included
Year-Quarter Fixed Effects		Included	Included
# Obs.		20,154	20,154
Adj. R^2		13.9%	32.8%

This table shows the regression results of the *Fog* index (10-Q MD&A) on the Compustat determinants from Li (2008) and period and industry fixed effect. Sample of quarterly observations from pre-EDGAR filings from 1987-1993. *Fog* is the Gunning (1952) *Fog* index calculated as 0.4*(avg. words per sentence + percent of complex words) of the MD&A section of a firm's 10-Q filing. The *Numbers/Words* ratio is calculated from the MD&A text of the 10-Q filing. *Operating earnings* are the contemporaneous quarterly Compustat operating earnings. *Operating earnings volatility* is the standard deviation of quarterly operating earnings for the last 12 quarters. *Size* is the natural logarithm of beginning of period market value of equity. *MTB* is the beginning of period market value of equity divided by its book value. section of firms' 10-Q filings. The aforementioned explanatory variables are winsorized at the 1% level. *Industry Fixed Effects* are based on Fama-French 17-industry definitions. All regressions are estimated with an intercept included, but the intercept is not reported. Robust t-statistics reported in [] parentheses. *** indicates significance at <0.01.

Table 8: The Association between the Prevalence of Numbers within MD&A and Quarterly Reported Profitability

Explanatory Variable	Pred. Sign	Dependent Variable: *Numbers/Words* in 10-Q MD&A	
Operating Earnings (q)	(+)	0.06***	0.05***
		[7.9]	[7.4]
Operating Earnings Variability	(-)	-0.07***	-0.05***
		[6.0]	[-4.8]
Size	(?)	-0.00	-0.02***
		[-0.3]	[-14.8]
MTB	(?)	-0.001**	0.001***
		[-2.3]	[3.3]
Fog	(-) Our Hypotheses H1 and H2		-0.006***
			[-75.3]
FF-17 Industry Fixed Effects		Included	Included
Year-Quarter Fixed Effects		Included	Included
# Obs.		20,154	20,154
Adj. R^2		4.5%	25.5%

This table shows the regression results of the *Numbers/Words* (10-Q MD&A) on Compustat determinants from Li (2008) and period and industry fixed effect. Sample of quarterly observations from pre-EDGAR filings from 1987-1993. The *Numbers/Words* ratio is calculated from the MD&A text of the 10-Q filing. *Fog* is the Gunning (1952) *Fog* index calculated as 0.4*(avg. words per sentence + percent of complex words) of the MD&A section of a firm's 10-Q filing. *Operating earnings* are the contemporaneous quarterly Compustat operating earnings. *Operating earnings volatility* is the standard deviation of quarterly operating earnings for the last 12 quarters. *Size* is the natural logarithm of beginning of period market value of equity. *MTB* is the beginning of period market value of equity divided by its book value. section of firms' 10-Q filings. The aforementioned explanatory variables are winsorized at the 1% level. *Industry Fixed Effects* are based on Fama-French 17-industry definitions. All regressions are estimated with an intercept included, but the intercept is not reported. Robust t-statistics reported in [] parentheses. *** indicates significance at <0.01, and ** indicates significance at <0.05.

Table 9: The Association between Analyst Following and 10-Q MD&A Readability

Explanatory Variable	Predicted Sign from Lehavy et al. (2011)	Dependent Variable is I/B/E/S *Analyst Following*		
Operating earnings (q)	(?)	-15.25*** [-7.9]	-14.85*** [-7.7]	-14.88*** [-7.7]
Operating Earnings Volatility	(+)	18.81*** [6.4]	18.47*** [6.3]	18.51*** [6.3]
Size	(+)	4.68*** [61.1]	4.68*** [62.2]	4.68*** [60.2]
MTB	(+)	-0.58*** [-10.1]	-0.58*** [-3.6]	-0.58*** [-2.8]
FOG	(+)	0.04** [2.5]		0.02 [0.8]
Numbers/Words	(-) From Our Hypothesis 4		-5.36*** [-3.6]	-4.71*** [-2.8]
FF-17 Industry Fixed Effects		Included	Included	Included
Year-Quarter Fixed Effects		Included	Included	Included
# Obs.		15,383	15,383	15,383
Adj. R^2		67.3%	67.3%	67.3%

This table shows the regression results of the I/B/E/S *Analysts Following* (# analysts issuing at least one forecast for the period) on Compustat determinants from Lehavy et al. (2011) and period and industry fixed effect. The data sample includes quarterly 10-Q filings from pre-EDGAR filings from 1987-1993. *Fog* is the Gunning (1952) *Fog* index calculated as 0.4*(avg. words per sentence + percent of complex words) of the MD&A section of a firm's 10-Q filing. The *Numbers/Words* ratio is calculated from the MD&A text of the 10-Q filing. *Operating earnings* are the contemporaneous quarterly Compustat operating earnings. *Operating earnings volatility* is the standard deviation of quarterly operating earnings for the last 12 quarters. *Size* is the natural logarithm of beginning of period market value of equity. *MTB* is the beginning of period market value of equity divided by its book value. section of firms' 10-Q filings. The aforementioned explanatory variables are winsorized at the 1% level. *Industry Fixed Effects* are based on Fama-French 17-industry definitions. All regressions are estimated with an intercept included, but the intercept is not reported. Robust t-statistics reported in [] parentheses. *** indicates significance at <0.01, and ** indicates significance at <0.05.