# Missing narratives: An analysis of biases in sample selection and variable choice in textual analyses*


Andrea Bafundi

*Universidad Carlos III de Madrid*

Beatriz García Osma

*Universidad Carlos III de Madrid*

Encarna Guillamón-Saorín

*Universidad Carlos III de Madrid*


August 2018


* **Corresponding author:** Andrea Bafundi. Department of Business Administration. Universidad Carlos III de Madrid. c/Madrid 126. 28903 Getafe, Madrid (SPAIN). E-mail. abafundi@emp.uc3m.es. Tel.: (34) 91 624 95 73. Fax: (34) 91 624 96 07.



We are grateful for helpful comments and discussion with Hyungjin Cho. We acknowledge financial support by the Spanish Ministry of Innovation and Science (ECO2013-48328, ECO2016-77579 and ECO2016-00501), FEDER (UNC315-EE-3636), and the Madrid Government Region (S2015/HUM-3417).

**Missing narratives: An analysis of biases in sample selection and variable choice in textual analyses**

**Abstract**

We study plausible biases in textual analysis studies of 10-K documents. The study of financial narratives using automated procedures is a relatively novel development in accounting and finance. Therefore, standardized methods to collect and systematically analyze these data are yet to be developed. We provide detailed step-by-step guidance on how to download and prepare these files to analyze, and study the biases introduced by a number of decisions regarding sample construction, data preparation, and variable choice. In particular, we focus on two widely studied properties of financial narratives: their Tone and Readability. We document that a number of these choices introduce significant biases into the samples studied, as well as induce differences in average observed Tone and Readability. Our results also indicate that a non-trivial proportion of the *Edgar* database population is missing from the textual analyses being conducted.

Keywords: *Narrative disclosure, Textual analysis, Readability, Tone, Sample selection, Biased samples.*

## 1. Introduction

The analysis of financial narratives has gained significant traction in recent years, with the development of a number of new proxies to measure qualitative attributes of corporate communication. Most of this research has focused thus far on two qualitative attributes: Tone (e.g. Henry and Leone 2016; Loughran and McDonald 2011), and Readability (e.g. Bonsall et al. 2017; Li 2008; Bonsall and Miller 2017), with a number of papers trying to understand the drivers and consequences of heterogeneity in narrative disclosures (Lo et al. 2017; Kim et al. 2017). While a number of commercial data providers permit access to accounting and finance data with a high assurance of construct validity, the research on narrative disclosure is still in its infancy, and little is known of the biases introduced by researchers through the use of non-standardized procedures.

We study two plausible biases that may impact existing research in the area and that are introduced by 1) sample selection and 2) variable definition choices. To do so, we proceed as follows. First, we review the existing literature on narrative disclosure to identify sample selection choices made by prior researchers, as well as the most commonly used measures of narrative disclosure (focusing on Tone and Readability). Second, we download from the *Edgar* database all available fillings and replicate the steps taken by prior studies, to identify any significant sample selection biases plausibly introduced by downloading and parsing procedures. Lastly, we compute a number of the most common Tone and readability measures, to understand whether the same firms are differently classified depending on the measures chosen.

In our final set of analyses, we build on our literature review and classify prior work on narrative disclosure into the main areas that have attracted the most interest thus far, dealing with firm performance, earnings quality, market reactions, and analysts' effects. Once these areas are identified, we download the main variables necessary to study associations between our narrative disclosure measures of Tone and Readability and critical constructs used in each of these lines of research. This process permits understanding whether any additional differences exist between the samples and variables studied in these primary areas of study.

Our focus is on 10-K reports, which are also the focus of most prior studies, albeit we also provide detailed evidence on other fillings available in *Edgar*. For our initial analyses, we download all 10-K documents available in the *Edgar* database, and come to an initial sample (Sample 0) following the steps detailed in Appendix A. Sample 0 represents the maximum group of observations from 10-K files that could be analysed without imposing any "*external*" constraints on the data, such as merging the files with commercial accounting and finance databases. This sample is the most extensive dataset that could be used, for example, by researchers interested in pure natural language processing of the documents available, perhaps to create novel measures of narrative disclosure quality, or by researchers in the fields of law, or psychology who would not be inclined to merge these files with accounting or financial data. This initial dataset is furtherly set to include only our section of interest of the 10-K (Sample I). Sample II is the result of matching *Edgar* with *Crsp* and *Compustat* conditional on a minimum presence of firm identifiers to identify firms across the databases such Central Index Code (cik), gvkey, fiscal year and month end). Sample III represents the maximum sample of 10-K files that could be used in a study in the field of accounting and finance, where we only impose as a restriction that minimum commonly used variables are available (such as total assets, revenues, and short-long term debt). Finally, we create different Sample IV, which include four independent sub-samples[1]. To create these four final samples, we impose the minimum requirement of variables to conduct analyses in the four

[1] We require only those variables necessary to run that study whose sub-sample belongs to.

identified fields that are the focus of prior literature: firm performance (Sample IV. Performance), earnings quality (Sample IV. Earnings), market to narrative disclosures (Sample IV. Market), and analysts' usage of qualitative information (Sample IV. Analysts).

<<INSERT TABLE 1 HERE>>

To identify the aforementioned four areas, we conduct a review of recent research. Our literature review is based on a systematic search of published academic articles related to textual analysis on narrative disclosures. We restrict the search to the leading journals in Accounting and Finance for the last decade.[2] In particular, we search by keywords associated with textual or narrative analysis (i.e., textual analysis, readability, language, narrative, and tone).[3] Table 1 Panel A details the number of articles found using this procedure. We review these articles and exclude those that do not refer to firm-issued documents as they focus, for example, on press news, analysts' reports, or the U.S. *Stock Exchange Commission* (SEC) comment letters. Articles that focus on textual analysis of firm-produced documents are reviewed to extract information on the period of analysis, number of observations, data sources, documents or sections of documents analysed (e.g., 10Ks, 10Q, press releases, MD&A, conference calls), measures of textual analysis (e.g., readability, tone, length, similarity) and the software employed to perform the analyses. As can be seen in Table 1 Panel B, a total of 56 articles are identified following this procedure. Table

[2] We focus on the period 1997-2017, and consider only published papers (leaving out those that are "in press") in the following journals: Journal of Accounting and Economics (JAE), Journal of Accounting Research (JAR), The accounting Review (TAR), Review of Accounting Studies (RAST), Contemporary Accounting Research (CAR), Accounting Organizations and Society (AOS) and European Accounting Review (EAR), Journal of Finance (JF), Journal of Financial Economics (JFE), and Review of Financial Studies (RFS).

[3] We search using the platform Science Direct which features sophisticat ed search and retrieval tools to facilitate the search of academic articles. This platform covers three of the journals included in our study (JFE, JAE, and AOS). We use the editor's website to perform the search for the rest of the journals: Wiley Online Library for JF, JAR and CAR, Springer for RAST, Taylor & Francis Online for EAR, American Accounting Association Library for TAR, and Oxford Academic for RFS.

1 Panel C also reveals the growing interest in the study of financial narratives, with ever increasing numbers of papers published throughout the decade. We detail the articles analyzed in Panel D. This review also serves to document the data gathering process of each study and the parsing procedures used.

Using our samples and a number of proxies for Tone and Readability, we obtain the following key findings. First, some little-discussed researcher-choices, such as whether to include or remove amended and late filings, or how to deal with duplicates, significantly affect sample sizes. Existing research is almost always silent on these first steps of the sample construction procedure, which can lead to significant differences in final samples. Second, by requiring that accounting and financial data is available in commercial databases, almost two-thirds of all valid observations are lost. We consider a valid observation one that contains full narrative data sufficient to create Tone and readability measures. Third, the narratives lost when merging the *Edgar* baseline dataset with *Compustat* and *Crsp* are significantly different from those that are retained. This problem is further compounded when additional data constraints are imposed on the data. Finally, we document that the choice of variable measurement may lead to different inferences. Different final outputs might represent a concern for researchers interested in the economic interpretation of the sign and size of their coefficients.

We contribute to the prior literature in a number of ways. First, we provide detailed step-by-step guidance on how to download and prepare *Edgar* fillings for analysis. Second, we document the biases introduced by a set of possible decisions regarding sample construction, data preparation, and variable choice. Our results indicate that a non-trivial proportion of the *Edgar* population is missing from the textual analyses being conducted. On average, these missing narratives appear significantly different from the ones that are retained in standard accounting and finance studies. In particular, we provide evidence of significant differences in two widely studied properties of financial narratives: their Tone and Readability. Further, we document that a number of the variable choices made by researchers also may induce differences in average observed Tone and Readability. Thus, our results question the generalizability of studies in narrative disclosure and also, set the question of what are these missing firms discussing in their annual reports. It appears that existing research may be ignoring much of what is being said.

The remainder of the paper is structured as follows. Section 2 briefly summarizes prior literature. Section 3 describes the method and data used in the study while section 4 presents our main results on sample selection biases in textual analyses. Section 5 identifies the major lines of research in the existing literature and further studies plausible biases introduced by demanding data merges of commercial databases. Finally, section 6 concludes.

## 2. Prior literature

The majority of research in financial and accounting journals is archival based, and it requires large amount of data usually provided by commercial databases. These underlying data deserve considerable attention as the validity and power of the results depends on the quality of the prepared dataset. An ample prior literature in accounting and finance studies and documents significant biases including heterogeneous database coverage that usually leads to smaller firms being less represented in samples as well as survivorship biases, errors in databases, or the use of different metrics to capture underlying constructs, such as industry composition.

One of the main problems described in the literature is the existence of database coverage issues. Databases such as *Crsp* and *Compustat*, while they are doubtless the most complete ones available, have been studied by prior research, documenting the existence of survivorship biases.

These studies find that companies excluded from databases are usually small (García Lara et al. 2006; Mutchler and Shane 1995), companies that subsequently are involved in bankruptcy or receive auditing qualified opinions and companies that are more likely to be audited by non-Big-Eight firms (Mutchler and Shane 1995). Similarly, a study of bias detected in *ExecuComp* shows that firms included in this database tend to be larger, more complex, followed by more analysts and have less concentrated institutional ownership than other firms (Cadman et al. 2010).

Another issue documented in prior work is the existence of differences in data across databases. Schwarz and Potter (2016) report a lack of overlap between the *SEC Mutual Fund Portfolios* and the *Crsp Mutual Funds Database Portfolios* and when they merge these two datasets with *Thomson Reuters* data, only 39% of portfolios overlap in all three sources for the same universe of funds. These differences are mainly due to voluntary reporting portfolios which may be included in *Thomson Reuters* but not reported to the SEC, and that may or may not be included in *Crsp*. These studies provide evidence that the database choice influences empirical results.

Industry classification is an important element in the methodology of accounting research. Researchers have generally used the *Standard Industrial Classification* (SIC) to classify companies into industry sectors but, as previously documented, significant bias can be introduced by the choice of database derived from the differences in data across databases, for example, Guenther and Rosman (1994) and Kahle and Walkling (1996) provide evidence of the bias introduced by the use of SIC codes from *Compustat* or *Crsp*, whereby, more than 36% of the classifications disagree at the two-digit level and nearly 80% at the four-digit level. Similarly, Krishnan (2003) examine the implications of using different industry classification systems by comparing the *North American Industry Classification System* (NAICS) and the SIC. Bhojraj et al. (2003) shows the differences from using the *Global Industry Classifications Standard* (GICS) system popular among financial practitioners and the Fama and French (1997) system used primarily by academics.

An issue that affects corporate governance studies is the ownership structure data and in which prior research has also demonstrated the effect of database choice and coverage (Anderson and Lee 1997). This prior work also documents the effect of survivor bias on the explanatory power of book-to-market equity, earnings yield and cash flow yield with respect to realized stock returns in the case of *Compustat* (Davis 1996) and on returns related to mutual funds in *Crsp* (Elton et al. 2001).

A particular concern related to the database coverage is the delisting bias. The issue highlighted in the literature is the presence of thousands of delisting returns in the database maintained by *Crsp*. Omitted delisting returns introduce bias to the studies as without delisting returns it is fairly difficult to accurately calculate the returns to a feasible portfolio (Shumway 1997). Delisting bias results in confounding empirical outcomes and affects mainly to NASDAQ rather than to *New York Stock Exchange* (NYSE) stocks. Research on this issue (Shumway 1997; Shumway and Warther 1999) reveals that correcting for the delisting bias eliminates the size effect considered as an economic phenomenon and first documented by Banz (1981) and Lamoureux and Sanger (1989) and lately by Fama and French (1995), and Berk (1995) among others.

Studies have also acknowledged the existence of errors in databases sufficient to change the nature of the data and suggest a method of quality control for competing databases (Rosenberg and Houglet 1974). Research has documented forecasts error metrics based on reported earnings numbers supplied by forecasts data providers such as *First Call*, *Zacks Investment Research*, *Crsp*, *Compustat* and *I/B/E/S* (Philbrick 1991; Canina et al. 1998; Rosenberg and Houglet 1974;

Abarbanell and Lehavy 2003; Ljungqvist et al. 2009) which leads to inconsistent inferences. Abarbanell and Lehavy (2003) identify two asymmetries in cross-sectional distributions of forecast error observations and demonstrate that analyst's tendency to commit systematic errors is not supported by broader analysis distribution of these errors. Elton et al. (2001) state that *Crsp* return data is biased upward and merger months are inaccurately recorded half of the time. Ljungqvist et al. (2009) find errors evidenced by the widespread changes present in the historical *I/B/E/S* analyst stock recommendation database including alterations of recommendations, additions and deletions of records and removal of analyst names from one download to the next in the period analyzed (2000-2007).

This literature review provides evidence of the high relevance of accurate and reliable data in research. The study of financial narratives using automated procedures is a relatively novel development in accounting and finance, but that has grown significantly and rapidly in the last decade (see Table 1 for a summary of the studies published in the main journals of accounting and finance). In this piece of research, we are interested in describing plausible biases in textual analysis studies of 10-K documents.

**3. Method and Data**

<<INSERT TABLE 2 HERE>>

The aim of this work is to demonstrate the existence of bias in textual analysis research. The literature review shows bias investigated and broadly documented in prior research. Our focus is the analysis of textual characteristics of 10Ks and the potential bias derived from the sample construction decisions, errors in files and coverage issues. To this end we access the data from *Edgar* provided by the SEC, and follow the procedure detailed in Appendix A. This procedure results in the extraction of 225,417 observations for the period 1994-2015 from downloading 161,131 individual 10-K reports. After cleaning duplicates, we obtain a sample of 159,338 10-K reports, corresponding to 33,466 unique firms. Over half of these 10-K reports use HTML tags (88,860). Table 2 Panel A summarizes the sample selection procedure. Nearly all of these 10-K reports contain the key items under study in prior work: Item 6 (Selected Financial Data) is available for 92% (146,302) of the 10-K reports, Item 7 (Management's discussion and analysis of financial condition and results of operations) for 92% (147,303) of the 10-K reports, Item 7A (Quantitative and qualitative disclosures about market risk) for 73% (116,168) of the 10-K reports, and Item 8 (Financial statements and supplementary data) for 93% (147,917) of the 10-K reports. These numbers are in line with prior research in the area, as reviewed in detail in Table 1.

At this step of the matching process, Sample I includes all the possible observations (147,303 firm/year observations for 31,405 unique companies) that include the presence of Item 7 from *Edgar* 10-K annual reports. Sample II accounts for the result of matching Sample I with *Crsp/Compustat Merged Fundamental Annual*. To access these data, we proceed as follows. From Wharton database, we download *Crsp/Compustat Merged Fundamental Annual* for both accounting and financial data. We select output data with the following settings: *Consolidated Level* (Consolidated), *Industry Format* (Standardized), *Population Source* (both Domestic and International), *Currency* (both U.S. and Canadian Dollar), and *Company Status* (both Active and Inactive). Since we extract data from a merged database, we rely on the primary links types provided in Wharton Database: *LC* (link research complete), *LU* (link is unsearched by Crsp), and *LS* (link valid for this security only).[4] On the downloaded data, we exclude those observations not

[4] Refer to the supporting manual for a broader description of these links. https://wrds-web.wharton.upenn.edu/wrds/support/Data/_001Manuals%20and%20Overviews/_002Crsp/ccm-overview.cfm

associated with a primary link marker (*Linkprim*) equal to "J" and "N" (joiner secondary issue of a company), and we keep only observations with a primary link "P," and "C". Table 2 Panel B provides details on the sample selection procedure when we match qualitative with quantitative data; it also shows the observations lost when we further impose data requirements.

To obtain Sample II with accounting and finance data, we follow two parallel methods that either exclude or include the usage of *Linking Table WRDS-SEC* (No Link Table and Yes Link Table respectively) also provided by *Wharton Database*. The *Linking Table* associates historical cik from Edgar with the gvkey variable of *Crsp* and *Compustat* databases respectively[5]. On the one hand, this method might alleviate the measurement error of associating qualitative data from *Edgar* with accounting and finance data. After merger and acquisition operations, companies might either cease to exist or change their identifiers. On the other hand, in *Crsp* and *Compustat* there are less missing observations for the gvkey variable compared to the cik, and this condition might improve the possible number of associable observations. Under both the methods, we also rely on both fiscal month and year to identify companies across the time. Sample III represents a subsample of Sample II after setting minimum requirement regarding accounting variables to run basic analyses (both lagged and forward observations).[6] Despite the fact that we only impose minimum data requirements, only almost half of the original sample is retained. On the one hand, Sample II is composed of either 72,934 (11,051 unique firms) or 70,489 (11,293 unique companies) firm-year observations depending on the usage of *Linking Tables*. On the other hand, Sample III is formed of either 68,494 (10,307 unique firms) or 66,083 (10,501 unique companies).

---

[5] To use the Linking Table, we merge Edgar data by cik code, and its output is furtherly matched with Crsp and Compustat adopting gvkey variable as the firm identifier.

[6] The following variables are extracted from "Variable List" of Wharton Database Crsp/Compustat MERGED: AT: Total Assets; cik: Central Identification Number; DLTT: Total Long Term Debt; DLC: Total Short Term Debt; FYEAR: Fiscal Year End; FYR: Fiscal Month End; REVT: Total Revenues.

From these two first subsamples, we observe how the number of total observations is lower when we do rely on *Linking Tables*. On the contrary, the number of unique firms is higher suggesting how this method is superior in capturing the same companies across years.

To conduct some of our additional analyses discussed above, we create our variations of Sample III: Sample IV. Performance, and Sample IV. Earnings respectively. As reported in Table 2 Panel B, we require a minimum accounting and finance data to replicate previous studies investigating the role of qualitative information on firm performance and earnings quality respectively. We also require that data on two additional databases is available: *Crsp/Compustat Merged Security Daily*, and *I/B/E/S* respectively. Table 2 Panel B provides details on this sample selection procedure, it also shows those observations lost after imposing data availability for daily prices and volumes from *Crsp/Compustat Daily Securities* (Sample IV. Market) and analysts' forecasts from *I/B/E/S* (Sample IV. Analysts). The latter is our most restrictive sample, as it requires that all data are available. This restricted sample is composed of 13,785 (13,250) 10-K reports.

For each of our samples, we compute common measures of the narrative disclosure of Item 7. We choose these proxies after careful consideration of the prior literature detailed in Table 1 Panel D. In particular, we focus our analyses on two narrative disclosures characteristics: Tone and Readability. As Table 1 Panel D reveals, these are the most common narrative characteristics studied. Our analyses reveal that many authors create their own word lists and strategies to systematically conduct content analyses in search for particular meaning, phrases or words. Given however how idiosyncratic these choices are, we focus on the aforementioned characteristics only. Although the literature on these areas is relatively young, it can be readily seen in Panel D that a number of proxies and lists exist to capture these constructs. We detail next our approach to measuring them.

<<INSERT FIGURE 2 HERE>>

Figure 2 Panel A and B show the variation of the various subsample from an alternative perspective. After having divided qualitative variables of Sample I into their deciles, we observe how the sample reduction affects more those observations included into higher deciles. The same observation is valid when we also consider Sample III as our reference point. Figure 2 Panel C and D include the variation of the fiscal month and year (conformed reported period) across the various subsamples. The higher number of observations seem to be lost during the recent years. This result is particularly evident for those studies focusing on performance (Sample IV. Performance), and earnings quality (Sample IV. Earnings) respectively.

### 3.1 Data restriction associated with main areas of research

**Performance:** Based on prior work that investigates firm performance in association with qualitative disclosures (Davis et al. 2012; Loughran and McDonald 2011; Allee and Deangelis 2015; Li 2008; Merkley 2014; Huang et al. 2014; Davis et al. 2015), we select a minimum number of variables for running this type of studies.[7] We merge *Edgar* created with *Crsp/Compustat MERGED* using cik(gvkey) fiscal month and year obtained from the Conformed Period of Report included in the 10-K reports. From the full 147,303 observations included in Sample I, we can match 46,288 (44,947) companies' observations (Sample IV. Performance) for 7,785 (7,985) unique firms with accounting and finance variables.

[7] In particular, the following variables are extracted from "Variable List" of Wharton Database Crsp/Compustat MERGED: AQC: Acquisitions; AT: Total Assets; CEQ: Total Common Ordinary Equity; IB: Income Before Extraordinary Items; MKVALT: Total Fiscal Market Value; OANCF: Net Cash Flow from Operating Activities; NI: Net Income; SSTK: Sale of Common and Preferred Stock

**Earnings Quality**. Based on prior work that investigates the links between earnings quality and firm narrative disclosures,[8] we select a minimum number of variables for running this type of studies (Frankel et al. 2016; Huang et al. 2014; Feldman et al. 2010; Lo et al. 2017). From the original 147,303 observations included in our Sample I (*Edgar* baseline), we can match 37,162 (36,183) companies' observations for 7,213 (7,360) unique firms with accounting and finance variables.

**Market Reaction.** Based on prior work that studies market reaction to firm narratives (Lawrence 2013; Loughran and McDonald 2011; Allee and Deangelis 2015; Koo et al. 2017; Baginski et al. 2016; Campbell et al. 2014; Hope et al. 2016; Segal and Segal 2016; Kothari et al. 2009; Drake et al. 2016; Loughran and McDonald 2013; Lee 2012; You and Zhang 2009; Kravet and Muslu 2013; Huang et al. 2014; Henry and Leone 2016; Miller 2010; Lee 2016; Brochet et al. 2016; Lundholm et al. 2014), we select a minimum number of variables for running this type of studies.[9] We merge *Edgar* created with Crsp/Compustat MERGED using cik FYR and FYEAR obtained from the Conformed Period of Report included in 10-K. We then merge matched observations with Crsp/Compustat SECURITY DAILY. The latter database contains daily prices and volumes for listed companies. To merge the two databases, we rely on Filing Date contained in the 10-K reports, on Permno and Permco numbers. We require a minimum number of variables

---

[8] The following variables are extracted from "Variable List" of Wharton Database Crsp: AOLOCH: Other Net Change in Assets and Liabilities; APALCH: Increase/Decrease in Accounts Payable and Accrued Liabilities; AT: Total Assets; IB: Income Before Extraordinary Items; INVT: Inventory; OANCF: Net Cash Flow from Operating Activities; PPEGT: Gross Value of Property, Plan, and Equipment; RECT: Total Receivable; REVT: Total Revenue; TXACH: Increase/Decrease in Income Taxes Accrued; XAD: Advertising Expense; XRD: Research and Development Expense

[9] The following variables are extracted from "Variable List" of Wharton Database Crsp/Compustat MERGED: AQC: Acquisitions; AT: Total Assets; CEQ: Total Common Ordinary Equity; IB: Income Before Extraordinary Items; LOC: Current ISO Country Code Headquarters Location; MKVALT: Total Fiscal Market Value; OANCF: Net Cash Flow from Operating Activities; SSTK: Sale of Common and Preferred Stock

to compute returns and volume for a three-day window on the filing date.[10] We are able to match 31,290 (29,990) observations that include 4,868 (4,963) unique companies.

**Analysts**. Based on prior work on analysts and firm narratives (Lehavy et al. 2011; Bozanic and Thevenot 2015; Allee and Deangelis 2015), we select a minimum number of variables for running this type of studies.[11] We merge the previous database created for computing market reaction with *I/B/E/S database*. We require to have at least one observation per company on the variables selected for a forecast horizon of one fiscal year. To merge *Crsp/Compustat* with *I/B/E/S*, we rely on the SAS code "*iclink*" from Wharton Database that allows to create a link table between the two databases. This procedure results in a final sample where we are able to match 13,785 (13,250) observations for 2,211 (2,224) unique companies.

### 3.2. Computing Narrative Disclosure Tone

In this study we focus on disclosure Tone and Readability. To evaluate disclosure Tone we need to transform the narrative into a numeric value that represents the Tone of the firm´s disclosures regarding performance. A generally used and accepted approach to measure Tone is to count the frequency of certain words contained in the disclosures and compute a score (Henry and Leone 2016). To assess the relative frequency and calculate the score a predefined list of "positive" or "negative" words is used.

To measure Tone, we review the existing literature to identify all the vocabularies that have been used to classify qualitative information (words) prepared by either companies or analysts. As

[10] The following variables are extracted from "Variable List" of Wharton Database Crsp, and from the help file provided by Crsp to compute return: AJEXDI: Adjustment Factor (Issue); PRCCD: Daily Price Close; TRFD: Daily Total Return Factor; CSHOC: Outstanding Shares; CSHTRD Daily Trading Volume

[11] EPS: Earnings per Share

is common in the literature, to compute Tone, we proceed with simple counts of the number of times specific words contained in the vocabularies appears in the text. To automatize the process, we use the 2015 version for academic use of the Linguistic Inquiry and Word Count Software (Larcker and Zakolyukina 2012).[12] In general, we extract the various vocabularies from which we compute Tone from articles, online appendixes, and software available. Whenever possible, we look for the sources of vocabulary's data.

First, to have the original list of Diction's vocabularies, we extract wordlists from its software. Diction 7.0[13] is a software that allows computing optimism (as a result of the differences between positive and negative words). However, it does not report a complete list of positive and negative words. To extract this list, we use both software's settings and the online help manual available for Diction 7.0. Optimism is defined as "*the difference between positive (Praise, Satisfaction, and Inspiration) and negative (Blame, Hardship, and Denial) wordlists*" (Diction 7.0 Manual, page 5). By merely summing the various lists extracted from log files of the software, it is possible to find duplicate words within the same groups of sentiment lists, i.e., there are words associated with more than one-word list.[14] Not to inflate results, we drop duplicate words since we are interested in the leading group of wordlists: optimistic and pessimistic, respectively. Second,

[12] The main advantage of this software is that it does not require programming skills for researchers that are required by alternative methods such as using Python and UNIX. In particular, it is possible to insert a list of words associated with a vocabulary, and obtain results for every file under analysis. When using alternative software, it is be necessary to control for case sensitive search (low and capital letter cases), and non-overlapping search.

[13] Diction 7.0 was the last version available of "Diction: The Text-analysis Program" when we conducted the analyses described in this paper. A link to the online help manual is present on the home page of Diction at "http://www.dictionsoftware.com/".

[14] For example, in the list that reports positive sentiment, both Inspiration and Satisfaction have the following words in common "charm," "comfort," and "courage." In negative sentiment, both Blame and Hardship share the same words "afraid," "biased," and "cursed." However, this overlap does not necessarily mean that results out of Diction 7.0 suffer from measurement error, as the software may correct for it. By extracting wordlists by software's log files, we might be exposed to measurement error if we do not control for these duplicates.

to have the original wordlists of General Inquirer from Harvard University, we select the vocabulary Harvard IV-4 categories.[15] In the file reported, it is possible to find duplicates associated with both positive and negative word lists. The presence of duplicates is mainly due to the different classification that the same word might have across contexts, and also by the fact that General Inquirer has included parallel wordlists in its vocabulary.[16] We keep unique words associated with positive and negative Tone.

<<INSERT TABLE 3 HERE>>

Finally, for the remaining vocabularies, we directly refer to both articles and online appendixes provided in prior research. In particular, to compute forward-looking statements, we rely on the list reported in Appendix B from Li (2010). For a causations' wordlist, we refer to Panel C provided by Dikolli et al. (2016). For a list of constraining words, we refer to Appendix C in Bodnaruk et al. (2015). For wordlists associated with litigiousness, strength, weakness, uncertainty, optimism, and pessimism, we refer to previous studies (Loughran and McDonald 2014, 2011).[17] Table 3 Panel A shows that for the average firm in our sample the reporting month is December (month 12). Panel A also provides evidence on the heterogeneous sizes of 10-K files, with a distribution that is heavily skewed to the right.

Table 3 Panel B provides descriptive statistics of Sample I for the calculations our Tone measures based on Diction (Diction Neg. and Diction Pos.), (Loughran and McDonald 2014,

[15] A complete list is present at the following link http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm by downloading the file named "General Inquirer Augmented Spreadsheet".

[16] In the Excel file provided by Harvard University, it is possible to appreciate the presence of duplicates if we compare the first with the last column that contains description over the word included in a specific vocabulary.

[17] Wordlists associated with these vocabularies are reported in https://www3.nd.edu/~mcdonald/Word_Lists.html. For our analysis, we rely on both lists from 2011 and 2014.

2011).[18] (LM Neg. and LM Pos.), and General Inquirer (Inquirer Neg. Inquirer Pos.). We report results obtained for the analysis of Item 7, similar findings are obtained for the other items. It can be readily seem that the use of General Inquirer is likely to return a more positive Tone, and also the highest negative Tone on average. This suggests that wordlist use is an important determinant of average Tone observed, as noted by Loughran and McDonald (2011). Table 3 Panel C provides descriptive statistics for our calculations of additional scores associated with word lists commonly used in prior work (see Table 1 Panel D), in particular, we calculate a score for constraining vocabulary (Constraining (Bodn.)), four of the word lists in Loughran and McDonald (2011) (litigious, strong, weak and uncertainty), a score for the use of causation (Causation (Dikolli)) and forward-looking (ForwardLook (Li)) vocabularies. While these scores are not comparable directly between themselves, they are in line with those reported in prior research, and will be used later to understand differences between Samples I, II, III and IV.

### 3.3. Computing Narrative Disclosure Readability

Readability is our second measure. Financial reporting clearly and accurately is a fundamental part of the scientific process, facilitating both the dissemination of knowledge and the reproducibility of results. The clarity of written language can be quantified using Readability formulas, which estimate the understandability of written texts. The assumption is, therefore, that better written documents include less ambiguity and lead to better corporate valuation which reflects on lower price volatility of the stocks after the filling of 10Ks (Loughran and McDonald 2014).

We focus on three indicators of Readability which have been primarily used in previous

[18] Wordlists associated with these vocabularies are reported in https://www3.nd.edu/~mcdonald/Word_Lists.html. For our analysis, we rely on both lists from 2011 and 2014.

studies: Gunning Fog index, Flesch reading ease score, and Flesch–Kincaid grade level (Brochet et al. 2016; Li 2008; Law and Mills 2015; De Franco et al. 2015). First, we compute the Fog index as the sum of words per sentence and percentage of complex words (Gunning 1952).[19] The score of the index is associated with a scale reporting the minimum number of years a reader would need to interpret information. Second, we calculate the Flesch reading ease score as the difference between number of words per sentence and syllables per words (Flesch 1948).[20] The index reinterprets the presence of "complex" words by measuring the total number of syllables. Readability is associated with a minimum educational level which readers might need to attain to understand a text. Lastly, we compute the Flesch-Kincaid grade as the sum of words per sentence and syllables per sentence (Kincaid et al. 1988).[21] Similar to the previous index, it associates Readability to U.S. grade school levels. However, the interpretation is different: higher values of both Fog index and Flesch–Kincaid grade level indicate lower Readability. In contrast, higher Flesch reading scores are associated with higher Readability. Consistent with Li (2008), we use Perl (package named Fathom::EG) to measure Readability. We extract the following variables for computing Readability indexes: number of words, number of sentences, the percentage of complex words, and total syllables per sentence respectively. Table 3 Panel D presents average values of these indexes. All scores are within the expected values reported in prior literature.

## 4. Main research results

As described in the above sections, we first construct three separate samples (Samples I, II, and

---

[19] [(Number of words per sentences + Percentage of complex words) * (0.4)]. Where percentage of complex words is computed as number of complex words over total words. A word would be "complex", if it is composed by three or more syllables.

[20] [206.835 – (1.015 * Number of words per sentences) – (84.6 * Syllables per words)]

[21] [-15.59 + (0.39 * Number of words per sentences) + (11.8 * Syllables per words)]

III) of 10-K reports. For each of these samples, we compute, at the individual 10-K report level three measures of Tone (Diction, LM and Inquirer) segregated by positive (Pos.) and negative (Neg.), four measures of Readability (Fog, Flesh, Flesch-Kincaid, No. of words), and seven measures of content analyses (constraining, litigious, strong, weak, uncertainty, causation and forward looking). Using these samples and measures, we study whether samples commonly used in prior research in accounting and finance (i.e., our Sample IV) can be generalized as compared to both Sample I (the full usable *Edgar*) and Sample III (the subsample with the minimum requirement of accounting data) and whether the narratives not commonly studied in prior research (henceforth, the 'missing narratives') are, on average, different from those that are under the increasing scrutiny of research.

<<INSERT TABLE 4 HERE>>

Table 4 compares the textual characteristics of the missing narratives with those usually studied in prior research in accounting and finance by comparing Sample II with Sample I. Panel A compares mean and median Tone, Panel B compares mean and median scores of commonly used content analyses variables, and Panel C of Readability measures. It can be readily seen that differences systematically exist. In fact, they are significantly different in all Panels. Missing narratives appear to contain more negative words on average than those that are kept in Sample II. They also show signs of having lower complexity in terms of Readability. This may indicate that these are smaller firms, less likely to use optimistic Tone in an opportunistic way and potentially more inclined to use simple language. Overall, the evidence in this table contains compelling evidence that the missing narratives are not alike the narratives commonly studied in accounting and finance, and suggest that sample selection procedures likely bias the findings, impeding generalizing the results to the general population of firms.

<<INSERT TABLE 5 HERE>>

Table 5 shows the differences between Sample III and Sample I. We observe a similar pattern compared with Table 4 suggesting that previous results are generalizable to those observations containing the minimum requirement of accounting data, and not just the match with *Crsp/Compustat* database.

## 5. Further analyses based on research lines

As noted above, we build on our literature review (see Table 1 Panels B to D) to identify the main lines of research that have been the focus of the literature in textual and narrative analysis thus far. Without aiming to be exhaustive, we identify four main lines of interest for our study. First, the association between aspects of narratives derived through textual analysis and the market reaction to those narratives is the predominant line of research thus far (Lawrence 2013; Loughran and McDonald 2011; Allee and Deangelis 2015; Lundholm et al. 2014; Brochet et al. 2016; Lee 2016; Miller 2010; Henry and Leone 2016; Huang et al. 2014; Kravet and Muslu 2013; You and Zhang 2009; Lee 2012; Loughran and McDonald 2013; Drake et al. 2016; Kothari et al. 2009; Segal and Segal 2016; Hope et al. 2016; Campbell et al. 2014; Baginski et al. 2016; Koo et al. 2017). This line of research addresses an important research question: *Do narratives have information content?* The second line of research that has also drawn significant interest relates to the association between recognition and disclosure quality, i.e., studies that analyse the links between earnings quality and textual analysis issues (Frankel et al. 2016; Lo et al. 2017; Feldman et al. 2010; Huang et al. 2014). Related to that line, a number of papers study the association between narratives and performance, which includes the studies of Li (2008), Huang et al. (2014), Merkley (2014), Loughran and McDonald (2011), Davis et al. (2015), Davis et al. (2012), Allee and Deangelis (2015). Finally, the issue of how narratives affect financial analysts' recommendations and data gathering processes has also attracted the attention of academic research (Bozanic and Thevenot 2015; Allee and Deangelis 2015; Lehavy et al. 2011). All these areas of interest impose further restrictions on the samples available for study. We detail those restrictions next.

### *5.2. Results from reduced datasets*

Similar to how we proceeded before, in Tables 6, 7, 8 and 9 we provide evidence of the new missing narratives. These are 10-K reports that are lost in the process of comparing Sample III (68,494 and 66,083 respectively) with the data required to conduct the analyses that are common in prior literature (Sample IV). Again, the results are split into three panels, Panel A shows differences for Tone, Panel B for usual content scores, and finally, Panel C for Readability.

<<INSERT TABLE 6 HERE>>

<<INSERT TABLE 7 HERE>>

Across the various tables in Panel A, it can be readily seen that again differences systematically exist between Sample III and the various Sample IV except Table 7. In this concrete case, the difference regarding negative Tone for both Diction and Inquirer tend to be less statistically significant compared with the other subsamples of Sample IV. On the magnitude of Tone for all these subsamples, Missing Narratives include observations that have lower both negative and positive words. However, the average reduction in both negative and positive words is asymmetrical suggesting that net optimism, computed as the difference between positive and negative words, might still be significantly different across paired samples (missing and non-missing narratives). Whether at the Tone level it is possible to appreciate a clear difference in terms of mean and median across missing and non-missing narrative, the same does not hold for other vocabularies and Readability. As reported in Panel B and C, mean and median tend to be more similar, and less significant differences compared with results of Panel A.

<<INSERT TABLE 8 HERE>>

<<INSERT TABLE 9 HERE>>

Overall, the evidence in these tables again indicates that the missing narratives are different from those commonly studied in accounting and finance, and suggest that sample selection procedures likely bias the findings, impeding generalizing the results to the general population of firms. Furthermore, we observe that by augmenting those restrictions of both accounting and finance data, the differences introduced by *Linking Tables* tend to disappear: across the various Sample IV mean and median seem to be similar.

**6. Discussion and Conclusions**

We study plausible biases in textual analysis studies of 10-K documents. The study of financial narratives using automated procedures is a relatively novel development in accounting and finance. Therefore, standardized methods to collect and systematically analyse these data are yet to be developed. We provide detailed step-by-step guidance on how to download and prepare these files for analyses, and study the biases introduced by a number of decisions regarding sample construction, data preparation, and variable choice. In particular, we focus on two widely studied properties of financial narratives: their Tone and Readability. We document that a number of these choices introduce significant biases into the samples studied, as well as induce differences in average observed Tone and Readability. Our results also indicate that a non-trivial proportion of the *Edgar* population is missing from the textual analyses being conducted.

We contribute to the prior literature in a number of ways. First, we provide detailed step-by-step guidance on how to download and prepare *Edgar* fillings for analysis. Second, we document the biases introduced by a number of decisions regarding sample construction, data preparation, and variable choice. Our results indicate that a non-trivial proportion of the *Edgar* population is missing from the textual analyses being conducted. On average, these missing narratives appear significantly different from the ones that are retained in common accounting and finance studies. In particular, we provide evidence of significant differences in two widely studied properties of financial narratives: their Tone and Readability. Further, we document that a number of the variable choices made by researchers also may induce differences in average observed Tone and Readability. Thus, our results question the generalizability of studies in narrative disclosure and also, set the question of what are these missing firms discussing in their 10-K reports? It appears that existing research may be ignoring much of what is being said.

## **Appendix A.** Detailed downloading and parsing procedures

We investigate the content of 10-K reports released by the Stock Exchange Commission (SEC). In this appendix, we provide detail on the procedure used to download and parse the data.

### I. Downloading Documents

***Step One: Access to Edgar Database***

*<<INSERT FIGURE 1 HERE>>*

The *Edgar* database stores all documents released by the SEC since 1993. Prior work usually focuses on: 10-K, 10-Q, 8-K, and 20-F. Appendix B provides descriptive evidence on the documents available for downloading in the *Edgar* database between 1994 and 2015.[22] It can be readily seen that sample sizes vary depending on the document type, with 8-K fillings being the more numerous. We show data both for the original files and their amendments. Figure 1 Panel A graphically shows time-trends in 10-K fillings, which are the focus of this study, while Panel B provides details on the percentage of amended documents. Amendments can be easily identified as they are labelled differently (/A documents). Unsurprisingly, late filings (NT documents) are less likely to be amended. Only 1.58% (0.88%) of late 10-Ks (10-Qs) fillings are amended, relative to 25.44% (8.24%) of on-time 10-Ks (10-Qs). To download 10-K reports, we connect to the *Edgar* website. In its archive, web links to SEC's documents are stored in files that are divided by year and by quarters respectively (master files).[23]

***Step Two: Download Master Files***

Since we ignore the exact quarter in which 10-K have been released, we download all the files

[22] We exclude 1993 for being a transitional year i.e. the first year *Edgar* was reporting companies' information.
[23] *Edgar* website: https://www.sec.gov/edgar.shtml; Archives: https://www.sec.gov/Archives/edgar/full-index/

containing web links for every quarter folder from 1994 to 2015 (both included). To automatize downloading, UNIX allows creating a batch file that recursively opens and downloads 10-K reports.[24] To use UNIX on Windows, we download and install Cygwin.[25] After installing and updating Cygwin, we can download all the weblinks' files automatically[26].

***Step Three: Select Documents***

Downloaded files are structured in tables with the same variables across quarters and years: Form Type, Company Name, cik, Date Filed, and File Name, respectively. Since our master files contain different types of documents released by the SEC, we extract only those links to documents for the Form Type "10-K." Then, we remove headings that include description file. Second, we select only those web links associated to a 10-K. To handle these two tasks, we download Powergrep 5.[27] By using Powergrep, from our files, we extract only those lines that contain web links for downloading "10-K"[28].

***Step Four: Download Documents***

Once all previous steps are complete, we have a full list of 10-K reports. To download them, we

[24] Parallel automatic approaches might be useful, but we decide to rely on UNIX, and more specifically on a command named "Wget" for being the most effective and user friendly method we found.

[25] Cygwin is a freeware software allows Windows' users to have a UNIX's interface without having to install the UNIX's operating system. After having downloaded Cygwin, it is important to update the software by including the wget package. Cygwin itself does not contain all the packages we might need to use for our goals.

[26] The command "wget" on the Cygwin interface (without quotes) downloads the master file of 1993 for the first quarter in our folder "downloads" with the name "masteridx_1993_1.txt". To make the command recursive, it would be sufficient to replace the year from the weblink.

[27] Powergrep is not a freeware software. However, it is possible to obtain the same functions of Powergrep from other software such as Python, as they share the same usage of regular expressions (regexes). The main advantage of Powergrep is being more user-friendly, and having windows commands that allow visualizing of software's output. Therefore, Powergrep is recommended for those users who do not have a proficient skill in software's coding.

[28] In the Library section of Powergrep, it is possible to find a command that allows extracting only those lines that contain a specific word: Collect lines containing a search term. In Powergrep, as a "Search type," we select "Literal text" and as a "Search" the word "|10-K|" (without quotes). This would prevent us from selecting other Form Types such 10-K/A, NT 10-K, 10-K405, and 10-K405/A that contain the same word 10-K.

follow the approach used to obtain the master files. The SEC associates a unique File Name to every document. This File Name partially contains the link to download the document but also some identifiers of the company. To have a complete match between the master files and the downloaded documents, we use part of the File Name as the filename for storing 10-K files.

## II. Parsing Documents

***Step Five: Recognize HTML vs. NON-HTML files***

We download all 10-K reports as TXT files. Although it would be possible to download HTML documents, HTML tags allow us to extract the various sections of 10-K reports. However, not all 10-Ks have HTML tags. We use Powergrep to recognize those files that contain HTML tags by using specific keywords, such as <a>, <body>, <dir>, <h1>, or <href>.

***Step Six: Extract Specific Sections***

We focus on four sections of the 10-K report: Item 6 (Selected Financial Data), Item 7 (Management's discussion and analysis of financial condition and results of operations), Item 7A (Quantitative and qualitative disclosures about market risk), and Item 8 (Financial statements and supplementary data).

For files without HTML tags, we use regular expressions (regexes) that allow capturing titles of the various sections.[29] For files with HTML tags, we use HTML tags' scheme for extracting titles' sections. First, we remove potential confounding titles that might be used as references in different

---

[29] E.g., "ITEM 6"; "^ITEM 6"; "^\s+ITEM 6"; "ITEM\s+6"; "^ITEM \s+6"; "^\s+ITEM\s+6". We rely on a regex that can capture titles of 10-K sections written in capital letters, at the beginning of the line, and after several spaces with all the possible combinations.

parts of the 10-K.[30] Second, we use some paths associated with the use of HTML to include titles in texts.[31] The approach includes and develops techniques used to exploit the HTML tags used by firms (Campbell et al. 2014). Contrary to Li (2008), we use the words included in the titles of the items under analysis only as the last step of our extraction (if the methodology of using HTML tags does not provide results). By using those words that define the title of our items, we face two possible scenarios. First, we have to edit in advance the 10-K files by confounding titles (Li 2008). This process would increase the time and resources devoted to parsing documents. Second, we would face the risk of extracting titles of sections which do not refer to the beginning of the section but instead refer to a continued section. We use those keywords suggested by Feldman et al. (2010) (item number, titles, surrounding language, and new item number) to extract both the beginning and end of our sections.

***Step Six (Bis): Robustness Tests***

We run some robustness tests to reduce the likelihood of extracting wrong sections. A possible error is associated with having selected either a title of contents' table or a section's reference. To assess this eventuality, a solution is either to count the number of words or to look at the file's size. Both low numbers of words and a small file's size are indicative of possible missteps. A further solution is to look at the presence of non-consecutive items.[32] Furthermore, we might not find some or even all items. This might be the case for smaller firms or if they consolidate information of one item into a further one.[33] The latter might be the case if a company decides only to provide

---

[30] E.g., "HREF" followed by "ITEM 6", and "HREF" followed by "Selected financial statement"

[31] E.g., "a name" followed by "ITEM 6", "style" and "ITEM 6", "link1" and "ITEM 6", "link2" and "ITEM 6", "font" and "ITEM 6", "size" and "ITEM 6".

[32] For example, the presence of Item 7A into the file that contains Item 6.

[33] The Item 7A is, on average, the section which is most time consolidated into Item 7.

a reference to the section.

### *Step Seven: Remove Tables*

Consistent with previous studies (Bonsall et al. 2017; Miller 2010; Li 2008), we remove all tables. We use a regex that accounts for how tables are reported.[34] We erase tables as self-defined by firms. However, we do not drop lines with special characters such as <s> and <c> (Loughran and McDonald 2011). HTML <s> identifies no longer useful text, <c> states for HTML code. We leave these tags intact, not to remove those parts of the 10-K that are lists of elements and might contain words associated with our vocabularies. Furthermore, applying the same regex even to non-HTML files is possible. Indeed, there is a consistent amount of non-HTML files for which tables are reported using HTML tags. However, it is important to highlight a possible limitation to this approach. In both types of files, we can extract and remove tables that are self-reported by the firm. Tables may be copied and pasted on the 10-K (this scenario is more probable for non-HTML tags that use tags only for inserting tables) without using HTML tags.

### *Step Eight: Generate Readable Files*

After removing tables from all the sections, we convert TXT files with HTML tags into readable text files. First, we convert TXT 10-Ks into HTML files.[35] This process generates readable HTML files without tags. Second, we convert HTML files into readable TXT files.[36]

---

[34] For example, in 10-K tables usually starts with “<table” and end with “/table>”.
[35] A batch file in Windows can copy and paste all the files in folder from txt format with HTML tags to HTML.
[36] In Cygwin, both “w3m” and “-dump” commands convert HTML into readable TXT files dropping HTML tags.

**Appendix B.** *Edgar* fillings with at least 150 observations in any year between 1994-2015.

| | SEC Filling | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Year** | **10-K** | **10-K/A** | **10-K405** | **10-K 405/A** | **10-Q** | **10-Q/A** | **20-F** | **20-F/A** | **8-K** | **8-K/A** | **NT 10-K** | **NT 10-K/A** | **NT 10-Q** | **NT 10-Q/A** |
| 1994 | 1,912 | 616 | 11 | 5 | 6,632 | 558 | 0 | 0 | 3,623 | 375 | 51 | 1 | 68 | 3 |
| 1995 | 2,218 | 933 | 1,018 | 190 | 14,131 | 1,692 | 0 | 0 | 6,339 | 798 | 187 | 6 | 373 | 7 |
| 1996 | 4,315 | 1,495 | 1,944 | 317 | 25,758 | 2,319 | 9 | 7 | 15,736 | 2,155 | 749 | 21 | 1,261 | 16 |
| 1997 | 6,698 | 2,152 | 3,201 | 538 | 29,024 | 2,036 | 39 | 4 | 24,143 | 2,858 | 1,533 | 36 | 1,757 | 16 |
| 1998 | 6,930 | 1,943 | 3,357 | 589 | 29,254 | 2,135 | 108 | 28 | 28,004 | 3,285 | 1,687 | 41 | 1,828 | 18 |
| 1999 | 6,761 | 1,798 | 3,361 | 610 | 28,701 | 2,426 | 144 | 21 | 27,946 | 2,941 | 1,815 | 36 | 1,940 | 36 |
| 2000 | 6,652 | 1,530 | 3,217 | 483 | 28,301 | 1,926 | 195 | 55 | 29,747 | 2,946 | 2,285 | 41 | 3,705 | 37 |
| 2001 | 6,248 | 1,578 | 3,000 | 494 | 26,012 | 1,722 | 281 | 57 | 35,333 | 2,805 | 2,493 | 49 | 4,046 | 38 |
| 2002 | 6,762 | 2,010 | 2,168 | 123 | 24,111 | 1,778 | 1,131 | 148 | 45,136 | 3,165 | 2,455 | 43 | 4,114 | 49 |
| 2003 | 8,468 | 2,021 | 0 | 0 | 21,876 | 1,694 | 1,031 | 189 | 67,746 | 3,742 | 2,310 | 34 | 3,632 | 33 |
| 2004 | 8,567 | 2,096 | 0 | 0 | 20,878 | 1,704 | 1,004 | 234 | 91,445 | 4,191 | 2,061 | 58 | 3,604 | 53 |
| 2005 | 9,017 | 2,180 | 0 | 0 | 20,676 | 1,946 | 1,028 | 331 | 116,353 | 5,497 | 2,540 | 37 | 4,377 | 41 |
| 2006 | 8,852 | 1,510 | 0 | 0 | 20,049 | 1,574 | 980 | 233 | 106,390 | 5,009 | 2,357 | 43 | 4,575 | 36 |
| 2007 | 8,574 | 1,470 | 0 | 0 | 20,054 | 1,065 | 883 | 214 | 101,389 | 4,479 | 2,278 | 17 | 4,062 | 32 |
| 2008 | 8,746 | 1,801 | 0 | 0 | 27,047 | 1,517 | 820 | 193 | 90,993 | 3,505 | 2,308 | 27 | 3,681 | 25 |
| 2009 | 9,839 | 2,320 | 0 | 0 | 27,822 | 2,195 | 768 | 188 | 82,043 | 3,119 | 2,122 | 36 | 3,422 | 18 |
| 2010 | 9,165 | 2,213 | 0 | 0 | 26,556 | 2,060 | 746 | 217 | 80,519 | 2,923 | 1,817 | 21 | 3,079 | 15 |
| 2011 | 8,840 | 1,995 | 0 | 0 | 25,677 | 3,539 | 748 | 226 | 78,904 | 3,998 | 1,704 | 21 | 3,179 | 25 |
| 2012 | 8,393 | 1,840 | 0 | 0 | 24,192 | 3,045 | 723 | 267 | 77,417 | 2,908 | 1,574 | 10 | 2,908 | 25 |
| 2013 | 8,105 | 1,765 | 0 | 0 | 23,260 | 1,829 | 702 | 197 | 76,440 | 2,931 | 1,542 | 19 | 2,701 | 12 |
| 2014 | 8,084 | 1,557 | 0 | 0 | 22,883 | 1,420 | 676 | 90 | 76,996 | 2,735 | 1,423 | 8 | 2,502 | 10 |
| 2015 | 7,985 | 1,258 | 0 | 0 | 22,174 | 1,095 | 691 | 81 | 76,482 | 2,410 | 1,332 | 7 | 2,239 | 8 |

We exclude 10-KT and NT 20-F (and their amendments). The maximum number of 10-KT (NT 20-F) fillings is 41 (120) in 2014 (2003).

**Appendix C.** Selection procedure to create Samples II to IV

**No link Table**

| | Sample I | | Sample II | | Sample III | | Sample IV. Performance | | | Sample IV. Earnings | | | Sample IV. Market | | | Sample IV. Analysts | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Deciles** | Obs. | % | Obs | % over Sample I | Obs | % over Sample I | Obs | % over Sample I | % over Sample III | Obs | % over Sample I | % over Sample III | Obs | % over Sample I | % over Sample III | Obs | % over Sample I | % over Sample III |
| **1** | 14,730 | 10 | 8,297 | 56.33 | 7,795 | 52.92 | 4,364 | 29.63 | 55.98 | 4,190 | 28.45 | 53.75 | 3,026 | 20.54 | 38.82 | 1,137 | 7.72 | 14.59 |
| **2** | 14,730 | 10 | 8,951 | 60.77 | 8,397 | 57.01 | 5,902 | 40.07 | 70.29 | 4,499 | 30.54 | 53.58 | 3,990 | 27.09 | 47.52 | 1,694 | 11.50 | 20.17 |
| **3** | 14,731 | 10 | 8,649 | 58.71 | 8,132 | 55.20 | 5,962 | 40.47 | 73.32 | 4,331 | 29.4 | 53.26 | 4,052 | 27.51 | 49.83 | 1,873 | 12.71 | 23.03 |
| **4** | 14,730 | 10 | 8,728 | 59.25 | 8,190 | 55.60 | 6,244 | 42.39 | 76.24 | 4,453 | 30.23 | 54.37 | 4,184 | 28.40 | 51.09 | 1,878 | 12.75 | 22.93 |
| **5** | 14,730 | 10 | 8,226 | 55.85 | 7,731 | 52.48 | 5,821 | 39.52 | 75.29 | 4,255 | 28.89 | 55.04 | 3,806 | 25.84 | 49.23 | 1,768 | 12.00 | 22.87 |
| **6** | 14,731 | 10 | 6,817 | 46.28 | 6,371 | 43.25 | 4,833 | 32.81 | 75.86 | 3,598 | 24.42 | 56.47 | 3,073 | 20.86 | 48.23 | 1,496 | 10.16 | 23.48 |
| **7** | 14,730 | 10 | 5,769 | 39.16 | 5,372 | 36.47 | 3,998 | 27.14 | 74.42 | 3,165 | 21.49 | 58.92 | 2,488 | 16.89 | 46.31 | 1,190 | 8.08 | 22.15 |
| **8** | 14,731 | 10 | 5,482 | 37.21 | 5,115 | 34.72 | 3,538 | 24.02 | 69.17 | 3,013 | 20.45 | 58.91 | 2,365 | 16.05 | 46.24 | 1,020 | 6.92 | 19.94 |
| **9** | 14,730 | 10 | 5,073 | 34.44 | 4,763 | 32.34 | 2,708 | 18.38 | 56.85 | 2,773 | 18.83 | 58.22 | 1,840 | 12.49 | 38.63 | 743 | 5.04 | 15.60 |
| **10** | 14,730 | 10 | 6,942 | 47.13 | 6,628 | 45.00 | 2,918 | 19.81 | 44.03 | 2,885 | 19.59 | 43.53 | 2,466 | 16.74 | 37.21 | 986 | 6.69 | 14.88 |
| **Total** | **147,303** | **100** | **72,934** | **49.51** | **68,494** | **46.50** | **46,288** | **31.42** | **67.58** | **37,162** | **25.23** | **54.26** | **31,290** | **21.24** | **45.68** | **13,785** | **9.36** | **20.13** |

**Yes Link Table**

| | Sample I | | Sample II | | Sample III | | Sample IV. Performance | | | Sample IV. Earnings | | | Sample IV. Market | | | Sample IV. Analysts | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Deciles** | Obs. | % | Obs | % over Sample I | Obs | % over Sample I | Obs | % over Sample I | % over Sample III | Obs | % over Sample I | % over Sample III | Obs | % over Sample I | % over Sample III | Obs | % over Sample I | % over Sample III |
| **1** | 14,730 | 10 | 7,984 | 54.20 | 7,484 | 50.81 | 4,242 | 28.80 | 56.68 | 4,071 | 27.64 | 54.4 | 2,918 | 19.81 | 38.99 | 1,097 | 7.45 | 14.66 |
| **2** | 14,730 | 10 | 8,659 | 58.78 | 8,119 | 55.12 | 5,719 | 38.83 | 70.44 | 4,388 | 29.79 | 54.05 | 3,832 | 26.01 | 47.20 | 1,622 | 11.01 | 19.98 |
| **3** | 14,731 | 10 | 8,412 | 57.10 | 7,900 | 53.63 | 5,825 | 39.54 | 73.73 | 4,256 | 28.89 | 53.87 | 3,899 | 26.47 | 49.35 | 1,815 | 12.32 | 22.97 |
| **4** | 14,730 | 10 | 8,448 | 57.35 | 7,911 | 53.71 | 6,040 | 41.00 | 76.35 | 4,327 | 29.38 | 54.7 | 3,983 | 27.04 | 50.35 | 1,802 | 12.23 | 22.78 |
| **5** | 14,730 | 10 | 7,981 | 54.18 | 7,491 | 50.86 | 5,676 | 38.53 | 75.77 | 4,175 | 28.34 | 55.73 | 3,651 | 24.79 | 48.74 | 1,717 | 11.66 | 22.92 |
| **6** | 14,731 | 10 | 6,616 | 44.91 | 6,160 | 41.82 | 4,683 | 31.79 | 76.02 | 3,507 | 23.81 | 56.93 | 2,927 | 19.87 | 47.52 | 1,438 | 9.76 | 23.34 |
| **7** | 14,730 | 10 | 5,563 | 37.77 | 5,172 | 35.11 | 3,851 | 26.14 | 74.46 | 3,078 | 20.9 | 59.51 | 2,363 | 16.04 | 45.69 | 1,130 | 7.67 | 21.85 |
| **8** | 14,731 | 10 | 5,281 | 35.85 | 4,914 | 33.36 | 3,429 | 23.28 | 69.78 | 2,934 | 19.92 | 59.71 | 2,240 | 15.21 | 45.58 | 969 | 6.58 | 19.72 |
| **9** | 14,730 | 10 | 4,909 | 33.33 | 4,605 | 31.26 | 2,631 | 17.86 | 57.13 | 2,698 | 18.32 | 58.59 | 1,766 | 11.99 | 38.35 | 713 | 4.84 | 15.48 |
| **10** | 14,730 | 10 | 6,636 | 45.05 | 6,327 | 42.95 | 2,851 | 19.36 | 45.06 | 2,749 | 18.66 | 43.45 | 2,411 | 16.37 | 38.11 | 947 | 6.43 | 14.97 |
| **Total** | **147,303** | **100** | **70,489** | **47.85** | **66,083** | **44.86** | **44,947** | **30.51** | **68.02** | **36,183** | **24.56** | **54.75** | **29,990** | **20.36** | **45.38** | **13,250** | **9.00** | **20.05** |

## References


Abarbanell, J., and R. Lehavy. 2003. Biased forecasts or biased earnings? The role of reported earnings in explaining apparent bias and over/underreaction in analysts' earnings forecasts. *Journal of Accounting and Economics* 36:105-146.

Allee, K. D., and M. D. Deangelis. 2015. The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research* 53 (2):241-274.

Anderson, R. C., and D. S. Lee. 1997. Ownership studies: The data source does matter. *Journal of Financial and Quantitative Analysis* 32 (3):311-329.

Baginski, S., E. Demers, C. Wang, and J. Yu. 2016. Contemporaneous verification of language: evidence from management earnings forecasts. *Review of Accounting Studies* 21 (1):165-197.

Banz, R. W. 1981. The relationship between return and the market value of common stocks. *Journal of Financial Economics* 9:103-126.

Berk, J. 1995. A critique of size-related anomalies. *Review of Financial Studies* 8:275-286.

Bhojraj, S., C. M. C. Lee, and D. K. Oler. 2003. What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41 (5):745-774.

Bodnaruk, A., T. Loughran, and B. McDonald. 2015. Using 10-K Text to Gauge Financial Constraints. *Journal of Financial and Quantitative Analysis* 50 (4):623–646.

Bonsall, S. B., A. J. Leone, B. P. Miller, and K. Rennekamp. 2017. A plain English measure of financial reporting readability. *Journal of Accounting and Economics* 63 (2):329-357.

Bonsall, S. B., and B. P. Miller. 2017. The impact of narrative disclosure readability on bond ratings and the cost of debt. *Review of Accounting Studies* 22 (2):608-643.

Boone, A. L., I. V. Floros, and S. A. Johnson. 2016. Redacting proprietary information at the initial public offering. *Journal of Financial Economics* 120 (1):102-123.

Bozanic, Z., and M. Thevenot. 2015. Qualitative Disclosure and Changes in Sell-Side Financial Analysts' Information Environment. *Contemporary Accounting Research* 32 (4):1595-1616.

Brochet, F., M. Loumioti, and G. Serafeim. 2015. Speaking of the short-term: disclosure horizon and managerial myopia. *Review of Accounting Studies* 20 (3):1122-1163.

Brochet, F., P. Naranjo, and G. Yu. 2016. The capital market consequences of language barriers in the conference calls of non-U.S. firms. *The Accounting Review* 91 (4):1023-1049.

Brown, S. V., and J. W. Tucker. 2011. Large-sample evidence on firms' year over year MD&A modifications. *Journal of Accounting Research* 49 (2):309-346.

Bushman, R. M., B. E. Hendricks, and C. D. Williams. 2016. Bank competition: Measurement, decision-making, and risk-taking. *Journal of Accounting Research* 54 (3):777-826.

Cadman, B., S. Klasa, and S. Matsunaga. 2010. Determinants of CEO pay: A comparison of ExecuComp and non-ExecuComp firms. *The Accounting Review* 85 (5):1511-1543.

Campbell, J. L., H. Chen, D. S. Dhaliwal, H.-m. Lu, and L. B. Steele. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies* 19 (1):396-455.

Canina, L., R. Michaely, R. Thaler, and K. Womack. 1998. Caveat Compounder: A warning about using the daily CRSP equal-weighted index to compute long-run excess returns. *The Journal of Finance* 53 (1):403-416.

Davis, A. K., W. Ge, D. Matsumoto, and J. L. Zhang. 2015. The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies* 20 (2):639-673.

Davis, A. K., J. M. Piger, and L. M. Sedor. 2012. Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research* 29 (3):845-868.

Davis, J. L. 1996. The cross-section of stock returns and survivorship bias: Evidence from delisted stocks. *Quarterly Review of Economics & Finance* 36 (3):365-375
De Franco, G., O.-K. Hope, D. Vyas, and Y. Zhou. 2015. Analyst Report Readability. *Contemporary Accounting Research* 32 (1):76-104.
Dikolli, S. S., T. Keusch, W. J. Mayew, and T. D. Steffen. 2016. A Linguistic-Based Approach to Measuring Innate Executive Traits: The Case of CEO Integrity: Duke Univesity.
Drake, M. S., D. T. Roulstone, and J. R. Thornock. 2016. The usefulness of historical accounting reports. *Journal of Accounting and Economics* 61 (2–3):448-464.
Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64 (2):221-245.
Elton, E. J., M. J. Gruber, and C. Blake, R. . 2001. A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and morningstar mutual fund databases. *Journal of Finance* 56 (6):2415-2430
Fama, E. F., and K. R. French. 1995. Size and book-to-market factors in earnings and returns. *Journal of Finance* 50:131-155.
Fama, E. F., and K. R. French. 1997. Industry costs of equity. *Journal of Financial Economics* 43 (2):153-193.
Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. Management’s tone change, post-earnings announcement drift and accruals. *Review of Accounting Studies* 15 (4):915-953.
Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (3):221-233.
Frankel, R., J. Jennings, and J. Lee. 2016. Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics* 62 (2–3):209-227.
García Lara, J. M., B. García Osma, and B. Gill de Albornoz Noguer. 2006. Effects of database choice on international accounting research. *Abacus* 42 (3/4):426-454.
Goel, S., J. Gangolly, S. R. Faerman, and O. Uzuner. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* 7:25–46.
Guay, W., D. Samuels, and D. Taylor. 2016. Guiding through the Fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics* 62 (2–3):234-269.
Guenther, D. A., and A. J. Rosman. 1994. Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics* 18 (1):115-128.
Gunning, R. 1952. *the technique of clear writing*. New York: McGraw-Hill.
Henry, E. 2006. Market reaction to verbal components of earnings press releases: event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* 3 (1):1-19.
Henry, E. 2008. Are investors influenced by how earnings press releases are written? *Journal of Business Communication* 45 (4):363-407.
Henry, E., and A. J. Leone. 2016. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review* 91 (1):153-178.
Hoberg, G., and V. Maksimovic. 2015. Redefining financial constraints: A text-based analysis. *The Review of Financial Studies* 28 (5):1312-1352.
Hoberg, G., and G. Phillips. 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies* 23 (10):3773-3811.
Hope, O.-K., D. Hu, and H. Lu. 2016. The benefits of specific risk-factor disclosures. *Review of Accounting Studies* 21 (4):1005-1045.
Huang, X., S. H. Teoh, and Y. Zhang. 2014. Tone management. *The Accounting Review* 89 (3):1083-1113.
Humpherys, S. L., K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix. 2011. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems* 50 (3):585–594.
Hwang, B.-H., and H. H. Kim. 2017. It pays to write well. *Journal of Financial Economics* 124 (2):373-394.

Irani, R. M., and D. Oesch. 2013. Monitoring and corporate disclosure: Evidence from a natural experiment. *Journal of Financial Economics* 109 (2):398-418.

Kahle, K. M., and R. A. Walkling. 1996. The impact of industry classifications on financial research *Journal of Financial and Quantitative Analysis* 31 (3):309-335.

Kim, J., Y. Kim, and J. Zhou. 2017. Languages and earnings management. *Journal of Accounting and Economics* 63 (2):288-306.

Kincaid, J. P., R. Braby, and J. E. Mears. 1988. Electronic Authoring and Delivery of Technical Information. *Journal of Instructional Development* 11 (2):8-13.

Kolk, A., D. Levy, and J. Pinkse. 2008. Corporate responses in an emerging climate regime: The institutionalization and commensuration of carbon disclosure. *European Accounting Review* 17 (4):719-745.

Koo, D. S., J. Julie Wu, and P. E. Yeung. 2017. Earnings Attribution and Information Transfers. *Contemporary Accounting Research* 34 (3):1547-1579.

Kothari, S. P., X. Li, and J. E. Short. 2009. The effect of disclosure my management, analysts and business press on cost of capital, return volatility, and analysts forecasts: A study using content analysis *The Accounting Review* 84 (5):1639-1670.

Kravet, T., and V. Muslu. 2013. Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies* 18 (4):1088-1122.

Krishnan, J. a. E. P. 2003. The North American industry classification system and its implications for accounting research. *Contemporary Accounting Research* 20 (4):685-717.

Lamoureux, C. G., and G. C. Sanger. 1989. Firm size and turn-of-the-year effects in the OTC/Nasdaq market. *Journal of Finance* 44:1219-1245.

Lang, M., and L. Stice-Lawrence. 2015. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics* 60 (2–3):110-135.

Larcker, D., and A. Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50 (2):495-540.

Law, K. K. F., and L. F. Mills. 2015. Taxes and Financial Constraints: Evidence from Linguistic Cues. *Journal of Accounting Research* 53 (4):777-819.

Lawrence, A. 2013. Individual investors and financial disclosure. *Journal of Accounting and Economics* 56 (1):130-147.

Lee, J. 2016. Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls. *The Accounting Review* 91 (1):229-250.

Lee, Y.-J. 2012. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research* 29 (4):1137-1170.

Lehavy, R., F. Li, and K. Merkley. 2011. The effect of annual report readability on analyst following and the properties of their earningsforecasts. *The Accounting Review* 86:1087-1115.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2/3):221-247.

Li, F. 2010. The information content of forward-looking statements in corporate filings - A naive bayesian machine learning approach. *Journal of Accounting Research* 48 (5):1049-1102.

Li, F., R. Lundholm, and M. Minnis. 2013. A measure of competition based on 10-K filings. *Journal of Accounting Research* 51 (2):399-436.

Li, F., M. Minnis, V. Nagar, and M. Rajan. 2014. Knowledge, compensation, and firm value: An empirical analysis of firm communication. *Journal of Accounting and Economics* 58 (1):96-116.

Ljungqvist, A., C. Malloy, and F. Marston. 2009. Rewriting history. *Journal of Finance* 64 (4):1935-1960.

Lo, K., F. Ramos, and R. Rogo. 2017. Earnings management and annual report readability. *Journal of Accounting and Economics* 63 (1):1-25.

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

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

Loughran, T., and B. McDonald. 2014. Measuring readability in financial disclosures. *Journal of Finance* 69:1643-1671.

Lundholm, R., R. Rogo, and J. L. Zhang. 2014. Restoring the tower of babel: how foreign firms communicate with US investors. *The Accounting Review* 89 (4):1453-1485.

Mayew, W. J., M. Sethuraman, and M. Venkatachalam. 2015. MD&A disclosure and the firm's ability to continue as a going concern. *The Accounting Review* 90 (4):1621-1651.

Merkley, K. 2014. Narrative disclosure and earnigns performance: Evidence from R&D disclosures. *The Accounting Review* 89 (2):725-759.

Miller, B. 2010. The effects of reporting complexity on small and large investor trading. *The Accounting Review* 85:2107-2143.

Mukherjee, A., M. Singh, and A. Žaldokas. 2017. Do corporate taxes hinder innovation? *Journal of Financial Economics* 124 (1):195-221.

Mutchler, J., and P. Shane. 1995. A comparative analysis of firms included in and excluded from the NAARS database. *Journal of Accounting Research* 33 (1):193-202.

Newman, M., J. Pennebaker, D. Berry, and R. J. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29 (5):665–675.

Philbrick, D. R. a. W. E. R. 1991. Using value line and IBES analyst forecasts in accounting research. *Journal of Accounting Research* 29 (2):397-417.

Purda, L., and D. Skillicorn. 2015. Accounting cariables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research* 32 (3):1193-1223.

Rosenberg, B., and M. Houglet. 1974. Error rates in CRSP and compustat data bases and their implications. *Journal of Finance* 29 (4):1303-1310.

Schwarz, C. G., and M. E. Potter. 2016. Revisiting mutual fun portfolio disclosure. *Review of Financial Studies* 29 (12):3519-3544.

Segal, B., and D. Segal. 2016. Are managers strategic in reporting non-earnings news? Evidence on timing and news bundling. *Review of Accounting Studies* 21 (4):1203-1244.

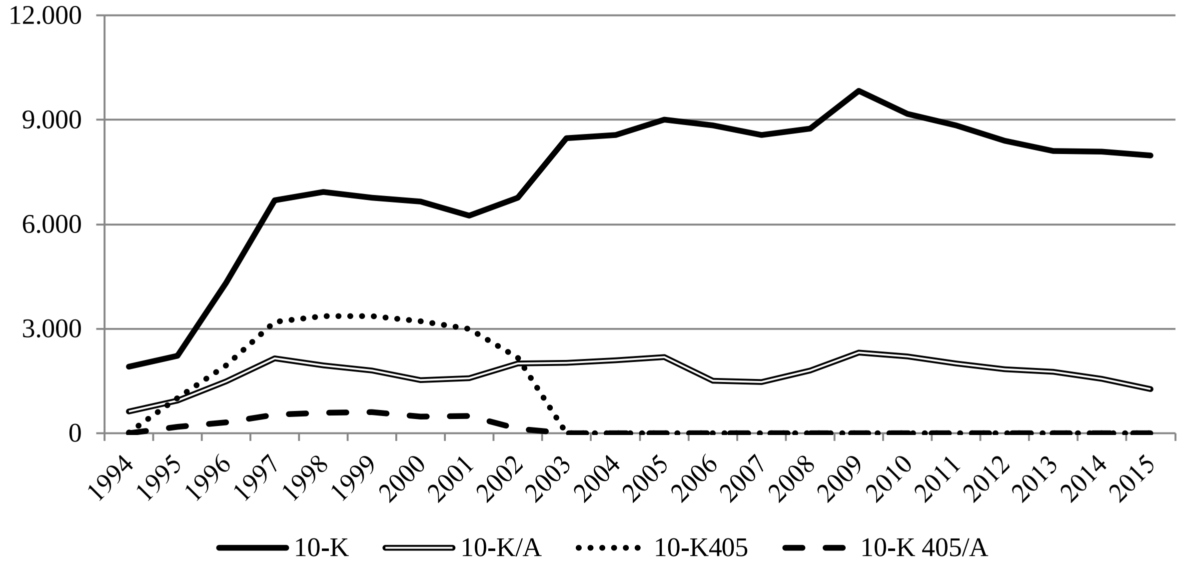
Shumway, T. 1997. The delisting bias in CRSP data. *Journal of Finance* 52 (1):327-340.

Shumway, T., and V. A. Warther. 1999. The delisting bias in CRSP's Nasdaq data and its implications for interpretation of the size effect. *Journal of Finance* 54 (6):2361-2379.

You, H., and X. Zhang. 2009. Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies* 14:559-586.

**Figure 1.** Edgar Fillings

*Panel A.* Trends in 10-K fillings in the period 1994-2015




*Panel B.* Percentage of amended *Edgar* fillings in the period 1994-2015

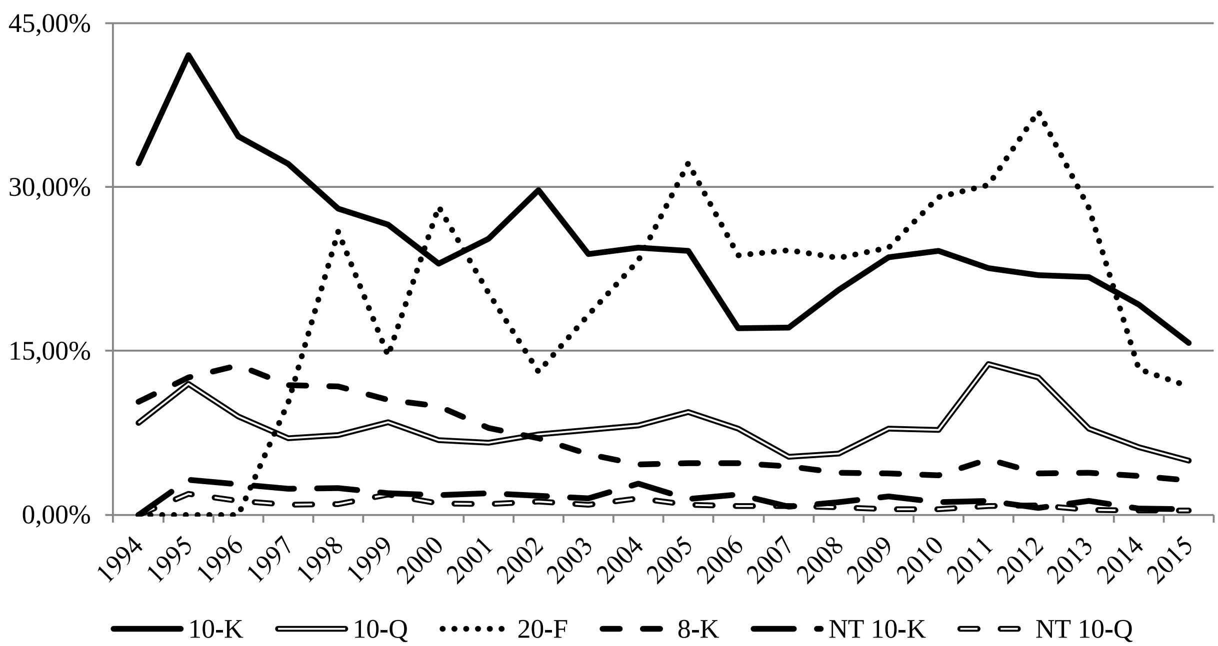
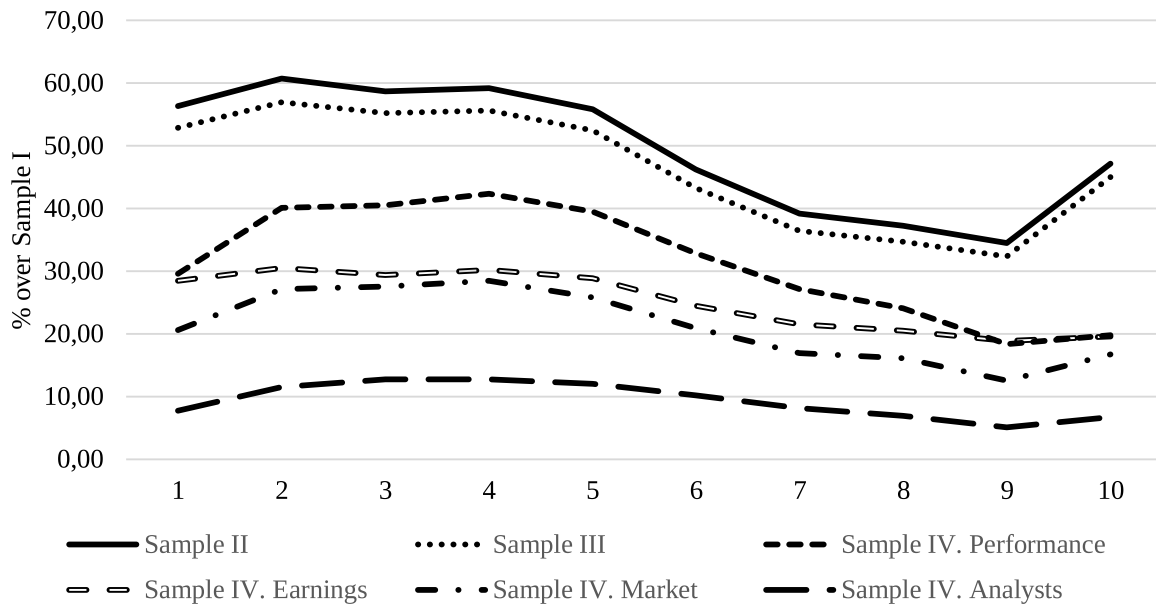
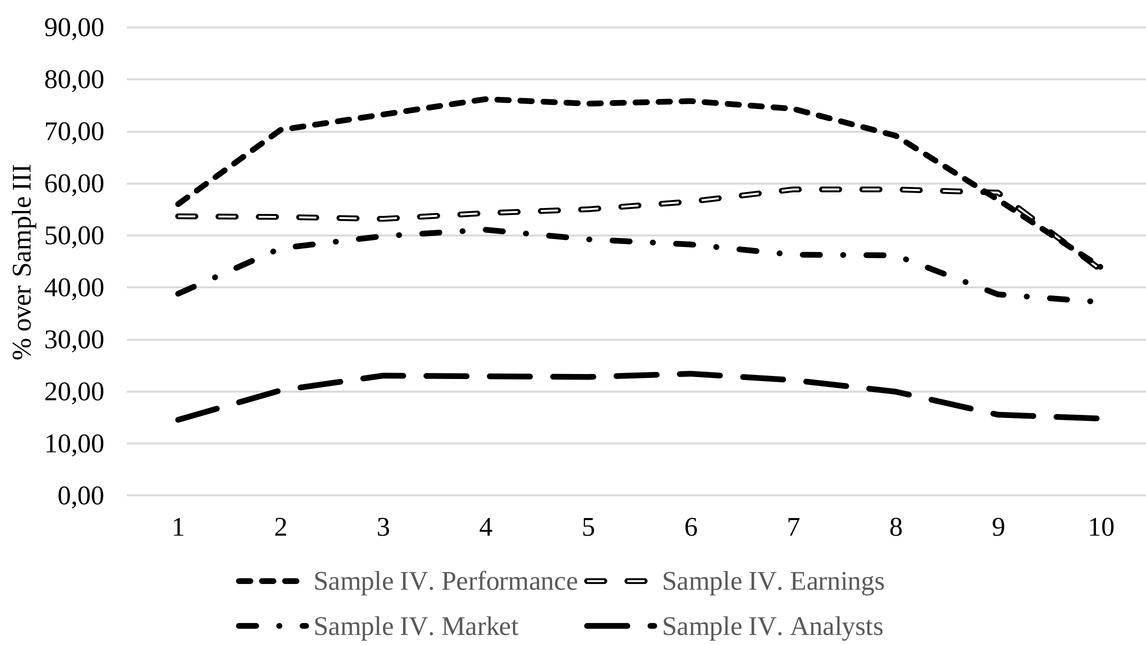


**Figure 2.** Deciles evolution of observations across samples

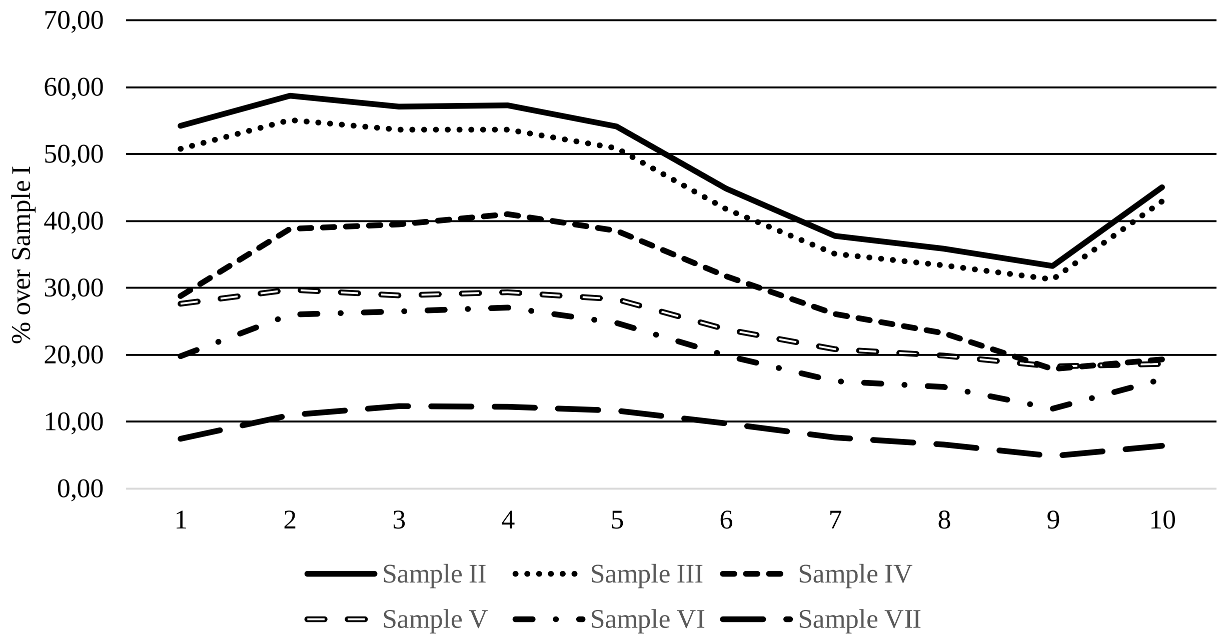
*Panel A.* Deciles evolutions without Link Table *to Crsp/Compustat over Sample I*




*Panel B.* Deciles evolutions without Link Table *to Crsp/Compustat over Sample III*



*Panel C.* Deciles evolutions with Link Table *to Crsp/Compustat over Sample I*




*Panel D.* Deciles evolutions with Link Table *to Crsp/Compustat over Sample III*

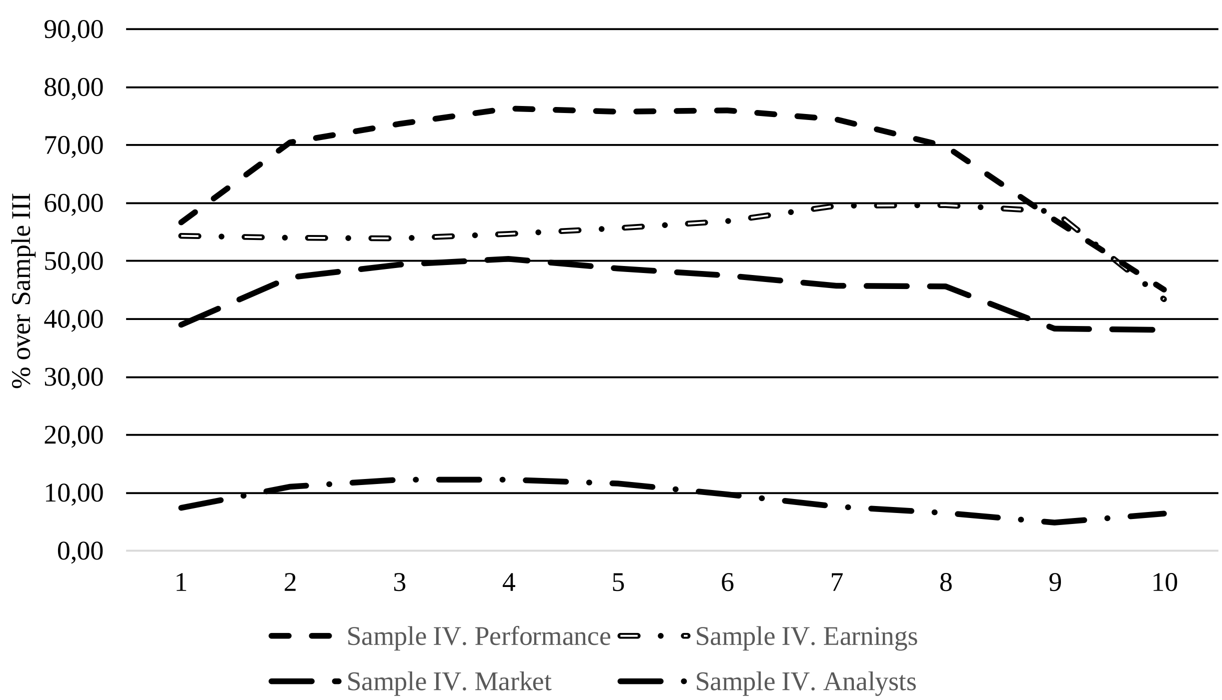


**Table 1.** Summary of Literature Review

*Panel A.* Total articles containing keywords related with firm narratives (2007-2017)

| *Keyword* | **JFE** | **JF** | **RFS** | **JAE** | **JAR** | **TAR** | **CAR** | **AOS** | **RAST** | **EAR** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Textual analysis | 19 | 39 | 2 | 16 | 29 | 3 | 19 | 34 | 16 | 1 | 178 |
| Readability | 96 | 69 | 1 | 23 | 29 | 11 | 28 | 8 | 13 | 2 | 280 |
| Language | 81 | 441 | 16 | 40 | 58 | 7 | 86 | 187 | 19 | 30 | 965 |
| Narrative | 19 | 102 | 0 | 10 | 33 | 3 | 40 | 124 | 16 | 17 | 364 |
| Tone | 20 | 85 | 3 | 20 | 43 | 6 | 45 | 66 | 17 | 5 | 312 |

*Panel B.* Total article published by leading Journals (2007-2017)

| | **JFE** | **JF** | **RFS** | **JAE** | **JAR** | **TAR** | **CAR** | **AOS** | **RAST** | **EAR** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Articles published | 5 | 2 | 2 | 10 | 8 | 10 | 6 | 1 | 10 | 2 | 56 |

*Panel C.* Total article published by year (2007-2017)

| | **2008** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Articles published | 2 | 2 | 4 | 4 | 5 | 5 | 7 | 8 | 12 | 7 | 56 |

Journal of Accounting and Economics (JAE), Journal of Accounting Research (JAR), The accounting Review (TAR), Review of Accounting Studies (RAST), Contemporary Accounting Research (CAR), Accounting Organizations and Society (AOS) and European Accounting Review (EAR), Journal of Finance (JF), Journal of Financial Economics (JFE), and Review of Financial Studies (RFS).

*Panel D.* Articles details

| | | | Sample | | Measures | | |
|---|---|---|---|---|---|---|---|
| **Authors** | **Journal** | **Year** | **Period** | **No. Observations** | **Readability** | **Tone** | **Others** |
| Hwang & Kim | JFE | 2017 | 2003-2013 | 92 funds; 6,507 monthly-yrs. | Readability | . | . |
| Mukherjee, Singh & Žaldokas | JFE | 2017 | 1990-2006 | 47,632 firm-yrs | . | . | Press releases on "New Products" |
| Boone, Floros, & Johnson | JFE | 2016 | 1996-2011 | 2,199 IPOs, 875 redacting firms | . | . | Registration statements uses term "confidential" |
| Irani & Oesch | JFE | 2013 | 1994-2005 | 39,384 firm-yrs | Fog index and number of words, Li (2008) | . | . |
| Loughran & McDonald | JFE | 2013 | 1997–2010 | 1,887 IPOs | . | Uncertain, weak, negative, positive, legal, and strong. | . |
| Li | JAE | 2008 | 1994-2004 | 55,719 firm-yrs | Fog index and average length of sentence | . | . |
| Lang & Stice-Lawrence | JAE | 2015 | 1998-2011 | 15,000 firms, 87,608 obs., 42 non-US countries | Fog index | . | Length, boiler plate, comparability and complexity |
| Frankel, Jennings, & Lee | JAE | 2016 | 1994-2013 | 71847 firm-yrs | Fog index | . | File size, length |
| Drake, RoulsTone & Thornock | JAE | 2016 | 2003-2012 | 24,617 firm-yrs | | | total number of words (table characters) |
| Lo, Ramos, & Rogo | JAE | 2017 | 2000-2012 | 4,855 firms, 26,967 firm-yrs | Fog index | . | . |
| Bonsall IV, Leone, & Miller | JAE | 2017 | 1994-2011 | 46,424 firm-yrs | Bog index, Fog index | . | Plain English index, sentence length, passive voice, weak verbs, overused and complex words, jargon |
| Guay, Samuels, &Taylor | JAE | 2016 | 1995-2012 | 72,366 firm-yrs | REadIndex (Flech, LIX, RIX, Fog index, ARI, SMOG) principal component. Readability | . | Length 10K |
| Lawrence | JAE | 2013 | 1994- | 95,107 firm-yrs | Fog index (Li, 2008) | | |
| Dyer, Lang & Stice-Lawrence | JAE | 2017 | 1996-2013 | 10,452 firms, 75,991 firm-yrs | Fog index | | Boilerplate, length, redundant words, sticky words |

| Authors | Journal | Year | Period | No. Observations | Readability | Tone | Others |
|---|---|---|---|---|---|---|---|
| Li, Minnis, Nagar, & Rajan | JAE | 2014 | 2003-2007 | 17,419 conference calls | . | . | length of text, number of comments |
| Cho, Roberts, & Patten | AOS | 2010 | 2002 | 190 firms | . | Optimism (Diction dictionary) | certainty (Diction dictionary) |
| Loughran & McDonald | JF | 2011 | 1994-2008 | 50,115 10Ks, 37,287 MD&A | | Uncertain, weak, negative, positive, legal, and strong. Harvard list. | |
| Loughran & McDonald | JF | 2014 | 1994-2011 | 66,707 firm-yrs | Fog index | | length, words per sentence, % complex, common, financial words |
| Gruning | EAR | 2011 | 2005-2008 | 127,895 firm-yrs | . | . | AIMD (Artificial Intelligence measurement of disclosure). N-Gram |
| Kolk, Levy, & Pinkse | EAR | 2008 | 2003-2007 | FT500 firms/380 responding firms | . | . | Carbon disclosure project (CDP 5) |
| Lundholm, Rogo, & Zhang | TAR | 2014 | 2000-2012 | 3,499 (1,582) foreign firm-yrs (press releases) and 37,344 (21,976) from US | Fog index | | length |
| Brochet, Naranjo, & Yu | TAR | 2016 | 2002-2010 | 25,830 conference calls | Fog index | . | length |
| Lehavy, Li, & Merkley | TAR | 2011 | 1995-2006 | 57,642 firm-yrs | Fog index | . | . |
| Miller | TAR | 2010 | 1995-2006 | 13,000 10K reports | Fog index | . | . |
| Huang, Teoh, & Zhang | TAR | 2014 | 1997-2007 | 14,475 firm-yrs | . | Loughran and McDonald (2011), Henry (2008) Harvard General Inquiry | . |
| Lee | TAR | 2016 | 2002-2011 | 40,820 conference calls | . | . | cosine-similarity |
| Mayew, Sethuraman, & Venkatachalam | TAR | 2015 | 1995-2012 | 45,265 (460) Non-(bankrupt) firm-yrs | | Loughran and McDonald (2011) | |
| Merkely | TAR | 2014 | 1996-2007 | 22,482 firm-yrs | . | . | Dictionary of common R&D keywords |

| Authors | Journal | Year | Period | No. Observations | Readability | Tone | Others |
|---|---|---|---|---|---|---|---|
| Henry & Leone | TAR | 2016 | 2004-2012 | 143,972 earnings announcements | . | Henry (2006, 2008), Diction, General Inquirer, Loughran and McDonald (2011) | . |
| Kothari, Li, & Short | TAR | 2009 | 1996-2001 | 889 firms and 5,350 firm-yrs | . | General Inquirer | Market, firm, organizational, reputational, performance, & regulatory risks |
| Hoberg, & Phillips | RFS | 2010 | 1997-2006 | 50,104 firm-yrs | . | . | cosine-similarity |
| Hoberg & Maksimovic | RFS | 2014 | 1997-2009 | 48,512 10Ks | . | . | cosine-similarity, boiler plate, constraints scores |
| Feldman, Gvindaraj, Kivnaat, Segal | RAST | 2010 | 1993-2007 | 153,988 10K, 10Q | . | Loughran and McDonald (2011) | . |
| Segal, & Segal | RAST | 2016 | 2005-2013 | 335,328 8Ks | . | Loughran and McDonald (2011) | . |
| Kravet, & Muslu | RAST | 2013 | 1994-2007 | 28,110 firm-yrs | . | . | risk-related sentences |
| Hope, Hu, & Lu | RAST | 2016 | 2001-2004 | 14,865 firm yrs | . | . | specificity |
| Davis, Ge, Matsumoto, & Zhang | RAST | 2015 | 2002-2009 | 25 firms, 121 CEOs & CFOs | . | Diction, Henry (2006, 2008), Loughran and McDonald (2011) | . |
| Campbell, Chen, Dhaliwal, Lu, & Steele | RAST | 2014 | 2005-2009 | 9,076 10Ks | . | . | risk-related list |
| Brochet, Loumioti, & Serafeim | RAST | 2015 | 2002-2008 | 70,042 conference calls | . | . | short-term keywords. |
| Bonsall IV, & Miller | RAST | 2017 | 1994-2014 | 3,659 10Ks | Bog index Bonsall et al. (2016) | Forward-looking (Bo-zanic et al. 2015) Tone, Uncertainty (Loughran and McDonald, 2011) | Risk disclosures (Kravet and Muslu 2013; Campbell et al. 2014). |
| Baginski, Demers, Wang, Yu | RAST | 2016 | 1997-2006 | 1,764 mngt earnings forecasts | . | Diction, Loughran and McDonald (2011), | . |
| You, & Zhang | RAST | 2009 | 1995-2005 | 123,449 10Ks | complexity | . | . |

| Authors | Journal | Year | Period | No. Observations | Readability | Tone | Others |
|---|---|---|---|---|---|---|---|
| Bozanic & Thevenot | CAR | 2015 | 1984-2012 | 1,838 firm yrs 160 unique firms. | Fog index, complex words | . | similarity (Brown and Tucker, 2010) Diversity (Goel et al. 2010, Humphreys et al. 2011) |
| Purda, & Skillicorn | CAR | 2015 | 1994-2006 | 240 firms, 4,895 10Ks and 10Q | . | Uncertainty, negativity (Loughran and McDonald 2011) | litigation (Loughran and McDonald 2011), deception (Newman et al. 2003) |
| Lee | CAR | 2012 | 2001-2007 | 60,161 earnings announcements | Fog index, Length | . | . |
| Koo, Wu, & Yeung | CAR | 2017 | 2001-2007 | 1,765 mngt forecasts | . | . | attribution phrases |
| Davis, Piger, & Sedor | CAR | 2012 | 1998-2003 | 23,017 firm-quarters | | Diction | |
| Davis, & Tama-Sweet | CAR | 2012 | 1998-2003 | 11, 826 firm-quarters | . | Diction | . |
| Li, Lundholm, & Minnis | JAR | 2013 | 1995-2009 | 33,492 firm-yrs | . | Tone | length, competition keywords |
| Li | JAR | 2010 | 1997-2007 | 145,479 firms quarters | Fog index | LIWC and Diction | Risk (Li 2010 ) Forward-looking Diction, General Inquirer and LIWC |
| Law, & Mills | JAR | 2015 | 1994-2011 | 5,418 firm-years (2,340 firms) | Fog index, N. of words, Flesch Reading Ease, Kincaid Readability | Loughran and McDonald (2011) | Negotiation, Uncertainty, Strong, litigious, Constraints words |
| Larcker, & Zakolyukina | JAR | 2012 | 2003-2007 | 29,663 conference calls | . | LIWC, emotions list | References, Calculation Cognitive Process, Other Cues from LIWC and self-constructed list |
| Bushman, Hen-dricks & Williams | JAR | 2016 | 1996-2012 | 14,633 bank-quarters | . | . | competition (Li et al. 2013) |
| Brown & Tucker | JAR | 2011 | 1997-2006 | 28,279 firm years | . | . | similarity cosine |
| Allee, & Deangelis | JAR | 2015 | 2004-2014 | 73,201 transcripts | . | Loughran and McDonald (2011) | . |
| Loughran & McDonald | JAR | 2016 | | | Fog index, Flesch, Flesch-Kincaid, Common, Financial | Loughran and McDonal 2011, Henry 2008, Harvard GI, Dition | Uncertainty, weak and strong modal, cosine similarity, Naïve Bayes |

**Table 2.** Sample selection procedure

*Panel A.* Procedure to create Sample I

| | Firms (cik) | Observations | % |
|---|---|---|---|
| Total observations from 10-K files | | 225,417 | |
| Possible duplicates | | (64,286) | |
| | | | |
| 10-K fillings in *Edgar* | | 161,131 | |
| Missing Conformed Period of Report | | (1,759) | |
| Duplicates (cik, FYEAR, FYR) | | (913) | |
| **Final Observations (Sample 0)** | **33,466** | **159,338** | |
| Presence of HTML tags | | 88,860 | 56% |
| Item 6 | | 146,545 | 92% |
| Item 7 | | 147,303 | 92% |
| Item 7A | | 116,168 | 73% |
| Item 8 | | 147,917 | 93% |

*Panel B.* Selection procedure to creates Sample from I to IV

| **Sample name** | **Sample type** | **10K reports** | | | **Description** |
|---|---|---|---|---|---|
| | | **No Link Table** | **Yes Link Table** | | |
| Sample I | Subsample of Sample 0 | 147,303 | | Item 7 observations | All the 10K containing Item 7 section. |
| Sample II | Subsample of Sample I | 72,934 | 70,489 | (1) Merge with CRSP/COMPUSTAT | Observations containing Central Index Key (cik), Fiscal Year (fyear), Fiscal Month (fyr), link to Compustat (linkprim equals to "P", and "C"), and Stock ownership code (no subsidiary). |
| Sample III | Subsample of Sample I | 68,494 | 66,083 | (2) Merge with CRSP/COMPUSTAT | Current and lagged values of Total assets (at), Revenues (revt), Short term debt (dlc), and Long term debt (dltt). |
| Sample IV. Performance | Subsample of Sample III | 46,288 | 44,947 | Performance | Current and lagged values of Net income (ni), Income Before Extraordinary Items (ib), Acquisition (aqc), Sale of Common and Preferred Stock (sstk), Common Ordinary Equity (ceq), Net Cash Flow from Operating Activities (oancf), Market Value (mkvalt). |
| Sample IV. Earnings | Subsample of Sample III | 37,162 | 36,183 | Earnings | Current and lagged values of Before Extraordinary Items (ib), Net Cash Flow from Operating Activities (oancf), Advertising Expenses (xad), Research and Development Expenses (xrd), Net Change Assets and Liabilities (aoloch), Accounts Payable (ap), Invetories (invt), PPE (ppegt), Receivables (rect), Accrued Income Taxes (txach). |
| Sample IV. Market | Subsample of Sample III | 31,290 | 29,990 | Market | Current and lagged values of Income Before Extraordinary Items (ib), Common Ordinary Equity (ceq), Market Value (mkvalt), Net Cash Flow from Operating Activities (oancf), Sale of Common, Preferred Stock (sstk), and adjustements for computing firms returns. |
| Sample IV. Analysts | Subsample of Sample III | 13,785 | 13,250 | Analysts | Current and lagged values of Income Before Extraordinary Items (ib), Common Ordinary Equity (ceq), Market Value (mkvalt), Net Cash Flow from Operating Activities (oancf), Sale of Common and Preferred Stock (sstk), and earnings per share in U.S. currency. |

We report samples obtained from merging qualitative and quantitative data from Edgar database and CRSP/Compustat of Warthon database. We use cik code, Fiscal year (fyear), and Fiscal month (fyr) as reported both in annual reports and CRSP/Compustat as the three identifiers between Edgar and CRSP/Compustat database. To extract cik and fiscal variables from annual reports, we refer to the master file provided by the SEC, and "Conformed Period of Report".

**Table 3.** Descriptive statistics of Tone and Readability measures for Sample I (N=147,303).

*Panel A. Average file size and number of years available*

| VARIABLES | Mean | Median | Sd | Min | P25 | P75 | Max |
|---|---|---|---|---|---|---|---|
| Reporting month | 10 | 12 | 3 | 1 | 9 | 12 | 12 |
| File size (KB) | 3,262 | 541 | 123,434 | 0 | 155 | 1,321 | 19,339,000 |

*Panel B. Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer.*

| VARIABLES | Mean | Median | Sd | Min | P25 | P75 | Max |
|---|---|---|---|---|---|---|---|
| Diction Neg. | 0.0147 | 0.0116 | 0.0197 | 0 | 0.0084 | 0.0155 | 0.2727 |
| Diction Pos. | 0.0091 | 0.0099 | 0.0061 | 0 | 0.0055 | 0.0128 | 0.0909 |
| LM Neg. | 0.0124 | 0.011 | 0.0143 | 0 | 0.0066 | 0.0152 | 0.375 |
| LM Pos. | 0.0049 | 0.005 | 0.0035 | 0 | 0.0027 | 0.0069 | 0.0426 |
| Inquirer Neg. | 0.0179 | 0.0202 | 0.0103 | 0 | 0.0143 | 0.0245 | 0.0926 |
| Inquirer Pos. | 0.1201 | 0.1201 | 0.0452 | 0 | 0.1022 | 0.1386 | 0.9375 |

*Panel C. Content analysis of other attributes.*

| VARIABLES | Mean | Median | Sd | Min | P25 | P75 | Max |
|---|---|---|---|---|---|---|---|
| Constraining | 0.0055 | 0.0054 | 0.0049 | 0 | 0.0027 | 0.0073 | 0.0976 |
| LM Litigious | 0.0061 | 0.0044 | 0.0071 | 0 | 0.0023 | 0.0072 | 0.1315 |
| LM Strong | 0.0024 | 0.0021 | 0.0022 | 0 | 0.001 | 0.0032 | 0.06 |
| LM Weak | 0.0033 | 0.0028 | 0.0036 | 0 | 0.0013 | 0.004 | 0.1112 |
| LM Uncertainty | 0.0112 | 0.0114 | 0.0077 | 0 | 0.0074 | 0.0147 | 0.2223 |
| Causation (Dikolli) | 0.0341 | 0.0302 | 0.0161 | 0 | 0.0262 | 0.0357 | 0.2 |
| Forward Look (Li) | 0.0075 | 0.0072 | 0.0061 | 0 | 0.0039 | 0.01 | 0.1667 |

*Panel D. Readability measures*

| VARIABLES | Mean | Median | Sd | Min | P25 | P75 | Max |
|---|---|---|---|---|---|---|---|
| Fog Index | 18.9637 | 18.5429 | 3.8728 | 0 | 17.022 | 20.5724 | 88.8067 |
| Flesch Index | 22.606 | 26.7765 | 14.3684 | -178.0791 | 20.0605 | 31.2249 | 115.1627 |
| Flesch Kincaid index | 14.4483 | 14.2818 | 2.9151 | -2.1218 | 13.0137 | 15.6375 | 82.5664 |

**Table 4.** Tone and Readability differences between Sample I and Sample II.

*Panel A. Differences in Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer between Missing Narratives and Sample II*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 74,369) | | Sample II (Obs. 72,934) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (76,814) | | Sample II (Obs. 70,489) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Diction Neg. | 0.0185 | 0.0124 | 0.0109 | 0.0111 | 75.20 | 51.36 | 0.00 | 0.0182 | 0.0122 | 0.011 | 0.0112 | 70.84 | 44.45 | 0.00 |
| Diction Pos. | 0.0084 | 0.0091 | 0.0098 | 0.0105 | -44.66 | -50.91 | 0.00 | 0.0084 | 0.0092 | 0.0098 | 0.0105 | -44.29 | -48.89 | 0.00 |
| LM Neg. | 0.0137 | 0.011 | 0.011 | 0.0111 | 36.44 | -3.01 | 0.01 | 0.0136 | 0.0109 | 0.0111 | 0.0111 | 33.97 | -5.02 | 0.00 |
| LM Pos. | 0.0044 | 0.0044 | 0.0054 | 0.0054 | -54.32 | -63.14 | 0.00 | 0.0044 | 0.0044 | 0.0054 | 0.0054 | -55.67 | -64.26 | 0.00 |
| Inquirer Neg. | 0.0164 | 0.0191 | 0.0194 | 0.0211 | -56.92 | -53.34 | 0.00 | 0.0164 | 0.0192 | 0.0195 | 0.0211 | -57.38 | -52.58 | 0.00 |
| Inquirer Pos. | 0.1168 | 0.1192 | 0.1236 | 0.1208 | -28.86 | -22.22 | 0.00 | 0.117 | 0.119 | 0.1236 | 0.121 | -28.24 | -23.17 | 0.00 |

*Panel B. Differences in Content analysis of other attributes between Missing Narratives and Sample II*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 74,369) | | Sample II (Obs. 72,934) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (76,814) | | Sample II (Obs. 70,489) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Constraining | 0.005 | 0.0048 | 0.006 | 0.0058 | -42.29 | -54.53 | 0.00 | 0.005 | 0.0049 | 0.006 | 0.0058 | -39.19 | -51.57 | 0.00 |
| LM Litigious | 0.0051 | 0.0038 | 0.0071 | 0.0049 | -56.28 | -76.33 | 0.00 | 0.0052 | 0.0038 | 0.0071 | 0.0049 | -51.20 | -71.97 | 0.00 |
| LM Strong | 0.0026 | 0.0022 | 0.0022 | 0.0021 | 28.04 | 1.92 | 0.00 | 0.0025 | 0.0021 | 0.0023 | 0.0021 | 21.82 | -4.20 | 0.52 |
| LM Weak | 0.003 | 0.0025 | 0.0036 | 0.0029 | -34.53 | -40.79 | 0.00 | 0.003 | 0.0025 | 0.0036 | 0.0029 | -35.96 | -43.54 | 0.00 |
| LM Uncertainty | 0.0103 | 0.0106 | 0.012 | 0.012 | -41.61 | -50.45 | 0.00 | 0.0103 | 0.0107 | 0.0121 | 0.012 | -44.36 | -52.97 | 0.00 |
| Causation (Dikolli) | 0.0361 | 0.0303 | 0.032 | 0.0302 | 48.83 | 8.77 | 0.32 | 0.036 | 0.0303 | 0.032 | 0.0302 | 48.46 | 9.86 | 0.08 |
| Forward Look (Li) | 0.0076 | 0.007 | 0.0074 | 0.0073 | 4.52 | -14.54 | 0.00 | 0.0075 | 0.007 | 0.0075 | 0.0074 | -1.03 | -19.74 | 0.00 |

*Panel C. Differences in Readability measures of other attributes between Missing Narratives Sample II*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 74,369) | | Sample II (Obs. 72,934) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (76,814) | | Sample II (Obs. 70,489) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Fog Index | 19.454 | 19.386 | 18.463 | 17.945 | 49.52 | 88.64 | 0.00 | 19.430 | 19.333 | 18.456 | 17.940 | 48.58 | 86.56 | 0.00 |
| Flesch Index | 20.345 | 25.290 | 24.911 | 27.904 | -61.77 | -62.62 | 0.00 | 20.393 | 25.327 | 25.017 | 27.955 | -62.52 | -63.49 | 0.00 |
| Flesch Kincaid index | 14.619 | 14.680 | 14.275 | 13.906 | 22.70 | 58.54 | 0.00 | 14.612 | 14.680 | 14.270 | 13.903 | 22.54 | 57.36 | 0.00 |

**Table 5.** Tone and Readability differences between Sample I and Sample III.

*Panel A. Differences in Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer between Missing Narratives and Sample III*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 78,809) | | Sample III (68,494) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 81,220) | | Sample III (Obs. 66,083) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Diction Neg. | 0.0181 | 0.0123 | 0.0109 | 0.0111 | 71.22 | 49.67 | 0.00 | 0.0178 | 0.0122 | 0.011 | 0.0112 | 67.16 | 43.23 | 0.00 |
| Diction Pos. | 0.0085 | 0.0092 | 0.0098 | 0.0105 | -39.94 | -46.11 | 0.00 | 0.0085 | 0.0093 | 0.0098 | 0.0105 | -39.53 | -44.23 | 0.00 |
| LM Neg. | 0.0136 | 0.0109 | 0.0111 | 0.0112 | 33.42 | -6.08 | 0.00 | 0.0134 | 0.0108 | 0.0111 | 0.0112 | 31.12 | -7.98 | 0.00 |
| LM Pos. | 0.0044 | 0.0045 | 0.0054 | 0.0054 | -50.95 | -60.21 | 0.00 | 0.0044 | 0.0045 | 0.0054 | 0.0055 | -52.42 | -61.48 | 0.00 |
| Inquirer Neg. | 0.0165 | 0.0192 | 0.0194 | 0.0212 | -53.73 | -52.73 | 0.00 | 0.0166 | 0.0193 | 0.0195 | 0.0212 | -54.33 | -52.21 | 0.00 |
| Inquirer Pos. | 0.1171 | 0.1191 | 0.1237 | 0.1209 | -27.95 | -22.39 | 0.00 | 0.1172 | 0.119 | 0.1237 | 0.1211 | -27.53 | -23.53 | 0.00 |

*Panel B. Differences in Content analysis of other attributes between Missing Narratives and Sample III*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 78,809) | | Sample III (68,494) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 81,220) | | Sample III (Obs. 66,083) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Constraining | 0.005 | 0.0049 | 0.0061 | 0.0058 | -42.04 | -53.45 | 0.00 | 0.005 | 0.0049 | 0.006 | 0.0058 | -38.98 | -50.65 | 0.00 |
| LM Litigious | 0.0051 | 0.0038 | 0.0072 | 0.005 | -57.02 | -75.15 | 0.00 | 0.0052 | 0.0039 | 0.0072 | 0.005 | -51.96 | -70.93 | 0.00 |
| LM Strong | 0.0026 | 0.0022 | 0.0022 | 0.0021 | 32.95 | 8.96 | 0.00 | 0.0025 | 0.0022 | 0.0022 | 0.0021 | 27.26 | 3.36 | 0.00 |
| LM Weak | 0.003 | 0.0026 | 0.0036 | 0.0029 | -29.06 | -34.50 | 0.00 | 0.003 | 0.0026 | 0.0036 | 0.0029 | -30.38 | -37.05 | 0.00 |
| LM Uncertainty | 0.0105 | 0.0107 | 0.012 | 0.012 | -38.04 | -46.70 | 0.00 | 0.0104 | 0.0108 | 0.0121 | 0.012 | -40.72 | -49.17 | 0.00 |
| Causation (Dikolli) | 0.0358 | 0.0302 | 0.0321 | 0.0303 | 43.73 | 5.64 | 0.30 | 0.0357 | 0.0302 | 0.0321 | 0.0302 | 43.33 | 6.50 | 0.68 |
| Forward Look (Li) | 0.0076 | 0.0071 | 0.0073 | 0.0073 | 10.51 | -7.47 | 0.00 | 0.0076 | 0.007 | 0.0074 | 0.0073 | 5.47 | -12.18 | 0.00 |

*Panel C. Differences in Readability measures of other attributes between Missing Narratives Sample III*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 78,809) | | Sample III (68,494) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 81,220) | | Sample III (Obs. 66,083) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Fog Index | 19.392 | 19.2851 | 18.4708 | 17.9393 | 45.86 | 83.59 | 0.00 | 19.3696 | 19.2401 | 18.4648 | 17.9339 | 44.90 | 81.69 | 0.00 |
| Flesch Index | 20.6934 | 25.5254 | 24.8065 | 27.8675 | -55.37 | -56.87 | 0.00 | 20.7352 | 25.5605 | 24.9052 | 27.9167 | -55.98 | -57.66 | 0.00 |
| Flesch Kincaid index | 14.5912 | 14.68 | 14.284 | 13.906 | 20.20 | 54.49 | 0.00 | 14.5849 | 14.68 | 14.2805 | 13.9033 | 19.96 | 53.44 | 0.00 |

**Table 6.** Tone and Readability differences between Sample III and Sample IV. Performance

*Panel A. Differences in Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer between Missing Narratives and Sample IV.Performance*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 22,206) | | Sample IV (Obs. 46,288) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 21,136) | | Sample IV (Obs. 44,947) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Diction Neg. | 0.0092 | 0.0096 | 0.0117 | 0.0116 | -53.43 | -51.85 | 0.00 | 0.0093 | 0.0098 | 0.0117 | 0.0116 | -50.11 | -49.13 | 0.00 |
| Diction Pos. | 0.0083 | 0.009 | 0.0105 | 0.0109 | -52.16 | -49.74 | 0.00 | 0.0083 | 0.009 | 0.0105 | 0.0109 | -48.34 | -46.58 | 0.00 |
| LM Neg. | 0.0091 | 0.009 | 0.012 | 0.0118 | -53.73 | -54.18 | 0.00 | 0.0092 | 0.0091 | 0.012 | 0.0118 | -50.69 | -51.36 | 0.00 |
| LM Pos. | 0.0048 | 0.005 | 0.0057 | 0.0056 | -33.46 | -31.60 | 0.00 | 0.0048 | 0.005 | 0.0057 | 0.0056 | -31.49 | -29.51 | 0.00 |
| Inquirer Neg. | 0.0164 | 0.0192 | 0.0208 | 0.0219 | -59.02 | -48.37 | 0.00 | 0.0166 | 0.0192 | 0.0208 | 0.0219 | -56.49 | -46.85 | 0.00 |
| Inquirer Pos. | 0.1293 | 0.127 | 0.1209 | 0.1189 | 27.11 | 35.67 | 0.00 | 0.1293 | 0.1272 | 0.1211 | 0.1191 | 26.41 | 34.46 | 0.00 |

*Panel B. Differences in Content analysis of other attributes between Missing Narratives and Sample IV.Performance*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 22,206) | | Sample IV (Obs. 46,288) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 21,136) | | Sample IV (Obs. 44,947) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Constraining | 0.0053 | 0.0042 | 0.0064 | 0.0063 | -28.97 | -71.13 | 0.00 | 0.0053 | 0.0042 | 0.0064 | 0.0063 | -28.56 | -69.07 | 0.00 |
| LM Litigious | 0.0082 | 0.0045 | 0.0068 | 0.0051 | 23.00 | -13.20 | 0.00 | 0.0081 | 0.0045 | 0.0067 | 0.0051 | 21.98 | -12.45 | 0.00 |
| LM Strong | 0.002 | 0.0018 | 0.0023 | 0.0022 | -25.32 | -36.57 | 0.00 | 0.002 | 0.0018 | 0.0023 | 0.0022 | -23.27 | -34.03 | 0.00 |
| LM Weak | 0.0033 | 0.0023 | 0.0037 | 0.0031 | -14.06 | -51.71 | 0.00 | 0.0033 | 0.0024 | 0.0038 | 0.0031 | -14.26 | -49.04 | 0.00 |
| LM Uncertainty | 0.0101 | 0.0103 | 0.0129 | 0.0125 | -51.79 | -55.12 | 0.00 | 0.0102 | 0.0105 | 0.0129 | 0.0125 | -49.01 | -52.44 | 0.00 |
| Causation (Dikolli) | 0.0342 | 0.0309 | 0.0311 | 0.03 | 34.14 | 20.58 | 0.00 | 0.034 | 0.0308 | 0.0312 | 0.03 | 31.37 | 18.00 | 0.00 |
| Forward Look (Li) | 0.0057 | 0.0056 | 0.0081 | 0.0078 | -63.45 | -69.50 | 0.00 | 0.0058 | 0.0057 | 0.0081 | 0.0078 | -60.03 | -66.14 | 0.00 |

*Panel C. Differences in Readability measures of other attributes between Missing Narratives Sample IV.Performance*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 22,206) | | Sample IV (Obs. 46,288) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 21,136) | | Sample IV (Obs. 44,947) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Fog Index | 18.9672 | 18.2276 | 18.2327 | 17.8539 | 23.86 | 21.03 | 0.00 | 18.9552 | 18.2163 | 18.2342 | 17.8518 | 23.13 | 20.11 | 0.00 |
| Flesch Index | 21.8842 | 26.7944 | 26.2085 | 28.1963 | -41.96 | -26.00 | 0.00 | 22.0588 | 26.8847 | 26.2437 | 28.2235 | -40.02 | -24.37 | 0.00 |
| Flesch Kincaid index | 14.6061 | 14.064 | 14.1295 | 13.851 | 19.31 | 14.29 | 0.00 | 14.6015 | 14.0584 | 14.1296 | 13.8514 | 18.89 | 13.63 | 0.00 |

**Table 7.** Tone and Readability differences between Sample III and Sample IV Earnings

*Panel A. Differences in Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer between Missing Narratives and Sample IV.Earnings*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 31,332) | | Sample IV (Obs. 37,162) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 29,990) | | Sample IV (Obs. 36,183) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Diction Neg. | 0.0107 | 0.0111 | 0.011 | 0.0112 | -5.23 | -3.26 | 0.02 | 0.0108 | 0.0111 | 0.0111 | 0.0112 | -5.31 | -3.40 | 0.01 |
| Diction Pos. | 0.0099 | 0.0109 | 0.0097 | 0.0102 | 4.31 | 13.99 | 0.00 | 0.0099 | 0.0108 | 0.0097 | 0.0102 | 4.67 | 14.10 | 0.00 |
| LM Neg. | 0.0108 | 0.0112 | 0.0113 | 0.0111 | -8.88 | -5.38 | 0.25 | 0.0109 | 0.0112 | 0.0113 | 0.0112 | -8.95 | -5.37 | 0.30 |
| LM Pos. | 0.005 | 0.0053 | 0.0056 | 0.0056 | -24.79 | -22.04 | 0.00 | 0.0051 | 0.0053 | 0.0057 | 0.0056 | -24.31 | -21.45 | 0.00 |
| Inquirer Neg. | 0.0184 | 0.0207 | 0.0202 | 0.0216 | -25.97 | -21.67 | 0.00 | 0.0184 | 0.0207 | 0.0203 | 0.0216 | -25.94 | -21.61 | 0.00 |
| Inquirer Pos. | 0.1229 | 0.121 | 0.1242 | 0.1209 | -4.47 | -3.33 | 0.71 | 0.1229 | 0.1211 | 0.1244 | 0.1211 | -4.86 | -3.47 | 0.86 |

*Panel B. Differences in Content analysis of other attributes between Missing Narratives and Sample IV.Earnings*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 31,332) | | Sample IV (Obs. 37,162) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 29,990) | | Sample IV (Obs. 36,183) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Constraining | 0.0062 | 0.006 | 0.0059 | 0.0057 | 9.33 | 5.50 | 0.00 | 0.0062 | 0.0059 | 0.0059 | 0.0057 | 8.75 | 5.02 | 0.00 |
| LM Litigious | 0.0078 | 0.005 | 0.0068 | 0.005 | 17.26 | 5.34 | 0.59 | 0.0077 | 0.005 | 0.0067 | 0.005 | 17.61 | 5.31 | 0.94 |
| LM Strong | 0.0019 | 0.0019 | 0.0024 | 0.0023 | -44.77 | -45.36 | 0.00 | 0.0019 | 0.0019 | 0.0025 | 0.0023 | -43.74 | -45.15 | 0.00 |
| LM Weak | 0.0034 | 0.0028 | 0.0038 | 0.003 | -13.69 | -21.84 | 0.00 | 0.0034 | 0.0028 | 0.0038 | 0.003 | -15.01 | -23.08 | 0.00 |
| LM Uncertainty | 0.0116 | 0.0118 | 0.0123 | 0.0121 | -14.61 | -12.75 | 0.00 | 0.0116 | 0.0119 | 0.0124 | 0.0122 | -14.80 | -13.34 | 0.00 |
| Causation (Dikolli) | 0.0324 | 0.03 | 0.0319 | 0.0304 | 6.79 | -5.83 | 0.00 | 0.0324 | 0.03 | 0.0318 | 0.0304 | 6.42 | -5.93 | 0.00 |
| Forward Look (Li) | 0.0065 | 0.0067 | 0.008 | 0.0078 | -40.91 | -41.43 | 0.00 | 0.0066 | 0.0068 | 0.0081 | 0.0078 | -41.04 | -41.79 | 0.00 |

*Panel C. Differences in Readability measures of other attributes between Missing Narratives Sample IV.Earnings*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 31,332) | | Sample IV (Obs. 37,162) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 29,990) | | Sample IV (Obs. 36,183) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Fog Index | 18.6088 | 17.9383 | 18.3545 | 17.9402 | 8.76 | 3.92 | 0.91 | 18.6159 | 17.9336 | 18.3399 | 17.9344 | 9.42 | 4.19 | 0.99 |
| Flesch Index | 23.7103 | 27.6643 | 25.7308 | 28.0537 | -20.67 | -12.14 | 0.00 | 23.7567 | 27.7041 | 25.8543 | 28.1002 | -21.22 | -12.38 | 0.00 |
| Flesch Kincaid index | 14.383 | 13.8811 | 14.2005 | 13.9279 | 7.85 | 1.65 | 0.00 | 14.3901 | 13.8773 | 14.1899 | 13.9244 | 8.53 | 2.02 | 0.01 |

**Table 8.** Tone and Readability differences between Sample III and Sample IV Market

*Panel A. Differences in Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer between Missing Narratives and Sample IV.Market*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 37,204) | | Sample IV (Obs. 31,290) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 36,093) | | Sample IV (Obs. 29,990) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Diction Neg. | 0.0104 | 0.0108 | 0.0115 | 0.0114 | -23.93 | -21.56 | 0.00 | 0.0105 | 0.0109 | 0.0115 | 0.0114 | -21.59 | -19.58 | 0.00 |
| Diction Pos. | 0.0092 | 0.0099 | 0.0104 | 0.0111 | -30.69 | -35.74 | 0.00 | 0.0093 | 0.0099 | 0.0104 | 0.011 | -27.93 | -33.18 | 0.00 |
| LM Neg. | 0.0105 | 0.0106 | 0.0118 | 0.0118 | -24.10 | -27.09 | 0.00 | 0.0106 | 0.0106 | 0.0117 | 0.0118 | -21.64 | -24.83 | 0.00 |
| LM Pos. | 0.0052 | 0.0053 | 0.0055 | 0.0056 | -11.48 | -13.34 | 0.00 | 0.0053 | 0.0054 | 0.0055 | 0.0056 | -9.97 | -11.57 | 0.00 |
| Inquirer Neg. | 0.0187 | 0.0206 | 0.0203 | 0.0218 | -22.83 | -22.68 | 0.00 | 0.0188 | 0.0207 | 0.0203 | 0.0218 | -20.54 | -21.21 | 0.00 |
| Inquirer Pos. | 0.1268 | 0.1237 | 0.1199 | 0.1184 | 23.99 | 30.18 | 0.00 | 0.1268 | 0.1238 | 0.12 | 0.1186 | 23.27 | 28.57 | 0.00 |

*Panel B. Differences in Content analysis of other attributes between Missing Narratives and Sample IV.Market*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 37,204) | | Sample IV (Obs. 31,290) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 36,093) | | Sample IV (Obs. 29,990) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Constraining | 0.0057 | 0.0053 | 0.0065 | 0.0063 | -21.07 | -43.54 | 0.00 | 0.0057 | 0.0053 | 0.0064 | 0.0063 | -20.30 | -41.55 | 0.00 |
| LM Litigious | 0.0072 | 0.0048 | 0.0072 | 0.0052 | 1.07 | -15.81 | 0.00 | 0.0072 | 0.0048 | 0.0072 | 0.0052 | 0.06 | -14.65 | 0.00 |
| LM Strong | 0.0023 | 0.0021 | 0.0021 | 0.0021 | 11.99 | 6.82 | 0.00 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 12.80 | 8.86 | 0.00 |
| LM Weak | 0.0036 | 0.0028 | 0.0036 | 0.003 | -1.68 | -14.51 | 0.00 | 0.0036 | 0.0028 | 0.0036 | 0.003 | -1.80 | -12.19 | 0.00 |
| LM Uncertainty | 0.0114 | 0.0116 | 0.0127 | 0.0124 | -26.00 | -26.75 | 0.00 | 0.0115 | 0.0117 | 0.0128 | 0.0124 | -23.74 | -24.20 | 0.00 |
| Causation (Dikolli) | 0.0327 | 0.0304 | 0.0314 | 0.0301 | 14.56 | 8.27 | 0.00 | 0.0325 | 0.0303 | 0.0315 | 0.0301 | 11.92 | 5.93 | 0.00 |
| Forward Look (Li) | 0.0071 | 0.007 | 0.0076 | 0.0075 | -14.94 | -18.56 | 0.00 | 0.0072 | 0.0071 | 0.0076 | 0.0075 | -12.16 | -15.37 | 0.00 |

*Panel C. Differences in Readability measures of other attributes between Missing Narratives Sample IV.Market*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 37,204) | | Sample IV (Obs. 31,290) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 36,093) | | Sample IV (Obs. 29,990) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Fog Index | 18.5771 | 18.0294 | 18.3444 | 17.8452 | 8.02 | 10.30 | 0.00 | 18.555 | 18.019 | 18.3562 | 17.8428 | 6.78 | 9.52 | 0.00 |
| Flesch Index | 24.2843 | 27.7743 | 25.4275 | 27.9664 | -11.67 | -3.88 | 0.00 | 24.4533 | 27.8328 | 25.4492 | 27.9977 | -10.05 | -3.14 | 0.01 |
| Flesch Kincaid index | 14.3358 | 13.962 | 14.2225 | 13.8431 | 4.87 | 6.19 | 0.00 | 14.3221 | 13.955 | 14.2304 | 13.8423 | 3.90 | 5.75 | 0.00 |

**Table 9.** Tone and Readability differences between Sample III and Sample IV Analysts

*Panel A. Differences in Tone as measured by Diction, Loughran and McDonald (2011) and Inquirer between Missing Narratives and Sample IV.Analysts*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 54,709) | | Sample IV (Obs. 13,785) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 52,833) | | Sample IV (Obs. 13,250) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Diction Neg. | 0.0107 | 0.011 | 0.0115 | 0.0114 | -14.38 | -13.44 | 0.00 | 0.0108 | 0.0111 | 0.0116 | 0.0115 | -12.90 | -12.17 | 0.00 |
| Diction Pos. | 0.0096 | 0.0103 | 0.0104 | 0.011 | -15.96 | -16.90 | 0.00 | 0.0096 | 0.0103 | 0.0104 | 0.0109 | -14.87 | -16.11 | 0.00 |
| LM Neg. | 0.0109 | 0.011 | 0.0117 | 0.0116 | -13.05 | -13.47 | 0.00 | 0.011 | 0.0111 | 0.0118 | 0.0116 | -12.27 | -12.82 | 0.00 |
| LM Pos. | 0.0053 | 0.0054 | 0.0057 | 0.0057 | -15.13 | -15.19 | 0.00 | 0.0053 | 0.0054 | 0.0058 | 0.0057 | -14.67 | -14.61 | 0.00 |
| Inquirer Neg. | 0.0191 | 0.021 | 0.0206 | 0.0218 | -16.21 | -14.67 | 0.00 | 0.1255 | 0.1228 | 0.1165 | 0.1157 | 25.10 | 32.99 | 0.00 |
| Inquirer Pos. | 0.1254 | 0.1226 | 0.1166 | 0.1157 | 24.57 | 33.37 | 0.00 | 0.0192 | 0.021 | 0.0206 | 0.0218 | -15.55 | -14.55 | 0.00 |

*Panel B. Differences in Content analysis of other attributes between Missing Narratives and Sample IV.Analysts*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 54,709) | | Sample IV (Obs. 13,785) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 52,833) | | Sample IV (Obs. 13,250) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Constraining (Bodn.) | 0.006 | 0.0057 | 0.0064 | 0.0063 | -9.83 | -23.68 | 0.00 | 0.006 | 0.0057 | 0.0064 | 0.0062 | -9.64 | -22.95 | 0.00 |
| LM Litigious | 0.0073 | 0.0049 | 0.007 | 0.0053 | 4.59 | -10.53 | 0.00 | 0.0072 | 0.0049 | 0.0069 | 0.0052 | 3.92 | -10.31 | 0.00 |
| LM Strong | 0.0022 | 0.0021 | 0.0022 | 0.0022 | -4.20 | -7.66 | 0.00 | 0.0022 | 0.0021 | 0.0023 | 0.0022 | -3.93 | -6.47 | 0.00 |
| LM Weak | 0.0035 | 0.0028 | 0.004 | 0.0031 | -12.55 | -24.34 | 0.00 | 0.0035 | 0.0028 | 0.004 | 0.0031 | -12.43 | -22.97 | 0.00 |
| LM Uncertainty | 0.0116 | 0.0118 | 0.0133 | 0.0128 | -25.93 | -28.30 | 0.00 | 0.0117 | 0.0118 | 0.0133 | 0.0128 | -24.30 | -26.98 | 0.00 |
| Causation (Dikolli) | 0.0323 | 0.0303 | 0.0314 | 0.0302 | 8.60 | 0.85 | 0.80 | 0.0322 | 0.0302 | 0.0315 | 0.0303 | 7.23 | -0.56 | 0.38 |
| Forward Look (Li) | 0.0071 | 0.0071 | 0.0082 | 0.008 | -24.47 | -28.42 | 0.00 | 0.0072 | 0.0071 | 0.0082 | 0.008 | -23.08 | -26.83 | 0.00 |

*Panel C. Differences in Readability measures of other attributes between Missing Narratives Sample IV.Analysts*

| | No Link Table | | | | | | | Yes Link Table | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing Narratives (Obs. 54,709) | | Sample IV (Obs. 13,785) | | T-test | Wilcoxon rank-sum | Nonparametric medians | Missing Narratives (Obs. 52,833) | | Sample IV (Obs. 13,250) | | T-test | Wilcoxon rank-sum | Nonparametric medians |
| VARIABLES | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p | Mean | Median | Mean | Median | T-stat | Z-stat | Fisher's p |
| Fog Index | 18.5062 | 17.9757 | 18.3304 | 17.8242 | 4.87 | 6.96 | 0.00 | 18.4987 | 17.9693 | 18.3296 | 17.8228 | 4.64 | 6.75 | 0.00 |
| Flesch Index | 24.6043 | 27.8983 | 25.6091 | 27.7701 | -8.25 | -0.05 | 0.04 | 24.7229 | 27.9596 | 25.6324 | 27.7887 | -7.38 | 0.56 | 0.01 |
| Flesch Kincaid index | 14.289 | 13.9126 | 14.2643 | 13.8795 | 0.85 | 1.54 | 0.12 | 14.2851 | 13.9088 | 14.2621 | 13.8761 | 0.79 | 1.51 | 0.15 |