MICHAEL LEE-CHIN & FAMILY INSTITUTE
FOR STRATEGIC BUSINESS STUDIES

Information Supporting Investor Valuations: Evidence from a Comparative Content Analysis of Analyst Reports and Form 10-K

Working Paper Series in Strategic Business Valuation
WP 2024-02

Mary E. Barth *
Graduate School of Business, Stanford University
mbarth@stanford.edu

Ken Li
DeGroote School of Business, McMaster University
lik130@mcmaster.ca

Charles G. McClure
Booth School of Business, University of Chicago
charles.mcclure@chicagobooth.edu

* Corresponding author

We thank seminar participants at New York University, University of Hawaii, University of Macau, University of Melbourne, and Washington University in St. Louis for helpful comments and suggestions. We also thank the Stanford University Graduate School of Business Data, Analytics, and Research Computing for research support and Boyie Chen for excellent research assistance. We appreciate the support provided by the Stanford University Graduate School of Business, the Social Sciences and Humanities Research Council of Canada, and the Liew Faculty Fellowship at the University of Chicago Booth School of Business.

DeGroote SCHOOL OF BUSINESS EDUCATION WITH PURPOSE | McMaster University

March 2024

Information Supporting Investor Valuations: Evidence from a Comparative Content Analysis of Analyst Reports and Form 10-K

Abstract

We address whether financial reports include information supporting investor valuations. We view analyst reports (AR) as reflecting this information and employ topic modeling to compare the contents of AR and Form 10-K. Our main findings follow. (i) Form 10-K focuses heavily on financial reporting, whereas AR focuses most on performance, followed by analysis, and business. However, AR discusses financial reporting almost as much as business, which suggests financial reporting is a crucial component of AR. The proportion of AR and Form 10-K dedicated to each category of topics has been converging since 2005. (ii) For topics within the performance category, AR and Form 10-K both focus most on revenues and margins. As expected, Form 10-K focuses more than AR on earnings and expenses, whereas AR focuses more on ratios, target prices, recommendations, and adjusted earnings. How AR and Form 10-K discuss performance topics diverged early in our sample. (iii) Form 10-K's MD&A section focuses more than AR on performance, analysis, and business, which suggests MD&A discussion resembles AR discussion. (iv) AR and Form 10-K are more similar for loss firms. Additionally, technology firms exhibit similar differences in AR and Form 10-K as nontechnology firms. Together, our findings reveal financial reports include information supporting investors' valuations, which is inconsistent with financial reports lacking relevance.

Keywords: Capital Markets; Financial Reporting; Textual Analysis; Analyst Reports; Value Relevance.

JEL classification: G10, G18, M40, M41.

Authors:
Mary E. Barth, Ken Li, and Charles G. McClure

Information Supporting Investor Valuations: Evidence from a Comparative Content Analysis of Analyst Reports and Form 10-K

1. Introduction

The question we address is whether financial reports include information supporting investor valuations of firms. This question is fundamental to financial reporting because financial reports are designed to provide investors with information to help them make decisions about providing capital to a firm. Yet, Form 10-K, a firm's primary financial report, has been criticized for lacking information relevant to investors in estimating the value of the firm. We presume that equity analyst reports (AR) contain information supporting investor valuations because AR describe the analysis that justifies the analyst's forecasts of the firm's performance and recommendations regarding whether investors should buy, hold, or sell the firm's stock. Thus, we address our research question by analyzing and comparing the contents of equity analyst reports and Form 10-K. Although we identify differences in AR and Form 10-K contents, our findings reveal many similarities, which calls into question the criticism that financial reports lack relevance.

We analyze and compare the contents of AR and Form 10-K for three reasons. First, equity analysts comprise a large class of users of financial reports who have an equity investor perspective. Although analysts are not investors per se, they provide investors with analysis supporting their valuations and investment advice, and their recommendations are often a major input for institutional investors. Second, whereas the content of Form 10-K must comply with U.S. Securities and Exchange Commission (SEC) regulations and U.S. GAAP, the content of AR is determined by the analyst who writes it. Thus, AR content reveals what information supports the analyst's valuation analysis. Third, related prior research finds that AR content reflects

information investors find relevant to valuing a firm's equity. Thus, AR content is a natural benchmark for analyzing Form 10-K content.

To address our research question, we assemble a sample of 26,757 equity analyst reports that provide fundamental analysis and Forms 10-K for 4,335 firms from 1997 to 2018. We then use a neural network-based word embedding topic modeling approach, Word2vec, to identify topics discussed in the combined AR and Form 10-K corpus. Word2vec identifies words with similar meanings, which enables us to cluster words in the corpus into topics based on their meanings. We organize these topics into six broad categories: Performance, Analysis, Business, Financial Reporting, Regulatory, and Other. We then calculate the proportion of words in AR and Form 10-K in each category and each topic within each category. We use these proportions to compare the contents of AR and Form 10-K and to determine how the contents of these two document types change over time.

We also compare the contents of AR and two prominent sections of Form 10-K, Management's Discussion and Analysis (MD&A) and Financial Statements (FS). These two sections are designed to provide investors with different types of information. MD&A provides management's analysis of the firm's resources and operations, whereas FS reports on the firm's financial condition and performance. As a result, we expect the categories discussed in AR to resemble those discussed in MD&A more than those discussed in FS. We also compare the similarity of AR and Form 10-K contents for economically dissimilar types of firms to provide insights into how the contents vary depending on a firm's circumstances. The dissimilar firm types we consider are profit versus loss, large versus small, non-technology versus technology, and non-financial versus financial.

Regarding categories discussed in AR, we find that AR discusses Performance the most, followed by Analysis and Business, which is consistent with analysts focusing on interpretation and analysis. Perhaps surprisingly, although Financial Reporting is the fourth most-discussed category in AR, it is discussed almost as much as Business, which is the third. This means that analysts devote almost as much of their reports to discussing Financial Reporting as they devote to discussing the firm's Business. As expected, we find that analysts devote considerably less of their reports to Regulatory topics. Taken together, these findings reveal that, contrary to claims about the lack of relevance, Performance and Financial Reporting are crucial components of AR.

Comparisons of categories for AR and Form 10-K reveal that the documents discuss most categories to similar extents. The largest difference relates to Financial Reporting—Form 10-K discusses Financial Reporting almost twice as much as AR. The large difference in Financial Reporting suggests that the concern regarding information overload primarily arises from this category.

Regarding how the contents of these two documents change over time, we find aggregate differences between the discussion categories in AR and Form 10-K decreased and then increased in the early part of our sample. However, beginning in 2005, we find almost monotonic convergence in the aggregate discussion of categories. We also find AR and Form 10-K converge for all five categories, with the greatest convergence in Performance. These findings suggest AR and Form 10-K are becoming more similar over time, which is inconsistent with the notion that Form 10-K is becoming more irrelevant for analysts.

Despite the similarities between AR and Form 10-K discussion by category, we find differences in the topics they discuss within each category. The topics discussed in AR and Form 10-K differ most for Performance. In particular, AR has considerable discussion of

Revenues/Margins, Ratios, Earnings, Target Prices, and Recommendations, which are consistent with analysts focusing on stock price performance and accounting performance. Although Revenues/Margins also is the most-discussed Performance topic for Form 10-K, we find a large difference between AR and Form 10-K in Performance topics because the second, fourth, and fifth-most discussed topics in AR—Ratios, Target Prices, and Recommendations—are the three least-discussed Performance topics in Form 10-K. The second and third most-discussed topics in Form 10-K are Earnings and Expenses.

Notably, we find that Earnings is the third most-discussed topic in AR, whereas Adjusted Earnings is the seventh and Cash Flows is the least. These findings are inconsistent with analysts viewing earnings as irrelevant and focusing on alternative performance measures, including cash flows. Together with finding that Performance is the most-discussed category in AR, these findings are inconsistent with analysts viewing quantitative information as irrelevant.

Regarding Analysis and Business, we find that AR and Form 10-K generally discuss similar topics and discuss each topic to a similar extent. Most notably, for both AR and Form 10-K, the most-discussed Analysis topic is Markets and Industries followed by Trends and Forecasts, and the most-discussed Business topic is Products and Markets. The similarity of the Analysis and Business topics between AR and Form 10-K implies that, despite the focus of Form 10-K on financial reporting, Form 10-K contains considerable discussion of Analysis and Business, which are also focuses of analysts.

Regarding Financial Reporting, we find that AR and Form 10-K discuss Fair Value in similar proportions, but this is not the case for the remaining Financial Reporting topics. AR focuses more than Form 10-K on Investments, Equity, Interest Rates, Issuances, Mergers & Acquisitions (M&A), Loans, and Debt and Notes, and Form 10-K focuses more than AR on the

other 11 topics in this category. Regarding Regulatory, we find that AR focuses more than Form 10-K on Responsibilities, Risks, Audit Report, and Disclaimers, whereas Form 10-K focuses more than AR on Forms and Exhibits, Legal Language, Uncertainties, and Internal Controls. Our finding that AR contains considerable discussion of risk suggests that the assertion that risk disclosures are uninformative is overstated.

We also examine whether the proportion of topics within categories changes over time. This analysis can reveal whether AR and Form 10-K become more aligned in their discussion of each category. Except for Performance, we find the proportion of topics within categories has converged. For Performance, we find a large divergence in topics in the early part of our sample that remained steady thereafter. The early divergence in Performance topics is primarily driven by four topics: Recommendations, Target Prices, Earnings, and Revenues/Margins. This finding suggests how AR and Form 10-K discuss Performance has become more dissimilar, especially in earlier in our sample period.

Focusing our AR content comparisons on MD&A and FS rather than the full Form 10-K reveals additional insights. Notably, unlike the full Form 10-K, we find that MD&A and FS both contain more discussion of Performance than AR. MD&A also contains more discussion of Analysis than AR, and the more extensive discussion of Business in Form 10-K is attributable to MD&A. The fact that MD&A contains more discussion of these three categories is notable because these categories typify the role of analysts and are the three most-discussed categories in AR. We also find that both MD&A and FS discuss Financial Reporting more than AR, although—as one would expect—the difference is more pronounced for FS. In fact, these differences are the largest among the categories, which indicates MD&A and, particularly, FS discussions about Financial Reporting are least consistent with AR discussion. Perhaps

surprisingly, both MD&A and FS contain less discussion of Regulatory than AR, not more, as is the case for the full Form 10-K.

Focusing our AR and Form 10-K content comparisons on economically dissimilar firms also reveals additional insights. In particular, we find that AR focuses more on Performance for Profit, Non-Technology, and Financial firms, whose values are more likely attributable to existing operations than growth options. AR focuses more on Business for Loss, Small, Technology, and Non-Financial firms, whose businesses likely are less well-understood. We find that Form 10-K focuses more on Performance and Financial Reporting for Profit, Large, Non-Technology, and Financial firms, whose value is better reflected in financial accounting. We also find that Form 10-K has more discussion of Business for Loss and Technology firms, whose businesses likely are less well-understood.

Our comparisons of the proportion of categories discussed in AR and Form 10-K for economically dissimilar firms reveal greater similarity in AR and Form 10-K for Loss firms in that Loss firms have more similar proportions of discussion of Performance, Analysis, Financial Reporting, and Regulatory. However, we do not find that Technology firms have more dissimilar AR and Form 10-K content than Non-Technology firms in that AR and Form 10-K are more similar in their discussion of Financial Reporting and less similar in Performance and Business. This result may be surprising because much of the criticism directed toward financial accounting relates to its incompletely reflecting businesses of firms in the New Economy, such as Technology firms. These findings suggest that the criticism of accounting's irrelevance in the New Economy may be overstated.

This study contributes to the literature by addressing whether financial reports contain information supporting investor valuations using a large-scale comparative content analysis of

AR and Form 10-K. To our knowledge, ours is the first study to do so. Despite some notable differences, we find many similarities in the contents of AR and Form 10-K, and these differences are decreasing over time. Thus, taken together, our findings are inconsistent with financial reports lacking relevance.

The paper proceeds as follows. Section 2 relates our study to prior research and describes our predictions. Section 3 develops the research design, Section 4 describes the sample, and Section 5 reports the findings. Section 6 presents additional analyses, and Section 7 concludes.

2. Related literature and predictions

2.1 Analyst reports

Financial analysts are sophisticated market participants who provide new information about firms' values and interpret previously released information (Asquith, Mikhail, and Au 2005). Prior research finds that information provided by sell-side analysts is a major input into investors' valuations and the market reacts to analyst forecasts, price targets, and recommendations (Schipper 1991; Womack 1996; Howe, Unlu, and Yan 2009). These findings imply that analyst-provided information is correlated with investors' valuation assessments. Importantly for our study, analysts do not provide forecasts and recommendations in isolation but instead issue reports that explain the analysis and details supporting them. Because there are no standards or requirements for the content of analyst reports, each report reveals what information supports the analyst's valuation analysis.[1]

[1] Analysts might explain the basis for their forecasts and recommendations using discussion topics similar to those in financial reports because they believe readers will find the reports more persuasive, not because the analyst developed the forecasts and recommendations using those topics. However, finding that analysts explain their forecasts and recommendations using topics typically found in financial reports is consistent with financial reports including information supporting investor valuations.

Early research examining the content of analyst reports, often based on a small sample, finds the reports tend to focus on performance measures, especially accounting-based measures that often deviate from U.S. GAAP (Govindarajan 1980; Previts, Bricker, Robinson, and Young 1994). However, subsequent research finds that analysts do not justify their analysis solely on accounting-based measures. Breton and Teffler (2001) finds that analysts discuss soft information, such as business strategy and management, more frequently than accounting-based information. Bradshaw (2002) finds that this soft information is especially relevant when analysts have negative views of the firm.

More recent literature uses text-based methods to analyze large samples of analyst reports to infer the role of analysts and to test whether the text can predict firm outcomes. Huang, Lehavy, Zang, and Zheng (2018) estimates that nearly one-third of analyst reports is devoted to new information, whereas the remaining two-thirds interprets already-known data. However, Martineau and Zoican (2021) notes that the information content of analyst reports varies as a function of the number of analysts following the firm. Relatedly, Huang, Zang, and Zheng (2014) finds that the tone of analyst reports predicts returns and earnings, which suggests the reports' exposition offers insights into a firm's valuation. Huang, Tan, Wang, and Yu (2023) uses keywords to identify analysts' use of discounted cash flow models and discussions of cash flow and discount rate information, and finds that analysts are more likely to discuss cash flow and discount rate information for firms with more uncertainty. Li, Mai, Shen, Yang, and Zhang (2024) use a ChatGPT-based method to examine the text of analyst reports to identify analysts' views of corporate culture. This research motivates our analysis of the textual content of analyst reports.

2.2 Relevance of accounting information and disclosures and predictions

An extensive literature examines the value relevance of accounting information.[2] Several studies conclude that accounting information has lost its relevance (Brown, Lo, and Lys 1999; Lev and Zarowin 1999; Core, Guay, and Van Buskirk 2003; Balachandran and Mohanram 2011; Lev and Gu 2016). This evidence is based primarily on the low and declining associations between firms' equity market values and accounting amounts, particularly earnings. However, this conclusion is subject to three qualifications. First, this literature typically limits the assessment of value relevance to amounts presented in financial statements. Second, the approach measures correlations, which are not designed to determine whether investors use a particular amount in their decision-making.

Third, these studies assume a valuation model that links prices and accounting amounts, such as Ohlson (1995), whereas investors use a variety of valuation models (Joos and Plesko 2005). Barth, Li, and McClure (2023) uses machine learning to mitigate some of these issues and concludes that accounting information has not yet lost its relevance. These studies often rely on a subset of accounting amounts and often impose a functional form, which limits the extent to which and how accounting amounts map into equity prices. Our comparative content analysis of AR and Form 10-K offers a new approach to assess the relevance of accounting by focusing on information analysts use to support their valuation assessments, which is broader than just accounting amounts.

Form 10-K includes not only a firm's financial statements, but also discussion of its business and financial condition.[3] As a result, Form 10-K provides equity investors with substantial information on which to base their investment decisions. However, the SEC, practitioners, and the academic literature observe that Form 10-K has become longer, more

[2] See Barth, Beaver, and Landsman (2001) and Holthausen and Watts (2001) for summaries.
[3] https://www.investor.gov/introduction-investing/investing-basics/glossary/form-10-k.

repetitive, more boilerplate, and not necessarily more informative (Dyer, Lang, and Stice-Lawrence 2017).[4] Survey evidence suggests a majority of investment professionals are dissatisfied with current financial reporting because it is not very useful for some analyses (Drake, Hales, and Rees 2019; Cascino, Clatworthy, Osma, and Imam. 2021). To address these issues, the SEC has taken steps to simplify disclosures, including the Plain English Initiative and changes to Regulation S-K.[5] Despite these changes, concerns about disclosure remain.[6]

Prior research identifies undesirable characteristics of mandatory disclosures, such as low readability, that obfuscate low earnings and increase crash risk (Loughran and McDonald 2014; Kim, Wang, and Zhang 2018). Furthermore, firms whose annual reports contain more boilerplate have lower liquidity, analyst following, and institutional ownership (Lang and Stice-Lawrence 2015). This evidence suggests the content and exposition of Form 10-K affect investors' decisions.

Our comparison of the contents of AR and Form 10-K does not imply that we expect the topics discussed in AR and Form 10-K to be the same. Whereas Form 10-K focuses more on presenting the firm's current financial condition and performance, AR focuses more on analyzing the valuation implications of that performance (Previts et al. 1994; Asquith et al. 2005). Thus, we expect AR to discuss performance and analysis more than Form 10-K. However, our expectations might not be borne out if accounting performance measures are not

[4] Related studies often remove or consider separately topics and words that are repetitive, have no particular meaning, and the writer did not include by choice. These words are sometimes referred to as boilerplate. For example, Dyer et al. (2017) defines boilerplate words as 4-word phrases shared by at least 75% of firms in a given year, and Huang et al. (2014) removes disclaimers, required disclosures, and glossaries. Our Word2vec topic modeling approach identifies boilerplate as topics, which we include in the Other category. See Section 3.2. In an additional analysis, we remove the portion of AR most likely to contain standardized disclosures about the brokerage firm that employs the analyst, which are a form of boilerplate. Untabulated findings reveal inferences that are consistent with those based on the tabulated analyses. See Section 6.1.

[5] https://www.sec.gov/plainwriting.

[6] See, for example, https://www.wsj.com/articles/the-109-894-word-annual-report-1433203762.

relevant to analysts, as predicted by studies that find accounting information has limited use (Drake et al. 2019; Cascino et al. 2021).

Analysts also devote considerable attention to discussing a firm's business (Breton and Teffler 2001), which leads us to expect AR discusses business more than Form 10-K. However, MD&A also allows a firm to discuss its business, particularly as it relates to changes in the firm's performance. Thus, it is an empirical question whether AR or Form 10-K discusses a firm's business more.

The criticism that Form 10-K contains information that is of little use to investors suggests that Form 10-K discusses some topics more than AR (Dyer et al. 2017; Drake et al. 2019). Much of this criticism relates to requirements of accounting standards and regulation. Thus, we expect that Form 10-K contains more discussion of financial reporting and regulation than AR. However, if analysts find particular topics helpful—such as particular detailed financial or risk disclosures—it is possible that AR contains more, or the same amount of, discussion for those topics as Form 10-K (De Franco, Wong, and Zhou 2011; Cazier, McMullin, and Treu 2014; Hope, Hu, and Lu 2016; Christensen, Floyd, Liu, and Maffett 2017).

3. Research design

3.1 Overview

Addressing our research question requires us to analyze the contents of AR and Form 10-K and construct metrics to compare them. We begin by identifying the topics discussed in AR and Form 10-K and organizing these topics into broad categories. We then calculate the proportions of AR and Form 10-K devoted to each topic and category. By using these proportions to analyze the content of AR, we provide insights into the information that supports investor valuations and how the information in Form 10-K differs from the information

supporting investor valuations. We also provide insights into how these differences have changed over time.

We also compare the contents of AR and the Management's Discussion and Analysis (MD&A) and Financial Statement (FS) sections of Form 10-K and compare the contents of MD&A and FS to each other. We do so because these two sections are designed to provide investors with different types of information. MD&A is intended to provide management's analysis of the firm's resources and operations (Li 2010). As such, it bears some resemblance to AR. Also, MD&A is perhaps the most important and read section of Form 10-K (Tavcar 1998). We examine the FS section because financial statements, including the accompanying footnotes, provide many of the inputs to investor valuation models. We also compare the content of AR and Form 10-K for several groups of economically dissimilar firms to provide insight into the similarities and differences between these documents depending on a firm's circumstances.

3.2 Identifying topics and topic categories discussed in AR and Form 10-K

We identify topics discussed in AR and Form 10-K by implementing a neural network-based word embedding topic model, Word2vec. See the Appendix for details. In brief, because our focus is on understanding closely related words that form topics in AR and Form 10-K, three features of Word2vec make it well-suited for our analysis. First, Word2vec allows for, but does not require, the exclusion of common words. This is a desirable feature because understanding the extent of discussion of topics that include common words such as *income* and *cash* is important to our analysis. Second, Word2vec ensures that each word appears in only one topic. Thus, Word2vec provides a straightforward way to identify the topic being discussed. Third, Word2vec identifies words as similar based on the idea that words that co-occur with similar

neighboring words have similar meanings. Accordingly, Word2vec can identify words such as *earnings* and *net_income* as similar words.[7]

We identify topics in six steps. First, we obtain AR from Thomson Reuters Investext, which collects millions of AR from over 1,600 investment banks and sell-side research firms. Our focus is on the content of equity research reports that provide fundamental analysis. Thus, we eliminate AR from brokers that primarily provide robo reports, event transcripts, proxy and governance reports, credit rating reports, and company descriptions. We remove these reports because either they only reiterate company disclosures and are likely machine generated (e.g., so-called "robo" reports), which makes the reports less representative of what informs analysts' valuations, or they relate to non-equity valuations (e.g., credit reports). After eliminating these reports, we randomly select one AR for each firm-year. Restricting our analyses to one report per firm year ensures that our topic modeling approach is computationally feasible. However, this limitation hinders our ability to conduct analyses within a single firm-year.[8] For this sample, we convert AR from PDF to text for analysis.

Second, we obtain Form 10-K from the EDGAR database. We remove HTML from each Form 10-K based on the approach in Loughran and McDonald (2016). We also identify the MD&A and FS sections of each Form 10-K based on the approach in Dyer at al. (2017).

Third, because AR is shorter, on average, than Form 10-K, we scale the AR corpus to be the same length as the Form 10-K corpus, so that both document types receive equal weight

[7] See the Appendix for an explanation of why we use Word2vec rather than Latent Dirichlet Analysis (LDA) and Bidirectional Encoder Representations from Transformers (BERT), which are other commonly used textual analysis methods.

[8] Chen, Cheng, and Lo (2010) finds that AR content can vary in terms of interpretation and discovery throughout the fiscal year. To construct a sample of AR that reflects this variation, we do not constrain the random selection procedure to particular points in the firm's fiscal year.

when determining topics.[9] We combine AR and Form 10-K into a single corpus so that our topics span both document types. Training Word2vec on only one document type would disregard topics that are exclusive to the other type.[10] In addition, if AR and Form 10-K use different words to discuss the same topic, training on the combined corpus permits Word2vec to group those different words into the same topic.[11]

To focus our analysis on meaningful words and phrases, we eliminate uninteresting words such as punctuation symbols and single-character words. We also remove the 50 most common words in the corpus—such as *the*, *of*, and *and*—unless the words could relate to a topic of potential substantive interest. In addition, we treat common multi-word phrases (e.g., *net income*) as a single word to allow topic modeling to learn the meanings of both words and phrases, which may have different meanings from individual words. Specifically, we treat as single words sequences of up to 16 words that appear, on average, more than once per AR. For example, we treat the word sequence *earnings per share* as the single word *earnings_per_share*. The result of these steps is a combined AR and Form 10-K corpus that underlies our analyses.

Fourth, we train the topic model on the combined corpus in that Word2vec converts each word to a vector in a 100-dimensional vector space in which words with similar (dissimilar) meanings are close together (far apart) in the vector space. Fifth, we cluster these word vectors into 100 clusters that minimize the within-cluster sums of mean squared vector differences.

[9] By equal weighting the document types, we ensure the resulting topics represent the average context of each word. This weighting scheme prevents the topics from being excessively influenced by a single document type when there are variations in context between the two types.

[10] For example, suppose the only topic in Form 10-K is expenses and the only topic in AR is revenue. If we train Word2vec on only the AR corpus, Word2vec would lack the knowledge that Form 10-K discusses expenses; it only would be able to identify that Form 10-K discusses something other than revenue. However, by combining the two document types, we gain the understanding that Form 10-K discusses expenses and AR discusses revenue.

[11] As a hypothetical example, assume AR uses *earnings* and Form 10-K uses *net_income* to mean the same thing. Word2vec would identify these two words as similar if they are used in the same contexts (i.e., they tend to have similar neighboring words).

Each cluster represents a topic. Together, these topics include all words in the combined AR and Form 10-K corpus. Sixth, we label topics and, following Dyer et al. (2017), organize them into categories to ease interpretation. We focus our analyses on the five most interpretable categories—Performance, Analysis, Business, Financial Reporting, and Regulatory.[12] We classify topics into these categories because they reflect the focus of prior research related to analyst reports and the concerns of academics, standard setters, and regulators regarding the relevance to investors of Form 10-K (see Section 2).

3.3 Quantifying the extent of category and topic discussion

We focus on two types of differences between AR and Form 10-K discussion. The first is the proportion of each document type's discussion devoted to each of the five topic categories. The second is the proportion of discussion within each category that is devoted to each topic in each document type.

We determine the amount a document discusses a particular category by scaling the number of words in that category by the total number of words. Thus, the proportion that document *i* in year *t* discusses category *c* is as follows.

$$Category\ Proportion_{c,i,t} = \frac{\#\ of\ Words\ in\ c_{i,t}}{\#\ of\ Words\ in\ i_t} \quad (1)$$

To determine the amount a document discusses a topic within a category, we scale the number of words in that topic by the number of words in the category to which that topic belongs. Thus, the within category *c* topic weight of topic *j* in document *i* in year *t* is as follows.

$$Topic\ Proportion_{j,i,t} = \frac{\#\ of\ Words\ in\ j_{i,t}}{\#\ of\ Words\ in\ c_{i,t}} \quad (2)$$

[12] These five categories include 59 of the 100 topics Word2vec identifies. We include the remaining 41 topics in the Other category but do not analyze them because they have no discernable topic focus. See Section 5.1.

To determine the category or topic proportions across documents, such as within a document type (e.g., AR), we average the proportions across firm-year observations.

The proportions from Equations (1) and (2) provide the basis for us to analyze the content of AR and to compare it (i) to that of Form 10-K, (ii) to those of the MD&A and FS sections of Form 10-K (i.e., Items 7 and 8 of Form 10-K) and to compare the contents of MD&A and FS to each other, and (iii) for several groups of economically dissimilar firms. We test whether the discussion of a particular topic or category differs between two sets of documents (e.g., AR versus Form 10-K) using a t-test of the difference in proportion means for each document set.

3.4 Differences in AR and Form 10-K contents for economically dissimilar firms

We compare the contents of AR and Form 10-K for four types of economically dissimilar firms to provide insights into how much content varies depending on a firm's circumstances. Two comparisons reflect differences potentially applicable to all firms—profit versus loss and large versus small—and two reflect differences associated with industry membership—non-technology versus technology and non-financial versus financial.

We compare content for profit and loss firms because loss firms are more difficult to value (Joos and Plesko 2005; Darrough and Ye 2007). We compare content for large and small firms because the sources of their equity value likely differ. For example, large (small) firms are more likely to derive value from operations (growth options). We compare content for non-technology and technology firms because technology firms are emblematic of the New Economy. These firms typically have substantial intangible assets, whose value generally is not recognized in financial statements. We compare content for non-financial and financial firms because most financial firms' assets are financial, which are recognized and measured differently

from other assets. This is why financial firms often are excluded from or examined separately in most accounting and finance research (Fama and French 2006).

To assess differences between these types of firms, we estimate the following equations:

$$Proportion_{c,i,t} = \beta_0 + \beta_1 Loss_{i,t} + \beta_2 Small_{i,t} + \beta_3 Tech_{i,t} + \beta_4 Financial_{i,t} + \epsilon_{j,i,t} \quad (3)$$

Proportion is either the amount of discussion for category *c* within a document type (e.g., Form 10-K) or the absolute difference in the amount of discussion for category *c* between AR and Form 10-K. *Loss*, *Small*, *Tech*, and *Financial* are indicator variables that equal one if firm *i* in year *t* reports a loss, is small, operates in a technology industry, or is a financial firm, and zero otherwise. Loss firms report negative earnings. Small firms have equity market value below the sample median. Technology firms have three-digit SIC codes 283, 357, 360-368, 481, 737, and 873 (Francis and Schipper 1999; Core et al. 2003; Barth et al. 2023). Financial firms are members of the Fama-French 48 industries of banking (44), insurance (45), and trading (47).

4. Sample

Our sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018. 1997 is the first full year the SEC required all firms to submit Form 10-K electronically on EDGAR and 2018 is when access to AR on Thomson Investext became limited. We require sample firms to have, in a particular firm-year, data in Compustat and CRSP, at least one fundamental analysis AR available from Investext, and a Form 10-K for which we can identify the MD&A and FS sections using the approach in Dyer et al. (2017). See the Appendix for details.

Untabulated statistics reveal that the average AR contains 3,418 words, which is substantially fewer than the average Form 10-K, which contains 65,951 words.[13] The average numbers of words in MD&A and FS are 10,162 and 15,345.

5. Findings

5.1 Topics discussed in analyst reports and Form 10-K

As Section 3.2 explains, we organize the topics Word2vec identifies in AR and Form 10-K into categories—Performance, Analysis, Business, Financial Reporting, Regulatory, and Other. Table A.1, Panels A through F, in the Appendix presents the topics discussed in each category, listed in the order of the most to the least discussed category in AR. Each row of the table presents the 15 most frequently occurring words in each topic, listed in order of the highest to the lowest frequency in the combined AR and Form 10-K corpus. Untabulated statistics reveal that the number of dictionary words in each topic ranges from 1 to 273, and the total dictionary size is 7,279 words.[14]

Panel A of Table A.1 reveals that Performance comprises eight topics: Revenues/Margins, Ratios, Earnings, Target Prices, Recommendations, Adjusted Earnings, Expenses, and Cash Flows. Regarding words in each topic, Revenues/Margins includes words such as *sales*, *revenues*, and *margins* and Expenses includes words such as *costs*, *expenses*, and

[13] The average number of words in Form 10-K is larger than the 45,349 words in Dyer et al. (2017) for two reasons. First, our sample is more recent. Dyer et al. (2017) finds Form 10-K has become longer during that study's sample period. Untabulated statistics reveal both AR and Form 10-K also become longer during our sample period. The average annual increase in the average number of words in AR (Form 10-K) is 126 (1,310), which is 3.7% (2.0%) of the mean number of words during our sample period. The increase in Form 10-K words is similar to the median increase reported in Dyer et al. (2017) of 1,588 words (= (50,000 – 23,000)/17). Thus, the increase in number of words applies to AR as well. Untabulated statistics also reveal that for both AR and Form 10-K, all topic categories have become longer. Taken together, these statistics are consistent with the amount of information required to value firms increasing over time (Barth et al. 2023). Second, our sample firms are followed by analysts, which means that, on average, they are larger than the Dyer et al. (2017) sample firms. AR and Form 10-K typically are longer for larger firms. Untabulated statistics for our sample reveal the mean number of words in AR and Form 10-K for Large (Small) firms are 4,011 and 74,383 (2,824 and 57,518).

[14] Dictionary words are the unique words included in our analysis.

fees. Earnings includes words such as *earnings* and *net_income* and Adjusted Earnings and Cash Flows include words such as *ebitda*, *adjusted*, and *cash_flow*.

Panel B reveals that Analysis comprises nine topics: Markets and Industries, Trends and Forecasts, Positives and Negatives, Estimates and Assessments, Increases and Decreases, Successes and Challenges, Views and Opinions, Economic Environment, and Affect and Effect. These topics include words that relate to interpretations for changes. Panel C reveals that Business comprises 16 topics. These topics include words referring to (i) general business and operations such as *services*, *technology*, *solutions*, *products*, *operations*, *customers*, *business*, *management*, and *strategy*; (ii) industry-specific business descriptions such as *clinical*, *patients*, *treatment*, *production*, *oil*, *gas*, *insurance*, *benefits*, and *coverage*; and (iii) specific business details such as *new_york*, *delaware*, *located*, *project*, *expansion*, and *area*.

Panel D reveals that Financial Reporting comprises 18 topics. These topics include words referring to (i) accounting standards such as *accounting* and *standards*; (ii) various assets and liabilities such as *assets*, *impairment*, *goodwill*, *property*, *equipment*, *lease*, *liabilities*, *payable*, and *reserves*; and (iii) financial statements such as *financial*, *statement*, and *balance_sheet*. Panel E reveals that Regulatory comprises eight topics. The topics in this category relate to risk and uncertainty and regulatory or legal language. This category includes words such as *risk*, *adverse*, *liability*, *rights*, *obligation*, *responsible*, *damages*, and *internal_control_over_financial_reporting*.

Untabulated statistics reveal the Other category comprises 41 topics that range from 0.00% to 10.07% of the total category weight. For parsimony, Panel F presents only the top 15 topics, which untabulated statistics reveal comprise more than 70% of the total category weight. We do not label or analyze the topics in this category because they appear to be aggregations of

generic words such as *it*, *only*, *part*, *when*, and *however*, or have no discernable topic focus such as *company*, *corporation*, *group*, and *llc* (see footnote 4).

5.2 Discussion of topic categories

5.2.1 Proportion of document devoted to each category

Table 1 presents the findings relating to AR and Form 10-K category discussions.[15] Figure 1 displays the average proportion of the document devoted to each category. Table 1 reveals that, on average, analysts devote 14.11% of AR to Performance, which is the most-discussed category. Analysts devote 12.90% and 11.71% of AR to Analysis and Business. These proportions indicate that, as expected, AR contains considerable analysis of a firm's performance and business activities, which reflects analysts' interpretation and analysis role (Huang et al. 2018).

Table 1 also reveals that, on average, analysts devote almost as much of AR to Financial Reporting as they do to Business (11.71% versus 11.41%). This finding suggests financial reporting is a crucial component of AR. By contrast, analysts devote only 6.75% of AR to Regulatory, which is not surprising because regulatory topics are less likely to be relevant to analysts' valuation analysis. Table 1 reveals that 43.11% of AR comprises words with no discernible topic focus. Table 1 also reveals that, in aggregate, AR and Form 10-K differ in their discussion of the categories by 15.43%.[16]

More importantly for our research question, the findings reveal differences between AR and Form 10-K content for all categories, with some being expected and some unexpected. Table 1 reveals that, as expected, AR discusses Performance and Analysis more than Form 10-K (14.11% and 12.90% versus 10.08% and 10.03%), and Form 10-K discusses Regulatory more

[15] Table 1, and Tables 2 to 6 that follow, present statistics related to tests of the significance of these differences.
[16] This amount is the sum of the absolute differences in category proportions, divided by two.

than AR (10.26% versus 6.75%). Although these differences are significant, they are not as large economically as perhaps some would expect.

Unexpectedly, Form 10-K discusses Business more than AR (12.84% versus 11.71%). This finding suggests Form 10-K has more discussion of firms' operations, despite the general perception that Form 10-K largely provides quantitative accounting information. The largest difference between AR and Form 10-K relates to Financial Reporting. Although we expect Form 10-K to discuss Financial Reporting more than AR, Form 10-K discusses this category almost twice as much (22.20% versus 11.41%). This large Financial Reporting difference suggests the concern of Form 10-K information overload primarily arises from this category.

5.2.2 Categories trend analyses

Figure 2 shows how the difference in AR and Form 10-K content changes over our sample period. Panel A plots—for each year—the sum of the absolute differences in categories across the two document types, divided by two. In the early years of our sample, the difference decreases then increases, but largely varies between 16.0% and 17.0%. However, beginning in 2005, there is an almost monotonic decrease in the difference that settles at approximately 14.5%. Untabulated findings from estimating the relation between the average absolute difference for each year and a time trend over our entire sample period yields a coefficient (t-statistic) on the time trend of –0.12% (–7.14), which reveals an average convergence of 12 basis points per year. Estimating this relation for sample years after 2004 yields a larger coefficient (t-statistic) of –0.19% (–11.82). These findings are perhaps unexpected because they suggest the discussions in AR and Form 10-K have been converging, which is inconsistent with financial accounting losing its relevance.

We next determine which categories contribute to this convergence of AR and Form 10-K discussion. We do so by first plotting the average absolute differences in the proportion of each document type devoted to each category over our entire sample period. A decrease (increase) in these differences indicates that the extents to which AR and Form 10-K discuss these categories are converging (diverging) over time. We plot these differences in Panel B of Figure 2, which shows the proportions for all categories are converging, with the greatest convergence in Performance. To quantify the extent of convergence of the category proportions, we estimate the relation between the absolute difference in category proportions and a time trend, where a negative (positive) relation indicates convergence (divergence) in that category's discussion over time. Untabulated findings reveal the coefficient on each category is negative and significant, with coefficients ranging from –0.18% to –0.03% (t-stats. from –13.23 to –2.79). These findings reveal that the overall convergence in panel A is not attributable to a single or a subset of categories. Rather, the findings reveal convergence of all discussion categories.

We next investigate the reason behind the convergence between AR and Form 10-K's category discussion. Specifically, we examine whether this convergence is attributable to one document increasingly resembling the other or whether both documents are converging toward a midpoint.

The untabulated findings reveal AR exhibits significantly decreasing (significantly increasing) trends in Performance, Analysis, and Business (Financial Reporting and Regulatory) discussion of –0.13%, –0.10%, –0.12% (0.04% and 0.12%) relative to the mean levels of discussion in Table 1 of 14.11%, 12.90%, and 11.71% (11.41% and 6.75%). By contrast, Form 10-K exhibits significantly increasing (decreasing) trends for Performance and Analysis (Regulatory) of 0.08% and 0.16% (–0.01%) relative to the mean levels of discussion in Table 1

of 10.08% and 10.03% (10.26%). We find no significant trend for Business and Financial Reporting. These findings suggest that for Performance, Analysis, and Regulatory categories, the proportions of AR and Form 10-K discussions are converging toward each other. By contrast, for Financial Reporting, AR is converging to the proportion of Form 10-K discussion.[17]

5.3 Discussion of topics within each category

5.3.1 Proportion of category discussion devoted to each topic

Table 2, Panels A through E, presents the findings relating to discussion topics within each category.[18] Figure 3, Panels A through E, displays these findings. Each panel presents the average proportion of each topic relative to the category discussion in that document type. These findings provide insight into the topics on which AR and Form 10-K focus when discussing a particular category.

Regarding Performance, Panel A reveals that AR contains considerable discussion of all topics, except for Expenses and Cash Flows. Specifically, the AR proportions are only 4.58% and 2.30% for Expenses and Cash Flows, whereas the proportions range from 26.33% to 9.58% for the other Performance topics. In contrast, when discussing Performance, Form 10-K focuses on three topics, Revenues/Margins, Earnings, and Expenses. The Form 10-K proportions for these three topics are 35.22%, 29.69%, and 18.17%, whereas the proportions for other topics range from 5.85% to 0.90%. These findings reveal that within their discussions of Performance, although Revenues/Margins is the largest topic in both documents, AR and Form 10-K focus on different topics. In fact, Panel A reveals that, in aggregate, the extent to which AR and Form 10-

[17] We find an insignificant decline for Business in Form 10-K. Because Business declines for AR, which has a lower mean proportion of Business than Form 10-K, there is no clear reason for its convergence.

[18] A small number of documents do not discuss any topic within a particular category. In those rare cases, we remove the document when compiling the within-category statistics for the category not discussed. Untabulated statistics reveal that less than 1%, 1%, 1%, 1%, and 5% of documents have no discussion of Performance, Analysis, Business, Financial Reporting, and Regulatory topics. Tables 2, 4, and 6, and Figures 2 and 4 do not present topics in the Other category because those topics have no discernable topic focus. See Section 3.2.

K discuss Performance topics are 39.53% different, which is the largest difference among the five categories.

Regarding Analysis, Panel B reveals that the top five topics discussed in AR are Markets and Estimates, Trends and Forecasts, Positives and Negatives, Estimates and Assessments, and Increases and Decreases (proportions range from 27.68% to 10.36%). AR contains less discussion of Successes and Challenges, Views and Opinions, Economic Environment, and Affect and Effect (proportions range from 7.46% to 1.61%). Panel B also reveals that Form 10-K focuses on the same topics, except for Positives and Negatives. In fact, in aggregate, the discussions of Analysis topics in AR and Form 10-K are only 10.19% different, which is the second smallest difference among the five categories. This small difference is unexpected, given that analysis is a key analyst role.

Regarding Business, Panel C reveals that the topics discussed in AR and Form 10-K are similar. For both documents, the top three topics are Products and Markets, Business and Operations, and Services and Technology (proportions range from 16.96% to 12.28% for AR and from 16.41% to 11.65% for Form 10-K). Perhaps surprisingly, Form 10-K does not contain less discussion of these topics, in aggregate. In fact, it contains slightly more. The four largest differences are that AR has more discussion of Cities and International Locations and Energy (Diff. = 3.11 pp. and 2.19 pp.) and Form 10-K has more discussion of Insurance and Healthcare and Business and Operations (Diff. = –4.12 pp. and –2.08 pp.).[19] In addition, in aggregate, the discussions of Business in AR and Form 10-K are only 9.07% different, which is the smallest difference among the five categories.

[19] pp. refers to percentage points.

Regarding Financial Reporting, Panel D reveals that AR focuses more than Form 10-K on Investments, Equity, Interest Rates, Issuances, M&A, Loans, and Debt and Notes. Except for Investments (Diff. = 12.96%), the differences generally are not economically large (Diff. ranges from 2.71% pp. to 0.42%). Regarding Regulatory, Panel E reveals that AR focuses more on Risks, Responsibilities, Audit Report, and Disclaimers (Diff. ranges from 9.96 pp. to 2.69 pp.), and Form 10-K focuses more on Forms and Exhibits, Legal Language, Uncertainties, and Internal Controls (Diff. ranges from 12.33 pp. to 1.18 pp.). However, Panel D (Panel E) also reveals that, in aggregate, the discussions of Financial Reporting (Regulatory) in AR and Form 10-K are 20.95% (21.93%) different, which is the third (second) largest difference among the five categories. These findings complement those in Table 1 and suggest Form 10-K's greater discussion of Financial Reporting and Regulatory is concentrated in particular topics.

5.3.2 Topics within each category trend analyses

We next provide insights into differences over time in AR and Form 10-K discussions of topics within each category by examining how the differences between AR and Form 10-K in the topics discussed within each category change over our sample period. Figure 4 plots the average difference in topic proportions, within category, for each year. Figure 4 reveals modest decreases in the topic discussion differences for all categories except for Performance. For Performance, we find approximately a 20 pp. increase in the difference of the topics in the earlier part of our sample. Untabulated analysis in which we regress Performance topic weights on a time trend reveals this increase is largely attributable to four topics, Recommendations, Target Prices, Earnings, and Revenues/Margins (coefs. = 0.45%, 0.40%, 0.25%, and 0.09%; t-stats. = 3.50, 5.83, 7.05, and 3.02). Beginning in 2005, only the Earnings topic exhibits a significantly positive trend (coef. = 0.17%; t-stat. = 4.81). Taken together, the analyses underlying Figure 4

reveal that for all categories except Performance, AR and Form 10-K category discussions have been converging since 2005. For Performance, AR and Form 10-K discussions have been diverging over the entire sample period, but most of the divergence occurred prior to 2005.

5.4 Differences in topic discussion in MD&A and financial statements

Table 3 presents the findings from comparing categories in AR and the MD&A and FS sections of Form 10-K. Figure 5 displays the findings. Table 3 reveals that, unlike the full Form 10-K, both MD&A and FS contain more discussion of Performance than AR (17.76% and 16.03% versus 14.11%). MD&A also contains more discussion of Analysis and Business than AR (14.16% and 13.28% versus 12.90% and 11.71%), but FS contains less (8.61% and 8.65%).[20] That MD&A contains more discussion of Performance, Analysis, and Business is surprising because these three categories typify the role of analysts and Table 1 reveals these are the top three categories discussed in AR.

Table 3 also reveals, as expected, that MD&A and FS contain more discussion of Financial Reporting than AR (20.04% and 30.59% versus 11.41%). Also, as one might expect, the difference is most pronounced for FS. In fact, the Financial Reporting differences between AR and the MD&A and FS sections of Forms 10-K are the largest among the five categories (8.63 pp. and 19.18 pp.). Thus, again, Financial Reporting discussions, particularly in FS, appear over-emphasized, relative to the discussion in AR. In addition, both MD&A and FS discuss Regulatory issues less than AR (5.93% and 6.37% versus 6.75%). Despite this difference, and perhaps unexpectedly, these findings reveal that the discussion in MD&A is more similar to that in AR than in FS. In aggregate, the AR and MD&A and FS sections of Form 10-K discussions of the categories are 15.11% and 21.11% different, and the MD&A and FS discussions are only

[20] Table 3 also reveals that the more extensive discussion of Business in Form 10-K revealed by Table 1 is attributable to MD&A, not FS (13.28% and 8.65% versus 11.71%).

11.90% different. These findings reveal MD&A and FS are more similar to each other than AR is to either MD&A or FS. However, AR is more similar to MD&A than to FS.

Table 4, Panels A through E, presents the findings relating to topics for each category. Regarding Performance, Panel A reveals that the pattern of each topic's discussion in both MD&A and FS mirrors that of the full Form 10-K in Table 2, Panel A. For example, the top three Performance topics for both sections and the full Form 10-K are Revenues/Margins, Earnings, and Expenses. However, Panel A also reveals that MD&A (FS) contains more discussion of Revenues/Margins (Earnings and Expenses) than the full Form 10-K. Strikingly, almost half (48.35%) of MD&A's discussion of Performance is devoted to discussing Revenues/Margins, whereas AR and FS devote only 26.33% and 25.92%. The largest differences between MD&A and FS relate to discussion of Revenues/Margins and Earnings (22.43 pp. and 13.64 pp.). Panel A reveals that, taken together, the discussions of Performance in MD&A and FS are 44.99% and 39.50% different from the discussion in AR. Although the difference for FS is similar to that for the full Form 10-K (39.53%), the difference for MD&A is considerably larger.

Regarding Analysis, Panel B reveals that the discussion in MD&A and FS largely are similar to that in the full Form 10-K as presented in Table 2, Panel B. The most notable difference is that MD&A contains more discussion of Increases and Decreases than FS and the full Form 10-K contain (24.13% versus 6.04% and 10.36%). Again, this likely reflects the requirement for firms to explain in MD&A large changes in financial statement amounts. The second largest difference is that FS contains more discussion of Estimates and Assessments than MD&A (20.89% versus 12.21%). As with the corresponding difference in Table 2, Panel B, this

likely reflects requirements for firms to disclose information about estimates included in financial statements.

Regarding Business, Panel C reveals patterns for MD&A and FS that are similar to those in Table 2 for the full Form 10-K. The only notable finding is that the greater discussion of Insurance and Health Care for Form 10-K in Table 2 is largely attributable to FS rather than MD&A (15.25% and 8.47% versus 9.09%), which likely reflects firms needing to report information on their pension plans.

Regarding Financial Reporting, Panel D also reveals patterns for MD&A and FS that are similar to those in Table 2 for the full Form 10-K. Two notable findings are that FS contains more discussion of Equity than MD&A (11.98% versus 7.49%) and MD&A contains more discussion of Credit and Financing than FS (11.07% versus 6.14%). Regarding Regulatory, Panel E reveals that the top two topics for MD&A are Risks and Uncertainties (28.80% and 20.07%), whereas for FS they are Forms and Exhibits and Responsibilities (25.06% and 20.71%). For AR, the top two topics are Risks and Responsibilities (23.80% and 23.19%). These findings, combined with those in Table 3, reveal that at the category level, MD&A resembles AR more than FS does. However, because the topics within categories differ, how MD&A and AR discuss these categories differ.

5.5 Differences in topic discussion for economically dissimilar firms

Table 5 presents regression summary statistics from estimating Equation (3), which relates to the category discussion in AR and Form 10-K between economically dissimilar firms. Panel A reports coefficients when the category proportion in AR is the dependent variable. This panel reveals that AR for Loss firms contains more discussion of Business and Regulatory and less discussion of Performance and Analysis than for Profit firms (coefs. = 0.48% and 0.48%; –

1.09% and –0.93%). AR of Small firms contains more discussion of Performance and Analysis and less discussion of Regulatory (coefs. = 1.74% and 0.46%; –0.64%). Table 5 also reveals that for Technology firms, AR contains less discussion of Financial Reporting (coef. = –0.49%). For Financial firms, AR contains more discussion of Performance and Financial Reporting, and less discussion of Analysis, Business, and Regulatory (coefs. = 0.37% and 2.99%; –0.57%, –1.93%, and –0.22%). The Financial Reporting coefficient of 2.99% is the largest.

Several of these findings are intuitive. AR is less focused on Performance for Loss firms because Profit firms' values are more likely attributable to existing operations than to growth options. AR is more focused on Business for Loss firms, which likely is because these firms' businesses are less well-understood. AR is more focused on Regulatory for Loss firms, which could be because some of these firms are awaiting regulatory approval or have losses related to adverse regulation. AR is more (less) focused on Financial Reporting for Financial (Technology) firms, whose equity value is more (less) likely to be reflected in financial reports.

Panel B uses the category proportion in Form 10-K as the dependent variable. This panel reveals that Form 10-K of Loss firms discuss Performance and Financial Reporting less than Profit firms but Business and Regulatory more than Profit firms (coefs. = –0.80% and –0.60%; 0.56% and 0.32%). These results for Loss firms are similar to Panel A, which suggests that Form 10-K, despite needing to comply with U.S. GAAP, adjusts in a similar fashion as AR, which is unregulated. This panel also shows that Form 10-K for Small firms discuss Performance and Financial Reporting less than Form 10-K for Large firms, but discuss Regulatory topics more (coefs. = –0.64% and –0.28%; 0.18%). Less discussion of Performance and Financial Reporting for Small firms may occur because some accounting-based performance measures and financial reporting do not apply to smaller firms.

We also find that Form 10-K for Technology firms have less discussion of Performance and Financial Reporting, but more discussion of Analysis and Business than other firms (coefs. = –0.47% and –0.93%; 0.25% and 0.51%). These differences in Form 10-K likely reflect that Technology are emblematic of the New Economy, which is only incompletely reflected in accounting. Thus, Technology firms may not have business models for which Performance and Financial Reporting is particularly applicable. We also find that Form 10-K for Financial firms discuss all categories less than Form 10-K for Non-Financial firms except for Financial Reporting. The greater discussion of Financial Reporting in Form 10-K for Financial firms likely reflects the fact that these firms tend to hold more marketable securities and loans.

Panel C reports coefficient estimates of equation (3) when the absolute difference in the discussion between AR and Form 10-K is the dependent variable. Because this difference is unsigned, a negative (positive) coefficient implies AR and Form 10-K has a larger (smaller) difference in the proportion of the document dedicated to a particular category. Panel C reveals a greater similarity in AR and Form 10-K discussion for Loss firms in that all the coefficients on the Loss indicator, except for Business and Other, are negative and significant (coefs. range from –0.14% to –0.63%). By contrast, the coefficients on the Small indicators are all positive and significant, except for Financial Reporting and Other (coefs. range from 0.20% to 1.44%). This result suggests a greater difference for Small firms between what analysts include in AR and what Form 10-K includes.

Perhaps surprisingly, we find the difference for Technology firms is not substantively different from Non-Technology firms. We find significantly positive coefficients for Performance and Business (coefs. = 0.33% and 0.33%), significantly negative coefficients for Financial Reporting and Other (coefs. = –0.44% and –0.37%). The Analysis and Regulatory

coefficients are insignificantly different from zero. This result perhaps is surprising because we would expect AR for Technology firms to discuss issues that are less related to what is in Form 10-K because these firms represent the New Economy and their business are only incompletely reflected in accounting. Thus, our finding of no consistent difference in the discussion for Technology firms suggests the criticism that accounting is irrelevant for Technology may be overblown.

In untabulated analyses, we also examine the extent to which AR and Form 10-K discussions devoted to the 100 topics differ across subsamples of dissimilar firms. These analyses reveal fewer differences in topics within AR and Form 10-K between the subsamples than within the subsamples between AR and Form 10-K. These findings suggest that although there are differences in AR between dissimilar firms, the differences are small relative to differences between AR and Form 10-K. Perhaps more strikingly, the difference in topics discussed in AR between dissimilar firms is smaller than the difference in topics discussed in Form 10-K. This is striking because AR largely is unregulated, whereas Form 10-K largely requires the same information for all firms.

6. Additional analyses

6.1 Analysis portion of analyst reports

AR contains standardized brokerage disclosures such as information relating to the brokerage firm and legal disclaimers relating to the information in the AR. These disclosures are a form of boilerplate because they do not relate to the analyst's analysis (Huang et al. 2014). Thus, we examine separately the portion of AR that is unlikely to include these disclosures. Our review of a subsample of AR reveals that when short (long) brokerage disclosures exist, they typically appear at the beginning (end) of the document. Thus, we identify and remove the first

10% and last 75% of the sentences in each AR and analyze the remaining content. We refer to this truncated portion of the AR as the analysis portion.

We first assess whether the analysis portion of AR is more likely to contain fundamental analysis and less likely to contain standardized brokerage disclosures. Specifically, we identify three topics in the Other category that appear unrelated to the analyst's analysis and determine whether these topics are discussed more in the full AR than in the analysis portion. These topics are (i) brokerage names, with words such as *ubs*, *morgan_stanley*, *jp_morgan*, and *investment_banking*; (ii) subsidiaries and affiliates, with words such as *subsidiaries*, *affiliates*, *persons*, *person*, and *subsidiary*; and (iii) broker regulations, with words such as *regulated*, *authority*, *financial_services*, *uk*, and *united*. Untabulated findings reveal that each of these topics is discussed more in the full AR than in the analysis portion (2.62%, 1.11%, and 0.85% versus 1.03%, 0.38%, and 0.11%).

More importantly for our research question, untabulated findings based on the analysis portion of AR reveal inferences consistent with those revealed by our tabulated findings. In particular, untabulated findings reveal that the analysis portion contains more discussion of Performance than the full AR (17.76% versus 14.11%). This finding is consistent with Table 1 in that Performance is the most-discussed category in AR. Regarding Performance topics, the analysis portion has more discussion of Revenues/Margins, Ratios, and Adjusted Earnings, a similar discussion of Expenses, and less discussion of Earnings (35.15%, 22.09%, 11.37%, 4.40%, and 11.52% versus 26.33%, 16.85%, 9.58%, 4.58%, and 14.71%). These findings are consistent with Table 2, Panel A, in that among Performance topics, AR focuses more on Revenues/Margins and less on Expenses than Form 10-K.

In addition, untabulated findings reveal that the analysis portion contains less discussion of Financial Reporting and Regulatory than the full AR (9.93% and 4.10% versus 11.41% and 6.75%). Regarding Financial Reporting topics, the analysis portion discusses Investments less than the full AR (10.27% versus 19.34%). Regarding Regulatory topics, the analysis portion discusses Risks and Uncertainties more than the full AR (30.17% and 19.65% versus 23.80% and 10.97%). These findings are consistent with analyst reports focusing less on Financial Reporting and Regulatory, except for Risks. Finding consistent results based on the analysis portion of AR and based on the full AR, suggests our inferences are not dependent on the particular portion of AR we use in our main analyses.

6.2 AR category discussions and analyst forecast errors

We next determine whether AR category discussions are associated with higher or lower analyst forecast errors. Specifically, we estimate the relation between absolute analyst forecast error (AFE) and five indicator variables that equal one for each of the five topic categories and zero otherwise. The relation also includes return on assets, market value of equity, and Fama-French 48 industry fixed effects as controls for firm characteristics that could affect our inferences. AFE is the absolute difference between earnings for the year and the most recent mean consensus earnings forecast before fiscal year end, scaled by fiscal year end share price.

Untabulated findings reveal a significantly negative relation between AFE and AR Analysis discussion (t-stat. = –2.95) and significantly positive relations between AFE and AR Business, Financial Reporting, and Regulatory discussions (t-stats. range from 2.32 to 2.89). The relation between AFE and AR Performance discussion is not significant. These findings indicate that when AR contains more Analysis discussion the analyst earnings forecast is

significantly more accurate. When AR contains more discussion of Business, Financial Reporting, and Regulatory the forecast is significantly less accurate.[21]

7. Conclusion

We address whether financial reports include information supporting investor valuations. Addressing this question is fundamental because financial reports are designed to provide investors with information to help them make their decisions about providing capital to a firm, but financial reports have been criticized for lack of relevance to investors. To address our research question, we analyze and compare the contents of Form 10-K, which is the firm's primary financial report, and equity analyst reports (AR), which reveal information supporting investor valuations.

Using a word-embedding topic model, we identify the topics discussed in AR and Form 10-K and categorize the topics into broad categories on which we focus our analysis: Performance, Analysis, Business, Financial Reporting, and Regulatory. Most notably, we find that Performance, Analysis, and Business are the three most-discussed topic categories in AR. Surprisingly, we find that Financial Reporting topics are discussed almost as much as Business, which reveals that financial reporting is a crucial component of AR and, thus, supports investor valuations. However, our evidence also is consistent with Form 10-K over-emphasizing Financial Reporting in that Form 10-K discusses this category almost twice as much as AR.

Over our sample period, we find a steady decline in the difference of categories discussed in AR and Form 10-K. This result is inconsistent with accounting losing its relevance with the rise of the New Economy. We also find the extent to which AR and Form 10-K discuss all five categories are converging over time.

[21] Although we include controls for some firm characteristics, analyst forecast errors could be associated with others.

Regarding Performance, we find that revenues and margins is the most-discussed topic in both AR and Form 10-K. However, AR and Form 10-K discussions of the other Performance topics differ considerably. In particular, we find that Form 10-K focuses more on earnings and expenses, whereas analyst reports focus more on ratios, target prices, recommendations, and adjusted earnings. Surprisingly, analyst reports focus more on earnings than on adjusted earnings or cash flows. These findings are inconsistent with analysts viewing earnings as irrelevant and focusing on alternative performance measures, including cash flows.

We find the difference in the proportion AR and Form 10-K discuss topics in each category has declined except within the Performance category. Thus, how AR and Form 10-K discuss most of the categories is becoming increasingly similar. For Performance, we observe a large divergence in topic proportions in the early part of our sample and remained steady thereafter. This result suggests how AR and Form 10-K discuss Performance became more dissimilar.

We also compare AR to two sections of Form 10-K: the MD&A and the financial statements section. We find that the discussion in AR is more similar to Form 10-K's MD&A section than to the discussion in Form 10-K's financial statements section. In addition to both Form 10-K sections discussing Performance more than AR, MD&A discusses Analysis and Business more than AR. This finding is notable because these categories typify the roles of analysts and suggest the MD&A resembles many aspects of AR.

Our comparisons of differences in AR and Form 10-K content for economically dissimilar firms reveal that the differences depend on firms' circumstances. We find that AR and Form 10-K are most similar for Loss firms. Surprisingly, we find AR and Form 10-K for

Technology firms are not substantively different from each other compared to Non-Technology firms. This result is inconsistent with accounting being more irrelevant in the New Economy.

Additional analyses reveal two insights. First, our inferences are not dependent on the particular portion of AR we use in our main analyses. Second, when AR contains more Analysis discussion the analyst earnings forecast is more accurate, and when AR contains more discussion of Business, Financial Reporting, and Regulatory the forecast is less accurate.

Despite some notable differences, we find many similarities in the contents of Form 10-K and AR. Taken together, our findings are inconsistent with financial reports lacking relevance to investor valuations.

References

Asquith, P., M.B. Mikhail, and A.S. Au. 2005. Information content of equity analyst reports. *Journal of Financial Economics* 75(2): 245-282.

Balachandran, S., and P. Mohanram. 2011. Is the decline in value relevance of accounting driven by increased conservatism? *Review of Accounting Studies* 16 (2): 272–301.

Barth, M.E., W.H. Beaver, and W.R. Landsman. 2001. The relevance of the value relevance literature for financial accounting standard setting: Another view. *Journal of Accounting and Economics* 31: 77-104.

Barth, M.E., K. Li, and C.G. McClure. 2023. Evolution in value relevance of accounting information. *The Accounting Review* 98(1): 1-28.

Bradshaw, M.T. 2002. The use of target prices to justify sell‐side analysts' stock recommendations. *Accounting Horizons* 16(1): 27-41.

Breton, G., and R.J. Taffler. 2001. Accounting information and analyst stock recommendation decisions: A content analysis approach. *Accounting and Business Research* 31(2): 91-101.

Brown, S., K. Lo, and T. Lys. 1999. Use of R^2 in accounting research: Measuring changes in value relevance over the last four decades. *Journal of Accounting and Economics* 28 (2): 83–115.

Cascino, S., M.A. Clatworthy, B.G. Osma, J. Gassen, and S. Imam. 2021. The Usefulness of Financial Accounting Information: Evidence from the Field. *The Accounting Review* 96(6): 73-102.

Cazier, R.A., J.L. McMullin, and J.S. Treu. 2021. Are lengthy and boilerplate risk factor disclosures inadequate? An examination of judicial and regulatory assessments of risk factor language. *The Accounting Review* 96(4): 131-155.

Chen, X., Q. Cheng, and K. Lo. 2010. On the relationship between analyst reports and corporate disclosures: Exploring the roles of information discovery and interpretation. *Journal of Accounting and Economics* 49(3): 206-226.

Christensen, H.B., E. Floyd, L.Y. Liu, M. Maffett. 2017. The real effects of mandated information on social responsibility in financial reports: Evidence from mine-safety records. *Journal of Accounting and Economics.* 64(2-3): 284-304.

Collins, D.W., E.L. Maydew, and I.S. Weiss. 1997. Changes in the value-relevance of earnings and book values over the past forty years. *Journal of Accounting and Economics* 24: 39-67.

Core, J.E., W.R. Guay, and A. Van Buskirk. 2003. Market valuations in the New Economy: An investigation of what has changed. *Journal of Accounting and Economics* 34: 43-67.

Damodaran, A. (2009). Valuing young, start-up and growth companies: estimation issues and valuation challenges. *Unpublished Working Paper*.

Darrough, M., and J. Ye. 2007. Valuation of loss firms in a knowledge-based economy. *Review of Accounting Studies* 12(1): 61-93.

De Franco, G., M.H.F. Wong, and Y. Zhou. 2011. Accounting adjustments and the valuation of financial statement note information in 10-K filings. *The Accounting Review* 86(5): 1577-1604.

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint:1810.04805.

Drake, M., J. Hales, and L. Rees. 2019. Disclosure Overload? A Professional User Perspective on the Usefulness of General Purpose Financial Statements. *Contemporary Accounting Research* 36(4): 1935-1965.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of Form 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64: 221-245.

Fama, E.F., and K.R. French. 2006. Profitability, investment and average returns. *Journal of Financial Economics* 82(3): 491-518.

Francis, J., and K. Schipper. 1999. Have financial statements lost their value relevance? *Journal of Accounting Research* 37(2): 319-352.

Govindarajan, V. 1980. The objectives of financial statements: An empirical study of the use of cash flow and earnings by security analysts. *Accounting, Organizations and Society* 5(4): 383-392.

Holthausen, R.W., and R.L. Watts. 2001. The relevance of the value relevance literature for financial accounting standard setting. *Journal of Accounting and Economics* 31: 3-75.

Hope, O.K., D. Hu, and H. Lu. 2016. The benefits of specific risk-factor disclosures. *Review of Accounting Studies* 21: 1005-1045.

Howe, J.S., E. Unlu, and X. Yan. 2009. The predictive content of aggregate analyst recommendations. *Journal of Accounting Research* 47(3): 799-821.

Huang, A.H., R. Lehavy, A.Y. Zang, and R. Zheng. 2018. Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science* 64(6): 2833-2855.

Huang, A.H., A.Y. Zang, and R. Zheng. 2014. Evidence on the information content of text in analyst reports. *The Accounting Review* 89(6): 2151-2180.

Huang, S., H. Tan, X. Wang, and C. Yu. 2023. Valuation uncertainty and analysts' use of DCF models. *Review of Accounting Studies* 28: 827-861.

Investor Responsibility Research Center Institute (IRRC). 2016. *The corporate risk factor disclosure landscape*.

Joos, P., and G.A. Plesko. 2005. Valuing loss firms. *The Accounting Review* 80(3): 847-870.

Kim, C., K. Wang, and L. Zhang. 2019. Readability of 10‐K reports and stock price crash risk. *Contemporary Accounting Research* 36(2): 1184-1216.

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.

Lev, B., and F. Gu. 2016. *The End of Accounting and the Path Forward for Investors and Managers*. John Wiley & Sons, Inc. Hoboken, NJ, USA.

Lev, B., and P. Zarowin. 1999. The boundaries of financial reporting and how to extend them. *Journal of Accounting Research* 37 (2): 353–385.

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

Li, K., F. Mai, R. Shen, and X. Yan. 2021. Measuring corporate culture using machine learning. *Review of Financial Studies*, 34(7): 3265-3315.

Li, K., F. Mai, R. Shen, C. Yang, and T. Zhang. 2024. Dissecting corporate culture using generative AI - Insights from analyst reports. *Unpublished working paper.*

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. 2014. Measuring readability in financial disclosures. *Journal of Finance* 69(4): 1643-1671.

Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54(4): 1187-1230.

Martineau, C., and M. Zoican. 2021. Measuring information in analyst reports: A machine learning approach. *Unpublished working paper.*

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv*: 1301-3781.

Ohlson, J.A. 1995. Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research* 11(2): 661-687.

Previts, G.J., R.J. Bricker, T.R. Robinson, and S.J. Young. 1994. A content analysis of sell-side financial analyst company reports. *Accounting Horizons* 8(2): 55-70.

Rehurek, R., and P. Sojka. 2010. Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*: 45-50.

Schipper, K. 1991. Analyst forecasts. *Accounting Horizons* 5(4): 105-121.

Tavcar, L.R. 1998. Make the MD&A more readable: Certified Public Accountant. *CPA Journal* 68(1): 10.

Womack, K.L. 1996. Do brokerage analysts' recommendations have investment value? *Journal of Finance* 51(1): 137-167.

Appendix

Overview of Word2vec and why we use it

To identify the information in equity analyst reports (AR) and Form 10-K, we use a neural network-based word embedding topic modeling approach, Word2vec. Word2vec has been used in a variety of fields, including speech recognition, machine translation, question answering, and finance. For example, in finance, Li, Mai, Shen, and Yan (2021) uses Word2vec to develop a dictionary to measure corporate culture.

Word2vec groups together words with similar meanings. This grouping approach is appropriate for our setting because in accounting different words often are used when discussing the same topic. For example, pairwise *costs* and *expenses*, *repurchase* and *buyback*, and *eps* and *earnings per share* likely relate to the same topic. Word2vec identifies words as having similar meanings if the words co-occur with similar neighboring words. For example, Word2vec would identify *earnings* and *net income* as having similar meanings if they both occur frequently in the context of "We had record _____ this quarter." We find that Word2vec produces coherent topics in that it includes in the same topic words such as *properties* and *leasing*.

An alternative topic-analysis method is Latent Dirichlet Allocation (LDA). LDA has three differences from Word2vec that make it less appropriate for our research question. First, LDA requires the exclusion of common words because including common words can prevent the estimator from converging (Dyer et al. 2017). If we had to remove the most common 50 words in our corpus, we would be unable to analyze words such as *income* and *cash*, which are important in accounting.[22]

[22] Dyer et al. (2017) eliminates the 0.1% most common words in that study's corpus, which is 49 words.

Second, in LDA words can appear in multiple topics because LDA probabilistically associates words with latent topics. For example, in Dyer et al. (2017), the word *earnings* appears in the topics Foreign Currency Exchange, Derivatives, Performance Growth, Summary of Consolidated Results, and Pension Retirement Plans. Consequently, using LDA would make it difficult to analyze *earnings* as a separate topic. In contrast, Word2vec groups together words with similar meanings into topics, and each word appears in only one topic (Mikolov, Chen, Corrado, and Dean 2013). This enables us to calculate how much discussion relates to each topic, even topics that include the words most commonly used in AR and Form 10-K.

Third, LDA identifies broad topics whereas Word2vec identifies more specific topics. For example, in LDA, *earnings* appears in multiple topics in the context of other words related to multiple themes such as *foreign*, *currency*, *derivatives*, *consolidated*, *benefit*, and *pension*. In Word2vec, *earnings* appears in one topic with more specific and closely related words such as *net*, *operating*, *income*, *loss*, *tax*, and *net_income*.

Another method to represent text is BERT (Bidirectional Encoder Representations from Transformers; Devlin, Chang, Lee, and Toutanova 2018). An advantage of BERT over Word2vec is that BERT can take more context from words and sentences into account. However, BERT text representations typically are based on sentences and paragraphs. Thus, the BERT representation of a particular word may be different based on other words in the sentence or paragraph. As a result, as with LDA, a word can appear in multiple topics in BERT whereas, as explained above, an advantage of Word2vec is that it includes a word in only one topic, which facilitates interpretation of topics.

Implementation of Word2vec

We implement the Word2vec approach in six steps. First, we obtain AR from Thomson Reuters Investext. We eliminate AR from brokers that primarily provide robo reports, event transcripts, proxy and governance reports, credit rating reports, and company descriptions. To do this, we focus on the top 100 brokers by number of reports in our sample and review a random sample of five reports per broker. We keep reports from only brokers that primarily provide fundamental analysis. We randomly select one AR for each firm-year to ensure our topic modeling approach is computationally feasible. We convert AR from PDF to text for analysis. For PDF reports that are not machine readable, we use optical character recognition to convert the PDF to text.

Second, we obtain Form 10-K from the EDGAR database and parse the text to remove HTML. This is necessary because in the EDGAR database Form 10-K contains HTML, which is not subject to our analyses. We base our approach for removing HTML on Loughran and McDonald (2016). In particular, we begin by removing all document segments that contain GRAPHIC, ZIP, EXCEL, JSON, PDF, and XML. We then parse HTML code—for example, we replace \ and \ with a blank space—and remove SEC headers and footers and HTML tags. We also remove XBRL, but we retain text that appears in tables. Because AR generally does not contain HTML, this step is not necessary before we analyze AR.

In this step, we also extract two sections of Form 10-K, namely Management's Discussion & Analysis (MD&A) and financial statements (FS). We base our approach for identifying the MD&A and FS sections on Dyer et al. (2017). In particular, we identify all instances of logical references to the MD&A and FS sections and impose a minimum section length of 2,000 characters. If multiple instances exist, we choose the longest.

Third, we process the text for analysis. Because AR is shorter, on average, than Form 10-K, we scale the AR corpus to be the same length as the Form 10-K corpus. That is, we duplicate the AR corpus R times, where R is the ratio of Form 10-K total word count to AR total word count, truncated to an integer. We then combine the scaled AR corpus and the Form 10-K corpus into a single combined corpus. Next, we convert uppercase letters to lowercase and eliminate punctuation symbols, single-character words, and words that are not all alphabetical characters. We then remove the 50 words that appear most frequently in the combined corpus such as *the*, *of*, and *and*, unless the words could relate to a topic of potential substantive interest, such as *income* and *cash*. We do not focus on word stems because many accounting words with different meanings are based on the same stem.[23]

We concatenate as a single word—using the underscore symbol—sequences of up to 16 words (i.e., 16-grams) that appear, on average, more than once per AR. For example, because the word sequences *net income* and *earnings per share* appear, on average, more than once per AR, we treat them as the single words *net_income* and *earnings_per_share*. To manage vocabulary size, we exclude words that appear, on average, less than once per 25 firm-year observations.

Fourth, we train the Word2vec model on the combined AR and Form 10-K corpus. We use the gensim library, which is an open-source Python package, to implement Word2vec (Rehurek and Sojka 2010). We follow the standard implementation and use a word window size of five and a vector size of 100. The word window specifies how many words before and after a given word Word2vec considers when determining the word's meaning. The vector size specifies how many dimensions of meaning Word2vec considers. See Mikolov et al. (2013) for

[23] Stem refers to the root of the word. For instance, *sales* and *sale* have the same root, *sale*. However, whereas *sales* often relates to revenue, *sale* can relate to disposition of an asset or a line of business.

details. This step effectively converts each word to a vector in a 100-dimensional vector space, in which words with similar (dissimilar) meanings are close together (far apart).

Fifth, we use K-means clustering to cluster the word vectors into clusters that minimize the within-cluster sums of mean squared vector differences. We instruct the model to create 100 clusters because untabulated statistics reveal that the rate of reduction in the sums-of-squares levels off at close to 100 clusters (Dyer et al. 2017).[24] Each cluster represents a topic. Together, these topics include all the words in the combined AR and Form 10-K corpus.

Sixth, for ease of discussion, we label and, following Dyer et al. (2017), organize the topics into topic categories. We identify five categories—Performance, Analysis, Business, Financial Reporting, and Regulatory—on which we focus our analysis—and Other. For each of these categories, Table A.1, Panels A through F, presents the topics listed in order of the most to least discussion in the AR category. It also presents the 15 most frequently occurring words in each topic, listed in order of the frequency with which they occur in the combined AR and Form 10-K corpus.

[24] The reduction in sums-of-squares achieved by increasing the number of topics from 10 to 20 (90 to 100) is 4.6 (only 1.7) times the reduction in sums-of-squares achieved by increasing the number of topics from 90 to 100 (190 to 200).

Table A.1
Topic Categories, Topics, and Words

Panel A: Performance Topic Category

Topic	# Words	Words
Revenues/Margins	101	million was sales revenue were due revenues margin fiscal respectively segment compared year_ended_december million_million margins
Ratios	108	ratio pe na dec fy cap source_company mm book usd mil avg leverage figure nm
Earnings	96	net operating income loss tax earnings net_income losses foreign currency income_taxes taxes income_tax gains gain
Target prices	53	price investors return stocks analysts sp total_return target index coverage_universe stock_price peer stars reit ranking
Recommendations	38	rating ratings sector outperform underperform rated past_months none usd_usd nr stewardship weight next_months ib medium
Adjusted earnings	59	eps per_share ebitda diluted adjusted flow shares_outstanding gaap basic earnings_per_share weighted_average pro forma calculated calculation
Expenses	27	costs cost expense expenses fees deferred recorded amortization recognized charges depreciation interest_expense incurred charge operating_expenses
Cash flows	10	cash cash_flow cash_flows net_cash free_cash cash_cash investing cash_equivalents cashflow undiscounted

Table A.1 (continued)
Topic Categories, Topics, and Words

Panel B: Analysis Category

Topic	# Words	Words
Markets and industries	137	market these new certain more industry number well those high some customer most specific generally
Trends and forecasts	83	current expected future performance results rates conditions level past relative levels actual periods forecast volatility
Positives and negatives	179	strong continued demand low overall activity positive negative upside recovery solid despite incremental pressure relatively
Estimates and assessments	82	estimates estimate valuation analysis estimated model using assumptions reporting analyses moat determine assessment methodology quantitative
Increases and decreases	28	growth increase increased primarily higher lower decrease down decline increases reduction reduced decreased increasing improvement
Successes and challenges	63	than one likely recent competitive important favorable recently consistent stable attractive difficult volatile successful critical
Views and opinions	43	if believe expect view opinion believes think assurance anticipate believed evidence uncertain confidence ultimate probable
Economic environment	67	economic environment caused economy damage weakness political weather delays lost challenges pressures problems downturn deterioration
Affect and effect	9	effect impact affect affected impacted affecting influenced impacting affects

Table A.1 (continued)
Categories, Topics, and Words

Panel C: Business Category

Topic	# Words	Words
Products and markets	82	us products companies operations customers markets clients facilities sources businesses segments relationships components firms arrangements
Business and operations	110	business management development corporate key marketing program support resources strategy programs opportunities research_development focus personnel
Services and technology	203	services systems technology system line software solutions technologies enterprise network communications delivery applications provider storage
Banking	78	bank state federal trade local national commission banking banks department deposit canadian agency residents association
Energy	96	capital energy investment_banking_services lp sipc bluematrix wholly dominion noninvestment midstream petroleum gp resource incs qualifying
Goods and retail	62	product retail consumer selling category distributor specialty brand goods food wholesale brands channel circulation merchandise
Insurance and healthcare	79	benefit insurance benefits coverage plans health policy pension professional care retirement medical healthcare savings severance
Prices and production	80	prices production contract oil gas volume capacity natural_gas supply volumes fuel commodity coal raw crude
States and regions	92	new_york delaware located american south america california north texas central west city pacific district florida
Cities and international locations	44	branch office street london floor offices square east box road tel avenue tower po tokyo
Manufacturing and materials	125	manufacturing materials industrial produced industries water generation steel parts paper semiconductor automotive waste natural manufacture
Transportation and utilities	109	service power orders electric transportation utility utilities air aircraft replacement vehicle star fleet terminal vehicles
Hotels and entertainment	100	country home stores store trademarks owners franchise centers entertainment family names gaming restaurants hotel territory
Drugs and clinical trials	118	clinical patients treatment phase drug trial stage study trials response candidates patient cancer studies launch
Projects and pipelines	74	project expansion area projects pipeline operation drilling region feet exploration location wells field proved plants
Partnerships and ventures	66	license partners partnership ii joint iii venture pharmaceuticals licensing arrangement exclusive partnerships royalty pharmaceutical ventures

Table A.1 (continued)
Categories, Topics, and Words

Panel D: Financial Reporting Category

Topic	# Words	Words
Investments	56	securities investment investments trading transactions exchange contracts agreements funds private fund financial_instruments shortterm instruments derivative
Equity	59	stock shares equity share common_stock common outstanding dividend dividends class shareholders stockholders offering issuance proceeds
Financial statements	41	financial statement see consolidated table_contents included reported statements financial_statements results_operations table thousands millions discussed balance_sheet
Employees	86	employees executive you member who employee director committee termination employment directors board your participant board_directors
Interest rates	71	interest rate term longterm yield fixed interest_rate life discount forward premium deposits interest_rates maturity contractual
Issuances	36	issued purchase distributed acquired sold held owned purchased offered accounted treasury valued traded closed trades
M&A	62	acquisition sale control transaction acquisitions closing restructuring merger combination consolidation completion integration ma disposal disposition
Credit and financing	69	credit loan principal obligations payment payments senior financing facility liquidity default commitments credit_facility collateral borrowings
Compensation	44	plan compensation option units options exercise incentive granted grant awards award restricted restricted_stock stock_options vesting
Regulations	79	requirements regulatory regulations laws compliance government environmental safety states protection restrictions jurisdictions fda practices governmental
Loans	73	loans commercial portfolio trust mortgage real_estate finance residential bonds lending underwriting servicing lien housing portfolios
Litigation	96	legal claims against action decision settlement judgment court litigation claim actions patent final proceedings patents
PPE	44	property equipment lease properties construction leases maintenance tenant building plant improvements land space rent intellectual
Debt and notes	5	debt note notes indebtedness debentures
Liabilities	49	liabilities amounts accounts balance payable accrued reserves receivable reserve equivalents receivables premiums balances inventories contingent
Intangibles/impairment	39	assets asset impairment goodwill inventory allowance carrying deferred_tax intangible_assets useful intangibles impaired lives longlived allowances
Fair value	9	value fair_value method fair values blackscholes splits basecase optionpricing
Accounting standards	38	accounting standard guidance standards sfas adoption principles poors recognition fasb presentation asu sharebased asc financial_reporting

Table A.1 (continued)
Categories, Topics, and Words

Panel E: Regulatory Category

Topic	# Words	Words
Risks	82	change risk changes potential factors risks recommendations circumstances views events matters decisions uncertainty positions exposure
Responsibilities	147	but upon andor respect except opinions otherwise without liability rights event relevant extent right obligation
Forms and exhibits	80	under form exhibit filed defined accordance pursuant covered dated connection amended intended among mentioned described
Legal language	62	agreement section act herein applicable terms item law rules regulation code securities_exchange_commission solicitation schedule rule
Uncertainties	60	material result significant adverse adversely financial_condition cause materially reduce significantly substantial lead occur limit aware
Audit report	43	registered firm public independent subject_change_without_notice preparation audit manager counsel llp sarbanesoxley redistribution author audited oversight
Disclaimers	50	responsible personal nor damages directly_indirectly arising conflict conflicts breach acts whatsoever liable fraud unauthorized violation
Internal controls	9	internal_control_over_financial_reporting holdneutral internal_control sponsoring treadway ratingsib coso controlintegrated unqualified

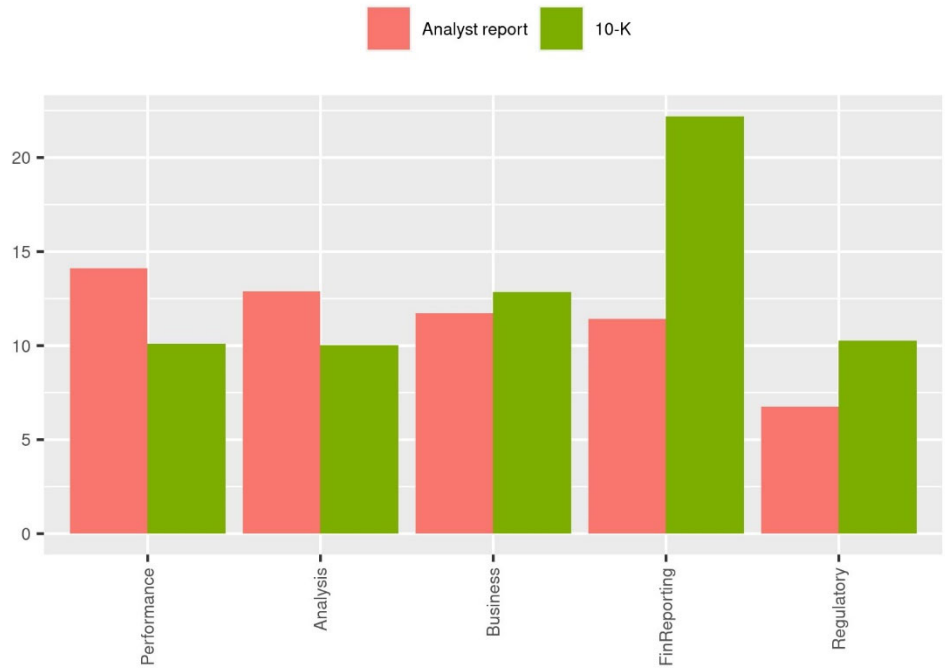
Table A.1 (continued)
Categories, Topics, and Words

Panel F: Other Category

Topic	# Words	Words
1	273	no it been also only part additional addition above when while however they currently further
2	113	year period during quarter years within time following prior first after end months last three
3	80	may will could available should required would do can does ability order reasonable must necessary
4	88	morningstar ubs morgan_stanley jp_morgan investment_banking jefferies deutsche_bank credit_suisse rbc_capital_markets morgan_stanley_research capital_markets sp_capital wells_fargo_securities_llc iq piper
5	116	total each over amount average basis annual up per approximately portion base percent unit less
6	186	company corporation group llc between co corp holdings ag rights_reserved usa bancorp chase associates dean
7	85	companys com buy equity_research price_target neutral target_price recommendation update history mar overweight sep jun underweight
8	121	limited general international united_states global third individual core primary separate single institutional major whole world
9	60	information page below disclosure disclosures please important_disclosures certification additional_information please_see information_data discussion list description definitions
10	89	including based related used about associated source regarding includes provides represents based_upon makes generated derived
11	133	sell provide make continue hold receive perform pay complete maintain remain become seek obtain meet
12	144	analyst cfa views_expressed registrant vice_president research_analyst mark certify president john michael james david former associate
13	120	provided made there effective paid received given determined prepared approved deemed established completed obtained indicated
14	200	had since making record remains maintaining known appears transition maintained revised created experienced assumes added
15	95	their through use into offer existing own access providing improve develop drive operate ensure manage

This table lists the topic categories, topics within each category, and the 15 top words for each topic estimated by Word2vec. For the Other category, Panel F lists the top 15 of the category's 41 topics. For each topic, words are listed in order of their frequency of occurrence in AR and Form 10-K.

Figure 1
Analyst Report and Form 10-K Discussion by Topic Category



This figure presents the average discussion of each topic category in analyst reports and Form 10-K, as a proportion of the total discussion in each document. See Table A.1, Panels A to E, in the Appendix for the discussion topics in each category and the most frequent words in each topic. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Figure 2
Difference in Category Discussion in Analyst Report and Form 10-K by Year

Panel A: Aggregate Difference

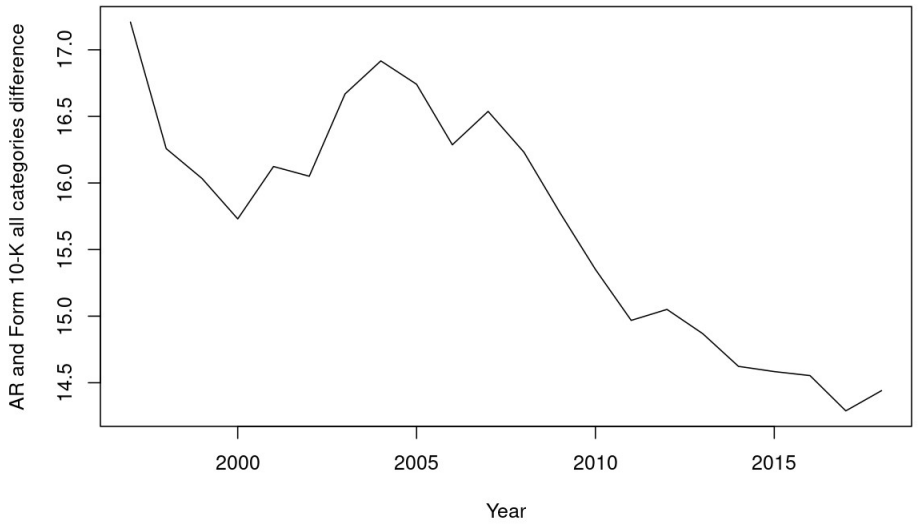
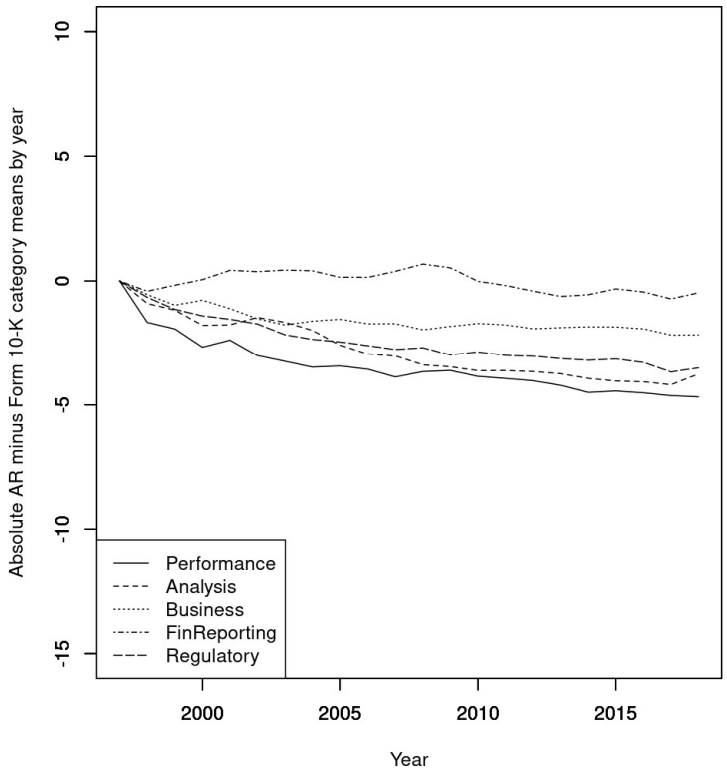


Figure 2 (continued)
Difference in Category Discussion in Analyst Report and Form 10-K by Year

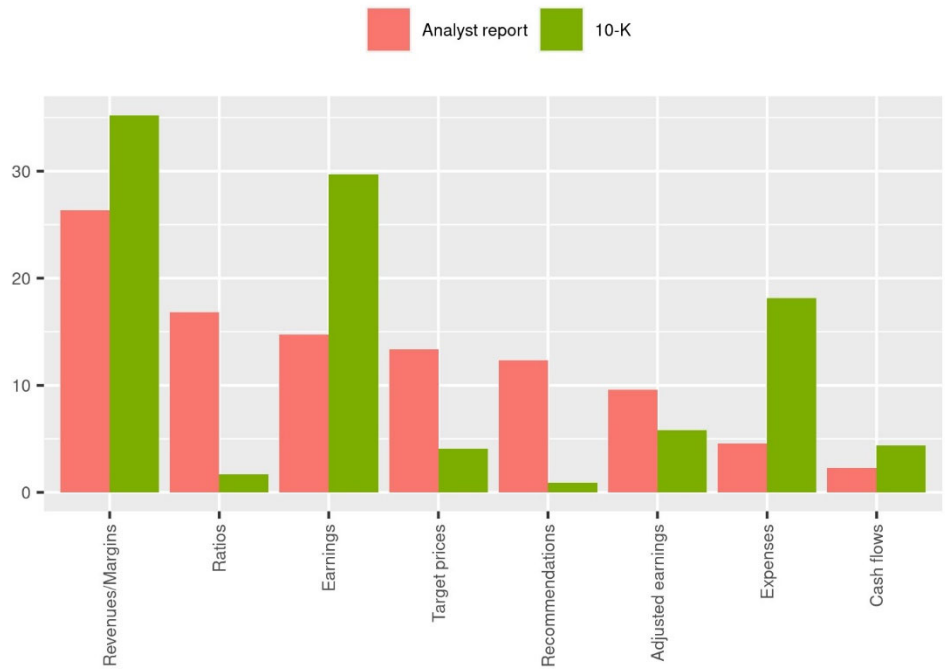
Panel B: Category Differences



This figure presents the difference in category discussions in Form 10-K and AR, averaged by year. Panel A reports the aggregate difference and is the sum of absolute differences in all categories, divided by two. Panel B reports the differences for each category and is the average absolute difference for each category by year. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Figure 3
Analyst Report and Form 10-K Discussion by Topic

Panel A: Performance Category



Panel B: Analysis Category

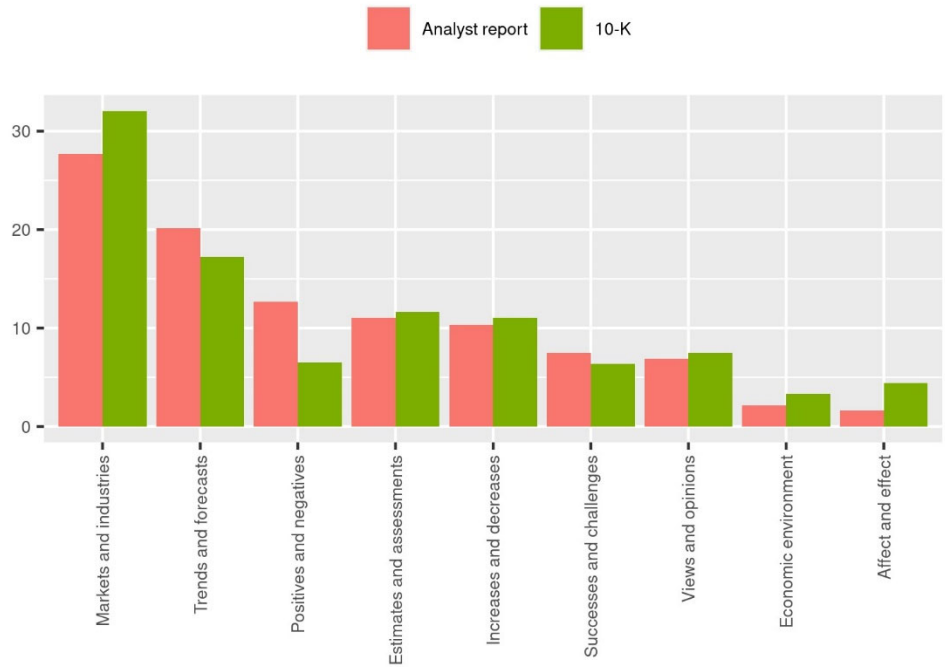
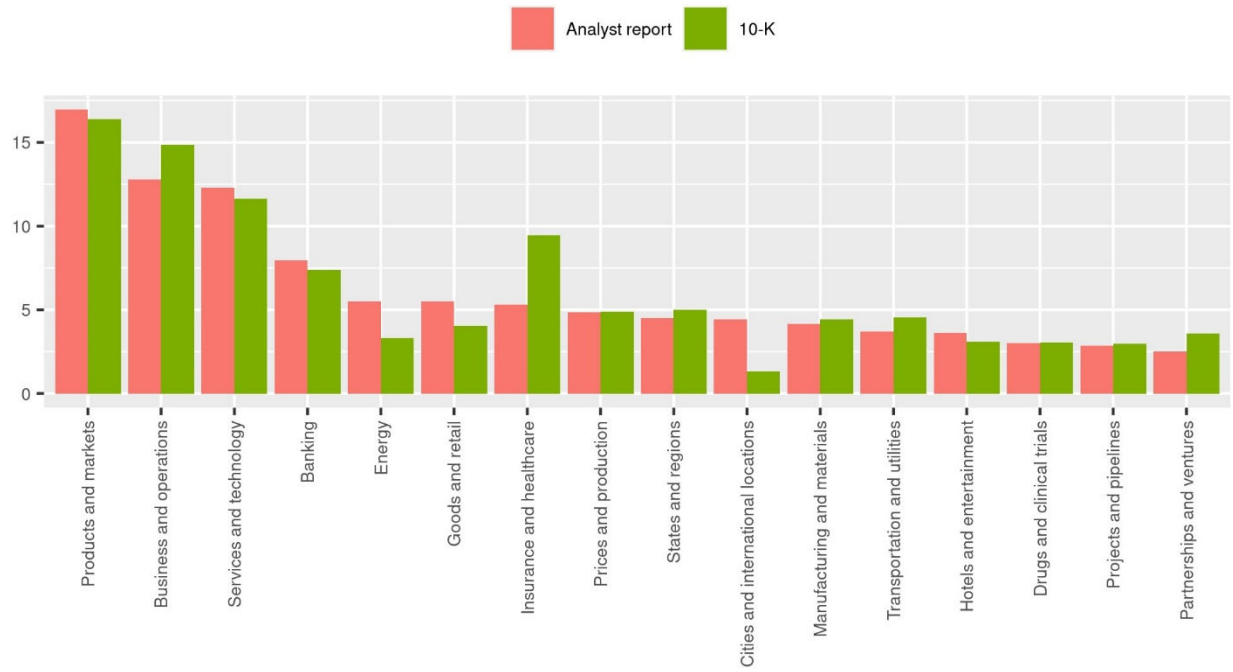


Figure 3 (continued)
Analyst Report and Form 10-K Discussion by Topic

Panel C: Business Category



Panel D: Financial Reporting Category

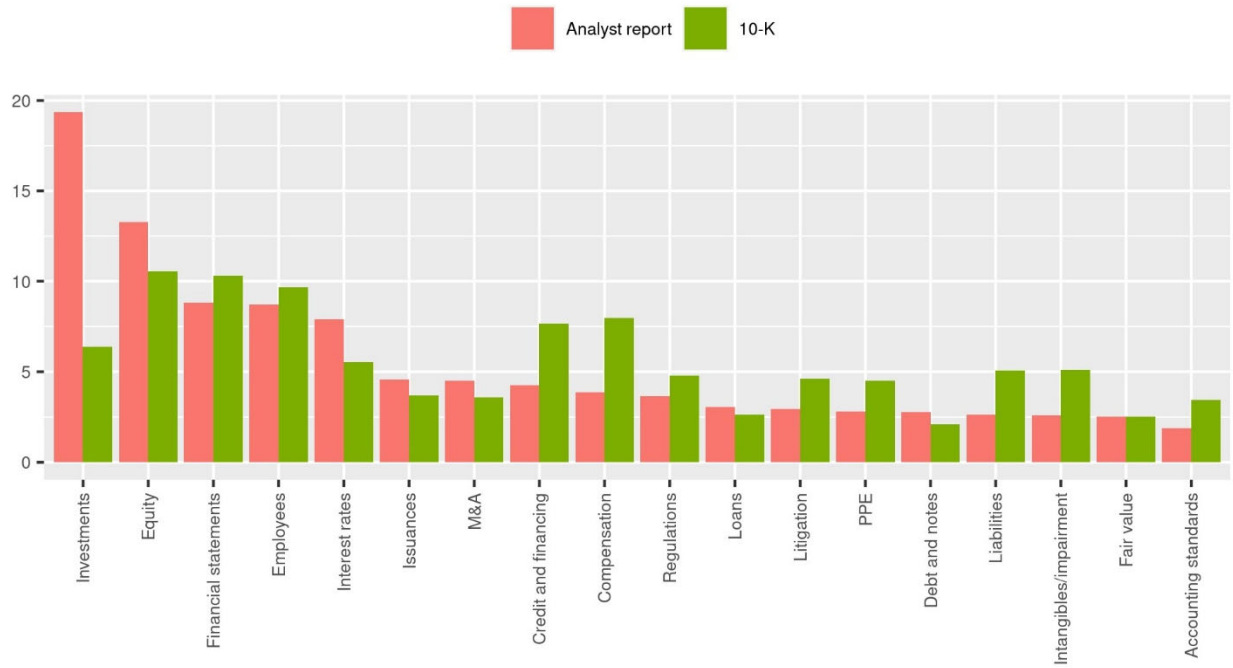
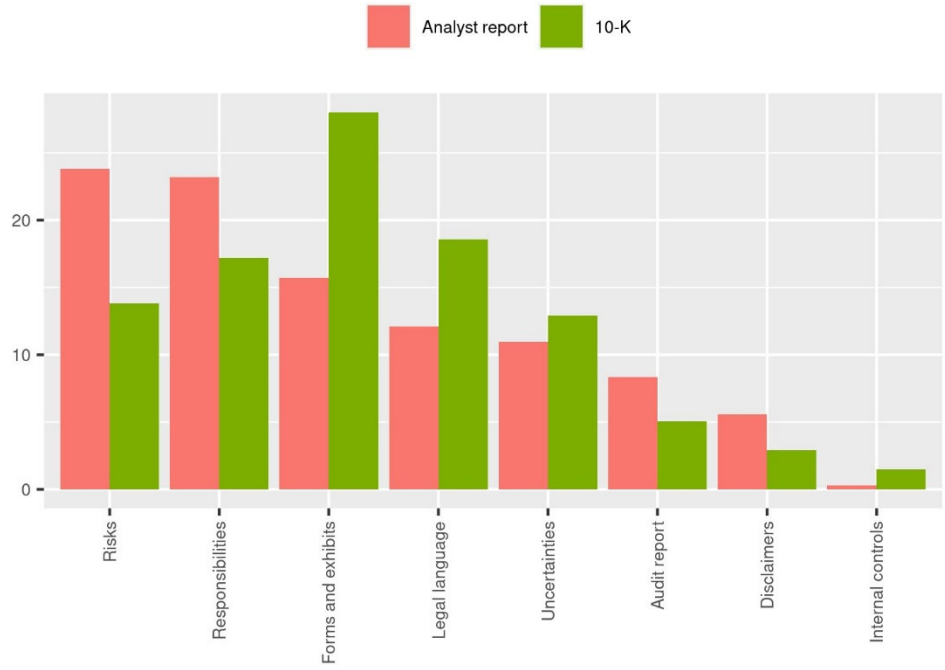


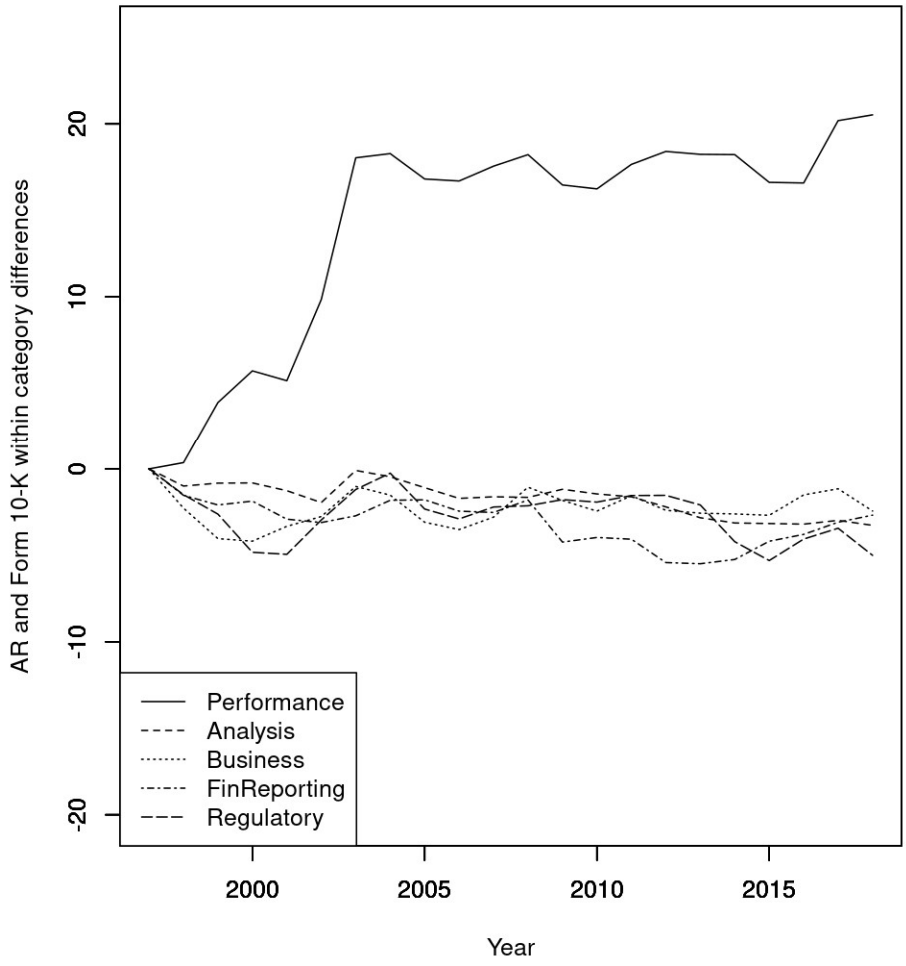
Figure 3 (continued)
Analyst Report and Form 10-K Discussion by Topic

Panel E: Regulatory Category



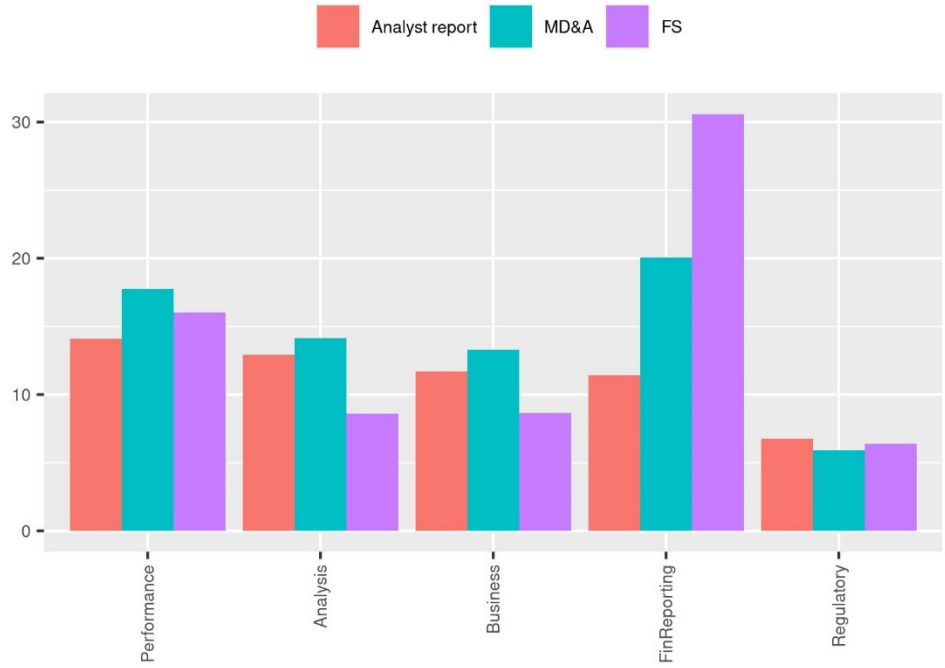
This figure presents the average discussion of each topic within each topic category in AR and Form 10-K, as a proportion of the total discussion in each document. Panels A through E present topics for the Performance, Analysis, Business, Financial Reporting, and Regulatory categories. See Table A.1, Panels A through E, for the most frequent words in each topic for each category. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Figure 4
Trend in the Difference in Analyst Report and Form 10-K Topic Discussion within Category



This figure presents the average absolute difference in the topics within each topic category in AR and Form 10-K. See Table A.1, Panels A through E, for the most frequent words in each topic for each category. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Figure 5
Analyst Report, MD&A, and Financial Statement Discussion by Topic Category



This figure presents the average discussion of each topic category in analyst reports (AR) and the Management's Discussion and Analysis (MD&A) and Financial Statement (FS) sections of Form 10-K, as a proportion of the total discussion in each document. See Table A.1, Panels A through E, for the discussion topics in each category and the most frequent words in each topic. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Table 1
Analyst Report and Form 10-K Discussion by Topic Category

	AR	10-K	Diff.	t-stat.
Performance	14.11	10.08	4.03***	18.24
Analysis	12.90	10.03	2.87***	9.59
Business	11.71	12.84	–1.13***	–9.31
FinReporting	11.41	22.20	–10.79***	–101.18
Regulatory	6.75	10.26	–3.51***	–24.34
Other	43.11	34.59	8.53***	23.30
All Categories	100.00	100.00	15.43	

This table presents the average discussion of each topic category in analyst reports (AR) and Form 10-K as a proportion of the total discussion in each document. The All Categories row in the Diff. column represents the sum of the absolute differences in all categories, divided by two. See Table A.1, Panels A through F, for the discussion topics in each category and the most frequent words in each topic. Tabulated amounts are in percentage points. *, **, and *** indicate significance at the 10, 5, and 1 percent levels, based on standard errors double clustered by firm and by year. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Table 2
Analyst Report and Form 10-K Discussion by Topic

Panel A: Performance Category

	AR	10-K	Diff.	t-stat.
Revenues/Margins	26.33	35.22	–8.89***	–17.03
Ratios	16.85	1.69	15.16***	59.37
Earnings	14.71	29.69	–14.99***	–52.52
Target prices	13.33	4.10	9.23***	19.02
Recommendations	12.32	0.90	11.41***	21.09
Adjusted earnings	9.58	5.85	3.73***	25.03
Expenses	4.58	18.17	–13.59***	–128.54
Cash flows	2.30	4.37	–2.07***	–46.35
All Performance Topics	100.00	100.00	39.53	

Panel B: Analysis Category

	AR	10-K	Diff.	t-stat.
Markets and industries	27.68	32.05	–4.37***	–21.58
Trends and forecasts	20.16	17.21	2.95***	13.36
Positives and negatives	12.65	6.48	6.18***	61.93
Estimates and assessments	11.02	11.64	–0.62***	–2.79
Increases and decreases	10.36	11.05	–0.69***	–5.09
Successes and challenges	7.46	6.40	1.06***	18.23
Views and opinions	6.90	7.48	–0.59***	–6.53
Economic environment	2.16	3.29	–1.13***	–21.95
Affect and effect	1.61	4.40	–2.79***	–54.20
All Analysis Topics	100.00	100.00	10.19	

Table 2 (continued)
Analyst Report and Form 10-K Discussion by Topic

Panel C: Business Category

	AR	10-K	Diff.	t-stat.
Products and markets	16.96	16.41	0.55	1.37
Business and operations	12.80	14.88	–2.08***	–13.09
Services and technology	12.28	11.65	0.63***	6.51
Banking	7.96	7.36	0.60**	2.51
Energy	5.51	3.32	2.19***	13.29
Goods and retail	5.51	4.06	1.46***	19.52
Insurance and healthcare	5.32	9.45	–4.12***	–27.92
Prices and production	4.86	4.90	–0.04	–0.43
States and regions	4.51	5.01	–0.50***	–4.80
Cities and international locations	4.42	1.32	3.11***	15.62
Manufacturing and materials	4.15	4.43	–0.28***	–4.17
Transportation and utilities	3.71	4.55	–0.84***	–13.17
Hotels and entertainment	3.62	3.09	0.53***	11.93
Drugs and clinical trials	3.02	3.04	–0.02	–0.35
Projects and pipelines	2.84	2.96	–0.12**	–2.49
Partnerships and ventures	2.51	3.57	–1.06***	–14.49
All Business Topics	100.00	100.00	9.07	

Table 2 (continued)
Analyst Report and Form 10-K Discussion by Topic

Panel D: Financial Reporting Category

	AR	10-K	Diff.	t-stat.
Investments	19.34	6.38	12.96***	54.36
Equity	13.26	10.55	2.71***	11.96
Financial statements	8.82	10.29	–1.47***	–9.94
Employees	8.72	9.66	–0.94***	–4.25
Interest rates	7.88	5.51	2.37***	15.95
Issuances	4.57	3.70	0.87***	11.64
M&A	4.51	3.58	0.94***	8.69
Credit and financing	4.24	7.65	–3.41***	–36.25
Compensation	3.87	7.96	–4.09***	–34.17
Regulations	3.64	4.78	–1.14***	–12.89
Loans	3.05	2.64	0.42***	6.42
Litigation	2.94	4.61	–1.67***	–26.40
PPE	2.81	4.49	–1.68***	–19.37
Debt and notes	2.77	2.08	0.68***	12.84
Liabilities	2.61	5.08	–2.47***	–31.56
Intangibles/impairment	2.58	5.09	–2.51***	–19.46
Fair value	2.52	2.52	–0.00	–0.02
Accounting standards	1.88	3.45	–1.57***	–7.13
All Financial Reporting Topics	100.00	100.00	20.95	

Table 2 (continued)
Analyst Report and Form 10-K Discussion by Topic

Panel E: Regulatory Category

	AR	10-K	Diff.	t-stat.
Risks	23.80	13.84	9.96***	18.16
Responsibilities	23.19	17.20	5.99***	20.01
Forms and exhibits	15.70	28.03	–12.33***	–37.69
Legal language	12.11	18.58	–6.47***	–14.33
Uncertainties	10.97	12.92	–1.95***	–3.40
Audit report	8.35	5.06	3.30***	28.41
Disclaimers	5.59	2.90	2.69***	11.38
Internal controls	0.29	1.47	–1.18***	–11.25
All Regulatory Topics	100.00	100.00	21.93	

This table presents the average level of discussion of topic in each category in analyst reports (AR) and Form 10-K, as a proportion of the total discussion within the category. See Table A.1, Panels A through E, for the discussion topics in each category and the most frequent words in each topic. The All Category Topics row in the Diff. column represents the sum of the absolute differences in each topic, divided by two. Tabulated amounts are in percentage points. *, **, and *** indicate significance at the 10, 5, and 1 percent levels, based on standard errors double clustered by firm and by year. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Table 3
Analyst Report, MD&A, and Financial Statement Discussion by Topic Category

				AR – MD&A		AR – FS		MD&A – FS	
	AR	MD&A	FS	Diff.	tstat.	Diff.	tstat.	Diff.	tstat.
Performance	14.11	17.76	16.03	–3.65***	–12.68	–1.92***	–12.60	1.73***	8.75
Analysis	12.90	14.16	8.61	–1.26***	–6.95	4.29***	14.06	5.54***	38.36
Business	11.71	13.28	8.65	–1.57***	–14.41	3.06***	20.42	4.63***	33.54
FinReporting	11.41	20.04	30.59	–8.63***	–45.51	–19.18***	–101.03	–10.55***	–35.52
Regulatory	6.75	5.93	6.37	0.83***	6.82	0.38***	3.58	–0.44***	–7.82
Other	43.11	28.83	29.74	14.29***	42.48	13.38***	73.62	–0.91***	–5.26
All Categories	100.00	100.00	100.00	15.11		21.11		11.90	

This table presents the average discussion of each topic category in analyst reports (AR) and the Management's Discussion and Analysis (MD&A) and Financial Statement (FS) sections of Forms 10-K as a proportion of the total discussion in each document. The All Categories row in the Diff. columns represent the sum of the absolute differences in all categories, divided by two. See Table A.1, Panels A through F, for the discussion topics in each category and the most frequent words in each topic. Tabulated amounts are in percentage points. *, **, and *** indicate significance at the 10, 5, and 1 percent levels, based on standard errors double clustered by firm and by year. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Table 4
Analyst Report, MD&A, and Financial Statement Discussion by Topic

Panel A: Performance Category

				AR – MD&A		AR – FS		MD&A – FS	
	AR	MD&A	FS	Diff.	tstat.	Diff.	tstat.	Diff.	tstat.
Revenues/Margins	26.33	48.35	25.92	–22.02***	–72.73	0.42	0.58	22.43***	39.81
Ratios	16.85	1.30	1.16	15.55***	58.57	15.69***	60.90	0.14***	4.33
Earnings	14.71	22.90	36.54	–8.19***	–25.21	–21.83***	–88.49	–13.64***	–76.31
Target prices	13.33	2.22	2.99	11.11***	23.34	10.34***	20.80	–0.77***	–19.79
Recommendations	12.32	0.60	0.37	11.71***	22.15	11.95***	22.18	0.23***	8.02
Adjusted earnings	9.58	2.97	8.48	6.61***	37.50	1.10***	3.44	–5.51***	–12.79
Expenses	4.58	17.62	19.71	–13.05***	–92.31	–15.13***	–130.03	–2.09***	–17.05
Cash flows	2.30	4.04	4.83	–1.74***	–31.53	–2.53***	–40.57	–0.79***	–10.73
All Performance Topics	100.00	100.00	100.00	44.99		39.50		22.80	

Panel B: Analysis Category

				AR – MD&A		AR – FS		MD&A – FS	
	AR	MD&A	FS	Diff.	tstat.	Diff.	tstat.	Diff.	tstat.
Markets and industries	27.68	24.60	29.04	3.08***	20.03	–1.36***	–3.94	–4.44***	–17.20
Trends and forecasts	20.16	16.64	21.95	3.52***	17.42	–1.79***	–6.00	–5.32***	–29.66
Positives and negatives	12.65	6.76	6.15	5.89***	51.71	6.51***	51.40	0.62***	5.97
Estimates and assessments	11.02	12.21	20.89	–1.19***	–3.29	–9.88***	–27.37	–8.69***	–32.94
Increases and decreases	10.36	24.13	6.04	–13.77***	–57.93	4.32***	30.62	18.09***	66.15
Successes and challenges	7.46	4.80	4.24	2.67***	41.49	3.22***	45.42	0.55***	12.58
Views and opinions	6.90	5.04	7.01	1.86***	26.24	–0.11	–0.92	–1.97***	–19.81
Economic environment	2.16	2.25	1.12	–0.09	–0.89	1.04***	19.40	1.13***	11.90
Affect and effect	1.61	3.57	3.56	–1.97***	–28.30	–1.95***	–12.76	0.02	0.17
All Analysis Topics	100.00	100.00	100.00	17.02		15.09		20.40	

Table 4 (continued)
Analyst Report, MD&A, and Financial Statement Discussion by Topic

Panel C: Business Category

				AR – MD&A		AR – FS		MD&A – FS	
	AR	MD&A	FS	Diff.	tstat.	Diff.	tstat.	Diff.	tstat.
Products and markets	16.96	17.45	16.90	–0.49**	–2.01	0.06	0.17	0.55***	2.71
Business and operations	12.80	16.62	12.35	–3.83***	–27.75	0.44***	2.80	4.27***	37.89
Services and technology	12.28	10.99	10.28	1.29***	9.13	2.01***	18.90	0.71***	5.09
Banking	7.96	5.15	8.00	2.81***	14.75	–0.04	–0.25	–2.85***	–41.00
Energy	5.51	3.94	4.29	1.58***	8.73	1.22***	5.87	–0.36***	–5.52
Goods and retail	5.51	5.66	3.86	–0.15	–1.50	1.65***	19.07	1.80***	21.75
Insurance and healthcare	5.32	8.47	15.25	–3.14***	–19.52	–9.93***	–37.45	–6.79***	–31.66
Prices and production	4.86	6.94	4.98	–2.08***	–17.39	–0.12	–0.97	1.97***	17.73
States and regions	4.51	2.80	4.36	1.71***	15.77	0.14	1.33	–1.57***	–19.22
Cities and international locations	4.42	0.76	0.91	3.66***	22.26	3.51***	21.30	–0.15***	–5.61
Manufacturing and materials	4.15	4.07	3.51	0.08	1.06	0.64***	7.96	0.56***	9.97
Transportation and utilities	3.71	4.49	4.07	–0.78***	–8.19	–0.36***	–3.36	0.42***	7.45
Hotels and entertainment	3.62	3.30	2.55	0.32***	4.22	1.07***	18.37	0.75***	11.72
Drugs and clinical trials	3.02	2.67	1.80	0.35***	4.56	1.22***	12.46	0.88***	11.01
Projects and pipelines	2.84	3.28	2.27	–0.43***	–7.47	0.57***	10.09	1.00***	20.90
Partnerships and ventures	2.51	3.41	4.61	–0.90***	–11.95	–2.09***	–22.03	–1.19***	–15.40
All Business Topics	100.00	100.00	100.00	11.80		12.54		12.90	

Table 4 (continued)
Analyst Report, MD&A, and Financial Statement Discussion by Topic

Panel D: Financial Reporting Category

				AR – MD&A		AR – FS		MD&A – FS	
	AR	MD&A	FS	Diff.	tstat.	Diff.	tstat.	Diff.	tstat.
Investments	19.34	7.18	6.73	12.16***	42.16	12.61***	48.70	0.45**	2.31
Equity	13.26	7.49	11.98	5.76***	22.89	1.28***	4.08	–4.48***	–22.37
Financial statements	8.82	11.33	13.59	–2.51***	–17.54	–4.77***	–26.99	–2.26***	–10.63
Employees	8.72	2.60	3.26	6.11***	67.84	5.46***	61.35	–0.66***	–12.09
Interest rates	7.88	7.71	5.87	0.17	0.93	2.01***	10.40	1.84***	34.09
Issuances	4.57	4.23	3.99	0.34***	3.61	0.57***	8.16	0.24***	4.98
M&A	4.51	5.05	3.01	–0.53***	–5.59	1.51***	13.82	2.04***	16.07
Credit and financing	4.24	11.07	6.14	–6.83***	–42.13	–1.90***	–14.44	4.93***	60.38
Compensation	3.87	4.12	9.27	–0.24*	–1.86	–5.39***	–32.41	–5.15***	–32.99
Regulations	3.64	3.36	1.58	0.27	1.44	2.06***	16.14	1.79***	17.74
Loans	3.05	3.42	2.01	–0.36***	–5.04	1.04***	12.37	1.41***	25.21
Litigation	2.94	3.24	2.75	–0.30***	–2.96	0.19***	2.81	0.49***	5.29
PPE	2.81	4.56	3.90	–1.75***	–23.88	–1.09***	–14.55	0.66***	11.52
Debt and notes	2.77	2.91	2.68	–0.15*	–1.91	0.09*	1.77	0.24***	4.66
Liabilities	2.61	6.00	7.25	–3.39***	–32.24	–4.64***	–48.48	–1.25***	–13.09
Intangibles/impairment	2.58	8.07	7.37	–5.50***	–25.75	–4.79***	–34.51	0.71***	5.99
Fair value	2.52	2.82	3.88	–0.31***	–3.22	–1.37***	–23.53	–1.06***	–19.25
Accounting standards	1.88	4.83	4.75	–2.95***	–7.79	–2.87***	–8.98	0.08	0.42
All Financial Reporting Topics	100.00	100.00	100.00	24.82		26.82		14.86	

Table 4 (continued)
Analyst Report, MD&A, and Financial Statement Discussion by Topic

Panel E: Regulatory Category

				AR – MD&A		AR – FS		MD&A – FS	
	AR	MD&A	FS	Diff.	tstat.	Diff.	tstat.	Diff.	tstat.
Risks	23.80	28.80	19.27	–5.00***	–9.75	4.53***	8.06	9.53***	60.02
Responsibilities	23.19	15.59	20.71	7.60***	22.36	2.48***	9.46	–5.12***	–33.86
Forms and exhibits	15.70	19.88	25.06	–4.19***	–15.27	–9.36***	–25.57	–5.17***	–10.82
Legal language	12.11	11.36	11.62	0.75***	3.25	0.49	1.45	–0.26	–1.32
Uncertainties	10.97	20.07	10.85	–9.09***	–22.21	0.12	0.38	9.21***	15.24
Audit report	8.35	2.99	7.65	5.37***	48.50	0.71***	4.44	–4.66***	–28.99
Disclaimers	5.59	1.28	2.37	4.31***	16.44	3.22***	14.19	–1.08***	–21.78
Internal controls	0.29	0.03	2.48	0.26***	10.71	–2.19***	–10.61	–2.45***	–11.27
All Regulatory Topics	100.00	100.00	100.00	18.28		11.55		18.75	

This table presents the average discussion of each topic of each category in analyst reports (AR) and the Management's Discussion and Analysis (MD&A) and Financial Statement (FS) sections of Forms 10-K as a proportion of the category discussion in each document. See Table A.1, Panels A through E, for the discussion topics in each category and the most frequent words in each topic. The All Category Topics row in the Diff. columns represent the sum of the absolute differences in each topic, divided by two. Tabulated amounts are in percentage points. *, **, and *** indicate significance at the 10, 5, and 1 percent levels, based on standard errors double clustered by firm and by year. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Table 5
Analyst Report and Form 10-K Discussion by Topic Category: Firm Type Differences

Panel A: Proportion of discussion for categories within AR

	Loss		Small		Tech		Financial	
	Coef.	t-stat.	Coef.	t-stat.	Coef.	t-stat.	Coef.	t-stat.
Performance	–1.09***	–8.05	1.74***	16.45	0.10	0.81	0.37***	2.76
Analysis	–0.93***	–10.96	0.46***	5.74	–0.06	–0.50	–0.57***	–5.76
Business	0.48***	4.27	–0.00	–0.01	0.26*	1.69	–1.93***	–16.69
FinReporting	0.03	0.41	–0.11	–1.57	–0.49***	–6.10	2.99***	22.27
Regulatory	0.48***	8.99	–0.64***	–9.30	–0.11	–1.60	–0.22***	–3.86
Other	1.03***	8.78	–1.45***	–9.74	0.29**	2.24	–0.63***	–4.34

Panel B: Proportion of discussion for categories within Form 10-K

	Loss		Small		Tech		Financial	
	Coef.	t-stat.	Coef.	t-stat.	Coef.	t-stat.	Coef.	t-stat.
Performance	–0.80***	–7.11	–0.64***	–9.19	–0.47***	–5.91	–0.12	–1.31
Analysis	–0.07	–1.05	0.01	0.13	0.25***	3.41	–0.15*	–1.84
Business	0.56***	4.06	–0.10	–1.30	0.51***	3.94	–2.80***	–20.83
FinReporting	–0.60***	–3.99	–0.28***	–3.54	–0.93***	–9.58	4.82***	29.51
Regulatory	0.32***	6.18	0.18***	4.80	–0.02	–0.30	–0.69***	–10.16
Other	0.59***	4.81	0.82***	7.22	0.65***	6.05	–1.07***	–8.16

Panel C: Differences in the proportion of category discussion between AR and Form 10-K

	Loss		Small		Tech		Financial	
	Coef.	t-stat.	Coef.	t-stat.	Coef.	t-stat.	Coef.	t-stat.
Performance	–0.21***	–2.89	1.44***	20.56	0.33***	3.56	0.15	1.28
Analysis	–0.54***	–6.36	0.20**	2.41	–0.18	–1.60	–0.39***	–4.64
Business	0.09	1.42	0.24***	4.09	0.33***	5.23	–0.75***	–11.20
FinReporting	–0.63***	–5.16	–0.16**	–2.19	–0.44***	–3.91	1.82***	10.16
Regulatory	–0.14***	–3.05	0.69***	10.08	0.06	0.88	–0.41***	–6.06
Other	0.35***	3.01	–1.89***	–11.75	–0.37**	–2.55	0.28*	1.78

This table presents results of regressing the level of discussion for each category in analyst reports (Panel A), in Form 10-K (Panel B), and the absolute difference in category discussion between AR and From 10-K (Panel C) on indicator variables for firm type. The firm types are loss firms, small firms, technology firms, and financial firms. Loss firms report negative earnings. Small firms have market value of equity below the sample median. Technology firms have a primary three-digit SIC code of 283, 357, 360-368, 481, 737, or 873. Financial firms are in the Fama-French 48 industries of banking (44), insurance (45), and trading (47). Table values represent coefficients and t-statistics. *, **, and *** indicate significance at the 10, 5, and 1 percent levels, based on standard errors double clustered by firm and by year. The sample comprises 26,757 firm-year observations for 4,335 firms from 1997 to 2018.

Michael Lee-Chin & Family Institute for Strategic Business Studies

Working Paper Series in Strategic Business Valuation

This working paper series presents original contributions focused on the theme of creation and measurement of value in business enterprises and organizations.

DeGroote School of Business at McMaster University
1280 Main Street West
Hamilton, Ontario, L8S 4M4

www.degroote.mcmaster.ca

DeGroote SCHOOL OF BUSINESS EDUCATION WITH PURPOSE | McMaster University