Measuring Firm Complexity

Tim Loughran
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.8432 *voice*
Loughran.9@nd.edu

Bill McDonald
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.5137 *voice*
mcdonald.1@nd.edu

August 29, 2019

Abstract

In business research, firm size is both ubiquitous and readily measured. Another firm-related construct, complexity, also is frequently relevant as a control variable, but difficult to measure and not well-defined. Typically, measures such as the number of firm segments or the readability of a firm's financial filings are used as proxies for some aspect of complexity. We argue that most extant measures of complexity are mismeasured or not widely available. We propose a text-based solution as an omnibus measure of this multidimensional concept and use audit fees, which primarily are driven by size and complexity, as the empirical framework for evaluation.

Key words: Firm complexity; audit fees; textual analysis; Form 10-K.

We thank Jeffrey Burks, Tony Cookson, Hermann Elendner, Margaret Forster, Jerry Langley, Jamie O'Brien, Marcelo Ortiz, Bill Schmuhl, and seminar participants at the Digital Innovation in Finance Conference, Humboldt University Summer Camp, Future of Financial Information Conference, and Baylor University for helpful comments.

1. Introduction

Joseph Blitzstein's mantra, in his popular statistics class at Harvard, is that "conditioning is the soul of statistics." In accounting and financial research, size is almost always used as a control variable to condition a regression examining some dependent variable of economic interest. In most applications, the theoretical basis for including size is neither explicit nor precise; it is self-evident that the economic magnitude of a firm is likely to affect most posited relations between various characteristics of a business. Lacking a specific theoretical basis, size is typically measured either as the market capitalization of a firm's publicly traded stock or as total assets, with both measures log-transformed due to their power-law like distributions.

Complexity, although to some extent correlated with size, measures an entirely distinctive and important aspect of a firm. Because a firm's complexity can be considered from many different perspectives and because it is difficult to measure, complexity is not a prominent variable in regression specifications. Complexity at the firm level can be viewed, for example, in the context of management hierarchy, product logistics, accounting ambiguities, financial engineering, or information dissemination.

After attempting to better define complexity in our application, we propose a text-based measure as an omnibus proxy of this broad construct. The measure addresses the multidimensional aspects of complexity and can be readily produced with current technologies for all firms filing Form 10-Ks with the Securities and Exchange Commission (SEC). Extending the methods of Loughran and McDonald (2011), we create a list of 255 words most likely to increase the complexity of firms and use counts of these words in annual reports as our complexity proxy.

Firm size and complexity are two first-order characteristics determining audit fees (Hay, Knechel, and Wong, 2006). Thus, we use audit fees as our arbiter of the success for the measure. Because "readability" is frequently used as a measure of selected dimensions of complexity, we spend some time discussing the limitations of this approach. We also consider other variables historically used as complexity measures in the context of predicting audit fees. We find that our text-based measure of complexity is statistically significant and dominates the plethora of variables that are sometimes used to explain audit fees.

2. Background

Attempts to measure the complexity of a publicly-traded firm are numerous in prior accounting and finance research. Typically, "firm complexity" can be thought of in terms of accounting, business, information, and reporting complications. As a proxy of firm-level complexity, previous papers have used items like the number of Compustat segments, whether or not the firm has foreign sales, the Fog Index, the number of words contained in the annual report, initialization of derivative usage, the number of methods listed in the revenue recognition disclosure, and the intangible asset percentage. Most of these measures are reasonable proxies for some aspects of firm complexity. Yet increasingly, many publicly-traded U.S. firms have global sales and engage in derivative usage, making these attributes less differentiating. A transparent, omnibus measure of firm-level complexity is missing from the literature.

Our paper takes a different approach in measuring firm-level complexity. We create a list of 255 words that proxy for the complexity of a firm's business model, management, and operations. Like the word lists created by Loughran and McDonald (2011), our complexity list is produced by examining actual word usage in U.S. annual reports (i.e., Form 10-K). Any word most likely implying business or information complexity is placed on the word list. From the 2000-2016 10-K sample, the most commonly appearing tokens on our complexity word list are: *subsidiaries, lease, acquisition, foreign, impairment, contracts,* and *subsidiary*. These words capture the complexity of the firm from the perspective of investors trying to estimate future cash flows or an auditor attempting to prepare financial statements. The average count of complexity words in an annual report during our sample period is 875.8, while the median count is 710.

A nice feature of our word list is that it combines many of the prior attempts to measure various aspects of firm complexity into a single multifaceted item. Thus, firm-level discussion in the annual report relating to M&A activity (*acquired, merger,* and *takeovers*), corporate events (*bankruptcies, partnership,* and *restructure*), legal issues (*sublease, subtenants, impairments,* and *royalties*), accounting terms (*accrete*, *carryforwards,* and *leaseback*), international operations (*foreign*, *global,* and *worldwide*), and derivatives (*derivatives, hedge,* and *unexercised*) are included in our word list.

Specifically, we gauge the complexity of a company by how often the firm uses language in its annual report from our new word list. Higher counts of words associated with complex events, transactions, and intricate business practices should be linked with larger levels of firm-level complexity. To establish the effectiveness of our word list in capturing firm-level complexity we have two main tests. Our primary test focuses on audit fees, because both firm size and complexity are considered dominant explanatory variables among an endless list of candidates that have been found to be significant in that context (see Simunic (1980) and Hay, Knechel, and Wong (2006)). Second, we examine the linkage between the number of Compustat segments, a variable frequently used to measure complexity, and firm-level complexity.

We use data from Audit Analytics, Compustat, and the SEC's EDGAR web site to analyze audit fees and the number of segments. The selected firm-level control variables are total assets, a dummy for a top 5 auditor, an S&P 500 dummy, and a dummy variable if the company has negative earnings. The average (median) audit fee is $1.87 million ($0.69 million) for our final audit sample containing 52,325 firm-year observations during the 2000-2016 time period.[1] Due to missing data, the sample size drops to 44,063 for the segment analysis. The average number of segments for our sample is 5.29 with a median value of 4.

Our specific prediction is that a higher count of complexity words in lagged annual reports should be positively associated with higher subsequent reported audit fees and a larger number of segments. Consistent with our expectation, we find that firms with more complex business, information, and reporting complexities, after controlling for client size and risk attributes, do indeed pay significantly higher audit fees and have a higher number of segments.

Within the audit fee literature, one weakness of some prior complexity measures is their non-public nature. For example, some papers have used the subjective rating of firm-level complexity provided by the actual or an experimental audit team.[2] This confidential information is obviously unavailable to the general public. Another common empirical measure of complexity is a firm's number of subsidiaries, which is not readily available for U.S. data (but is available for European firms). Similarly, the number of business segments is available for only some firms on Compustat. In all these cases, any sample using a traditional measure of complexity as a control variable will be constrained by data availability. One positive aspect of our methodology is its transparent nature; the measure of firm-level complexity is created from public annual reports and all of our complexity words are reported in this paper.

[1] Although Form 10-K documents are updated daily and available immediately on the EDGAR website—a feature we emphasize as a positive aspect of our measure—the analysis is limited to the years 2000-2016 due to the availability of Audit Analytics data.

[2] Using data for 249 U.S. audits conducted by a large accounting firm, O'Keefe, Simunic, and Stein (1994) find a strong linkage between audit fees and perceived complexity of the client firm. Prawitt (1995), in a survey experiment, uses environmental complexity manipulation to gauge how supervisors assign specific auditors in more challenging situations.

The prior literature has lacked an omnibus measure of the various dimensions of complexity associated with a firm. Using the language appearing in annual reports, we fill this gap in the literature by creating a list of 255 complexity words. This lexicon should capture the various aspects of business complexity that critically impact, for example, investors attempting to forecast future cash flows or auditors preparing the financial statements. Unlike some of the previous proxies for complexity, our measure is available for all U.S. firms filing on EDGAR.[3]

3. Literature Review

3.1. Measures of Firm-level Complexity

The literature measures the various aspects of firm complexity in quite a number of different ways. Here are just a few of the various measures: number of words in the Form 10-K; Fog Index; Hirfindahl-Hirschman indices measuring within firm industry and geographic concentration; the number of reported business segments (available on Compustat); if a firm reports foreign sales; the number of words from the revenue recognition disclosure in the Form 10-K; existence of a foreign currency translation; the number of methods listed in the revenue recognition disclosure in the Form 10-K; the number of special purpose entities reported; intangible assets as a percentage of total assets (i.e., accounting asset goodwill created from mergers); count of XBRL tags; and the initiation of derivative usage.

As a quick measure of firm-level informational complexity, numerous papers have used the word count in the annual report. Obviously, as managers provide more text describing their company's future or past operations, investors should have increased difficulty incorporating all the annual report disclosures into stock prices. For example, You and Zhang (2009) use the median

[3] We make available for all firms filing all 10-K/Q and their variants (e.g., 10-K405, 10-KSB, etc.) our tabulated measure for 1994-2018 at https://___.___.___.

10-K word count to categorize companies into low/high complexity groups. Bloomfield (2008) argues that firms facing adversity will have lengthier annual reports to explain their losses or other difficulties to investors. A number of papers have used the number of words in an annual report as a proxy for informational complexity (see Lehavy, Li, and Merkley (2011), Loughran and McDonald (2014), and Dyer, Lang, and Stice-Lawrence (2017)).

The number of Compustat business segments and a dummy variable equaling one if the firm has foreign sales have been used to identify complex firms (see Doyle, Ge, and McVay (2007), Ge and McVay (2005), and Ashbaugh-Skaife, Collins, and Lafond (2009)). Others have tabulated a revenue-based Hirfindahl-Hirschman index, calculated as the sum of the squares of each segment's sales as a percentage of the total firm revenue (see Bushman, Chen, Engel, and Smith (2004)). Hoitash and Hoitash (2018) report that a simple count of 10-K accounting items disclosed in eXtensible Business Reporting Language (XBRL) is a good proxy for a firm's accounting reporting complexity. Their XBRL data covers 10-K filings for only fiscal years 2011-2014. We discuss this variable in the robustness section of our empirical results.

The fractional percentage of intangible assets relative to total assets is also sometimes used as a measure of complexity (Gomes, Gorton, and Madureira (2007)) as is the initiation of derivative usage (Chang, Donohoe, and Sougiannis (2016)). The more text and recognition methods used in explaining to shareholders in the annual report how revenue is determined is linked to the chances a firm restates its reported revenue. Thus, Peterson (2012) uses the amount of text and a count of the recognition methods as a proxy for firm-level complexity.

3.2. Readability

Another firm specific variable of complexity used in the literature is the Fog Index. The Fog Index is a combination of two variables: average sentence length (in words) and complex words (fraction of words with more than two syllables). This readability measure estimates the number of years of formal education needed to comprehend the text in an initial reading. Since Li (2008) reports that the median Fog Index value for annual reports is 19.24, this implies that the reader needs slightly more than an MBA level of education to understand the document in a first reading. Although Jones and Shoemaker (1994) sharply criticize and Loughran and McDonald (2014) empirically discredit the use of the Fog Index, a number of accounting papers have continued to use it as a readability/complexity measure (see Lawrence (2013), Li and Zhang (2015), and Lo, Ramos, and Rogo (2017)).

Even if we ignore the empirical results of Loughran and McDonald (2014), where the dominant words driving readability scores are virtually all relatively common business words, the objective of the most frequently used measure—the Fog Index—is not at all clear.[4] Any reading of a sample of 10-Ks makes evident that writing style, in terms of vocabulary and density, is not something that varies much at all in the cross-section of firms. And, if it did, it would still not be clear what the objective was for readability, i.e., surely you would not want to minimize the score.

Leuz and Wysocki (2016) emphasize that it is impossible to disentangle the documents from the business, leading Loughran and McDonald (2016) to conclude that the broader topic of

[4] Word counts have a power-law distribution, much like market capitalization, where a small subset of words account for a major portion of the total counts. Table IV of Loughran and McDonald (2014) shows that 52 words from the thousands of complex words appearing in 10-Ks account for more than 25% of the complex word count in the Fog Index. All of the words are relatively common business terms, with the first five being *financial, company, interest, agreement,* and *including*.

complexity might be a more appropriate way of addressing the attribute intended to be captured by readability measures.

3.3. Audit Fees

Hay, Knechel, and Wong (2006) provide a survey of research on auditing and note that empirical research has clearly identified size and complexity as central components in determining audit fees. They consider 147 papers with 186 distinct independent variables. In their meta-analysis, size is shown to be the dominant factor in determining audit fees, typically accounting for around 70% of the variation in fees. In their itemization of independent variables used in explaining audit fees, Hays et al. (2006) specify size as the only pre-determined variable. Obviously, the higher are total assets, the more effort the auditor would likely expend to prepare the financial statements, thus the higher the audit fees should be. Another common measure of firm size is a dummy variable indicating membership in the S&P 500 Index (Chaney and Philipich (2002)). The empirical auditing literature clearly verifies that larger firms pay more in audit fees.

Second in their discussion of fee attributes is complexity. They identify 33 metrics in prior research used to proxy complexity. Typically complexity is measured by the number of subsidiaries or segments. They conclude that complexity is clearly relevant and the strongest results are for the number-of-subsidiaries proxy. For measuring risk they find that the most effective measure is a combination of inventory and receivables divided by total assets. For other independent variables such as profitability, leverage, and ownership form, the results are mixed.

Although early work suggests that top-tier auditors charge less in fees due to economies of scale (Simunic (1980)), the more recent evidence is that top 4, 5, 6, or 8 auditors are associated with significantly higher fees (Palmrose (1986) and Hogan and Wilkins (2008)). The reputation of auditors should have significant value that warrants increased compensation for their services (Balvers, McDonald, and Miller (1988)). Since auditors are potentially exposed to increased litigation risk if their client goes bankrupt, numerous papers have included a dummy variable for negative net income (Carcello et al. (2002), Whisenant, Sankaraguruswamy, and Raghunandan (2003), and Hogan and Wilkins (2008)). Hay, Knechel, and Wong (2006, page 171) note that "… the most recent results suggest that the existence of a loss for a client has become an increasingly important driver of audit fees." For perceived business risk, Bell, Landsman, and Shackelford (2001) used confidential survey data and found a positive linkage between client business risk and number of audit hours needed to prepare the financial statements.

Some of the prior evidence finds that financial institutions tend to pay less in audit fees than other industries. Part of the lower fees are driven by banks having limited receivables, inventory, and intellectual-based assets (Hay, Knechel, and Wong (2006)). However, the financial meltdown of 2008 dramatically exposed bank auditors to enormous client risk. Thus, regressions with audit fees as the dependent variable should incorporate industry dummies of the clients as controls.

4. Complexity and Its Measure

Many disciplines in both the natural sciences and social sciences consider complexity as an important attribute of systems they study. In some cases, such as computational complexity theory, the term is relatively precisely defined (see, for example, Goldreich (2010)), whereas in others, such as management (see, for example, Snowden and Boone (2007)), the term is more descriptive. To better delineate complex systems, the term is frequently juxtaposed with "complicated" systems. Although there is not a bright line separating complex from complicated systems, complicated systems are ones where, although having many layers, the layers themselves are capable of being understood to a degree of reasonable precision. Complex systems are more so characterized by unpredictability and nonlinear interactions, making them much more difficult to separate out for more comprehensive understanding.

A car is complicated, as it can be understood primarily as the sum of its components (e.g., engine, drive train, suspension, steering, etc.), whereas traffic, because it involves interactions dictated by the diversity of human behavior, is complex. The Latin derivatives of the two terms provide additional insight, with complicated coming from "complicare" which means "to fold together", while complex comes from "cum plectere" which means "to intertwine together." Unfolding a system to better understand its components is far easier than unbraiding.

Whether the perspective is management or analyst, a complicated system can be broken down into potentially predictable components and this makes the mapping of forward-looking strategies more straightforward. Alternatively, the more complex a system, the more difficult it is to disentangle its components, and because the interaction between the components can be chaotic, predicting outcomes is much more challenging. Because there are many dimensions of the firm that will impact its complexity and the nature of those interactions, prior measures of complexity have been relatively confined to specific aspects of the firm. We attempt to provide an omnibus measure of complexity by defining a set of words that most likely signal a layer of complexity in the firm and then counting the frequency of their occurrence in the company's Form 10-K filing, which describes the firm's operations, risks, governance, and finances.

When tabulating word counts, an often overlooked but critical input is how the terms will be weighted in the counts (see, for example, Manning and Schutze (2003)). Most frequently, papers in accounting and finance based on word counts use the proportion of words, that is, the word count normalized by the total number of words in the document as the term weighting method.[5] Importantly, in our measure of complexity we will use the raw counts of the words from our list. This embeds in the variable an indirect measure of document length, an attribute which has been used frequently to proxy for complexity. We discuss the specifics of the measure in the next section.

5. Data and Methodology

5.1. Merged EDGAR and Audit Analytics Data

As a first pass, we download all 10-K, 10-K405, 10KSB, 10-KSB, and 10KSB40 filings, excluding amended documents, from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website (www.sec.gov) and combine them with firm-level audit data from Audit Analytics. Although the EDGAR data is available and updated on a daily basis since 1994, the Audit Analytics data limits our sample to the years 2000-2016. Table 1 shows how the original sample of SEC filings and audit data is affected by our various data screens. The initial combination of the SEC 10-K and audit data is 66,814 firm-year observations.

As can be seen in the Table 1, the two data screens having the most impact on the sample are requiring a market value greater than zero as of the fiscal year end date (removing 6,590 observations) and requiring a value for net income (removing 6,900 observations). The audit fees are typically reported in a DEF 14A filing following the 10-K file date. The average number of calendar days for firms in our sample between the 10-K filing date to the disclosure of the audit fee information is 29.4 days. The median number of days is 28. Requiring the number of days between the 10-K filing date and the DEF 14A file date to be less than 180 days eliminates 164

[5] Loughran and McDonald (2011) provide an example of an alternative weighting, *tf-idf*, where words that are used infrequently have a bigger impact on the weighted counts.

observations. The final audit sample with available data is 52,325 firm-year observations during 2000-2016.

5.2. Complexity Word List

The complexity word list is created in a similar manner as the word lists created by Loughran and McDonald (2011). Only tokens appearing in the 2018 Loughran and McDonald Master Dictionary (https://sraf.nd.edu/textual-analysis/resources/) can potentially enter the sample. It should be noted that their Master Dictionary excludes proper nouns, single character letters, and acronyms.

Words the typical reader of the annual report would view as adding to the firm-level complexity for forecasting subsequent cash flows are included on our list. For example, significant annual report language describing *leases, intangible* assets, or *impairments* would make forecasting operating performance or the auditing of financial statements more challenging. To facilitate the ability of other researchers to use complexity words as a possible measure of firm-level complications and to be totally transparent, the entire list of 255 complexity words are reported in Table 2 and are available at https://___.___.___.

5.3. Most Frequent Appearing Complexity Words

In the textual analysis literature, it is always critical to show the reader which words have a disproportionate impact on the results. As argued by Loughran and McDonald (2016), transparency is critical to ensure the results from new word lists are not driven by misclassifications. For example, the commonly-used Harvard Dictionary misidentifies the tokens of *capital, depreciation, board, tax, liabilities,* and *vice* (as in *vice*-president) as capturing negative sentiment of a document. As noted by Loughran and McDonald (2011), these are common business terms which are unrelated to pessimistic tone in a business document. Also on the Harvard negative sentiment dictionary are tokens which clearly relate to specific industries, *crude* (oil industry) and *cancer* (pharmaceutical industry).

Although our complexity word list contains 255 different tokens, the top twenty most frequently occurring words account for almost 50% of the cumulative proportion. Table 3 reports the twenty most frequently occurring words on the complexity word list, the proportion of total, and the cumulative proportion. The token *subsidiaries* accounts for 4.93% of all the total complexity word counts. The next most frequent occurring complexity words are *lease, acquisition, foreign,* and *impairment*. All of the most commonly occurring words are obviously related to firm-level complexity. Conversely, of the 255 complexity tokens, the three words with the lowest frequency counts (not reported in Table 3) are *unrepatriated, conglomerate,* and *liquidates*. Misclassification of tokens does not appear to be an issue with our proposed word list.

5.4. Audit Fee Setting and Variable Definitions

As a first test of our complexity measure, we select the audit fee setting. As reported by Hay, Knechel, and Wong (2006), the audit fee literature is well-established. Clearly, the more complex the firm is, the more effort and time auditors need to expend preparing the firm's financial statements. More auditor effort will be directly related to higher auditor fees. The literature typically has the dependent variable as the natural log of total audit fees while the independent variables typically relate to firm size and other firm characteristics. The experimental variable (*Log(Complexity Count)* in our case) is added to regression after controlling for firm and auditor characteristics. Because of varying time and industry effects during the course of our sample period, calendar year and industry dummies will be controlled for.

Most of the firm-level characteristics are obtained from Audit Analytics. All of our control variables are known to investors *before* the disclosure of the audit fees. The well-established control variables from Audit Analytics include: *Total Assets* (as of the fiscal year end); *Top 5 Auditor Dummy* (a dummy variable set to one if the auditor is either PricewaterhouseCoopers, Ernst & Young, Deloitte & Touche, KPMG, or Arthur Andersen, else zero); *S&P 500 Dummy* (a dummy variable set to one if the firm is listed on the Standard & Poor's 500 Index, else zero); and *Loss Dummy* (a dummy variable set to one if the firm's net income is less than zero, else zero). From Compustat, we create the *Segments* variable (tabulation of all listed segments for each firm). Since some firms are missing segment information from Compustat, our sample size drops when *Segments* is used as a dependent variable.

5.5. Summary Statistics and Correlations

Panel A of Table 4 reports the summary statistics for our variables while Panel B of Table 4 presents the correlations. All variables have 52,325 firm-year observations except for *Segments* which has 44,063 unique observations. *Complexity Count* has a mean value of 875.8 compared to its median value of 710. Since the average number of words in the annual report is 50,752.6, the average fraction of complexity words is around 1.70%. The mean of *Audit Fees* is $1.87 million while the median value for *Total Assets* in our sample is $621 million. The dominance of the top-tier auditors is apparent since almost three quarters of the sample uses a top 5 auditor. The average number of *Segments* based on the Compustat data is 5.29.

In Panel B of Table 4, the correlations among the key variables is reported. *Complexity Count, Audit Fees, and Total Assets* are transformed by natural log. Consistent with our assertions, *Log(Complexity Count)* and *Log(Audit Fees)* are positively correlated (0.607).[6] More complexity words like *merger*, *leases, hedged*, and *global* appearing in the annual report is associated with higher subsequent auditor fees. *Log(Complexity Count)* is also positively linked with *Log(Total Assets)* and the *Top 5 Auditor Dummy*. Of all our reported correlations, the strongest relation (0.775) is between *Log(Audit Fees)* and *Log(Total Assets)*. This robust association is consistent with numerous prior papers.

To illustrate the time series pattern, Figure 1 reports the median *Complexity Count* and *Audit Fees* trend during our sample period. Both series move upward in tandem during 2000-2016. Since more complicated audits are charged higher fees, this evidence is consistent with the notion that our complexity word count captures the business and information complexity associated with firms. In the next section, we will see if this relation holds after controlling for firm characteristics.

[6] Although not reported in Panel B of Table 4, the correlation between *Log(Segments)* and *Log(Complexity Count)* is 0.265.

6. Empirical Findings

6.1. Log(Audit Fees) as the Dependent Variable

Can the text contained in an annual report capture firm-level complexity? As managers are forced to use more complex language to describe their firm's business and operating situation, are subsequent auditor fees higher? In Table 5, we estimate the regressions of *Log(Audit Fees)*:

$$
\begin{aligned}
\mathit{Log(Audit\ Fees)}_{i,t+1} = \alpha + \beta_1 \mathit{Log(Complexity\ Count)}_{i,t} + \beta_2\ \mathit{Log(Total\ Assets)}_{i,t} + \\
\beta_3 \mathit{Top\ 5\ Auditor\ Dummy}_{i,t} + \beta_4 \mathit{S\&P\ 500\ Dummy}_{i,t} + \qquad (1) \\
\beta_5 \mathit{Loss\ Dummy}_{i,t} + \varepsilon_{i,t}
\end{aligned}
$$

where *Log(Complexity Count)*$_{i,t}$ is the log of the count of words from our complexity word list. The dependent variable, *Log(Audit Fees)*$_{i,t+1}$, is the log dollar amount of audit fees reported after the Form 10-K filing date. As noted by the subscripts, our independent variables are all known at the time of the audit fee disclosure date.

There is a strong upward trend in audit fees during our time period. In 2000, the average audit fee was less than \$0.5 million while the average fees charged by auditors to prepare the financial statements increased to about \$2.6 million by 2016. The banking sector went from having typically cheaper average auditor fees to being the most expensive industry following the 2008 financial meltdown. Thus, regressions include an intercept, Fama and French (1997) 48-industry dummies, and calendar year dummies. The *t*-statistics are in parentheses with standard errors clustered by year and industry.

In the first column of Table 5, the only independent variable is *Log(Complexity Count)*. The coefficient on *Log(Complexity Count)* is positive (1.26) and statistically significant at the 1% level (*t*-statistic of 15.29). Higher complexity language usage in the annual report is associated with higher subsequent auditor fees. The R-squared value with only *Log(Complexity Count)* in the regression is 36.8%. The second column includes *Log(Complexity Count)* and industry and year dummies. In this regression, the R-squared value is raised to 52.5%. Thus, more than half of the variation in audit fees is explained by a simple count of complexity words, industry dummies, and year dummies.

In column (3), we only include the control variables in the regression. All of the controls have positive and statistically significant coefficient values. Not surprisingly, the variable with the largest *t*-statistic (34.21) is *Log(Total Assets)* while the coefficient with the lowest *t*-statistic is *S&P 500 Dummy* (4.60). Firms with higher total assets, a top 5 auditor, membership in the S&P 500 Index, or a negative net income are related to higher auditor fees. These results are consistent with the vast majority of the past audit fee literature (see Hay, Knechel, and Wong (2006)).

In the fourth column of Table 5, we include *Log(Complexity Count)* along with the control variables. In the presence of the control variables, the coefficient on *Log(Complexity Count)* remains significant (*t*-statistic of 10.45) with a value of 0.32. There is only a minor shift in the coefficient values for the control variables when included in the same regressions with the *Log(Complexity Count)* variable. The R-squared value for the last regression is 84.8%.

As a measure of firm size, some researchers have used *Log(Revenue)* or *Log(Market Value of Equity)* instead of *Log(Total Assets)*. As a robustness check, if *Log(Revenue)* replaces total assets in the last regression of Table 5, the coefficient on *Log(Complex)* is 0.43 with a *t*-statistic of 14.57. If *Log(Market Value of Equity)* is swapped in, the coefficient (0.52) on complexity count remains significant at the 1% level (*t*-statistic of 17.81).

6.2. Number of Segments as a Measure of Firm-Level Complexity

As a measure of complexity, some papers have used the number of Compustat segments (see Carcello et al. (2002), Hogan and Wilkins (2008), and Kim, Li, and Li (2015)). In our final table, *Log(Segments)* is the dependent variable. Companies with more segments should be considered more complex by outside investors attempting to forecast the firm's cash flows. As before, *Log(Total Assets), Top 5 Auditor Dummy, S&P 500 Dummy, Loss Dummy*, and industry/year dummies are included in each regression. The number of firm-year observations in the regressions drops to 44,063 due to missing segment data from Compustat.

In column (1) of Table 6, the coefficient on *Log(Complexity Count)* is positive (0.17) and statistically significant at the 1% level (*t*-statistic of 7.21). Higher counts of tokens from the complexity word list are associated with a larger number of Compustat segments. This finding reinforces our assertion that the complexity word list proxies for complexity of publicly-traded U.S. companies. Among the control variables, only *Log(Total Assets)* and *Loss Dummy* are significant. Firms with more assets and positive net income are associated with higher segment counts.

6.3. Robustness Tests

There are some extreme audit fee values in our sample. For example, American International Group (AIG) paid PricewaterhouseCoopers (PwC) an amazing $97,700,000 in audit fees in 2008 (see AIG's DEF 14A filed on 2008-04-04). Likewise, Bank of America paid PwC $96,600,000 in audit fees in 2012 (see their DEF 14A filing on 2012-03-28) and $95,600,000 to PwC in 2011 (see DEF 14A filing on 2011-03-30). Hand checking the extreme values of mostly banks, AIG, and General Electric, during or after the financial crisis, we were not able to identify any errors in the Audit Analytics data. The extreme values, although outliers, are correct.

Although the audit fee variable is log transformed to mitigate the impact of these skewed observations, it is useful to also consider the results if extreme audit fees, such as these, are filtered using a 1% level winsorization. If the winsorized value of *Log(Audit Fees)* is the dependent variable in the same regression as in column (4) of Table 5, the coefficient on *Log(Complexity Count)* drops slightly from 0.32 to 0.28, while the *t*-statistic increases from 10.45 to 10.82. Thus, our results are robust to the winsorization of the audit fee data.

Hoitash and Hoitash (2018) find that a simple count of 10-K accounting items disclosed in eXtensible Business Reporting Language (XBRL) is a good proxy for a firm's accounting reporting complexity. Is our *Complexity Count* variable still statistically significant after controlling for a count of XBRL tags in the 10-K filing? We obtained Hoitash and Hoitash's key variable, *Log(ARC)* (the natural log of the total number of distinct monetary XBRL tags in Item 8 of the 10-K filing) and find that it is positively correlated with *Log(Complexity Count)* (0.445).[7] Focusing on their sample for fiscal years 2011-2014, if ARC is added as a control variable in Table 5's column (4) regression, the coefficient on *Log(Complexity Count)*, even with the smaller matched sample size of 12,215 observations, remains positive (0.27) and statistically significant at the 1% level (*t*-statistic of 5.90) while the coefficient on *Log(ARC)* is positive and significant at only the 10% level (*t*-statistic of 1.85).

[7] The Hoitash and Hoitash (2018) XBRL data is obtained from http://www.xbrlresearch.com/.

7. Conclusion

Our paper provides the literature with an omnibus measure of firm-level complexity. The complexity word list is created by selecting language used by managers to describe their operations in the annual report. Some of the most commonly occurring words on the list are *lease, impairment, segment, collateral,* and *global*. The first setting selected to gauge the ability of the complexity word list to measure the firm-level difficulty for investors forecasting cash flows or for auditors preparing the financial statements is audit fees. Our secondary test uses the number of segments to analyze the strength of the complexity word list.

We find a strong association between the count of complexity language in the annual reports and subsequent audit fees. Increased discussion of *intangible* assets, *acquisitions*, *foreign* operations, or *subsidiaries* by managers is linked with significantly higher fees charged by the auditors. Our measure has the advantage of being completely transparent; the entire list of 255 complexity words is reported in the paper. When comparing one of the variables frequently used to measure complexity in the auditing literature, number of business segments, we find a strong linkage between the count of complexity words and the number of segments a company has.

Complexity is, and will likely remain, an amorphous yet important attribute of the firm. Similar to firm size, when examining firm-related economic phenomena, complexity is a characteristic that merits inclusion in a regression specification as a control variable. It is not unrelated to size, but it is a distinctly different aspect impacting the inputs and outputs of corporations. At the same time, complexity is multidimensional and not precisely prescribed by a specific economic theory. A firm's 10-K report characterizes virtually all aspects of the firm and provides a collection of terms that potentially captures the varied dimensions of complexity. We hope that our proposed omnibus measure provides a better way of assimilating the various dimensions of this important attribute.

References

Ashbaugh-Skaife, H., Collins, D.W., and Lafond, R. "The Effect of SOX Internal Control Deficiencies on Firm Risk and Cost of Equity." *Journal of Accounting Research* 47 (2009): 1-43.

Balvers, Ronald J., Bill McDonald, and Robert E. Miller. "Underpricing of new issues and the choice of auditor as a signal of investment banker reputation." *Accounting Review* 63 (1988): 605-622.

Bell, Timothy B., Wayne R. Landsman, and Douglas A. Shackelford. "Auditors' perceived business risk and audit fees: Analysis and evidence." *Journal of Accounting Research* 39 (2001): 35-43.

Bloomfield, Robert. "Discussion of "annual report readability, current earnings, and earnings persistence"." *Journal of Accounting and Economics* 45 (2008): 248-252.

Bushman, R., Chen, Q., Engel, E., and Smith. A. "Financial Accounting Information, Organizational Complexity and Corporate Governance Systems." *Journal of Accounting and Economics* 37 (2004): 167-201.

Carcello, Joseph V., Dana R. Hermanson, Terry L. Neal, and Richard A. Riley Jr. "Board characteristics and audit fees." *Contemporary Accounting Research* 19 (2002): 365-384.

Chaney, Paul K., and Kirk L. Philipich. "Shredded reputation: The cost of audit failure." *Journal of Accounting Research* 40 (2002): 1221-1245.

Chang, Hye Sun, Michael Donohoe, and Theodore Sougiannis. "Do analysts understand the economic and reporting complexities of derivatives?" *Journal of Accounting and Economics* 61, (2016): 584-604.

Doyle, J., Ge, W., and McVay, S. "Determinants of Weaknesses in Internal Control over Financial Reporting." *Journal of Accounting and Economics* 44 (2007): 193-223.

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence. "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation." *Journal of Accounting and Economics* 64, (2017): 221-245.

Fama, Eugene F., and Kenneth R. French. "Industry costs of equity." *Journal of Financial Economics* 43 (1997): 153-193.

Ge, W. and McVay, S. "The Disclosure of Material Weaknesses in Internal Control after the Sarbanes-Oxley Act." *Accounting Horizons* 19 (2005): 137-158.

Goldreich, Oded. P, *Np, and Np-Completeness: The Basics of Computational Complexity*, (2010), Cambridge University Press.

Gomes, A., Gorton, G., and Madureira, L. "SEC Regulation Fair Disclosure, Information, and the Cost of Capital." *Journal of Corporate Finance* 13 (2007): 300-334.

Hay, David C., W. Robert Knechel, and Norman Wong. "Audit fees: A meta-analysis of the effect of supply and demand attributes." *Contemporary Accounting Research* 23 (2006): 141-191.

Hogan, Chris E., and Michael S. Wilkins. "Evidence on the audit risk model: Do auditors increase audit fees in the presence of internal control deficiencies?" *Contemporary Accounting Research* 25 (2008): 219-242.

Hoitash, Rani, and Udi Hoitash. "Measuring accounting reporting complexity with XBRL." *Accounting Review* 93 (2018): 259-287.

Jones, Michael John, and Paul A. Shoemaker. "Accounting narratives: A review of empirical studies of content and readability." *Journal of Accounting Literature* 13 (1994): 142-184.

Kim, Yongtae, Haidan Li, and Siqi Li. "CEO equity incentives and audit fees." *Contemporary Accounting Research* 32 (2015): 608-638.

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

Lehavy, Reuven, Feng Li, and Kenneth Merkley. "The effect of annual report readability on analyst following and the properties of their earnings forecasts." *Accounting Review* 86 (2011): 1087-1115.

Leuz, Christian, and Peter Wysocki, "The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research." *Journal of Accounting Research* 54 (2016): 525-622.

Li, Feng. "Annual report readability, current earnings, and earnings persistence." *Journal of Accounting and Economics* 45 (2008): 221-247.

Li, Yinghua, and Liandong Zhang. "Short selling pressure, stock price behavior, and management forecast precision: Evidence from a natural experiment." *Journal of Accounting Research* 53 (2015): 79-117.

Lo, Kin, Felipe Ramos, and Rafael Rogo. "Earnings management and annual report readability." *Journal of Accounting and Economics* 63 (2017): 1-25.

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

Loughran, Tim and Bill McDonald. "Measuring Readability in Financial Disclosures." *Journal of Finance* 69 (2014): 1643-1671.

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

Manning, Christopher and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*: (2003), MIT Press, Cambridge, MA.

O'Keefe, Terrence B., Dan A. Simunic, and Michael T. Stein. "The production of audit services: Evidence from a major public accounting firm." *Journal of Accounting Research* 32 (1994): 241-261.

Palmrose, Zoe-Vonna. "Audit fees and auditor size: Further evidence." *Journal of Accounting Research* 24 (1986): 97-110.

Peterson, K. "Accounting Complexity, Misreporting, and the Consequences of Misreporting." *Review of Accounting Studies* 17 (2012): 72-95.

Prawitt, Douglas F. "Staffing assignments for judgment-oriented audit tasks: The effects of structured audit technology and environment." *Accounting Review* 70 (1995): 443-465.

Simunic, Dan A. "The pricing of audit services: Theory and evidence." *Journal of Accounting Research* 18 (1980): 161-190.

Snowden, David and Mary Boone, "A Leader's Framework for Decision Making." *Harvard Business Review* (2007): 68-77.

Whisenant, Scott, Srinivasan Sankaraguruswamy, and Kannan Raghunandan. "Evidence on the joint determination of audit and non-audit fees." *Journal of Accounting Research* 41 (2003): 721-744.

You, H. and Zhang, X.J. "Financial Reporting Complexity and Investor Underreaction to 10-K Information." *Review of Accounting Studies* 14 (2009): 559-586.

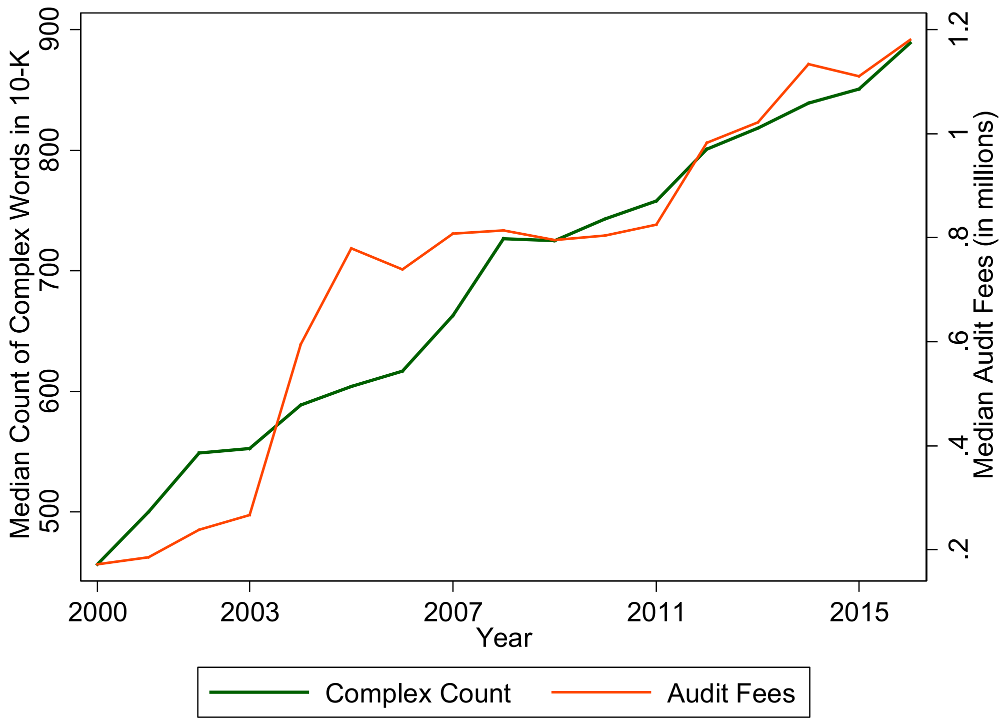

Figure 1. Time-series of the median *Complex Count* and *Audit Fees* during 2000-2016.

Table 1

Sample Creation

Table 1 reports the impact of various screens on the SEC EDGAR initial sample with available Audit Analytics data.

	Dropped	**Sample Size**
Combined Audit Analytics & 10-K SEC files 2000–2016		66,814
Drop if Market Value as of fiscal year end is missing	6,590	60,224
Drop if Total Audit Fees is missing	267	59,957
Drop if Total Assets is missing	557	59,400
Drop if Net Income is missing	6,900	52,500
Drop if number of Form 10-K words < 2,000	11	52,489
Drop if number of days is > 180 since 10-K/DEF 14A filings	164	52,325

Table 2
List of 255 Complexity Words

ACCRETE	CONVERSION	LEASING	RESTRUCTURE
ACCRETED	CONVERSIONS	LESSEE	RESTRUCTURED
ACCRETION	CONVERTIBILITY	LESSOR	RESTRUCTURING
ACCRETIVE	CONVERTIBLE	LICENSE	RESTRUCTURINGS
ACCRUAL	COUNTERPARTIES	LICENSEES	REVALUATION
ACCRUALS	COUNTERPARTY	LICENSES	REVALUE
ACCRUE	COVENANT	LICENSING	REVALUED
ACCRUED	COVENANTS	LIEN	REVOCABLE
ACCRUES	DEBENTURE	LIENS	REVOCATION
ACCRUING	DEBENTURES	LIQUIDATE	REVOCATIONS
ACQUIRE	DERIVATIVE	LIQUIDATED	REVOKE
ACQUIRED	DERIVATIVES	LIQUIDATES	REVOKED
ACQUIREE	EMBEDDED	LIQUIDATING	REVOKES
ACQUIRER	ENTITIES	LIQUIDATION	REVOKING
ACQUIRERS	EXERCISABLE	LIQUIDATIONS	ROYALTIES
ACQUIRES	EXERCISED	MERGE	ROYALTY
ACQUIRING	FLOATING	MERGED	SECURITIZATION
ACQUIROR	FOREIGN	MERGER	SECURITIZATIONS
ACQUISITION	FRANCHISE	MERGERS	SECURITIZE
ACQUISITIONS	FRANCHISED	MERGES	SECURITIZED
ADJUSTMENT	FRANCHISEE	MERGING	SEGMENT
ADJUSTMENTS	FRANCHISEES	NATIONALIZATION	SEGMENTS
AFFILIATE	FRANCHISES	NONMARKETABLE	SOVEREIGN
AFFILIATED	FRANCHISING	OUTSOURCE	SUBCONTRACT
AFFILIATES	FRANCHISOR	OUTSOURCED	SUBCONTRACTED
AFFILIATION	FUTURES	OUTSOURCES	SUBCONTRACTING
AFFILIATIONS	GLOBAL	OUTSOURCING	SUBCONTRACTOR
ALLIANCE	GLOBALIZATION	OVERCOLLATERALIZATION	SUBCONTRACTORS
ALLIANCES	GLOBALLY	PARTNER	SUBCONTRACTS
ASSUME	HEDGE	PARTNERED	SUBLEASE
ASSUMED	HEDGED	PARTNERING	SUBLEASED
ASSUMES	HEDGES	PARTNERS	SUBLEASES
ASSUMING	HEDGING	PARTNERSHIP	SUBLEASING
ASSUMPTION	IMBEDDED	PARTNERSHIPS	SUBLESSEE
ASSUMPTIONS	IMPAIR	PROMULGATE	SUBLESSEES
BANKRUPT	IMPAIRED	PROMULGATED	SUBLESSOR
BANKRUPTCIES	IMPAIRING	PROMULGATION	SUBLET
BANKRUPTCY	IMPAIRMENT	REACQUIRE	SUBLETTING
CARRYBACK	IMPAIRMENTS	REACQUIRED	SUBLICENSE
CARRYBACKS	IMPAIRS	REACQUISITION	SUBLICENSED
CARRYFORWARD	INFRINGE	RECAPITALIZATION	SUBLICENSEE
CARRYFORWARDS	INFRINGED	RECAPITALIZATIONS	SUBLICENSEES
COLLABORATE	INFRINGEMENT	RECAPITALIZE	SUBLICENSES
COLLABORATING	INFRINGEMENTS	RECAPITALIZED	SUBLICENSING
COLLABORATION	INFRINGES	RECLASSIFICATION	SUBSIDIARIES
COLLABORATIONS	INFRINGING	RECLASSIFICATIONS	SUBSIDIARY
COLLABORATIVE	INSOLVENCY	RECLASSIFIED	SUBSIDIZE
COLLABORATOR	INTANGIBLE	RECLASSIFY	SUBSIDIZED
COLLABORATORS	INTANGIBLES	RECLASSIFYING	SUBSIDY
COLLATERAL	INTERCONNECT	REISSUANCE	SUBTENANT
COLLATERALIZATION	INTERCONNECTED	REORGANIZATION	SUBTENANTS
COLLATERALIZE	INTERCONNECTING	REORGANIZATIONS	SWAP
COLLATERALIZED	INTERCONNECTION	REORGANIZE	SWAPS
COLLATERALIZING	INTERCONNECTIONS	REORGANIZED	TAKEOVER
COMPLEX	INTERCONNECTS	REORGANIZING	TAKEOVERS
COMPLEXITIES	INTERNATIONAL	REPATRIATE	UNEXERCISABLE
COMPLEXITY	INTERNATIONALLY	REPATRIATED	UNEXERCISED
CONGLOMERATE	LEASABLE	REPATRIATION	UNRECOGNIZED
CONTINGENCIES	LEASE	RESTATE	UNREPATRIATED
CONTINGENCY	LEASEBACK	RESTATED	VENTURES
CONTINGENT	LEASED	RESTATEMENT	WARRANT
CONTINGENTLY	LEASEHOLD	RESTATEMENTS	WARRANTS
CONTRACT	LEASEHOLDS	RESTATES	WORLDWIDE
CONTRACTS	LEASES	RESTATING	

Table 3
The Twenty Most Frequently Occurring Complexity Words Appearing in Annual Reports, 2000-2016

Complexity Word	Proportion of Total	Cumulative Proportion
Subsidiaries	4.93%	4.93%
Lease	4.09%	9.01%
Acquisition	3.89%	12.90%
Foreign	3.15%	16.05%
Impairment	3.14%	19.19%
Contracts	2.87%	22.06%
Subsidiary	2.78%	24.85%
Acquired	2.65%	27.50%
Segment	2.58%	30.08%
Contract	2.56%	32.63%
Collateral	2.03%	34.66%
Assumptions	1.99%	36.65%
Leases	1.71%	38.36%
Intangible	1.70%	40.06%
Acquisitions	1.63%	41.69%
International	1.63%	43.33%
License	1.59%	44.92%
Accrued	1.59%	46.50%
Partnership	1.48%	47.98%
Derivative	1.45%	49.44%

Table 4
Summary Statistics and Correlations

Complexity Count is the count of complexity words appearing in the annual report. *Audit Fees* are the auditor fees according to Audit Analytics. *Total Assets* are as of the end of the fiscal year. *Top 5 Auditor Dummy* is set to one if the auditor is among the top 5, else zero. *S&P Dummy* is set to one if the firm is on the S&P 500 Index, else zero. *Loss Dummy* is set to one if net income has a negative value, else zero. *Segments* is the number of reported "all segments" according to Compustat. All of the variables have an observation count of 52,325 except for *Segments* which has 44,063 observations. In Panel B, all the word count, audit fee, and total asset variables are in the log form. The sample period is 2000-2016.

Panel A: Summary Statistics

Variable Name	Mean	Median	Standard Deviation	10th Percentile	90th Percentile
Complexity Count	875.8	710	785	289	1,572
Audit Fees	$1.87 MM	$0.69 MM	$4.39 MM	$0.10 MM	$4.20 MM
Total Assets	$8,341 MM	$621 MM	$67,894 MM	$29 MM	$10,410 MM
Top 5 Auditor Dummy	0.73	1.00	0.45	0.00	1.00
S&P Dummy	0.12	0.00	0.32	0.00	1.00
Loss Dummy	0.32	0.00	0.47	0.00	1.00
Segments	5.29	4.00	3.90	1.00	10.00

Panel B: Correlations (N = 52,325)

	Complexity Count	*Audit Fees*	*Total Assets*	*Top Auditor*	*S&P 500*
Complexity	1.000				
Audit Fees	0.607	1.000			
Total Assets	0.529	0.775	1.000		
Top 5 Auditor	0.321	0.512	0.432	1.000	
S&P 500	0.234	0.457	0.473	0.216	1.000
Loss Dummy	-0.001	-0.209	-0.387	-0.105	-0.179

Table 5
Regressions with Log(Audit Fees) as the Dependent Variable

The dependent variable is *Log(Audit Fees)*. *Log(Complexity Count)* is the count of complexity words appearing in the annual report (i.e., Form 10-K). *Audit Fees* are the auditor fees according to Audit Analytics. *Total Assets* are as of the end of the fiscal year. *Top 5 Auditor Dummy* is set to one if the auditor is among the top 5, else zero. *S&P Dummy* is set to one if the firm is in the S&P 500 Index, else zero. *Loss Dummy* is set to one if net income has a negative value, else zero. All the regressions include an intercept, Fama and French (1997) 48-industry dummies, and calendar year dummies. The *t*-statistics are in parentheses with standard errors clustered by year and industry. The sample period is 2000-2016.

	Log(Audit Fees) (1)	Log(Audit Fees) (2)	Log(Audit Fees) (3)	Log(Audit Fees) (4)
Log(Complexity Count)	1.26	1.16		0.32
	(15.29)	(14.06)		(10.45)
Log(Total Assets)			0.46	0.40
			(34.21)	(32.33)
Top 5 Auditor Dummy			0.48	0.45
			(9.42)	(10.49)
S&P 500 Dummy			0.24	0.27
			(4.60)	(5.75)
Loss Dummy			0.20	0.11
			(8.31)	(6.32)
Intercept	Yes	Yes	Yes	Yes
Industry Dummies	No	Yes	Yes	Yes
Year Dummies	No	Yes	Yes	Yes
R-Squared	36.8%	52.5%	83.4%	84.8%
Sample Size	52,325	52,325	52,325	52,325

Table 6

Regression with Log(Segment) as the Dependent Variable

The dependent variable is *Log(Segment)*, the natural log of the number of segments for the firm according to Compustat. *Log(Complexity Count)* is the count of complexity words appearing in the annual report (i.e., Form 10-K). See Table 4 for the definitions of the other variables. The regression includes an intercept, Fama and French (1997) 48-industry dummies, and calendar year dummies. The *t*-statistics are in parentheses with standard errors clustered by year and industry. The sample period is 2000-2016.

	Log(Segment) (1)
Log(Complexity Count)	0.17
	(7.21)
Log(Total Assets)	0.11
	(10.95)
Top 5 Auditor Dummy	0.02
	(0.60)
S&P 500 Dummy	0.05
	(1.61)
Loss Dummy	-0.10
	(-2.67)
Intercept	Yes
Industry Dummies	Yes
Year Dummies	Yes
R-Squared	35.5%
Sample Size	44,063