Measuring Firm Complexity

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

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

December 8, 2022

ABSTRACT

In business research, firm size is both ubiquitous and readily measured. Complexity, another firm-related construct, is also relevant, but difficult to measure and not well defined. As a result, complexity is less frequently incorporated in empirical designs. Firm segment counts or the readability of a firm's financial filings are often used as proxies for some aspect of complexity. We argue that most extant measures of complexity are one-dimensional, have limited availability, and/or are frequently misspecified. Using both machine learning and an application specific lexicon, we develop a text solution that is based on widely available data, and that provides an omnibus measure of complexity. Three dependent variables are used that allow us to compare our measure with popular alternatives and to separate out the potential empirical overlap of size and complexity. Our proposed measure, used in tandem with 10-K file size, provides a useful proxy that dominates traditional measures.

JEL codes: D82; D83; G14; G18; G30; M40; M41
Key words: Firm complexity; textual analysis; Form 10-K; machine learning; lasso regression.

We thank Brad Badertscher, Jeffrey Burks, Tony Cookson, Nan Da, Hermann Elendner, Margaret Forster, Jerry Langley, Andrew Imdieke, Mikaela McDonald, Jamie O'Brien, Marcelo Ortiz, Jay Ritter, Bill Schmuhl, and seminar participants at the Digital Innovation in Finance Conference, Humboldt University Summer Camp, Future of Financial Information Conference, University of Notre Dame, University of Connecticut, Chinese University, Georgia State University, University of Colorado, Swiss Accounting Research Alpine Camp, International Research Symposium for Accounting Academics, Université Paris-Dauphine, and Baylor University for helpful comments.

1. Introduction

Joseph Blitzstein's mantra, in his popular statistics course at Harvard, emphasizes that "conditioning is the soul of statistics." In business research, company size is almost always used as a control variable to condition regressions examining some firm-related 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 company is likely to affect most posited relations between various company attributes. 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 falling within the penumbra of size, measures a distinct 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 usually not a prominent variable in regression specifications. At the firm level, complexity can be viewed in the context of organizational structure, product logistics, financial reporting, information dissemination, or financial engineering. Completely unbraiding firm size from complexity is impossible, but empirically it is helpful that the two constructs will in some cases be expected to have the same directional effect, while in others, their expected impact should diverge.

Although clearly an important attribute of a firm, complexity is a broad and amorphous concept that is difficult to quantify. That complexity is multifaceted suggests a one-dimensional quantitative measure might not capture the diverse firm characteristics embedded in its composition. For exactly this reason, we see this as an opportunity where textual analysis might uniquely add value in capturing the nuances of measuring complexity.

Historically, variables such as the number of firm segments, readability, diversity of XBRL tags, relative level of intangibles, presence of foreign sales, and firm age have been used when complexity is included as a conditioning variable in accounting and finance. We argue that all of these complexity proxies are limiting in at least one of three dimensions. First, many complexity proxies are limited in scope, focusing primarily on a single aspect of its measure. For example, XBRL diversity—as proposed by Hoitash and Hoitash (2018)—tends to isolate the accounting complexity of a firm. Second, many measures limit the sample size due to their availability in the various source datasets. Finally, we also argue that some of these alternatives are poorly measured.

In this paper, we will use 10-K filing word usage to create a measure of firm-level complexity.[1] Any word most likely implying business or information complexity is placed on the word list. Examples of the 374 complexity words on our list include *bankruptcies, counterparties, lawsuit, leases, swaps,* and *worldwide*.[2] 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 statement. Form 10-K filings have the advantage of being available for all firms with publicly traded securities. The 10-K filings are a credible source of firm-related text because they are an official record that, to the extent managers are not forthcoming or accurate in their revelations, can become the source of shareholder lawsuits, thus providing an incentive for management to be both honest and transparent.

In the prior textual analysis literature, researchers usually bifurcate on either using an indicative lexicon to identify targeted characteristics or using one of many machine learning techniques to identify topics or groupings of words that predict the characteristic of interest. We

[1] Throughout the paper we will use "10-K" to refer to 10-K, 10-K405, 10KSB, and 10KSB40 Securities and Exchange Commission (SEC) form types. We do not include amended filings.
[2] We will label our lexicon as "complexity words" in order to avoid confusion with the term "complex words" as used in the readability literature.

suggest a combination of both methods. We combine our relatively exhaustive list of 374 complexity words with a penalized regression method in an estimation sample to fit models using three dependent variables where the impact of complexity is well identified.

We first consider audit fees, where a long empirical literature on the topic clearly identifies firm size and complexity as two of the predominant variables explaining the dollar magnitude of audit fees. The empirical literature has clearly established that audit fees are positively impacted by complexity. The other two empirical frameworks are standardized unexpected earnings and stock return volatility. More complex firms should be associated with higher subsequent absolute earnings shocks and higher stock return volatility.

One of the difficulties in determining the efficacy of our proffered proxy is the overlap between the constructs of firm size and complexity. In the case of audit fees, we expect both firm size and complexity to have a positive impact, which could suggest, to the extent our measure is successful, that, due to multicollinearity, the measure is simply capturing size artifacts. In our favor, the correlation between our proposed measure and firm size is relatively low. More importantly, the expected impact of size and complexity should have opposite signs when focusing on absolute deviation of announced earnings from expected earnings and post-filing stock return variability. In these latter two cases, we expect size to have a negative effect, while complexity should have a positive effect.

In the first stage of the model estimation process, using a penalized regression method, we identify 53 words from the list of 374 potential candidates that are deemed most relevant in predicting the three dependent variables. All of the 53 words should add to the difficulty for auditors, analysts, and investors in projecting the future operations of the firm. We then compare the collective proportion of the selected words in the 10-K filing in competition with other complexity measures using a hold-out sample.

We find that our proposed measure performs well in all cases. As expected, we find that higher usage of our complexity words in a 10-K is associated with higher audit fees, the absolute value of unexpected earnings, and stock return volatility. Thus, our measure provides an omnibus proxy for complexity that is available for all publicly traded firms from 1996 to the current date.

One measure we include alongside our proposed complexity measure is the file size of the firm's 10-K document (i.e., annual report). File size was proposed in Loughran and McDonald (2014), where they show that the Fog Index is a poor measure of readability and then recommend gross file size as a reasonable proxy for the concept.[3] Gross file size includes pictures, spreadsheet files, and other non-text items that are converted from binary to text in order to comply with the filing guidelines. These insertions exponentially increase the size of the filing. Although Loughran and McDonald (2014) acknowledge this phenomenon, they use gross file size because it is highly correlated with net file size, where ASCII-encoded insertions, HTML, and XBRL have been removed.[4] Because cleaned 10-K files are now readily available on their website, typically net file size is used as the preferred measure. In a subsequent paper, Loughran and McDonald (2016) conclude that net file size, versus traditional measures of readability, likely goes beyond readability to capture some aspects of the "overall complexity of the firm" (p. 1198).

[3] They measure gross 10-K file size as the natural logarithm of file size in megabytes taken from the SEC's EDGAR (Electronic Data Gathering, Analysis, and Retrieval) "complete submission text file".

[4] ASCII (American Standard Code for Information Interchange) is one of the most common methods for encoding text data in computers. HTML (HyperText Markup Language) is the markup language used to display web pages. XBRL (eXtensible Business Reporting Language) is a markup language, required in SEC filings for the past decade, that facilitates computational parsing of business data.

Our study contributes to the literature primarily in two ways. First, we provide an example of combining the competing methods of lexicons and machine learning in textual analysis. Gentzkow, Kelly, and Taddy (2019) suggest that dictionary-based textual methods are most appropriate where there is prior information about the mapping of features to outcomes and where there is "no ground truth data on the actual level" of the construct being measured.

Often, when a machine learning technique is used to categorize words measuring a particular construct, tokens that are clearly inconsistent with the intended measure are identified. For example, in an early version of Ke, Kelly, and Xiu (2019), they identify "milk" and "banana" as positive and negative words, respectively, in measuring the sentiment of news articles.[5] Rudin (2019) and Stice-Lawrence (2022) emphasize that many machine learning approaches are essentially "black-box" methods, lacking economic interpretation and susceptible to "catastrophic" errors. One of the reasons we choose a penalized regression approach to identify the most appropriate subset of our initial word list is because of the relative transparency of the technique. In addition, by restricting the search space to a pre-selected collection of words, we avoid the potential errors associated with machine learning methods.

Our second contribution is creating a measure of complexity that is more all-encompassing, widely available, and straightforward to tabulate. As we will document, some aspect of complexity is frequently used in the literature as a control variable beyond the traditional and related measure of firm size. Unfortunately, the proxies for complexity are widely varied and many times limited in scope. Collectively and consistently in the results, we are able to show that our measure of complexity dominates alternative approaches and is not simply a redundant measure of size.

[5] Other examples include Lowry, Michaely, and Volkova (2020), Mai and Pukthuanthong (2021), and Akey, Grégoire, and Martineau (2022). A close examination of their word clouds reveals many tokens that are either clearly misidentified or not clearly linked to the underlying attribute. More concerning are studies that use machine learning to identify word categories but do not identify all of the words selected.

2. Background, Literature Review, and Prior Measures of Complexity

In this section, we will first attempt to better conceptualize complexity and then discuss some of the extant measures. Essentially our operational definition of complexity is any aspect of a firm that makes its valuation more difficult or ambiguous.

2.1. Complexity and its measure

Many disciplines in both the natural and social sciences consider complexity as an important attribute of systems they study. In some cases, such as computational complexity theory, the term's definition is relatively precise (see, for example, Goldreich (2010)), whereas in others, such as management (see, for example, Snowden and Boone (2007)), the definition 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, despite having many layers, the layers themselves are capable of being understood to a degree of reasonable precision.

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 an 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. We will not emphasize this distinction in the remainder of the paper, but along the spectrum from complicated to complex, we believe that, in the context of valuation, the system effects are more consistent with the notion of complexity and thus we will label the phenomenon as such.

2.2. Previous measures of firm-level complexity

2.2.1. 10-K File Size. As a simple proxy of firm-level informational complexity, numerous papers have used the file size or word count of annual reports. We will focus on net file size, since word count requires more parsing of the documents and is highly correlated with net file size (greater than 0.99 in our sample). Obviously, as managers provide more text describing their company's future or past operations, investors should have increased difficulty incorporating all of the annual report disclosures into stock prices. For example, You and Zhang (2009) use the median 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. Other papers using file size or total words as a proxy for informational complexity include Loughran and McDonald (2014), Bratten et al. (2017), Dyer, Lang, and Stice-Lawrence (2017), Ertugrul et al. (2017), Bao, Fung, and Su (2018), and Dou and Xu (2021). We will include the log of net file size as a control variable in all of our empirical models since, like our word-based measure, it is available for all firms filing a 10-K, it can be accurately and consistently measured, and it has repeatedly proven relevant in measuring some aspects of complexity.

2.2.2. Readability. Another firm-specific variable related to complexity and used frequently 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). The Fog Index estimates the number of years of formal education needed to comprehend a text in an initial reading. Li (2008) reports that the median Fog Index value for annual reports is 19.24, which implies that the reader needs slightly more than an MBA level of education to understand the document in a first reading.

Loughran and McDonald (2014) empirically discredit and question the fundamental premise of the Fog Index.[6] Word counts have a power-law distribution, much like market capitalization, where a small subset of words accounts for a major portion of the total counts. Table IV of Loughran and McDonald (2014) shows that 52 words, from the approximately 48,000 complex words appearing in 10-Ks, account for more than 25% of the total complex word count in the Fog Index. All of these 52 words are relatively common business terms, with the most frequently occurring being *financial*, *company*, *interest*, *agreement*, and *including*. Clearly such words will not challenge anyone reading a 10-K for investment purposes.

Even if we ignore the empirical results of Loughran and McDonald (2014), the objective of the Fog Index, and variants that have been proposed to this index, is not at all clear. 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. In fact, Loughran and McDonald (2014) show that increases in the use of financial jargon actually

[6] Jones and Shoemaker (1994) provide an early criticism of the Fog Index when used in evaluating business documents.

improve measures of valuation uncertainty. Attempts to use alternative readability measures such as Flesch-Kincaid or the Bog Index do not overcome this concern. Because of these criticisms, which question the measure at its most fundamental level, we do not consider the Fog Index as one of the alternative complexity measures in our empirical tests.

In spite of these limitations, a large number of papers have continued to use the Fog Index, or a variant of it, as a readability/complexity measure.[7] Clearly this is one aspect of complexity that would be useful to meaningfully measure. Leuz and Wysocki (2016) note that it is impossible to disentangle a firm's documents from its business, leading Loughran and McDonald (2016) to conclude that the broader topic of complexity might be a more appropriate way of addressing the attribute readability measures typically intend to capture.

2.2.3. Segments. Botosan, Huffman, and Stanford (2021) provide an excellent summary of the history and application of segment data both in practice and in research. One concern they express is the changing nature of segment reporting as regulatory regimes have evolved over time. The first requirement for segment reporting was SFAS 14 in 1977, which was changed with SFAS 131 in 1997. Although we believe that measurement consistency and availability is a significant problem for segment counts, given that the data available for our measure begins in 1996, we are not concerned about the regulatory differences.

More importantly for the use of segments are concerns with data availability, selection bias, and inconsistencies in reporting. Botosan et al. (2021) document in their table 3 that the percentage of publicly traded firms reporting at least one segment went from 85% in 1997 to 81% in 1999 and

---

[7] See Li and Zhang (2015), Guay, Samuels, and Taylor (2016), Bosnall and Miller (2017), Hwang and Kim (2017), Koo, Ramalingegowda, and Yu (2017), Lo, Ramos, and Rogo (2017), Beatty, Cheng, and Zhang (2019), Cassell, Cunningham, and Lisic (2019), Chen, Kim, Wei, and Zhang (2019), Glendening, Mauldin, and Shaw (2019), Kim, Wang, and Zhang (2019), and Gow, Larcker, and Zakolyukina (2021).

then dropped to 75% by 2017. Beyond this limitation, the missing data is concentrated in small and medium size firms, creating a bias in sample selection.

Across all firms, Botosan et al. (2021) note that only 50% reported more than one segment. In addition to the problem of some firms being less revealing in their segment disclosures, there is a lack of consistency in the disaggregation process that creates substantive measurement discrepancies. For example, in reporting geographical segments, some firms categorize segments based on region versus country or state (e.g., Asia versus the 48 countries in Asia, Europe versus the 44 countries in Europe, or Midwest versus the twelve states included in the midwestern U.S.).

In a 2018 report by the CFA Institute on segment reporting, they note that “segment reporting always makes the top of the list when it comes to comments by the U.S. Securities and Exchange Commission (SEC) calling out misapplication or questionable financial reporting practices” (p. 6).[8] Current U.S. accounting rules use a “management approach” to segment reporting that creates substantial discretion in how a company is partitioned for reporting purposes. The CFA report notes that professional investors are typically most concerned with over-aggregation by some firms. The SEC’s comments to companies most frequently concern the identification, aggregation, and changes in segments reported.

Because segment count is one of the more popular alternatives in accounting and finance for measuring complexity, below we provide some specific examples of measurement concerns:

- Amazon Web Services (AWS), which began operations in 2006 and was estimated to contribute approximately 52% of Amazon’s operating income in 2020, was not reported as a distinct segment until 2015. Marketwatch.com reported that the SEC attempted to get Amazon to disclose more information about AWS and Alexa products.[9]

[8] See https://www.cfainstitute.org/en/research/survey-reports/segment-disclosures-survey-report.

[9] See https://www.marketwatch.com/story/sec-tell-us-more-about-all-this-money-2018-04-19.

- Alphabet (Google), per Compustat, reports only four business segments and two geographic segments. Interestingly, Alphabet does not report YouTube as a separate segment.

- Manitex International, a manufacturer of lifting and loading products, started reporting geographic segments by country (versus region) and as a result goes from 17 geographic segments in fiscal year 2015 to 61 in fiscal year 2018.

- For DuPont, Compustat reports 4 and 22 geographic segments in fiscal years 2009 and 2010, respectively, even though DuPont's table reporting geographic information (from the corresponding 10-Ks) for the two periods is identical in terms of the countries identified.

- Compustat reports for General Motors 20 geographic segments in fiscal year 2013, which then declines and remains at 2 for fiscal years 2014-2021.

- Cummins goes from 17 geographic segments to 2 between fiscal years 2014 and 2015 as reported by Compustat. In 2014, their 10-K notes to the financial statements on segment information by geographic classification itemizes net sales by country. In 2015, they simply report "United States" and "International". At the same time, their "long-lived" assets are broken out into 17 countries in 2014 versus nine countries in 2015.

From these examples, it could be argued that the geographic segments, because of measurement inconsistencies, should be excluded from the counts. The business and operating segments, however, do not produce much variability within the firms. Of the 13,459 unique firms in our sample reporting segment data, more than half of the sample firms reported just one business or operating segment for all reporting periods, and more than 60% of the firms never changed the number of business or operating segments over all periods.

Thus, we argue that segment count, although popular in the literature, is a contaminated measure of complexity that can be significantly misspecified. In our subsequent results, we will see that segment count does not fare well across the various testing frameworks.

2.2.4. Other Measures of Complexity. We also consider other measures of complexity that are less dominant in the literature but appear with nontrivial frequency. We include firm age, a dummy variable for foreign sales, and the fractional percentage of intangible assets (i.e., goodwill, patents, and copyrights) relative to total assets, as variables that also have been used to identify complex firms (see Ge and McVay (2005), Doyle, Ge, and McVay (2007), Gomes, Gorton, and Madureira (2007), Cohen and Lou (2012), Harjoto, Laksmana, and Lee (2015), and Lee, Sun, Wang, and Zhang (2019)).

More recently Hoitash and Hoitash (2018) develop a measure of complexity that is a simple count of 10-K accounting items disclosed in the XBRL segments of a firm's 10-K. Although they label their measure as Accounting Reporting Complexity (*ARC*), their webpage (https://www.xbrlresearch.com) providing a repository for the data labels it as "a measure of firm complexity". Because of the SEC's implementation requirements for XBRL, their measure is broadly available beginning only in 2011.

This XBRL-based variable raises an important qualification for the measure we propose. Our measure is intended to broadly capture the construct of firm complexity. If a researcher is focusing on a specific aspect of complexity, for example in this case accounting complexity, then there is little question that domain specific measures, when available, would be more appropriate, or at least useful supplements to our proposed measure. We will see that although *ARC*, as would be expected, does well in the domain of audit fees, it is less successful in our other two frameworks for testing complexity. The complexity measure we develop attempts to improve on existing measures by providing a construct that is not sample limiting due to its availability and one that is multidimensional in its purview.

3. Empirical framework

3.1. Methods

Our proposed measure is based on the textual analysis of company 10-K filings. The textual analysis literature in accounting and finance is somewhat divided on the choice between machine learning methods versus dictionary-based methods for extracting useful information from text. We use a combination of both approaches.

In the first stage, we use a penalized regression technique to determine which words from a preselected list of promising candidates—described in the next section—show some validity in capturing the intended construct. Gentzkow, Kelly, and Taddy (2019) provide a useful summary on textual methods where they note that dictionary-based approaches are the most common method in the social science literature and are appropriate in cases where there is not "ground truth data" (p. 554). In the case of complexity, we do not have observations where the true state of complexity is actually measurable, which would provide a basis for a supervised learning model. They also note that penalized linear regressions are efficient for many prediction tasks in social sciences.

Among the penalized regression techniques, in our first stage, we specifically use lasso (least absolute shrinkage and selection operator) regressions to select from the initial list of candidate words those that are empirically consistent with the notion of complexity. Chinco, Clark-Joseph, and Ye (2019), in a paper predicting high-frequency short-term stock returns, provide an explanation of the technique and its advantages as a tool for reducing the dimensionality of a regression. Lasso regressions are similar to ridge regressions—both of them being penalized regression techniques—except that the penalty function for lasso is based on the sum of the absolute value of the coefficients versus the sum of the squared coefficients. By using the sum of the absolute values, the optimization will essentially force a variable's coefficient to zero if it is not deemed useful in minimizing the objective function.

The time series of data for the base 10-K sample is 1996-2021. Machine learning does not have an absolute rule about dividing a sample into model fitting and testing—typically the proportion of the training sample ranges from 50-70%, where models with larger numbers of parameters tend more to the higher values in the range. Within this range, we choose to split the sample into the 1996-2010 and 2011-2021 periods primarily due to *ARC* only becoming available in 2011 (*ARC* is not included in the model fitting regressions). We run the lasso regressions separately across all three of the dependent variables previously described. Because market capitalization and file size are available for the full sample and will be included as controls in all subsequent regressions, they are not subjected to elimination through the lasso objective. We want to see what value is added by the word-based measure beyond firm size and 10-K file size. The lasso objective will be used to select the most relevant words from the pre-selected list of potential complexity words.

In equation form, we have:

$$\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{2}\beta_j x_j - \sum_{k=1}^{L}\gamma_k z_k\right)^2 + \lambda\sum_{k=1}^{L}|\gamma_k|\right\}, \tag{1}$$

where β_0 is the regression intercept, x_j is a vector of length N containing the natural log of market capitalization for *j*=1 and the log of net file size for *j*=2, with β_j the corresponding regression coefficients. The proportion of the kth word appearing in a firm's 10-K filing from the initial lexicon of *L* words is represented by the vector z_k of length *N*, with its corresponding regression coefficient γ_k.[10] A hold-out sample is necessary to select the optimal weighting parameter, λ,

[10] In the actual estimation of this equation, we also include Fama-French (1997) 48 industry dummies and year dummies as non-penalized variables. For clarity, we have not included those in equation (1).

according to some model design criterion. Clearly from the equation, for a given λ, every non-zero coefficient on each word penalizes the minimization of the objective function. Note that when $\lambda=0$, the estimates converge to the ordinary least squares solution.

As with most machine learning methods, there are many variants for specifying and estimating the penalized regression. For example, we can include an additional penalty function that is the sum of the squared coefficients to essentially combine the lasso and ridge regression methods in what is labeled elastic net. Also, many different approaches can be used to select the appropriate weighting term, λ. To avoid overparameterization, our own selection bias, and in the interest of parsimony, we use the default Stata specification for estimating λ.[11] The second stage of the estimation process will take our complexity measure, detailed in the next section, and, using regressions, compare it with alternative measures of complexity using the three different dependent variables.

3.2. Our complexity measure

Prior measures of complexity have been confined to specific characteristics of the firm. We attempt to provide a more all-encompassing measure of complexity by initially identifying all words that we consider potentially linked to this attribute. Loughran and McDonald (2011) created their word lists by evaluating all tokens occurring in at least 5% of 10-K documents and selecting appropriate words for each of their sentiment lexicons. Following this approach, we create an

[11] See https://blog.stata.com/2019/09/09/an-introduction-to-the-lasso-in-stata/. Stata uses as its default K-fold cross-validation as its criterion, which is explained in their documentation.

initial list of candidate words by considering each word in a dictionary of approximately 86,000 words and assessing the likelihood that they might impact a firm's complexity.

For example, annual report language describing *leases, intangible* assets, *international* operations, or *acquisitions* would make forecasting operating performance or the auditing of financial statements more challenging. The list was then curated based on usage context samples in 10-K filings and by accounting professors and practitioners. This process produced our initial list of 374 candidate words. In the first stage of our estimation process, we will use a hold-out sample to determine which of these words are empirically consistent with the attributes of complexity.

The initial specification of the word list is intentionally generous, including all variants of root words that were deemed appropriate, since we will be statistically culling the list in the first stage. To avoid including rarely occurring words that can essentially become dummy variables for specific firms or industries, we require all words to appear in at least 5% of the 10-K documents. Examples of seldomly appearing words include *collateralizing, copyrightable,* and *reacquire*. The original list of 374 words is presented in Appendix A, with those words eliminated due to this criterion displayed using strikethrough.

To formulate our complexity measure, we only include those words selected in the lasso regressions whose estimated coefficients are positive across all three dependent variables in the model estimation sample. For the dependent variables, we would expect audit fees, the absolute value of unexpected earnings, and the standard deviation of stock returns to all be positively related to a firm's complexity. The final measure is then the sum of the proportional occurrence of each word identified from this process. Note that we do not use the specific regression parameters to weight this sum as we believe this would provide a false sense of precision.

3.3. Dependent variables tested

3.3.1. Audit Fees. Hay, Knechel, and Wong (2006) provide a comprehensive survey of auditing studies and note that empirical research has clearly identified size, complexity, and risk as central components in determining audit fees. They consider 147 papers with 186 distinct independent variables. In their meta-analysis, size is the dominant factor in determining audit fees, typically accounting for around 70% of the variation in fees. Obviously, larger firms require more billable hours of auditing. Another common measure of firm size is a dummy variable indicating membership in the S&P 500 Index (Chaney and Philipich (2002)). Not surprisingly, the empirical auditing literature verifies that larger firms pay more in audit fees.

Second, in their discussion of fee attributes, is complexity. Hay, Knechel, and Wong (2006) identify 33 metrics in prior research used to proxy complexity, with two of the most common being the number of segments or subsidiaries. They conclude that complexity is clearly relevant and the strongest results are for measures relating to how a firm is partitioned. Risk, as assayed in Hay, Knechel, and Wong (2006), focuses on the risk of error or specialized audit procedures, consistent with the models of Simunic (1980) and Stice (1991). The most common attributes used to measure this concept are relative levels of inventories and receivables, and they note that the combination of the two accounts seems to be more effective than considering them separately.[12]

Although early work suggests that top-tier auditors charge less in fees due to economies of scale (Simunic (1980)), more recent evidence finds that the 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

[12] Of the 129 analyses considered in Hay, Knechel, and Wong (2006), more than 71% use some combination of inventory and/or receivables as the proxy for risk.

(Balvers, McDonald, and Miller (1988)). Since auditors expose themselves to increased litigation risk if their client goes bankrupt, numerous papers have included a dummy variable for negative net income (e.g., Carcello, Hermanson, Neal, and Riley (2002) and Hogan and Wilkins (2008)). Hay, Knechel, and Wong (2006, p. 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."

Some of the prior evidence finds that financial institutions tend to pay less in audit fees than other industries. Part of this is 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 and substantially increased the average audit fee in this sector. Thus, regressions with audit fees as the dependent variable should incorporate both time and industry dummies as controls.

In sum, a large number of variables have been shown to be relevant in some context for predicting audit fees. For independent variables such as profitability, leverage, and ownership form, the results are mixed, with the significance of these candidates varying across samples and applications. Undoubtedly, at the margin, myriad variables affect the dollar amount auditing firms charge, but empirical studies to date identify size, complexity, and risk as the three dominant factors influencing audit fees.

3.3.2. Unexpected Earnings. Lehavy, Li, and Merkley (2011) relate 10-K readability to analyst following and various aspects of earnings forecasts. To the extent readability and complexity overlap—as they note in their discussion of readability—their hypothesis development for earnings forecast accuracy provides support for the positive relation between absolute earnings forecast errors and complexity that we test. Interestingly, they also emphasize that measures of document complexity do not address "overall complexity" and that this is a "particularly important" limitation. Their empirical results show, for various measures of analyst valuation imprecision, a positive relation with readability and a strong negative relation with size.

3.3.3. Post-filing Date Stock Return Volatility. Stock return volatility is frequently used to measure valuation uncertainty. Bloom (2014) provides a broad discussion of measuring uncertainty and uses stock return volatility as one of his primary proxies. In a widely cited study of investment dynamics, Bloom, Bond, and Reene (2007) use the standard deviation of daily stock returns over a one-year horizon as their measure of uncertainty "in an attempt to capture all relevant factors in one scalar measure" (p. 405).

Bond, Moessner, Mumtaz, and Syed (2005) show that the standard deviation of stock returns is correlated with analyst earnings forecasts and the dispersion of analyst forecasts, providing further justification for its use as a measure of valuation uncertainty. Chen, DeFond, and Park (2002) argue that stock return volatility "is consistent with greater uncertainty about future earnings" (p. 233). Kravet and Muslu (2013) look at the relation between company risk disclosures in their 10-K and stock return volatility, where they label return volatility as a measure of investor risk perception. Jiang, Lee, and Zhang (2005) consider the impact of information uncertainty on expected returns and use the standard deviation of daily stock returns as one of their measures of information uncertainty. They define information uncertainty as "value ambiguity" (p. 185).

Since prior research (see, Griffin (2003)) finds that the immediate impact of 10-K filings on stock returns is surprisingly modest, we will examine stock return volatility in the year after the filing date. The concept of complexity does not suggest any hypotheses concerning directional stock returns; however, our conceptualization of complexity defines it in terms of the ability to accurately value a firm. Consistent with prior applications of return volatility, we would expect the standard deviation of return to be higher for more complex firms.

Again, an important characteristic of both unexpected earnings and post-filing date stock return volatility is that we expect firm size to be negatively related to these variables, while we expect firm complexity to be positively related. Given that we expect some overlap between firm size and complexity, these two dependent variables allow us to parse out the differences.

4. Samples, data, and variables

In this section, we will define all of the variables used in the analysis and their data sources. A detailed derivation of the samples is available in the online data documentation and the variables are specifically defined in Appendix B. Because the availability of the variables varies substantially depending on the data source, we use the merged 10-K and CRSP data as the master data set and add where possible all of the other data sources to this base. We let the sample size vary with each regression depending on the data available for the variables included in each specification. The master data set, which has complete data for our complexity measure, firm size, and net file size consists of 120,994 firm/year observations for the period 1996-2021.

4.1. The three dependent variables

Audit fee data is taken from Audit Analytics, with data becoming available in fiscal year 2000. All of our variables will be measured through the end of 2021. We use the natural log of audit fees in the regressions and label the variable *Audit Fees*.

Unexpected earnings is calculated using the software available on Wharton Research Data Services (WRDS) authored by Denys Glushkov. We use method 3, which relies on IBES earnings estimates, to calculate the absolute value of the earnings forecast error, expressed as a percentage and winsorized at the 95th percentile. The variable is labeled *Abs(%Unexpected Earnings)* and it has data available for the filing years 1996-2021.

Return volatility is derived from CRSP data and is the standard deviation of the market-adjusted stock returns, expressed as a percentage, for a firm's stock over the 252-day interval following the 10-K filing date. The stock returns must be available for a minimum of 22 of the targeted 252 days for the observation to be included in the sample. The variable is labeled *StdDev Returns*.

4.2. Primary control variables

We include in all of the model estimation and holdout sample regressions our measure of complexity along with firm size and 10-K file size. Complexity, as previously detailed, is measured as the sum of the words selected in the first stage of the estimation process divided by the total number of words in the 10-K (expressed as a percentage). This variable is labeled *% Complexity* and is calculated using data from the SEC's EDGAR 10-K filings. We use the pre-parsed data available at https://sraf.nd.edu, which provides identifying information, net file size, SIC industry classifications, and word counts.[13] Each firm/year observation is identified by its CIK and, depending on the data being merged, the filing date or fiscal year. The earliest period relevant for all of our samples is dictated by the first year the SEC required periodic filings for all firms, which is 1996.

Firm size is measured in the regressions using the natural log of the market capitalization taken from CRSP. This data is available for the full 1996-2021 sample period. We use the

[13] The process of parsing the raw files down to a reasonable size is described in: https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/10x-stage-one-parsing-documentation/.

CRSP/Compustat link data to merge the CRSP data with the 10-K data. In some research, when examining audit fees, a firm's total assets is used as the proxy for firm size, but for consistency across the testing frameworks, we use market capitalization in all cases. The label for the firm size variable is *log(MktCap)*. Net file size represents the log transform of the net file size expressed in bytes and is labeled *log(NetFileSize)*. Since this variable is based on the 10-K filings, it is also available for the full 1996-2021 period.

4.3. Alternative measures of complexity

In addition to the primary control variables, we consider five additional measures of complexity that have been used as proxies for the concept. The variable we label *Segments* is taken from Compustat's segment data and is the total number reported for a given fiscal period corresponding to a firm's 10-K. Two other Compustat variables are *Foreign Income Dummy*, which is set to one if the pre-tax foreign income variable (PIFO) is not missing and non-zero, and *% Intangibles*, which is intangible assets divided by total assets. Intangible assets include items such as goodwill, patents, and copyrights. This variable is winsorized at the 95th percentile.

The variable labeled *Age* is the 10-K filing year minus the initial public offering year as reported by Compustat. When the latter item is not available, we use the year of the firm's initial listing on CRSP. *Segments*, *Foreign Income Dummy, % Intangibles,* and *Age* are all available for the filing years 1996-2021. We also consider the Hoitash and Hoitash (2018) *ARC* measure, which tabulates the number of unique XBRL tags in a firm's 10-K. A limitation of *ARC* is that it is only available beginning in 2011. We use the log transform of *ARC* in the regressions and label the variable *log(ARC)*.

4.4. Additional control variables

Five additional control variables are included in the full regression specifications for the first two dependent variables, where we have tried to select from broadly used firm characteristics. The first two variables we discuss are taken from Audit Analytics and the rest are from Compustat. The variables are: *Top-5 Auditor Dummy*, which is set equal to one if the auditor is either PricewaterhouseCoopers, Ernst & Young, Deloitte & Touche, KPMG, or Arthur Andersen; *S&P 500 Dummy*, which is set equal to one if the firm is in the S&P500 Index during that fiscal year; *Loss Dummy,* which is set equal to one if net income is negative; *% Leverage*, which is defined as (short-term debt + long-term debt)/total assets; and *% Inventory + Receivables*, defined as inventory plus receivables normalized by total assets. The latter two variables are winsorized at the 95th percentile.

5. Empirical results

5.1. Model estimation results

Using the data from filing years 1996-2010, we estimate equation (1) for each of the three dependent variables. From the initial word list of 374 potential complexity words, we first eliminate those that appear in fewer than 5% of the 10-Ks, leaving 198 candidate words. Using the lasso method will tend to push the less relevant word coefficients to zero. The more important constraint is taking the final coefficient estimates from the three lasso regressions (audit fees, unexpected earnings, and return volatility) and requiring a given word to have strictly positive coefficients across the three cases. After going through this filtering process, we are left with 53 words to be included in our final estimate of complexity from the original list of 374 (see table 1). For a given firm/year observation, we sum the counts for the 53 words and divide by the total number of words in the 10-K filing and then multiply by 100 to express as a percentage, producing the final estimate of *% Complexity*. We make this measure available for all 10-K Central Index Key (CIK) and year combinations from 1996-2021 at https://###.###.

As can be seen in table 1, because we have constrained the domain of the search process, all of the words (e.g., *derivative, global, litigation, repatriation,* and *ventures*) selected for the final measure appear to be reasonable proxies.[14]

5.2. Summary statistics

Summary statistics for all of the variables used in our analysis are presented in table 2 and are estimated over the full sample period. The median audit fee, adjusted for inflation, went from about $276,000 to $1,490,000 from the year 2000 to 2021. The top five firms paying the highest fees in 2020, with the exception of General Electric, were all financial firms.

Over the sample period, *Abs(%Unexpected Earnings)* seems most related to economic conditions, with a full period median of about 0.6% that increases during the great recession of 2008 and the COVID shock of 2020 to more than 1%. The two industries with both the highest median absolute earnings forecast error and highest median *StdDev Returns* were precious metals and pharmaceuticals. *StdDev Returns*, not surprisingly, also appears to move with economic cycles. As can be seen in table 2, the sample size varies widely depending on the specific variable. Note that we will only be using the dependent variables along with the three primary control variables in the first stage of the estimation process.

[14] Notice that the token *sovereign* makes the final cut. In business text, companies typically use *sovereign* to describe their exposure to European debt.

5.3. Sample results for % Complexity

The median value of *% Complexity* increases over the full sample period from 0.28 in filing year 1996, to a peak of 0.44 in 2014, and finishing at 0.42 in 2021. The five industries with the lowest average values of *% Complexity* are precious metals, non-metallic and industrial metal mining, insurance, banks, and candy and soda. Note that most of the firms in the Fama-French (1997) industry classifications categorized as banks are smaller state commercial banks and savings institutions. The five industries with the highest values of *% Complexity* are shipping containers, tobacco products, trading, chemicals, and real estate. The shipping containers category is dominated by Owens Illinois, a worldwide glass container manufacturer that has been in business since 1903.

Of the 20 firms with the highest average score over periods in which they appeared in the sample, 14 were in the broad area of finance. Smaller firms tended to have lower *% Complexity* scores, but if we consider only firms with a market capitalization greater than \$1 billion, the five firms with the lowest scores are Norfolk Southern—a railroad, Amerisafe—a provider of workers' compensation insurance for small and mid-sized firms, CoVel—a firm that applies artificial intelligence in health care, Casey's General Stores—a convenience store operating in 16 states, and AAON—an air conditioning and heating firm with two retail stores in Tulsa, Oklahoma.

Of some concern with these measures is whether there is a high degree of collinearity between the various proxies for complexity, and if there is a high correlation between our proposed measure and firm size. The correlation between *log(MktCap)* and *% Complexity* is less than 0.31, and between *log(NetFileSize)* and *% Complexity* it is less than 0.20. None of the complexity proxies considered, along with *log(MktCap),* has a correlation with the other complexity measures greater than 0.50 and the average correlation among these variables is 0.26.

5.4. Regression results

The regression results for the three dependent variables are presented in tables 3, 4, and 5. In addition to the coefficient estimates and *t*-statistics presented in the tables, each regression includes Fama-French (1997) 48-industry dummies and calendar year dummies. The standard errors used in calculating the *t*-statistics are clustered by year and CIK.

In the column (1) of each table (i.e., for each dependent variable), we first present the results for running the model from equation (1) directly on the estimation sample. Although this means that the inferential results are contaminated by the model fitting process, it provides a useful initial benchmark for comparison. The second column of each table runs the same regression as the first column to determine the out-of-sample effectiveness of the model derived from the first stage.

In addition to *% Complexity*, *log(MktCap)*, and *log(NetFileSize)*, we include all of the alternative measures of complexity in the last column of each table along with the additional control variables. The additional control variables are reasonable choices for both *log(AuditFees)* and *Abs(%Unexpected Earnings)*, but are not included in table 5 where the results for *StdDev Returns* are presented.

5.4.1. Audit Fees. Regression results for the dependent variable *log(Audit Fees)* are presented in table 3. The sample for the first stage model fitting process for *log(Audit Fees)* contained 46,318 observations for the filing-years 2000-2010.[15] Notably, the coefficient estimates remain significant

[15] Recall that the audit fee data begins in 2000, thus the benchmark year of 1996 is not used as the beginning date for this sample.

at similar levels in column (2) when we run the same model on the hold-out sample of filing years 2011-2021. In both cases, the coefficients for the three primary controls are positive and significant at the 0.01 level. As expected, larger and more complex firms have higher audit fees.

All of the five alternative complexity measures similarly are positive and significant, indicating that each of these variables seems to capture some unique aspect of complexity that impacts audit fees. Given that Hay, Knechel, and Wong (2006) identify 186 variables that have been empirically linked to audit fees, it is not surprising that all of the complexity measures do well in this context. For audit fees, the additional control variables indicate that top-5 auditors, firms with losses in the past fiscal year, and firms with relatively higher leverage, inventory, and receivables all generate higher auditing fees. The coefficient for *S&P 500 Dummy* was the only variable not significant in the regression. Interestingly, the r-square for the in-sample and out-of-sample models is essentially the same, which along with the consistency of coefficient estimates and standard errors, suggest that the fitted model does well out-of-sample.

5.4.2. Unexpected Earnings. Table 4 presents the second stage regression results for the dependent variable *Abs(UnexpectedEarnings)*. We expect complexity in this case to make valuation more challenging, thus increasing an analyst's absolute error in forecasting earnings. At the same time, we expect larger firms to, on average, have more stable and predictable earnings. For example, Lehavy, Li, and Merkley (2011) find size to be significant and negatively related to analyst dispersion and forecast accuracy.

The results are again presented in three columns, with the first column reporting the model with the three primary control variables based on the estimation sample and the second column running the same regression for the hold-out sample. The third column also considers the hold-out sample and adds both the alternative measures of complexity and additional controls to the regression. Interestingly, as we go from the base model including only the primary control variables in the estimation sample in column (1) to the same model in the hold-out sample of column (2), the r-square actually increases from 23.8% to 27.3%, again suggesting that the model is stable out-of-sample.

In all three of the columns of table 4, the estimated coefficients for the primary control variables align perfectly with expectation. Firm size, as measured by *log(MktCap)*, has estimated coefficients ranging from -1.072 to -0.814 with *t*-statistics all greater than -11.9 in magnitude. At the same time, across the three columns, *% Complexity* and *log(NetFileSize)* have positive coefficients in all cases with *t*-statistics greater than 8.3. All coefficient estimates for the primary control variables are significant at the 0.01 level.

Of most interest are the results for the alternative measures of complexity. In this case, *Segments* and *log(ARC)* have the correct sign but are not statistically significant. *Age* is the only alternative measure that is statistically significant and has the correct sign. Both *Foreign Income Dummy* and *% Intangibles* are significant but have the incorrect sign to the extent we consider them measures of complexity. In general, the alternative measures of complexity do not perform well as proxies of complexity in the context of valuation uncertainty that is measured by unexpected earnings.

Only three of the five additional controls are significant, with *Loss Dummy* and *% Leverage* having positive and significant coefficients. Given these variables are many times used as a proxies for risk, we would expect them to be positively related to *Abs(Unexpected Earnings)*. The variable *% Inventory + Receivables*, often used as a risk measure in the audit fee literature, in this case has a negative and significant estimated coefficient. While this variable would be expected to be positively related to audit fees—because even beyond their presumed relation to audit risk they require more billable hours to count and tabulate—in this case, higher levels of inventory and receivables could create a buffer in the sales to income calculation that reduces earnings forecast errors.

5.4.3. Post-filing Date Stock Return Volatility. Table 5 presents the regressions where *StdDev Returns* is the dependent variable using the same format as before. Once again, we expect firm size and complexity to have opposite signs. Larger firms, on average, have less volatile stock prices. However, because of valuation uncertainty, complex firms should have higher stock return volatility. Both *% Complexity* and *log(NetFileSize)* are positive and significant at the 0.01 level in all three specifications. At the same time, the coefficients across the three columns for *log(MktCap)* are negative and significant with *t*-statistics ranging from -14.15 to -18.25.

As noted before, we do not include the additional control variables for this dependent variable as they seem less relevant in this case. The alternative measures of complexity all fare poorly with negative estimated coefficients and two of them—*Segments* and *Age*—being negative and significant.

In sum, our measure of complexity paired with net file size seems to be empirically consistent with our priors about the three dependent variables. Given that we have no "ground truth" for measuring complexity, it is impossible to declare that the measures are unquestionably valid. In the case of file size, the link to complexity seems somewhat mechanical and thus the leap from this quantitative measure to the concept is not large. Because the vocabulary of *% Complexity* is constrained to words associated with firm complexity, we believe the logical linkage is also relatively clear for this variable.

5.5. Robustness

5.5.1. Private Firms. In table 6, we present alternative regressions to examine *% Complexity* in different contexts. Because 10-K and audit fee data are also reported for private firms with publicly traded debt, we can consider a restricted version of the second stage regressions. As before, the audit data only becomes available in 2000, but we consider the full 2000-2021 sample, since none of this data was used in the model derivation process. This also precludes including the alternative measures of complexity and control variables, but the year and industry fixed effects are still available. The results of this regression, with *% Complexity* and *log(NetFileSize)* as independent variables, are presented in column (1) of table 6. This selection process produces a sample size of 69,456 observations and once again both *% Complexity* and *log(NetFileSize)* are positive and significant at the 0.01 level. Private firms are often overlooked because of data availability, but these two measures of complexity are available for the large group of private firms that have publicly traded debt.

5.5.2. Excluding % Complexity. In the regressions with unexpected earnings and return volatility as the dependent variables, for both cases, the alternative measures of complexity performed poorly. An alternative interpretation would be that *% Complexity* was sufficiently correlated with these alternative measures so as to preclude their actual impact. In columns (2) and (3) of table 6, we reconsider both *Abs(UnexpectedEarnings)* and *SD_Ret* using the regressions specified in column (3) in both table 4 and 5, except *% Complexity* has been excluded. In both cases, these alternative complexity measures once again do not perform well. In the case of unexpected earnings, two of the coefficients are significantly negative, while *Segments* and *log(ARC)* are both positive and significant only at the 0.10 level. Only *Age* is significant at the 0.01 level and appearing with the correct sign. For *SD_Ret*, none of the alternative measures are significant, with the exception of *Age*, which in this case has the wrong sign. From these results, alternative measures of complexity that have arisen primarily in the context of audit fee research do not seem to perform well when used out of this original context.

5.5.3. Analyst Dispersion. In column (4) of table 6, we consider *Analyst Dispersion* as a variable that has frequently been used to measure valuation uncertainty (see Liu and Natarajan (2012) for a review of papers using analyst forecast dispersion) and a variable that was not used in deriving our complexity measure. Using the same estimation framework, the conclusions for analyst dispersion are very similar to those before. As expected, *log(NetFileSize)* and *% Complexity* are positively related to analyst dispersion while *log(MktCap)* is negatively related. All three of the coefficients are significant at the 0.01 level. In this regression, both *Age* and *log(ARC)* have the expected positive sign and are significant at the 0.01 and 0.05 levels, respectively. A concern of developing a model in the context of a specific framework is that it will not generalize to other applications. These results suggest, that at least in this case, the importance and impact of our complexity measure is sustained in a framework that differs from its initial development.

5.5.4. The Choice of Sample Partitioning. In order to identify the words ultimately included in our measure, we divided the sample based on the availability of data, i.e., *ARC*, one of the alternative complexity measures, only became available in 2011. Reasonable arguments could be made for making the dividing point at anywhere between 50-70% of the sample. If the collection of words selected from the 198 available (after eliminating those that occur infrequently) vary substantially depending on the split choice, we would be concerned about the stability of the measure. At the same time, we would not expect the list to be identical.

We reran the first stage process, this time splitting the sample in half (i.e., 1996-2008 and 2009-2021). This is an interesting split because it puts the final year at the peak of the Great Recession. The 2008 split produces 52 words versus 53 for the 2011 split. If we consider only the root form of the words, there are only 4 words appearing in the 2008 list that do not appear in the 2011 list—*acquirers*, *exercisable*, *futures*, and *interconnection*. Similarly, there are three words appearing in the 2011 list that do not appear in the 2008 list—*floating*, *reclassified*, and *segments*. Word usage and frequency will undoubtable change to some extent over time. With that considered, the degree of stability across these two sample choices is surprisingly high.

6. Conclusions

We use both machine learning and a lexicon to identify a list of words that attempt to capture the broad aspects of complexity. The initial complexity word list of 374 words is created by selecting words from management's description of their business, as detailed in a 10-K filing, that would typically be associated with greater complexity of a firm. Examples of our words are *carryforward, hedged, merging,* and *revaluation*. The data required for the measure is available at no cost for all firms with publicly traded debt or equity in the U.S. Although the file size of a firm's 10-K has been shown to perform well empirically and has the same availability, the measure by itself would not seem to capture all aspects of complexity. We propose using in tandem both file size and *% Complexity* when controlling for a firm's complexity.

The setting selected to gauge the proposed complexity measure relies on three economic variables where complexity should be relevant—audit fees, unexpected earnings, and return volatility. We find a strong association between the proportion of complexity language in the annual reports and the three dependent variables. Our complexity measure is consistently differentiated from firm size and five alternative complexity measures. The alternative complexity measures do not perform well once outside the realm of audit fees. Our results are robust to changes in the lasso regression sample specification and works well when evaluated using a variable (analyst dispersion) not used in the model derivation process.

Complexity is, and will likely remain, an amorphous yet important attribute of firms. Similar to firm size, when examining firm-related economic phenomena, complexity is a characteristic that frequently merits inclusion in a regression specification, typically as a control variable. It is related to size, but it is a distinctly different attribute affecting the inputs and outputs of corporations. At the same time, complexity is multidimensional and not precisely prescribed by a specific economic theory. Traditional quantitative measures of complexity are limited in the breadth of what they measure and in many cases the availability of data. A firm's 10-K report discusses in detail the business, operations, accounting, strategies, and other aspects of the firm, which, in turn, provides a collection of terms that potentially capture the varied dimensions of complexity. Perhaps measuring complexity provides a case where textual analysis can capture characteristics of a firm that are not well assayed by traditional quantitative measures. Any attempt to measure constructs such as this will be imperfect, but our proposed measure, along with file size, is widely available, multidimensional, and, importantly, appears to be empirically valid.

Appendix A. List of Potential Complexity Words*

~~ACCRUABLE~~
ACCRUAL
ACCRUALS
ACCRUE
ACCRUED
ACCRUES
ACCRUING
ACQUIRE
ACQUIRED
ACQUIREE
~~ACQUIREES~~
ACQUIRER
ACQUIRERS
ACQUIRES
ACQUIRING
~~ACQUIROR~~
~~ACQUIRORS~~
ACQUISITION
ACQUISITIONS
~~ACQUISITIVE~~
AFFILIATE
AFFILIATED
AFFILIATES
~~AFFILIATING~~
AFFILIATION
AFFILIATIONS
ALLIANCE
ALLIANCES
BANKRUPT
BANKRUPTCIES
BANKRUPTCY
~~BANKRUPTED~~
CARRYBACK
~~CARRYBACKS~~
CARRYFORWARD
~~CARRYFORWARDS~~
COLLABORATE
~~COLLABORATED~~
~~COLLABORATES~~
~~COLLABORATING~~
COLLABORATION
COLLABORATIONS
COLLABORATIVE
~~COLLABORATIVELY~~
~~COLLABORATOR~~
COLLABORATORS
COLLATERAL
COLLATERALIZATION
COLLATERALIZE
COLLATERALIZED
~~COLLATERALIZES~~
~~COLLATERALIZING~~
~~COLLATERALS~~
COMPLEX
COMPLEXITIES
COMPLEXITY
~~COMPLEXLY~~
~~CONGLOMERATE~~
~~CONGLOMERATES~~
CONTINGENCIES
CONTINGENCY
CONTINGENT
CONTINGENTLY
CONTRACT
CONTRACTED
~~CONTRACTHOLDER~~
~~CONTRACTHOLDERS~~
CONTRACTING
CONTRACTS
CONTRACTUAL
CONTRACTUALLY
~~CONTRACTUALS~~
~~CONTRACTURAL~~
CONVERSION
CONVERSIONS
~~CONVERTIBILITY~~
CONVERTIBLE
~~CONVERTIBLES~~
COPYRIGHT
~~COPYRIGHTABLE~~
~~COPYRIGHTED~~
~~COPYRIGHTING~~
COPYRIGHTS
COUNTERPARTIES
COUNTERPARTY
COVENANT
~~COVENANTED~~
~~COVENANTING~~
COVENANTS
DERIVATIVE
DERIVATIVES
EMBEDDED
ENTITIES
EXERCISABILITY
EXERCISABLE
~~EXERCISEABILITY~~
~~EXERCISEABLE~~
EXERCISED
FLOATING
FOREIGN
FRANCHISE
~~FRANCHISED~~
~~FRANCHISEE~~
~~FRANCHISEES~~
~~FRANCHISER~~
~~FRANCHISERS~~
FRANCHISES
~~FRANCHISING~~
~~FRANCHISOR~~
~~FRANCHISORS~~
FUTURES
GLOBAL
~~GLOBALIZATION~~
~~GLOBALIZE~~
~~GLOBALIZED~~
~~GLOBALIZING~~
GLOBALLY
HEDGE
HEDGED
HEDGES
HEDGING
~~IMBEDDED~~
INFRINGE
INFRINGED
INFRINGEMENT
~~INFRINGEMENTS~~
~~INFRINGER~~
~~INFRINGERS~~
INFRINGES
INFRINGING
~~INSOLVENCIES~~
INSOLVENCY
INSOLVENT
INTANGIBLE
INTANGIBLES
~~INTERCONNECT~~
~~INTERCONNECTED~~
~~INTERCONNECTEDNESS~~
~~INTERCONNECTING~~
INTERCONNECTION
~~INTERCONNECTIONS~~
~~INTERCONNECTS~~
INTERNATIONAL
~~INTERNATIONALIZATION~~
INTERNATIONALLY
LAWSUIT
LAWSUITS
~~LEASABLE~~
LEASE
~~LEASEABLE~~
LEASEBACK
~~LEASEBACKS~~
LEASED
LEASEHOLD
~~LEASEHOLDER~~
~~LEASEHOLDERS~~
~~LEASEHOLDS~~
~~LEASER~~
LEASES
LEASING
LESSEE
LESSEES
LESSOR
LESSORS
~~LICENCE~~
~~LICENCED~~
~~LICENCES~~
~~LICENCING~~
~~LICENSABLE~~
LICENSE
LICENSED
LICENSEE
LICENSEES
LICENSES
LICENSING
LICENSOR
LICENSORS
LIEN
~~LIENHOLDER~~
~~LIENHOLDERS~~
LIENS
LIQUIDATE
LIQUIDATED
~~LIQUIDATES~~
LIQUIDATING
LIQUIDATION
LIQUIDATIONS
LIQUIDATOR
~~LIQUIDATORS~~
~~LITIGATE~~
~~LITIGATED~~
~~LITIGATES~~
~~LITIGATING~~
LITIGATION
~~LITIGATIONS~~
~~LITIGIOUS~~
MERGE
MERGED
MERGER
MERGERS
~~MERGES~~
MERGING
NATIONALIZATION
~~NATIONALIZATIONS~~
~~NATIONALIZE~~
~~NATIONALIZED~~
~~NATIONALIZING~~
~~NONMARKETABLE~~
OUTSOURCE
OUTSOURCED
~~OUTSOURCER~~
~~OUTSOURCERS~~
~~OUTSOURCES~~
OUTSOURCING
PARTNER
~~PARTNERED~~
PARTNERING
PARTNERS
PARTNERSHIP
PARTNERSHIPS
PATENT
~~PATENTABILITY~~
PATENTABLE
PATENTED
~~PATENTEE~~
~~PATENTING~~
PATENTS
~~REACQUIRE~~
REACQUIRED
~~REACQUIRES~~
~~REACQUIRING~~
~~REACQUISITION~~
~~REACQUISITIONS~~
RECAPITALIZATION
RECAPITALIZATIONS
~~RECAPITALIZE~~
~~RECAPITALIZED~~
~~RECAPITALIZES~~
~~RECAPITALIZING~~
RECLASSIFICATION
RECLASSIFICATIONS
RECLASSIFIED
~~RECLASSIFIES~~
RECLASSIFY
~~RECLASSIFYING~~
~~REISSUANCE~~
~~REISSUANCES~~
~~REISSUE~~
REISSUED
~~REISSUES~~
~~REISSUING~~
~~REORGANISATION~~
REORGANIZATION
~~REORGANIZATIONAL~~
REORGANIZATIONS
~~REORGANIZE~~
REORGANIZED
~~REORGANIZES~~
~~REORGANIZING~~
REPATRIATE
REPATRIATED
~~REPATRIATES~~
~~REPATRIATING~~
REPATRIATION
~~REPATRIATIONS~~
RESTRUCTURE
RESTRUCTURED
~~RESTRUCTURES~~
RESTRUCTURING
RESTRUCTURINGS
REVALUATION
~~REVALUATIONS~~
~~REVALUE~~
REVALUED
~~REVALUES~~
~~REVALUING~~
~~REVOCABILITY~~
REVOCABLE
REVOCATION
~~REVOCATIONS~~
REVOKE
REVOKED
~~REVOKES~~
~~REVOKING~~
ROYALTIES
ROYALTY
~~SECURITIZABLE~~
SECURITIZATION
SECURITIZATIONS
~~SECURITIZE~~
SECURITIZED
~~SECURITIZER~~
~~SECURITIZERS~~
~~SECURITIZES~~
~~SECURITIZING~~
SEGMENT
~~SEGMENTAL~~
~~SEGMENTATION~~
~~SEGMENTATIONS~~
SEGMENTED
~~SEGMENTING~~
SEGMENTS
SOVEREIGN
~~SOVEREIGNS~~
~~SOVEREIGNTIES~~
~~SOVEREIGNTY~~
SUBCONTRACT
~~SUBCONTRACTED~~
~~SUBCONTRACTING~~
SUBCONTRACTOR
SUBCONTRACTORS
SUBCONTRACTS
SUBLEASE
SUBLEASED
~~SUBLEASEE~~
~~SUBLEASEHOLD~~
SUBLEASES
~~SUBLEASING~~
~~SUBLESSEE~~
~~SUBLESSEES~~
~~SUBLESSOR~~
~~SUBLESSORS~~
SUBLET
~~SUBLETS~~
~~SUBLETTING~~
~~SUBLETTINGS~~
~~SUBLICENSABLE~~
SUBLICENSE
~~SUBLICENSEABLE~~
~~SUBLICENSED~~
~~SUBLICENSEE~~
~~SUBLICENSEES~~
~~SUBLICENSES~~
~~SUBLICENSING~~
~~SUBLICENSOR~~
SUBSIDIARIES
SUBSIDIARY
SUBSIDIES
~~SUBSIDING~~
~~SUBSIDIZATION~~
~~SUBSIDIZE~~
~~SUBSIDIZED~~
~~SUBSIDIZERS~~
~~SUBSIDIZES~~
~~SUBSIDIZING~~
SUBSIDY
~~SUBTENANCIES~~
~~SUBTENANCY~~
~~SUBTENANT~~
~~SUBTENANTS~~
SWAP
SWAPS
~~SWAPTION~~
~~SWAPTIONS~~
TAKEOVER
TAKEOVERS
TRADEMARK
~~TRADEMARKED~~
~~TRADEMARKING~~
TRADEMARKS
UNEXERCISABLE
UNEXERCISED
UNRECOGNIZED
UNREMITTED
~~UNREPATRIATED~~
VENTURE
VENTURES
~~WARRANTEES~~
~~WARRANTIED~~
WARRANTIES
~~WARRANTING~~
~~WARRANTOR~~
WARRANTY
WORLDWIDE

* Words rendered with strikethrough appear in less than 5% of the filings and are not included in the model selection process.

Appendix B. Definitions of Variables

Panel A: Dependent Variables

log(Audit Fees)	The natural log of the dollar amount of audit fees disclosed after the Form 10-K filing date as reported by Audit Analytics.
Abs(Unexpected Earnings)	The absolute value of (Actual EPS minus median IBES EPS estimate) scaled by stock price.
StdDev Returns	The standard deviation for market-adjusted stock returns, expressed as a percentage, for one year of trading days following the 10-K filing date. A minimum of 22 trading day observations must be available for the calculation.
Analyst Dispersion	Following Lehavy, Li, and Merkley (2011), analyst dispersion is defined as the standard deviation of the individual analysts' forecasts in the first consensus annual earnings forecast issued after the 10-K filing date for the fiscal period following the 10-K filing, scaled by the filing-date share price. There must be at least two analysts in the forecasts to be included in the sample.

Panel B: Alternative Measures of Complexity

% Complexity	The count of words listed in table 1 that were retained based on the model selection process, divided by the total number of words appearing in the Form 10-K filing, times 100.
log(MktCap)	The market capitalization measured by CRSP price times shares outstanding on the trading day before the 10-K filing date.
log(Net file size)	The natural log of the net 10-K file size in bytes. Net file size reflects the removal of binary-encoded ASCII (e.g., pictures), HTML, XBRL, etc. The process for creating the pre-parsed 10-K files is described at: https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/10x-stage-one-parsing-documentation/.
Segments	The sum of Compustat business, geographic, operations, and state segments.
Foreign Income Dummy	Dummy variable set to one if the pre-tax foreign income variable (PIFO) is available (e.g., non-missing or non-zero), else zero. This variable is from Compustat.

% Intangibles	Intangible assets divided by total assets. Intangibles include items such as goodwill, patents, trademarks, and copyrights. This variable is winsorized at the 95th percentile and is from Compustat.
log(ARC)	The number of distinct monetary XBRL tags in Item 8 (Financial Statements and Supplementary Data) of a firm's SEC filing. *ARC* is documented in Hoitash and Hoitash (2018) and downloaded from their website (https://www.xbrlresearch.com).
Age	The 10-K filing year minus the year the initial public offering year as reported by Compustat. When the latter item is not available, we use the year of the firm's initial listing on CRSP.

Panel C: Additional Control Variables

Top-5 Auditor Dummy	Dummy variable set to one if the auditor is either PricewaterhouseCoopers, Ernst & Young, Deloitte & Touche, KPMG, or Arthur Andersen, else zero. This variable is from Audit Analytics.
S&P 500 Dummy	Dummy variable set to one if the firm is in the S&P 500 Index, else zero. This variable is from Audit Analytics.
Loss Dummy	Dummy variable set to one if net income as reported by Compustat has a negative value, else zero.
% Leverage	Defined as (short-term debt + long-term debt)/total assets. This variable is winsorized at the 95th percentile and is from Compustat.
% Inventory. + Receivables	Defined as (inventory + receivables)/total assets. This variable is winsorized at the 95th percentile and is from Compustat.

References

Akey, P., V. Grégoire, and C. Martineau. 2022. Price Revelation from Insider Trading: Evidence from Hacked Earnings News. Journal of Financial Economics 143, 1162-1184.

Balvers, R. J., B. McDonald, and R. E. Miller. 1988. Underpricing of New Issues and the Choice of Auditor as a Signal of Investment Banker Reputation. The Accounting Review 63, 605-622.

Bao, D., S. Fung, and L. Su. 2018. Can Shareholders Be at Rest after Adopting Clawback Provisions? Evidence from Stock Price Crash Risk. Contemporary Accounting Research 35, 1578-1615.

Beatty, A., L. Cheng, and H. Zhang. 2019. Are Risk Factor Disclosures Still Relevant? Evidence from Market Reactions to Risk Factor Disclosures Before and After the Financial Crisis. Contemporary Accounting Research 36, 805-838.

Bloom, N. Fluctuations in Uncertainty. 2014. Journal of Economic Perspectives 28, 153-176.

Bloom, N. S. Bond, and J. Reenan. 2007. Uncertainty and Investment Dynamics. Review of Economic Studies 74, 391-415.

Bloomfield, R. 2008. Discussion of "Annual Report Readability, Current Earnings, and Earnings Persistence." Journal of Accounting and Economics 45, 248-252.

Bond, S. R. Moessner, H. Mumtaz, and M. Syed. 2005. Microeconometric Evidence on Uncertainty and Investment, https://ifs.org.uk/docs/wpinvunc.pdf.

Bonsall, S., and B. Miller. 2017. The Impact of Narrative Disclosure Readability on Bond Ratings and the Cost of Debt. Review of Accounting Studies 22, 608-643.

Botosan, C., A. Huffman, and M. Stanford. 2021. The State of Segment Reporting by US Public Entities: 1976–2017. Accounting Horizons 35, 1-27.

Bratten, B., C. A. Gleason, S. A. Larocque, and L. F. Mills. 2017. Forecasting Taxes: New Evidence from Analysts. The Accounting Review 92, 1-29.

Carcello, J. V., D. R. Hermanson, T. L. Neal, and R. A. Riley Jr. 2002. Board Characteristics and Audit Fees. Contemporary Accounting Research 19, 365-384.

Cassell, C., L. Cunningham, and L. Lisic. 2019. The Readability of Company Responses to SEC Comment Letters and SEC 10-K Filing Review Outcomes. Review of Accounting Studies 24, 1252-1276.

Chaney, P. K., and K. L. Philipich. 2002. Shredded Reputation: The Cost of Audit Failure. Journal of Accounting Research 40, 1221-1245.

Chen, C., J. Kim, M. Wei, and H. Zhang. 2019. Linguistic Information Quality in Customers' Forward-Looking Disclosures and Suppliers' Investment Decisions. Contemporary Accounting Research 36, 1751-1783.

Chen, S., M. DeFond, and C. Park. 2002. Voluntary Disclosure of Balance Sheet Information in Quarterly Earnings Announcements. Journal of Accounting and Economics 33, 229-251.

Chinco, A., A. Clark-Joseph, and M. Ye. 2019. Sparse Signals in the Cross-Section of Returns. Journal of Finance 74, 449-492.

Cohen, L., and D. Lou. 2012. Complicated firms. Journal of Financial Economics 104, 383-400.

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

Dou, Y. and Z. Xu. 2021. Bank Lending and Corporate Innovation: Evidence from SFAS 166/167. Contemporary Accounting Research 38, 2017-2052.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation. Journal of Accounting and Economics 64, 221-245.

Ertugrul, M., J. Lei, J. Qiu, and C. Wan. 2017. Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing. Journal of Financial and Quantitative Analysis 52, 811-836.

Fama, E. F., and K. R. French. 1997. Industry Costs of Equity. Journal of Financial Economics 43, 153-193.

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

Gentzkow, M., Kelly, B., and Taddy, M. 2019. Text as Data. Journal of Economic Literature 57, 535-574.

Glendening, M., E. Mauldin, and K. Shaw. 2019. Determinants and Consequences of Quantitative Critical Accounting Estimate Disclosures. The Accounting Review 94, 189-218.

Goldreich, O. 2010. *P, Np, and Np-Completeness: The Basics of Computational Complexity*, Cambridge, England: Cambridge University Press.

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

Gow, I., D. Larcker, and A. Zakolyukina. 2021. Non-Answers During Conference Calls. Journal of Accounting Research 59, 1349-1384.

Griffin, P. 2003. Got Information? Investor Response to Form 10-K and Form 10-Q EDGAR Filings. Review of Accounting Studies 8, 433-460.

Guay, W., D. Samuels, and D. Taylor. 2016. Guiding through the Fog: Financial Statement Complexity and Voluntary Disclosure. Journal of Accounting and Economics 62, 234-269.

Harjoto, M.A., Laksmana, I., and Lee, R. 2015. The Impact of Demographic Characteristics of CEOs and Directors on Audit Fees and Audit Delay. Managerial Auditing Journal 30, 963-997.

Hay, D. C., W. R. Knechel, and N. Wong. 2006. Audit Fees: A Meta-analysis of the Effect of Supply and Demand Attributes. Contemporary Accounting Research 23, 141-191.

Hogan, C. E., and M. S. Wilkins. 2008. Evidence on the Audit Risk Model: Do Auditors Increase Audit Fees in the Presence of Internal Control Deficiencies? Contemporary Accounting Research 25, 219-242.

Hoitash, R., and U. Hoitash. 2018. Measuring Accounting Reporting Complexity with XBRL. The Accounting Review 93, 259-287.

Hwang, B., and H. Kim. 2017. It Pays to Write Well. Journal of Financial Economics 124, 373-394.

Jiang, G., C. Lee, and Y. Zhang. 2005. Information Uncertainty and Expected Returns. Review of Accounting Studies 10, 185-221.

Jones, M. J. and P. A. Shoemaker. 1994. Accounting Narratives: A Review of Empirical Studies of Content and Readability. Journal of Accounting Literature 13, 142-184.

Ke, Z., B. Kelly, and D. Xiu. 2019. Predicting Returns with Text Data. (No. w26186). National Bureau of Economic Research.

Kim, C., K. Wang, and L. Zhang. 2019. Readability of 10-K Reports and Stock Price Crash Risk. Contemporary Accounting Research 36, 1184-1216.

Koo, D., S. Ramalingegowda, and Y. Yu. 2017. The Effect of Financial Reporting Quality on Corporate Dividend Policy. Review of Accounting Studies 22, 753-790.

Kravet, T. and V. Muslu. 2013. Textual Risk Disclosures and Investors' Risk Perceptions. Review of Accounting Studies 18, 1088-1022.

Lee, C.M., Sun, S.T., Wang, R. and Zhang, R. 2019. Technological Links and Predictable Returns. Journal of Financial Economics 132, 76-96.

Lehavy, R., F. Li, and K. Merkley. 2011. The Effect of Annual Report Readability on Analyst Following and the Properties of their Earnings Forecasts. Accounting Review 86, 1087-1115.

Leuz, C., and P. Wysocki. 2016. The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research. Journal of Accounting Research 54, 525-622.

Li, F. 2008. Annual Report Readability, Current Earnings, and Earnings Persistence. Journal of Accounting and Economics 45, 221-247.

Li, Y., and L. Zhang. 2015. Short Selling Pressure, Stock Price Behavior, and Management Forecast Precision: Evidence from a Natural Experiment. Journal of Accounting Research 53, 79-117.

Liu, X., and R. Natarajan. 2012. The Effect of Financial Analysts' Strategic Behavior on Analysts' Forecast Dispersion. The Accounting Review 87, 2123-2149.

Lo, K., F. Ramos, and R. Rogo. 2017. Earnings Management and Annual Report Readability. Journal of Accounting and Economics 63, 1-25.

Loughran, T. and B. McDonald. 2011. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance 66, 35-65.

Loughran, T. and B. McDonald. 2014. Measuring Readability in Financial Disclosures. Journal of Finance 69, 1643-1671.

Loughran, T. and B. McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research 54, 1187-1230.

Lowry, M., R. Michaely, and E. Volkova. 2020. Information Revealed through the Regulatory Process: Interactions between SEC and Companies ahead of Their IPO. Review of Financial Studies 33, 5510-5554.

Mai, D. and K. Pukthuanthong. 2021. Economic Narratives and Market Outcomes: A Semi-supervised Topic Modeling Approach. htttps://ssrn.com/abstract_id=3990324.

Palmrose, Z. 1986. Audit Fees and Auditor Size: Further Evidence. Journal of Accounting Research 24, 97-110.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High States Decisions and Use Interpretable Models Instead. Nature Machine Intelligence 1, 206-215.

Simunic, D. A. 1980. The Pricing of Audit Services: Theory and Evidence. Journal of Accounting Research 18, 161-190.

Snowden, D. and M. Boone. 2007. A Leader's Framework for Decision Making. Harvard Business Review 85, 68-77.

Stice, J.D. 1991. Using Financial and Market Information to Identify Pre-Engagement Factors Associated with Lawsuits against Auditors. The Accounting Review 66, 516-533.

Stice-Lawrence, L. 2022. Practical Issues to Consider when Working with Big Data. Review of Accounting Studies 27, 1-8.

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

Table 1
List of 53 complexity words included after model selection.

ACCRUES	COUNTERPARTY	INTANGIBLES	OUTSOURCE	REVOCATION
AFFILIATES	COVENANT	INTERNATIONAL	PARTNERING	SECURITIZATIONS
BANKRUPTCIES	COVENANTS	LAWSUIT	RECLASSIFIED	SECURITIZED
CARRYBACK	DERIVATIVE	LAWSUITS	REPATRIATE	SEGMENTS
CARRYFORWARD	DERIVATIVES	LEASEHOLD	REPATRIATED	SOVEREIGN
CARRYFORWARDS	ENTITIES	LEASES	REPATRIATION	SUBLEASES
COLLATERAL	FLOATING	LESSORS	RESTRUCTURE	SUBSIDY
COLLATERIZATION	GLOBAL	LICENSING	RESTRUCTURED	SWAPS
COMPLEX	HEDGED	LITIGATION	RESTRUCTURING	VENTURES
CONVERTIBLE	HEDGES	MERGERS	REVALUATION	WORLDWIDE
COUNTERPARTIES	INFRINGEMENT	MERGING		

Table 2
Summary statistics.

Variable Name	Data Source	N	Mean	Median	Standard Deviation
<u>*Dependent Variables*</u>					
Audit Fees	Audit Analytics	89,633	\$1.83MM	\$0.68MM	\$4.33MM
Abs(%UnexpectedEarnings)	IBES/Compustat	71,092	2.33%	0.65%	3.93%
StdDev Returns	CRSP	119,909	3.85%	2.87%	3.49%
<u>Primary Control Variables</u>					
% Complexity	EDGAR	120,994	0.40%	0.37%	0.17%
MktCap	CRSP	120,994	\$3,673MM	\$299MM	\$21.92B
NetFileSize	EDGAR	120,994	395KB	336KB	280KB
<u>*Alternative Measures of Complexity*</u>					
Segments	Compustat	114,539	4.26	4.00	2.87
Foreign Income Dummy	Compustat	119,402	0.32	0.00	0.46
% Intangibles	Compustat	112,375	12.47%	3.55%	16.90%
Age	Compustat/EDGAR	120,994	15.86	11.00	15.67
ARC	xbrlresearch.com	38,780	351.14	332.00	161.12
<u>*Additional Control Variables*</u>					
Top-5 Auditor Dummy	Audit Analytics	89,633	0.72	1.00	0.45
S&P Dummy	Audit Analytics	89,633	0.28	0.00	0.45
Loss Dummy	Compustat	119,017	0.34	0.00	0.47
% Leverage	Compustat	118,737	22.29%	17.41%	20.79%
% Inventory + Receivables	Compustat	117,113	29.29%	23.99%	23.46%

This table reports summary statistics for the various samples. The Data Source column indicates the source of the variable or the source of the data from which the variable is derived. EDGAR data are available for filing years 1996-2021, and Compustat, IBES, and CRSP have corresponding data for all of these years. Audit Analytics is available for the period 2000-2021, and ARC for the period 2011-2021. Statistics are reported for the number of non-missing observations available in the final merged sample, where the master database is the merged EDGAR and CRSP data with complete data for market capitalization and net file size. In the subsequent regressions, the log of *Audit Fees*, *MktCap*, *Net FileSize*, and *ARC* are used. *Age* is expressed in years relative to the 10-K filing date. Detailed definitions of the variables are provided in Appendix B.

Table 3
Audit fee regressions.

	Estimation Sample (2000-2010)	*Hold-out Sample (2011-2021)*	
	(1)	(2)	(3)
% Complexity	1.653***	1.458***	0.720***
	(30.71)	(20.46)	(10.57)
log(MktCap)	0.346***	0.373***	0.265***
	(19.45)	(58.67)	(36.80)
log(NetFileSize)	0.603***	0.652***	0.391***
	(13.47)	(24.04)	(8.60)
Alternative Measures of Complexity			
Segments			0.031***
			(7.64)
Foreign Income Dummy			0.206***
			(11.86)
% Intangibles			0.004***
			(6.53)
Age			0.003***
			(5.76)
log(ARC)			0.437***
			(2.90)
Additional Controls Variables			
Top-5 Auditor Dummy			0.588***
			(24.38)
S&P 500 Dummy			(0.021)
			(1.13)
Loss Dummy			0.153***
			(10.40)
% Leverage			0.003***
			(4.82)
% Inventory + Receivables			0.005***
			(8.00)
Fixed Effects	Year/Industry	Year/Industry	Year/Industry
R-Squared	75.6%	75.3%	83.3%
Sample Size	46,318	43,315	36,416

This table examines the role of % *Complexity* in predicting *log(Audit Fees)*. The variables are defined in Appendix B. All of 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 CIK number.
*** indicates significance at the 0.01 level.

Table 4
Absolute value of standardized unexpected earnings regressions.

	Estimation Sample (1996-2010)	*Hold-out Sample (2011-2021)*	
	(1)	(2)	(3)
% Complexity	3.657***	2.681***	1.781***
	(13.46)	(9.87)	(8.34)
log(MktCap)	-1.072***	-1.034***	-0.814***
	(-16.75)	(-22.23)	(-11.91)
log(NetFileSize)	1.233***	1.484***	0.915***
	(9.96)	(17.57)	(12.57)
Alternative Measures of Complexity			
Segments			0.014
			(1.11)
Foreign Income Dummy			-0.255***
			(-2.91)
% Intangibles			-0.012***
			(-6.88)
Age			0.005***
			(2.50)
log(ARC)			0.215
			(1.11)
Additional Control Variables			
Top-5 Auditor Dummy			0.016
			(0.19)
S&P 500 Dummy			0.002
			(0.02)
Loss Dummy			2.410***
			(17.54)
% Leverage			0.019***
			(8.07)
% Inventory + Receivables			-0.007***
			(-2.29)
Fixed Effects	Year/Industry	Year/Industry	Year/Industry
R-Squared	23.8%	27.3%	34.9%
Sample Size	40,730	30,362	26,387

This table examines the role of % *Complexity* in predicting the absolute value of standardized unexpected earnings for the period following the 10-K filing date from which the *% Complexity* measure is derived. The variables are defined in Appendix B. All of 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 CIK number.
*** indicates significance at the 0.01 level.

Table 5

Post-filing date stock return volatility regressions.

	Estimation Sample (1996-2010)	*Hold-out Sample (2011-2021)*	
	(1)	(2)	(3)
% Complexity	2.577***	0.768***	0.969***
	(11.79)	(4.46)	(5.93)
log(MktCap)	-1.006***	-0.687***	-0.654***
	(-15.12)	(-18.25)	(-14.15)
log(NetFileSize)	0.689***	0.627***	0.697***
	(8.46)	(10.57)	(12.68)
Alternative Measures of Complexity			
Segments			-0.013**
			(-2.42)
Foreign Income Dummy			-0.017
			(-0.38)
% Intangibles			-0.001
			(-0.73)
Age			-0.004***
			(-3.06)
log(ARC)			-0.141
			(-1.12)
Fixed Effects	Year/Industry	Year/Industry	Year/Industry
R-Squared	37.9%	38.3%	39.5%
Sample Size	76,819	43,090	37,107

This table examines the role of % *Complexity* in predicting the post-filing date stock return volatility as measured by the standard deviation of daily market-adjusted returns for one year after the 10-K filing date. The variables are defined in Appendix B. All of 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 CIK number.
***, ** indicate significance at the 0.01 and 0.05 level, respectively.

Table 6
Robustness – alternative regressions.

		% Complexity Excluded		
	Audit Fees Non-CRSP 2000-2021	*Abs(Unexpected Earnings) 2011-2021*	*SD_Ret 2011-2021*	*Analyst Dispersion (2011-2021)*
	(1)	(2)	(3)	(4)
% Complexity	0.892***	---	---	0.474***
	(7.25)			(5.85)
log(MktCap)	---	-0.791***	-0.644***	-0.325***
		(-11.04)	(-13.92)	(-11.49)
log(NetFileSize)	1.621***	0.853***	0.690***	0.388***
	(42.57)	(10.91)	(13.84)	(8.57)
Alternative Measures of Complexity				
Segments		0.025*	-0.008	0.004
		(1.78)	(-1.40)	(0.88)
Foreign Income Dummy		-0.172**	0.023	-0.089***
		(-2.02)	(0.54)	(-3.25)
% Intangibles		-0.013***	-0.001	-0.005***
		(-7.45)	(-0.83)	(-7.51))
Age		0.005***	-0.004***	0.002***
		(2.67)	(-3.03)	(2.63)
log(ARC)		0.431*	-0.013	0.094**
		(1.80)	(-0.11)	(2.44)
Additional Controls	No	Yes	No	Yes
Fixed Effects	Year/Industry	Year/Industry	Year/Industry	Year/Industry
R-Squared	60.6%	34.5%	39.2%	49.4%
Sample Size	69,456	26,387	37,107	21,281

In column (1), results are presented using *log(AuditFees)* as the dependent variable for the sample of firms without publicly traded stock. Columns (2) and (3) present regressions for *Abs(UnexpectEarnings)* and *SD_Ret* excluding *% Complexity*. Column (4) results use analyst dispersion as the dependent variable. The variables are defined in Appendix B. All of 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 CIK number.
***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively.