Using Unstructured and Qualitative Disclosures to Explain Accruals

Richard Frankel
Olin Business School
Washington University in St. Louis
St. Louis, MO 63130-6431
frankel@wustl.edu

Jared Jennings
Olin Business School
Washington University in St. Louis
St. Louis, MO 63130-6431
jaredjennings@wustl.edu

Joshua Lee
College of Business
Florida State University
821 Academic Way
Tallahassee, FL 32306-1110
jalee@business.fsu.edu

May 2015

* We are grateful to Olin Business School and the Florida State University for financial support. We thank copy editor Tim Gray. This paper has benefited significantly from comments by seminar participants at Baruch College, Washington University in St. Louis, the University of Houston, and the University of North Carolina.

Using Unstructured and Qualitative Disclosures to Explain Accruals

Abstract:

We use MD&A disclosures to predict current-year firm-level accruals using support-vector regressions. We call these predictions *big-data accruals*. Our aim is to measure the explanatory power of MD&A disclosures for liquidity and critical accounting choices. We find that *big-data accruals* explain a statistically and economically significant portion of firm-level accruals and identify more persistent accruals. They have less incremental explanatory power when 10-K readability is low and more when earnings are more difficult to predict using fundamentals. In addition, we find that *big data accruals* are incrementally useful in predicting next period's cash flows. We apply our technique to conference calls and find that they have similar explanatory power for accruals. Our technique can be applied to a variety of unstructured and qualitative disclosures to assess their narrative content.

Keywords: Textual analysis; big data; accruals

JEL Codes: M40; M41

1. Introduction

We use support-vector regressions (Manela and Moreira, 2014) to assess the ability of unstructured and qualitative disclosures to explain firm fundamentals. In particular, we explain firm-level accruals using words and word-pairs from the Management's Discussion and Analysis (MD&A) section of the 10-K. We examine whether the predicted accruals add incremental explanatory power to the Dechow and Dichev (2002) model, as modified by McNichols (2002) and Ball and Shivakumar (2005). We also use support-vector regressions to examine whether the MD&A contains information useful for predicting future cash flows. To demonstrate the applicability of this method to other contexts and to provide an intuitive benchmark, we also apply our technique to conference call transcripts.

Many papers study narrative content. Some studies use pre-determined dictionaries to assess disclosure characteristics (e.g., uncertainty, tone, competition).[1] Others associate measures of readability, similarity, deception, or length with firm fundamentals.[2] Still others require researcher intervention to label characteristics identified by statistical techniques or to train the computer to identify characteristics.[3] In some cases the identified characteristics (e.g., tone) are used to explain firm fundamentals (e.g., performance). In contrast, our method automates the identification of patterns in the narrative that occur in conjunction with firm fundamentals. Therefore, it is amenable to circumstances where researchers have diffuse priors regarding the words associated with a quantifiable firm fundamental (e.g., future profitability, accruals). Its

[1] For example, Kavet and Muslu (2009); Allee and DeAngelis (2015); Davis, Piger, and Sedor (2012); Feldman, Govindaraj, Livnat, and Segal (2010); Kothari, Li, and Short (2009); Li (2010); Loughran and McDonald (2011); Muslu, Radhakrishnan, Subramanyam, and Lim, 2014; Rogers, Van Buskirk, and Zechman (2011); Tetlock, Saar-Tschansky, and Macskassy (2008); and Li, Lundholm, and Minnis (2013); Tetlock, 2007.

[2] For example, Bonsall, Leone, and Miller (2015); Brown and Tucker (2011); Lang and Stice-Lawrence (2014); Larker and Zakolyukina (2012); Lehavy, Li, and Merkley (2011); Li (2008); and Li, Minnis, Nagar, and Rajan (2014); Petersen, K., Schmardebeck, R., and Wilks (2015).

[3] For example, Bao and Datta (2014); Lang and Stice-Lawrence (2014); Campbell, Chen, Dhaliwal, Lu, and Steele (2014); and Kravet and Muslu, 2011; Li (2010). Core (2001) conjectures that researchers can reduce the labor intensity of this task by using natural language processing techniques imported from other fields.

reliance on automation also eases use across languages and contexts. Our method also offers a unconstrained measure of narrative usefulness that can serve as an alternative to information content measures based on textual characteristics (e.g., FOG, disclosure length, and scripting [Lee, 2015]). A weakness of our approach is that it can capture transitory relations alien to the impressions gleaned by readers. We use out-of-sample prediction to address this concern. Yet, common words chosen by managers facing similar circumstances (from hackneyed expressions to idiom) and the explanatory power of word count-based methods employed in existing research enliven belief in the potential efficacy of an automated approach.

We concentrate on how the annual report narrative explains accruals. Accruals represent the effect of accounting choices and estimates on performance measurement. The SEC (2002) notes that MD&A is "consistent with its purpose," if it includes "disclosure on liquidity and capital resources; and disclosure regarding critical accounting estimates." Accruals give a snapshot of activity at the intersection of accounting estimates and liquidity.

Our accrual prediction method has two steps. First, we use a prior-period sample to estimate the support-vector regressions with firm-level accruals as the dependent variable and the counts of all one- and two-word phrases in the MD&A as independent variables.[4] The number of sample observations limits the number of independent variables that can be included in an ordinary least squares regression. Because the number of unique words and phrases found in the MD&A disclosures exceeds the number of firm-year observations, we use support-vector regressions (SVR) to relax the constraint on the number of regressors. This procedure permits an estimated coefficient for each word and phrase in the MD&A. Second, we apply the estimated coefficients to the counts of the words and phrases of the MD&A in the current year to predict

[4] Though a full description of support-vector regression is beyond the scope of the current paper, we summarize the nature of this estimation procedure in Appendix A.

the current-year accruals. The resulting out-of-sample predicted value represents our prediction for firm-level accruals in the current year.

We allow the estimation method to identify industry and economy-wide patterns in the data by executing the above procedure twice. First, we identify industry-specific words relevant to accruals by estimating the support-vector regressions for all firms in the same industry using an estimation window that includes the prior five years. Second, we estimate support-vector regressions for all firms in the sample using an estimation window that includes the prior year, allowing us to identify economy-wide patterns. We obtain separate predicted values for firm-level accruals in the current year from each estimation and average the two predictions to obtain a composite predicted accrual variable, which we denote *big data accruals*.[5]

Using a sample of firm-year observations between 1994 and 2013, we find that *big data accruals* explain both a statistically and economically significant portion of accruals beyond that explained in the Dechow and Dichev (2002) accrual model, as modified by McNichols (2002) and Ball and Shivakumar (2005). The explanatory power of the model increases by approximately 24.1% after including *big data accruals,* relative to prior models.

We next examine the persistence of *big data accruals* relative to *nonbig data accruals*. To identify *nonbig data accruals*, we subtract *big data accruals* from aggregate accruals. Understanding the persistence of accruals aids equity valuation (Dechow, Ge, and Schrand, 2010). Moreover, our aim is to understand whether the MD&A offers excuses for accrual window dressing or whether it highlights accruals with greater value relevance. We find *big data accruals* are more persistent than *nonbig data accruals,* suggesting that support-vector analysis of the MD&A explains accruals relevant to financial analysis. We further examine the

[5] "Big data" refers to data sets with large numbers of variables, requiring the use of nontraditional statistical methods to identify patterns in the data.

persistence of the predicted portion of discretionary and nondiscretionary accruals using the MD&A.[6] We find that support vector regressions predict the more persistent portion of discretionary and non-discretionary accruals. These results suggest that the text in the MD&A can be useful in determining the persistence of total accruals as well as accrual components.

We perform three additional cross-sectional tests to explore the construct captured by our big data measure. We first find that *big data accruals* have less explanatory power after a class-action lawsuit. Research suggests that managers reduce disclosure following class-action lawsuits to limit the pool of disclosure that can be used against them in the legal proceedings or withhold information that they believe could trigger a future lawsuit (e.g., Rogers and Van Buskirk, 2009). Second, we find that *big data accruals* are less useful in explaining accruals when the MD&A is less readable—measured by word length (Li, 2008) and 10-K file size (Loughran and McDonald, 2015), but this result does not hold when using the FOG index (Li, 2008). Prior literature suggests that a manager produces less readable disclosures to obfuscate firm performance when it is poor (e.g., Bloomfield 2002; Li 2008). These results confirm that less readable disclosures have less narrative content. Third, we find that the narrative content of the MD&A in predicting accruals is greater when firm fundamentals are less useful in predicting earnings, which we proxy for using the standard deviation of seasonally adjusted earnings over the prior 16 quarters. We anticipate managers have stronger incentives to reduce information asymmetry through the MD&A when the firm's earnings history is less useful in predicting earnings, providing investors another tool to understand accruals when they are more complex or opaque.

[6] We divide working capital accruals into discretionary and nondiscretionary accruals by regressing the change in working capital accruals on cash flows in year *t*, cash flows in year *t-1*, revenue growth in year *t*, the level of property, plant, and equipment in year *t*, an indicator variable for negative cash flows in year *t*, and the interaction of this indicator variable with cash flows in year *t*. The explained portion of accruals is the nondiscretionary component of accruals and the residual is the discretionary component of accruals.

We perform several additional tests to further explore the usefulness of the MD&A in explaining current accruals. We first examine whether the MD&A provides information beyond the mere description of accrual components (e.g., receivables and inventory) in the financial statements. We re-estimate the support-vector regressions after removing words and phrases containing specific accrual titles such as “receivables” and “inventory” and find that the explanatory power of our *big data accruals* variable does not change. Thus the informativeness of the MD&A for explaining accruals is unlikely driven by direct references to the accrual components on the financial statements. Second, we test whether the MD&A contains information useful for explaining current accruals beyond its implications for future cash flows. Specifically, we include future cash flows in our accrual model and find evidence that *big data accruals* continue to increase the explanatory power of the model. This result suggests that the MD&A provides useful information beyond the mechanical reversal of accruals as described by Dechow and Dichev (2002). However, the increase in the explanatory power of this model is less than the increase in explanatory power when future cash flows are not included, suggesting that *big data accruals* can predict future cash flows. We further explore whether we can use support-vector regressions to predict future cash flows using the words and phrases in the MD&A. We find evidence consistent with the MD&A being useful to predict future cash flows.

We further expand our analysis to conference calls, finding that call narratives predict firm-level accruals. Moreover, conference calls provide incremental information to the MD&A. This evidence suggests that voluntary and mandatory disclosure can be useful in understanding firm-level accruals. In addition, this evidence suggests that support vector regression can be useful in assessing the narrative content of many different types of disclosure.

Caveats There are limits to the inferences that can be drawn from the statistical analysis of language in firm disclosures. The narrative content of the MD&A has been a concern to the Securities and Exchange Commission (SEC) for many years (Brown and Tucker, 2011; Li, Minnis, Lundholm, 2013). Though our results provide evidence that the MD&A disclosures can be informative, they may not provide either an upper bound or a lower bound on its narrative content to a human reader. MD&A is meant to be read as prose not decomposed into individual words and phrases.[7] While it is unlikely that support-vector regressions identify all the cues in the MD&A that a human reader would note, support-vector regressions do not become distracted or bored, nor can they be duped by verbosity and obfuscation. Therefore, statistical methods could identify narrative content in disclosures that humans would miss.

A question addressed by our investigation is whether the diversity of words in English limits the usefulness of our approach. Common phrases describing circumstances surrounding an increase in accruals or its components might not be present in the data at sufficient rates for support-vector regressions to identify meaningful relations between specific words and accruals. The exploration of this question represents another contribution of our research. Despite the diversity of English words, our results suggest that the statistical approach described here adds to the ability to explain a specific firm fundamental, such as firm-level accruals.

This paper is organized as follows. Section 2 discusses big data and its uses. Section 3 describes how we used big data statistical techniques to obtain a qualitative measure for unstructured and qualitative disclosures that is directly related to accrual generation. Section 4 describes the sample along with the empirical findings, and Section 5 addresses the persistence

[7] A mechanical algorithm isolating key words from the phrase "It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness…" might have far more difficulty understanding the situation described than a human reader. For example, a joint project by Stanford University and Google used 16,000 computer processors to teach itself to identify cats from YouTube videos (Autor, 2014). The program made errors a three-year-old could correct.

of *big data accruals*. Section 6 provides cross-sectional analyses, and Section 7, additional tests. Section 8 concludes.

2. Statistical Methods used in Textual Analysis

Accounting and finance researchers use many textual analysis techniques to understand the relation between the text in firm disclosures and firm fundamentals.[8] Early studies use indices (e.g., FOG index), pre-determined dictionaries, word counts, and disclosure length to identify specific textual characteristics (e.g., readability, tone), which are subsequently associated with firm fundamentals (e.g., future performance). For example, Li (2008) examines whether disclosure readability (measured using the Gunning Fog Index and disclosure length of the 10-K) is positively associated with earnings persistence and future earnings. In another paper, Kothari et al. (2009) examine whether more favorable disclosures (measured using word counts) are negatively associated with firm risk (e.g., cost of capital, stock return volatility). Other studies have continued to use indices, pre-determined dictionaries, word counts, and disclosure length to measure specific disclosure characteristics and relate them to firm fundamentals.[9]

Other accounting and finance papers categorize firm disclosures into topics using statistical learning methods (e.g., latent dirichlet allocation, Naïve Bayesian learning algorithms) and relate those topics to specific fundamentals. These learning algorithms typically require researcher judgment in determining the number and categorization of the topics identified by the learning algorithm. For example, Bao and Datta (2014) use latent dirichlet allocation (LDA) to

[8] The quantity of textual analysis studies prevents citation and description of all papers in this literature. We attempt to identify several papers representative of the existing literature.

[9] Some papers include Tetlock et al. (2008), Feldman et al. (2010), Lehavy et al. (2011), Loughran and McDonald (2011), Rogers et al. (2011), Larcker and Zakolyukina (2012), Price et al. (2012), Kravet and Muslu (2013), Li et al. (2013), Davis et al. (2014), Campbell et al. (2014), Li et al. (2014), Loughran and McDonald (2014), Muslu et al. (2014), Allee and DeAngelis (2015), and Bonsall et al. (2015).

assess the extent firm risk is discussed in the 10-K.[10] Li (2010) uses a Naïve Bayesian learning algorithm to estimate the tone of forward-looking statements in the MD&A and relates the tone of the forward-looking statements to future earnings.[11]

Support vector regressions (SVR) are another statistical learning method that can automate the identification of patterns in a narrative that occur in conjunction with firm fundamentals. In other words, SVR does not require researcher intervention in the form of identifying pre-determined dictionaries, manually categorizing topics or disclosures, or relying on researcher judgment to assess the narrative content of the disclosure in explaining firm fundamentals.[12] As a result, SVR has two advantages over other statistical methods: First, SVR can be applied to a variety of disclosures, languages, and contexts at a relatively low cost. Second, SVR can be applied to circumstances where researchers have diffuse priors regarding the words associated with a quantifiable fundamental. Based on the above, we believe that we are directly addressing the call by Core (2001) for textual analysis techniques that are less labor-intensive, less prone to researcher judgment, and that can be more easily applied to large samples.

In the finance literature, SVR has been used to examine firm risk and macro-economic uncertainty. Kogan et al. (2009) use SVR to dissect the MD&A and predict stock return volatility. Manela and Moreira (2014) use SVR to construct a text-based measure of uncertainty using the front-page of *The Wall Street Journal* starting in 1890 to estimate macro-economic

[10] LDA is an unsupervised learning method that estimates a set of topics from text and simultaneously assigns sentences to these topics (Bao and Datta, 2014). However, the topics identified through the LDA estimation process are not labeled in the estimation process. The researcher examines the sentences assigned to each topic and labels the topic based on his or her judgment. While automatic labeling is possible, Mei et al. (2007) suggests that automatic labeling is not appropriate in "cases where such labeling requires domain knowledge", such as financial knowledge (Bao and Datta, 2014).

[11] Li (2010) manually classifies 30,000 sentences of forward-looking sentences from MD&As based on the tone and content of the sentence. These sentences are then used to train the algorithm to categorize the tone and content of other forward-looking statements.

[12] See Appendix A for a general description of the SVR estimation process.

uncertainty during periods for which option-implied volatility is unavailable. See Section 3.2 for more detail on how we apply SVR in this study.

3. Empirical Model

We first review the literature's modeling of accruals based on financial variables. We then describe how big data and the associated statistical techniques can improve our understanding of this process. Our aim in using support-vector regressions is to assess the narrative content of the MD&A beyond current financial statement numbers, because the SEC emphasizes that "MD&A should not be a recitation of financial statements in narrative form or an otherwise uninformative series of technical responses to MD&A requirements." (SEC, 2002)

3.1. Accrual models and accounting variables

Dechow and Dichev (2002) suggest that the purpose of accruals is to "shift or adjust the recognition of cash flows over time so that the adjusted numbers (earnings) better measure firm performance." They provide an empirical model to measure the quality of accruals by examining how well past, current, and future cash flows map into accruals. They suggest that accrual quality increases in the ability of past, current, and future cash flows to explain accruals. Several other studies (e.g., McNichols 2002; Ball and Shivakumar 2005) have augmented the Dechow and Dichev (2002) model to improve the mapping of cash flows into accruals. We present the modified Dechow and Dichev (2002) model in Equation 1 below.

$$Accruals_{i,t} = \alpha_0 + \alpha_1\, CFO_{i,t-1} + \alpha_2\, CFO_{i,t} + \alpha_3\, CFO_{i,t+1} + \alpha_4\, Neg\ CFO_{i,t} + \alpha_5\, CFO_{i,t} * NEG\ CFO_{i,t} + \alpha_6\, \Delta Sales_{i,t} + \alpha_7\, PPE_{i,t} + IND + YEAR + \varepsilon_{i,t} \quad (1)$$

The variables in Equation 1 are defined as follows. The $Accruals_{i,t}$ variable equals working capital accruals, as defined by Dechow and Dichev (2002), for firm *i* in year *t* scaled by total assets in year $t-1$.[13] The $CFO_{i,t-1}$ ($CFO_{i,t}$) [$CFO_{i,t+1}$] variable equals operating cash flows in year $t-1$ (*t*) [$t+1$], scaled by total assets in year $t-2$ ($t-1$) [*t*]. Following Dechow and Dichev (2002), we expect positive coefficients on the $CFO_{i,t+1}$ and $CFO_{i,t-1}$ variables and a negative coefficient on the $CFO_{i,t}$ variable. The $Neg\ CFO_{i,t}$ variable and its interaction with the $CFO_{i,t}$ variable were introduced by Ball and Shivakumar (2005). The $Neg\ CFO_{i,t}$ variable equals 1 if the $CFO_{i,t}$ variable is less than zero. Ball and Shivakumar (2005) do not predict the coefficient on the $Neg\ CFO_{i,t}$ variable but do predict the coefficient on the interaction between the $CFO_{i,t}$ and $Neg\ CFO_{i,t}$ variable to be positive, suggesting that accrued losses are more likely in periods of negative cash flow. The $\Delta Sales_{i,t}$ and $PPE_{i,t}$ variables were introduced by McNichols (2002) and are included to identify changes to accrual generation that are due to sales growth and depreciation. The $\Delta Sales_{i,t}$ variable equals the change in sales from year $t-1$ to *t* less the change in receivables from year *t-1* to *t* scaled by total assets in year $t-1$. The $PPE_{i,t}$ variable equals property, plant, and equipment for firm *i* in year *t* scaled by total assets in year $t-1$. Following McNichols (2002), we expect a positive coefficient on the $\Delta Sales_{i,t}$ to reflect the growth in sales and a negative coefficient on the $PPE_{i,t}$ variable to reflect accruals that are related to the depreciation of the firm's tangible assets. We also include year and industry (GICS codes) fixed effects to control for unspecified year and industry effects, and we also cluster the standard errors by firm to adjust for serial-correlation (Petersen, 2009).

Since we aim to determine whether big data techniques can explain accruals in year *t*, we slightly modify the Dechow and Dichev (2002) model, as augmented by McNichols (2002) and

[13] Dechow and Dichev (2002) define working capital accruals as the following items from the statement of cash flows: $\Delta Accounts\ Receivable_{i,t} + \Delta Inventory_{i,t} – \Delta Accounts\ Payable_{i,t} – \Delta Taxes\ Payable_{i,t} + \Delta Other\ Assets\ (net)_{i,t}$.

Ball and Shivakumar (2005). To provide a benchmark that uses information available to investors at the time that accruals are reported, we eliminate the $CFO_{i,t+1}$ variable from Equation 1. See Equation 2 for our base accrual model.[14]

$$Accruals_{i,t} = \alpha_0 + \alpha_1\ CFO_{i,t-1} + \alpha_2\ CFO_{i,t} + \alpha_3\ Neg\ CFO_{i,t} + \alpha_4\ CFO_{i,t} * NEG\ CFO_{i,t} + \alpha_4\ \Delta Sales_{i,t} + \alpha_5\ PPE_{i,t} + IND + YEAR + \varepsilon_{i,t} \quad (2)$$

The R^2 calculated in Equation 2 serves as a benchmark for how well operating cash flows, the change in sales, and level of property, plant, and equipment explain working capital accruals. In the next section, we examine whether including predicted accruals based on big data techniques increases the explanatory power of the model relative to the base model described in Equation 2.

3.2. Support-Vector Regressions and Accruals

We use SVR to examine how specific words and short phrases in MD&A predict accrual levels. We provide a brief explanation of the estimation procedure for support-vector regressions in Appendix A, similar to that provided by Manela and Moreira (2014). We construct a firm-year level dataset with counts of all one- and two-word phrases included in each firm's MD&A, replacing highly frequent words (i.e., "stop words") such as "and" and "the" with an underscore symbol and removing words containing digits (Manela and Moreira 2014).[15] We also remove

[14] In additional robustness tests, we do not eliminate the $CFO_{i,t+1}$ variable from Equation 1 and find qualitatively similar results for all tests.

[15] We find similar results when we extend our analysis to include all one-, two-, three-, four-, and five-word phrases found in each firm's MD&A. A few examples of highly predictive phrases from this estimation include "increase _ primarily due _," "representing _ increase _ _," "compared _ _ net loss," and "increase _ _ compared." Relative to the adjusted R^2 of 0.201 from our primary analysis (Equation 3) using two-word phrases, the adjusted R^2 from our analysis using five-word phrases increases to 0.206, representing only a marginal increase in adjusted R^2. In addition, the number of unique phrases included in this analysis increases to 1.83 million, which significantly increases the required computation time. Given the marginal increase in explanatory power and the substantial increase in computational costs of expanding the examined set of phrases, we report our primary analyses using two-word phrases.

infrequent words and phrases by requiring each word and phrase to be included in the MD&A of at least 10 firms in each year. Our dataset comprises 71,847 observations with 296,329 unique words and phrases between 1994 and 2013, rendering traditional ordinary least squares estimation useless due the fact that the number of observations is less than the number unique words and phrases included in the MD&As.

We use the above dataset to predict accruals using the following SVR procedure. First, we estimate SVRs with working capital accruals as the dependent variable and the counts of each word/phrase as independent variables using firm-year observations from an estimation window prior to year *t*. This procedure allows for estimated coefficients for all words/phrases that are included as regressors. We then apply the estimated coefficients to the word/phrase counts in year *t* to obtain out-of-sample predicted values of working capital accruals in year *t*.

We estimate the above SVR procedure using two sets of estimation samples. First, we estimate the model to obtain estimated coefficients by industry, which we define as GICS codes, including data for the years $t-5$ to $t-1$. Performing the analysis at the industry level allows the estimation method to identify important words/phrases that are specific to firms in each industry for explaining accrual generation. We then apply the estimated coefficients to the words/phrases in year *t* to obtain predicted values for accruals in year *t*, which we label $\textit{BD Accruals – Industry}_{i,t}$.

Table 1 displays examples of words and phrases that explain significant variation in accruals. In Panel A, we present examples of words and phrases with positive and negative estimated coefficients from the SVR estimation for four industries: communications equipment (GICS = 452010), computers and peripherals (GICS = 452020), metals and mining (GICS = 151040), and biotechnology (GICS = 352010). Some words and phrases that are common across industry specifications include “loss” and “decreased.” The industry lists also include words and phrases that correspond to each industry. For example, the communications equipment industry list contains the phrase “_ equipment,” and the metals and mining industry includes the phrase “_ aluminum.” The *individual* words and phrases in Table 1 are not likely to be singly informative about accrual generation. However, support-vector regressions allow us to simultaneously examine all words and phrases included in the MD&A to predict accruals. In other words, a single word or short phrase may not significantly improve our ability to predict accruals, but a high word count of several hundred words or short phrases may significantly predict accruals. While understanding how individual words/phrases explain accruals is interesting, it does not provide the basis for assessing the construct validity or the “mechanical” nature of the method. Each word/phrase is only a small part of all the words/phrases that explain accruals. We therefore explore whether the explanatory power of SVR estimates coincides with our understanding of factors associated with the narrative content of disclosure. An important contribution of this paper is exploring whether seemingly robotic analysis can assess the narrative content of disclosure. As Samuel Johnson quipped about a dog waking on its hind legs, “It is not done well; but you are surprised to find it is done at all.”[16]

Second, we estimate the SVR model for all firm-year observations that are in year $t-1$ to estimate the coefficients on each word/phrase count. As in the industry estimation method described above, we apply the estimated coefficients to the word/phrase counts for each firm *i* in year *t* to obtain an out-of-sample predicted value for accruals in year *t*, which we label $BD\ Accruals - Year_{i,t}$. We calculate the predicted accruals at the yearly level across all firms in the sample to identify words and phrases that pertain to accrual generation and could be related to

[16] James Boswell, *The Life of Samuel Johnson*, LL. D. (London: John Sharpe, 1791), p. 140, https://books.google.com/books?id=VbtcAAAAcAAJ&pg=PA140&dq#v=onepage&q&f=false

macroeconomic factors. We note some influential words and phrases that are useful in predicting accruals at the yearly level are “receivable,” “net profit,” “net loss,” and “substantial doubt.” See Table 1 for a more complete description words that are useful in predicting accruals at the yearly level. After calculating the *BD Accruals – Industry*$_{i,t}$ and *BD Accruals – Year*$_{i,t}$ variables, we calculate a composite variable (*BD Accruals*$_{i,t}$) by averaging the two accrual prediction variables.

3.3. Full Accrual Model

We next examine whether *BD Accruals*$_{i,t}$ is correlated with actual accruals and whether its inclusion in the modified Dechow and Dichev (2002) model improves the overall explanatory power of the model. We do this by augmenting Equation 2 with the *BD Accruals*$_{i,t}$ variable. We present the revised accrual model below as Equation 3.

$$\begin{aligned} Accruals_{i,t} = \quad & \alpha_0 + \alpha_1\ CFO_{i,t-1} + \alpha_2\ CFO_{i,t} + \alpha_3\ Neg\ CFO_{i,t} + \alpha_4\ CFO_{i,t} * NEG\ CFO_{i,t} + \alpha_4 \\ & \Delta Sales_{i,t} + \alpha_5\ PPE_{i,t} + \alpha_6\ BD\ Accruals_{i,t} + IND + YEAR + \varepsilon_{i,t} \end{aligned} \quad (3)$$

If the *BD Accruals*$_{i,t}$ variable predicts accruals without error, we expect its coefficient to equal 1. The variable, however, may include error in its prediction, resulting in an attenuated coefficient that is less than 1. If the variable is useful in explaining firm-level accruals, then we expect the coefficient to be positive and significant. We also examine whether the *BD Accruals*$_{i,t}$ variable explains an economically significant portion of the variation in accruals by examining the percentage increase in R^2 from Equation 2 to Equation 3.

4. Empirical Results

4.1. Sample

Our sample includes firm-year observations between 1994 and 2013, resulting in 71,847 firm-year observations. We include all firm-year observations with sufficient data to calculate all variables included in Equation 3. We obtain all financial statement data from Compustat. We download the 10-Ks for all publicly traded firms from the SEC's EDGAR filing system and then use Python to extract the MD&A section from each report and count the number of words/phrases in each MD&A. We report descriptive statistics for our sample in Table 2. We note that the average firm included in our sample has negative accruals ($Accruals_{i,t}$) and big data accruals ($BD\ Accruals_{i,t}$). The average firm in our sample experiences 9.6% sales growth from year $t-1$ to t ($\Delta Sales_{i,t}$) and holds approximately 56.5% of its total assets as property, plant, and equipment ($PPE_{i,t}$). The average operating cash flow in year t ($CFO_{i,t}$) equals −0.063 and in year $t-1$ ($CFO_{i,t-1}$) is −0.085. Lastly, approximately 33.1% of the observations have operating cash flow that is less than zero.

We also compute univariate Pearson and Spearman correlations for all variables included in our primary analyses in Table 3. We note that the Spearman correlation between $BD\ Accruals_{i,t}$ and $Accruals_{i,t}$ equals 0.18 and is significant at the 1% level, providing preliminary support that big data techniques can predict accruals using the MD&A. We hesitate to draw too many conclusions using the univariate correlations given the high correlations between the $BD\ Accruals_{i,t}$ variable and the other independent variables included in Equation 3. For example, we find that $BD\ Accruals_{i,t}$ is positively correlated with $\Delta Sales_{i,t}$ and negatively correlated with $PPE_{i,t}$. Inconsistent with expectations, we find that $BD\ Accruals_{i,t}$ is positively correlated with the $CFO_{i,t}$ variable. Consistent with the prior literature, we also find that the Spearman correlation between *Accruals*$_{i,t}$ and *CFO*$_{i,t}$ is negative, *Accruals*$_{i,t}$ and *CFO*$_{i,t-1}$ is positive, and *Accruals*$_{i,t}$ and *ΔSales*$_{i,t}$ is positive.[17]

4.2. Accrual Model Results

We present the results using the modified Dechow and Dichev (2002) model using pooled and industry-specific regressions in Panels A and B of Table 4, respectively. In column 1 of Panel A, we present the results for the base modified Dechow and Dichev (2002) model (Equation 2) using the pooled regression. We use these results as a benchmark to determine the incremental explanatory power of big data techniques. Consistent with the prior literature, we find a positive and significant (1% level) coefficient on the *CFO*$_{i,t-1}$ variable and a negative and significant (1% level) coefficient on the *CFO*$_{i,t}$ variable. The interaction between *CFO*$_{i,t}$ and the *Neg CFO*$_{i,t}$ variable is positive and significant at the 1% level, suggesting that accrued losses are more likely in periods of negative cash flow. We also find a positive and significant (1% level) coefficient on the *ΔSales*$_{i,t}$ variable, suggesting that higher sales growth results in higher accruals. We find a negative and significant (1% level) coefficient on the *PPE*$_{i,t}$ variable, suggesting that high tangible asset levels result in higher depreciation accruals. We also note that the adjusted R^2 in the model is equal to 0.162, suggesting that approximately 16.2% of the variation in accruals is explained by the model detailed in Equation 2.[18]

[17] The Pearson correlation between the *CFO*$_{i,t}$ and *Accruals*$_{i,t}$ variables is positive and significant. Given that the Spearman correlation between the two variables is negative and significant, the positive Pearson correlation is likely due to extreme observations.

[18] We note that the R^2 reported in Table 4 is lower than the R^2 reported in Dechow and Dichev (2002). The difference in R^2s due to two factors. First, our sample covers a different time frame: we cover firm/year observations between 1994 and 2013, while Dechow and Dichev (2002) cover firm/year observations between 1987 and 1999. Second, we do not require the firm to have at least eight years of data to be included in the sample. When we do require the same time frame and at least eight years of data for each firm, we find a similar R^2 to that found by Dechow and Dichev (2002).

We present the results from Equation 3 in column 2 of Table 4. We find that the coefficient on the *BD Accruals*$_{i,t}$ variable equals 0.370 and significant at the 1% level.[19] This evidence suggests that using big data techniques to predict accruals can significantly improve understanding of accrual generation. We note that the coefficients on the other variables resemble the coefficients in column 1. The adjusted R^2 is equal to 0.201, which is approximately 24.1% larger than the base model presented in column 1, an increase that we believe is economically significant. This evidence suggests that analyzing unstructured and qualitative disclosures using SVR methods to predict a specific firm-level fundamental improves the ability to explain accruals relative to traditional accrual models.

In Panel B, we present the industry-specific regressions. In column 1 of Panel B, we present the results using the base Dechow and Dichev (2002) model. GICS codes yield 56 separate industry regressions. We find that the average and median coefficients are consistent with those in column of Panel A. The average (median) R^2 is equal to 0.253 (0.202). In column 2 of Panel B, we find that the mean (median) coefficient from the industry-specific regressions on the *Big Data Accruals*$_{i,t}$ variable equals 0.283 (0.237) and is significant at the 1% level. We find that the mean (median) R^2 increases 11.8% (17.3%). While these results are attenuated relative to the pooled regression results, we believe that these results are still economically significant.

5. Persistence of Big Data Accruals

[19] In untabulated results, we replace the *BD Accruals*$_{i,t}$ variable in equation 3 with the *BD Accruals – Industry*$_{i,t}$ and *BD Accruals – Year*$_{i,t}$ variables in separate regressions. Since we believe the *BD Accruals – Industry*$_{i,t}$ and *BD Accruals – Year*$_{i,t}$ variables identify slightly different aspects of accrual generation, we expect the coefficients on these variables to be lower and the R^2s in these regressions to be lower. As expected, we find that the coefficients on the *BD Accruals – Industry*$_{i,t}$ and *BD Accruals – Year*$_{i,t}$ variables to be less than the coefficient on the *BD Accruals*$_{i,t}$ variable. We also find that the adjusted R^2s for the regressions including the *BD Accruals – Industry*$_{i,t}$ variable and *BD Accruals – Year*$_{i,t}$ variable are less than the adjusted R^2 for the regression including the *BD Accruals*$_{i,t}$ variable.

We next examine whether *BD Accruals*$_{i,t}$ are more or less persistent than those accruals that are not predicted using big data techniques, providing additional evidence on the quality of *BD Accruals*$_{i,t}$. Dechow, Ge, and Schrand (2010) suggest that understanding the persistence of accruals is a valuable input for equity valuation. Prior to examining the persistence of accruals identified using big data techniques, we examine the persistence of operating cash flows and total accruals to establish a benchmark in evaluating those accruals identified using big data techniques. Equation 4 describes our base model, which is similar to that presented in Sloan (1996).

$$ROA_{i,t+1} = \alpha_0 + \alpha_1\, CFO_{i,t} + \alpha_2\, Accruals_{i,t} + \varepsilon_{i,t} \qquad (4)$$

To match the sample composition of Sloan (1996) and Richardson, Sloan, Soliman, and Tuna (2005), we require CRSP data when estimating Equation 4, reducing our sample to 54,110 firm-year observations. All variables are as previously defined with exception to the $ROA_{i,t+1}$ variable, which equals net income for firm *i* in year *t+1* scaled by total assets in year *t*. We also include year and industry fixed effects and cluster standard errors by firm. We present the results using Equation 4 in column 1 of Table 5. The coefficient on the $CFO_{i,t}$ variable equals 0.635 and is significant at the 1% level. Consistent with the prior literature, we find that the coefficient on $Accruals_{i,t}$ is significantly less than the coefficient on the $CFO_{i,t}$ variable (P-value < 0.001), suggesting that cash flows are more persistent than accruals. The $Accruals_{i,t}$ variable equals 0.472 and significant at the 1% level.

To test whether the accruals predicted using big data techniques are more or less persistent than those that are not predicted this way, we replace the $Accruals_{i,t}$ variable with those accruals that are predicted using big data techniques (*BD Accruals*$_{i,t}$) and those that are not

($Accruals_{i,t}$ less $BD\ Accruals_{i,t}$), which we call $Non\text{-}BD\ Accruals_{i,t}$. The results for the above specification are included in column 2 of Table 5. We find that the coefficient on the $BD\ Accruals_{i,t}$ variable equals 0.518, and the coefficient on the $Non\text{-}BD\ Accruals_{i,t}$ variable equals 0.466, both of which are significant at the 1% level. We also note that the coefficient on the $BD\ Accruals_{i,t}$ is significantly greater than the coefficient on the $Non\text{-}BD\ Accruals_{i,t}$ at the 1% level, suggesting that accruals predicted using big data techniques are more persistent.

We continue to analyze the persistence of accruals predicted using big data techniques by examining whether big data techniques are useful in identifying the portion of nondiscretionary and discretionary accruals that are more persistent, providing a valuable input for equity valuations. We first run the following model to establish a benchmark for the persistence of nondiscretionary and discretionary accruals.

$$ROA_{i,t+1} = \quad \alpha_0 + \alpha_1\ CFO_{i,t} + \alpha_2\ Disc\ Accruals_{i,t} + \alpha_3\ Non\text{-}Disc\ Accruals_{i,t} + \varepsilon_{i,t} \qquad (5)$$

All variables are as previously defined with exception to the discretionary ($Disc\ Accruals_{i,t}$) and nondiscretionary accruals ($Non\text{-}Disc\ Accruals_{i,t}$) variables. To calculate nondiscretionary accruals, we estimate Equation 2 for each GICS industry in year *t-1* and then apply the estimated coefficients to firm-year observations in year *t* to obtain a predicted value for the nondiscretionary component of accruals ($Non\text{-}Disc\ Accruals_{i,t}$). We compute discretionary accruals ($Disc\ Accruals_{i,t}$) by subtracting total accruals ($Accruals_{i,t}$) from nondiscretionary accruals ($Non\text{-}Disc\ Accruals_{i,t}$). We report our results in column 3 of Table 5. The coefficient on the $Non\text{-}Disc\ Accruals_{i,t}$ ($Disc\ Accruals_{i,t}$) equals 0.506 (0.507) and is significant at the 1% level.[20]

[20] The difference in coefficients between $Non\text{-}Disc\ Accruals_{i,t}$ and $Disc\ Accruals_{i,t}$ is not significant, which is inconsistent with Xie (2001).

After establishing a benchmark for the persistence of discretionary and nondiscretionary accruals, we examine whether big data techniques identify the more persistent portion of these accruals. To test this conjecture, we use support-vector regressions to identify the portion of discretionary and nondiscretionary accruals that can be predicted using the individual words and two-word phrases in the MD&A, similar to how we use big data techniques to predict aggregate accruals (*BD Accruals*$_{i,t}$). We first use SVR to predict industry/year discretionary and nondiscretionary accruals using firm-year observations for industry *j* between year $t-5$ and $t-1$. We use the coefficients from these regressions to predict discretionary and nondiscretionary accruals in year *t* for firm *i*. We also use SVR to predict discretionary and nondiscretionary accruals with all firm-year observations from year $t-1$. Using the coefficients from the support-vector regressions with all firm-year observations in year $t-1$, we form our second prediction of discretionary and nondiscretionary accruals in year *t* for firm *i*. Similar to how we predict *BD Accruals*$_{i,t}$, we average the predicted discretionary and nondiscretionary accruals using the industry/year and yearly support-vector regressions to calculate the composite discretionary accruals variable (*BD Disc Accruals*$_{i,t}$) and the composite nondiscretionary accruals variable (*BD Non-Disc Accruals*$_{i,t}$).

To obtain the nonpredicted components of discretionary (nondiscretionary) accruals, we subtract the *BD Disc Accruals*$_{i,t}$ (*BD Non-Disc Accruals*$_{i,t}$) from the *Disc Accruals*$_{i,t}$ (*Non-Disc Accruals*$_{i,t}$), which we call *Non-BD Disc Accruals*$_{i,t}$ (*Non-BD Non-Disc Accruals*$_{i,t}$). We include the predicted and nonpredicted components of discretionary and nondiscretionary accruals along with operating cash flow in the following equation to examine the differential persistence of the various accrual measures.

$$ROA_{i,t+1} = \quad \alpha_0 + \alpha_1\ CFO_{i,t} + \alpha_2\ BD\ Disc\ Accruals_{i,t} + \alpha_3\ Non\text{-}BD\ Disc\ Accruals_{i,t} + \alpha_4 \qquad (6)$$

$$BD\ Non\text{-}Disc\ Accruals_{i,t} + \alpha_5\ Non\text{-}BD\ Non\text{-}Disc\ Accruals_{i,t} + \varepsilon_{i,t}$$

We present the results of estimating Equation 6 in column 4 of Table 5. We find that the coefficient on the nonpredicted component of discretionary accruals (*Non-BD Disc Accruals*$_{i,t}$) is 0.501 and less than the coefficient on the predicted component of discretionary accruals (*BD Disc Accruals*$_{i,t}$), which equals 0.531, but is not statistically different. In addition, we find that the coefficient on the nonpredicted component of nondiscretionary accruals (*Non-BD Non-Disc Accruals*$_{i,t}$) equals 0.489 and is significantly less than (1% significance level) the coefficient on the predicted component of nondiscretionary accruals (*BD Non-Disc Accruals*$_{i,t}$), which equals 0.633. This evidence suggests that our method is able to identify the more persistent component of nondiscretionary accruals.

6. Cross-sectional Tests

We provide further evidence on the construct captured by our method. We first examine whether support vector regressions continue to predict accruals after litigation, when managers are more likely to reduce the level of public information provided to market participants. Second, we examine whether big data techniques are less useful when the MD&A is less readable. Finally, we examine whether managers provide additional information about accrual generation through the MD&A when earnings are more difficult to predict using fundamentals.

6.1. Litigation

In our first test, we examine whether the ability of big data techniques to predict accruals decreases when managers are less likely to provide useful disclosures. After the filing of a class action lawsuit, which typically alleges that managers have intentionally withheld or misstated their financial statements, managers have incentives to withhold additional information that can be used against them during legal proceedings or withhold information that might trigger a future lawsuit. As a result, managers likely provide less public information. Rogers and Van Buskirk (2009) provide evidence consistent with this conjecture in that managers provide less voluntary disclosure (e.g., management forecasts and conference calls) immediately following the filing of the class-action lawsuit.

If managers reduce their disclosures following litigation, we anticipate the MD&A to be less useful to investors following the filing of the lawsuit, resulting in a reduction in the usefulness of the big data predicted accruals. We use the following equation to test this conjecture.

$$\mathit{Accruals}_{i,t} = \alpha_0 + \alpha_1\, \mathit{CFO}_{i,t\text{-}1} + \alpha_2\, \mathit{CFO}_{i,t} + \alpha_3\, \mathit{Neg\ CFO}_{i,t} + \alpha_4\, \mathit{CFO}_{i,t} * \mathit{NEG\ CFO}_{i,t} + \alpha_5\, \Delta \mathit{Sales}_{i,t} + \alpha_6\, \mathit{PPE}_{i,t} + \alpha_7\, \mathit{Suit}_{i,t} + \alpha_8\, \mathit{BD\ Accruals}_{i,t} + \alpha_9\, \mathit{BD\ Accruals}_{i,t} * \mathit{Suit}_{i,t} + \varepsilon_{i,t} \tag{7}$$

We include firm-years one year prior to and one year immediately following the filing of 854 class action lawsuits.[21] We include an indicator variable *Suit*$_{i,t}$ equal to 1 for the year of the class action lawsuit and equal to 0 in the year prior to the filing of the lawsuit. We interact the *Suit*$_{i,t}$ variable with the *BD Accruals*$_{i,t}$ variable to examine whether the predictive ability of *BD Accruals*$_{i,t}$ decreases from the year before to the year of the lawsuit. We expect a negative coefficient on the interaction if the predictive ability of SVR decreases as managers provide less public information after the revelation of the lawsuit. All other variables included in Equation 7 are as previously defined.

[21] The litigation data are obtained from the Stanford Law School Securities Class Action Clearinghouse at http://securities.stanford.edu/index.html.

In Table 6, we report the results using Equation 7. We find a negative and significant (1% level) coefficient on the interaction between the $Suit_{i,t}$ and $BD\ Accruals_{i,t}$ variables, suggesting that managers provide less useful disclosures immediately following the filing of a lawsuit.

6.2. *Readability*

We also examine whether *big data accruals* are less useful when the MD&A is less readable. Li (2008) suggests that managers decrease the readability of the 10-K to obfuscate information about poor performance. If managers do this, we expect *big data accruals* to be less useful in understanding accrual generation. We anticipate that big data techniques will not be able to identify predictable trends in the words or phrases used in MD&As that are more difficult to understand. We use Equation 8 to examine whether big data techniques are less useful in predicting accruals when the MD&A is less readable.

$$Accruals_{i,t} = \alpha_0 + \alpha_1\ CFO_{i,t-1} + \alpha_2\ CFO_{i,t} + \alpha_3\ Neg\ CFO_{i,t} + \alpha_4\ CFO_{i,t} * NEG\ CFO_{i,t} + \alpha_5\ \Delta Sales_{i,t} + \alpha_6\ PPE_{i,t} + \alpha_7\ BD\ Accruals_{i,t} + \alpha_8\ Low\ Readability_{i,t} + \alpha_9\ BD\ Accruals_{i,t} * Low\ Readability_{i,t} + \varepsilon_{i,t} \quad (8)$$

We define MD&A readability in three ways: Fog index (Li 2008), word length (Li 2008), and file size (Loughran and McDonald 2015). Li (2008) develops a measure for 10-K readability called the Fog index, which is from the computational linguistics literature and is decreasing as the 10-K becomes more complex or difficult to read. Li (2008) also suggests that longer documents are harder to read.[22] Loughran and McDonald (2015) suggest that the 10-K document size is an additional measure of 10-K readability.[23] They suggest that more complicated and less

[22] Our analysis focuses on the MD&A portion of the 10-K, and thus our fog and word length readability measures are derived from the MD&A rather than from the full 10-K.

[23] Bonsall et al. (2015) suggest that 10-K file size is affected by content unrelated to the underlying text in the 10-K (e.g., HTML, XML, pdf, and jpeg file attachments). For this purpose, we restrict our file size measure to include

readable reports are larger. We perform three separate regressions—one for each readability proxy. The *Low Readability*$_{i,t}$ variable equals one if the Fog index, word length, or file size is above the sample median. To examine whether big data techniques are more useful in understanding accrual generation when the MD&A is less readable, we interact *BD Accruals*$_{i,t}$ with *Low Readability*$_{i,t}$, and expect a negative coefficient on the interaction.

We report these results in Table 7. Column 1 includes the results when the Fog index is used to calculate the *Low Readability*$_{i,t}$ variable, column 2 includes the results when the MD&A word length is used to calculate the *Low Readability*$_{i,t}$ variable, and column 3 includes the results when the file size of the 10-K is used to calculate the *Low Readability*$_{i,t}$ variable. Consistent with expectations, we find a negative and significant (1% level) coefficient on the interaction between the *BD Accruals*$_{i,t}$ and *Low Readability*$_{i,t}$ variables in column 2 and 3. This evidence is consistent with big data techniques capturing an information content concept similar to that measured by prior papers. These results reinforce the construct validity of prior measures as well as the present measure.

6.3. Predictability Using Fundamentals

We next examine whether support vector regressions are more useful in understanding accrual generation when earnings are more difficult to predict. We expect that managers provide more information about accruals in the MD&A when earnings are more difficult to predict using accounting fundamentals. If managers increase the informativeness of the MD&A when earnings are more difficult to predict using firm fundamentals, we expect *big data accruals* to be more useful in explaining accruals relative to the traditional accrual model. We use Equation 9 to

only the 10-K document and the associated EX-13 (i.e., annual report), if available. We then regress the file size variable on an indicator variable equal to 1 if the document is HTML and 0 if the document is text-based. The residual is our measure of 10-K file size.

examine the usefulness of support vector regressions when earnings are more difficult to forecast using simple accounting fundamentals.

$$Accruals_{i,t} = \alpha_0 + \alpha_1\ CFO_{i,t-1} + \alpha_2\ CFO_{i,t} + \alpha_3\ Neg\ CFO_{i,t} + \alpha_4\ CFO_{i,t} * NEG\ CFO_{i,t} + \alpha_5\ \Delta Sales_{i,t} + \alpha_6\ PPE_{i,t} + \alpha_7\ BD\ Accruals_{i,t} + \alpha_8\ High\ Earn\ Vol_{i,t} + \alpha_9\ BD\ Accruals_{i,t} * High\ Earn\ Vol_{i,t} + \varepsilon_{i,t} \quad (9)$$

Following Jennings, Lee, and, Matsumoto (2014), we measure the difficulty in predicting earnings using the volatility of seasonally adjusted quarterly earnings before extraordinary items scaled by lagged total assets over the last 16 quarters. The *High Earn Vol*$_{i,t}$ variable equals 1 when earnings volatility is above the *Earn Vol*$_{i,t}$ sample median value. We then interact *High Earn Vol*$_{i,t}$ with *BD Accruals*$_{i,t}$ to examine whether big data techniques are more useful in understanding accruals when earnings are more difficult to predict using accounting fundamentals. We report the results using Equation 9 in Table 8. Consistent with expectations, we find a positive and significant (1% level) coefficient on the interaction between the *High Earn Vol*$_{i,t}$ and *BD Accruals*$_{i,t}$ variables, suggesting that support vector regressions are more useful when managers are more likely to provide more information about earnings. [24]

7. Additional Tests

We perform several additional tests to further explore the usefulness of the MD&A in explaining current accruals and future cash flows. We examine whether the predictive ability of

[24] We interpret the coefficients on the interaction between *BD Accruals*$_{i,t}$ and each of our cross-sectional variables as an indication of the differential ability of *BD Accruals*$_{i,t}$ to explain accruals in these subsamples. However, we assume that the predictive ability of all other control variables included in the model does not differ based on the cross-sectional cut. We therefore restrict the sample to the top and bottom deciles of the cross-sectional variable of interest (e.g., word count, earnings volatility) and then standardize the dependent variable and all independent variables within each subgroup by ranking into deciles from 0 to 9 and dividing by 9 so that the variables range from 0 to 1. We then re-estimate the cross sectional tests in Tables 6, 7, and 8 and find similar results. However, the interaction term is only significant at the 10% level in the litigation and earnings volatility tests.

the MD&A is simply driven by descriptions of accrual components found in the financial statements. We examine the explanatory power of big data accruals relative to future cash flows. We directly predict future cash flows to explore whether the MD&A contains information useful for understanding future cash flows. We use SVR to analyze whether conference calls narratives explain current accruals. We examine the change in the narrative content of the MD&A over time.

7.1. Big Data Accruals Prediction Excluding Accrual-based Words

We first examine whether the MD&A provides information beyond the mere description of accrual components on the face of the financial statements. The SEC's MD&A guidance explicitly states, "MD&A should not be a recitation of financial statements in narrative form or an otherwise uninformative series of technical responses to MD&A requirements, neither of which provides this important management perspective." However, to the extent that managers simply restate the components of accruals found on the face of the financial statements, our analysis of the MD&A might mechanically predict accruals using these component descriptions. To alleviate this concern, we re-estimate the support-vector regressions, removing words and phrases containing specific accrual titles, such as asset*, liabilit*, receivable*, inventor*, payable*, and prepaid*. We call the resulting predicted accrual variable $\textit{BD Accruals (Excl Acc Phrases)}_{i,t}$. We then re-estimate Equation 3, replacing $\textit{BD Accruals}_{i,t}$ with $\textit{BD Accruals (Excl Acc Phrases)}_{i,t}$, and present the results in Panel A of Table 9. We find that the explanatory power of the model remains unchanged: the adjusted R^2 equals 0.201, which is identical to that reported in Table 4, and the coefficient on $\textit{BD Accruals (Excl Acc Phrases)}_{i,t}$ equals 0.368, which is just slightly lower than the 0.370 coefficient on $\textit{BD Accruals}_{i,t}$ reported in Table 4. Our results provide some assurance that the informativeness of the MD&A for explaining accruals is unlikely driven by direct references to the accrual components on the financial statements.

7.2. Big Data Accruals Prediction Relative to Future Cash Flows

Second, we examine whether the ability of the MD&A to predict current accruals is solely driven by its explanation of accruals that reverse in the following year as described by Dechow and Dichev (2002). Specifically, we re-estimate Equation 3, including cash flows for firm *i* in year *t+1* ($CFO_{i,t+1}$) and re-assess the explanatory power of the $BD\ Accruals_{i,t}$ variable. Panel B of Table 9 presents the results. The adjusted R^2 of the base model in column 1, excluding $BD\ Accruals_{i,t}$, equals 0.207, which is higher than the 0.162 adjusted R^2 of the base model reported in Table 4. Including the $BD\ Accruals_{i,t}$ variable in column 2 increases the adjusted R^2 to 0.235, which is a 13.5% increase relative to the base model reported in column 1. The coefficient on the $BD\ Accruals_{i,t}$ variable equals 0.310 and is significant at the 1% level, suggesting that the MD&A provides information useful for understanding accruals beyond their reversal in the following year. We note, however, that the percentage increase in the adjusted R^2 of the model (13.5%) is lower than the percentage increase in the adjusted R^2 of the model when future cash flows are not included (24.1%). This suggests that the MD&A at least partially explains cash flows that are realized in the following year.

7.3. Timeliness of Big Data Accruals

Our objective is to assess the narrative content of the MD&A in explaining the accrual generation process, regardless of the accrual's timeliness (i.e., whether the accrual persists from the prior period or is new to the current period). However, understanding whether $Big\ Data\ Accruals_{i,t}$ is able to identify accrual innovations can provide insight into the nature of the MD&A narrative. In particular, if accruals change little year-to-year and MD&A contains description that varies little, investors might find little timely insight by reading MD&A despite a high correlation between MD&A word counts and accrual levels.

To examine whether *Big Data Accruals*$_{i,t}$ identify timely accruals, we add the prior fiscal period accruals (*Accruals*$_{i,t-1}$) in Equation 3 as an independent variable, effectively converting our levels analysis to a changes analysis of accruals. Because there is positive serial-correlation in accruals, the portion of firm-level accruals explained by prior fiscal year accruals (*Accruals*$_{i,t-1}$) identifies the more persistent portion of accruals from the prior period. In Column 1 of Panel C in Table 9, we re-estimate the empirical model from Equation 2 (excluding *Big Data Accruals*$_{i,t}$) after including the *Accruals*$_{i,t-1}$ variable in the model.[25] Consistent with expectations, we find that the regression in Column 1 explains a much greater portion of accruals when compared to the portion of accruals explained in Column 1 in Table 4. In fact, the adjusted R^2 increases by approximately 24.7% after including *Accruals*$_{i,t-1}$ in the model specification.

Column 2 presents the results after including the *Big Data Accruals*$_{i,t}$ variable in the model. We note that the coefficient on the *Big Data Accruals*$_{i,t}$ variable continues to be positive and significant, suggesting that our method identifies a statistically significant portion of timely accruals. As expected, the increase in R^2 from Column 2 to Column 1 is approximately 4.0% and much smaller than the percentage increase in R^2 reported in Table 4. Since the percentage increase in R^2 is much greater in Table 4 than in Panel C of Table 9, this evidence suggests that our method is more effective at identifying and explaining the more persistent (i.e., not timely)

[25] We augment the Dechow and Dichev (2002) model with the *Accrual*$_{i,t-1}$ variable rather than examining the change in accruals for several reasons. First, we have an established model for explaining firm-level accruals. By including the prior fiscal year accruals, we are effectively changing our empirical model from a levels analysis to a changes analysis of accruals, without developing an explanatory model for accrual changes. Second, if we were to examine the change in accruals, we would effectively be assuming that the coefficient on the *Accruals*$_{i,t-1}$ variable is equal to one. By including the *Accruals*$_{i,t-1}$ variable as an additional independent variable, we allow the model to estimate the persistence of accruals from the prior period.

component of accruals. Nevertheless, the method in this paper explains a portion of timely accruals (i.e., changes to accruals).

7.4. Big Data Future Cash Flows Prediction

The SEC (2002) notes that the MD&A should provide information about the quality and variability of earnings and cash flow so that "investors can ascertain the likelihood that past performance is indicative of future performance." Therefore we also use support-vector regressions to assess the extent to which the MD&A contains information useful for predicting future cash flows. We replicate our support-vector regression procedure replacing $Accruals_{i,t}$ with $CFO_{i,t+1}$ to obtain an out-of-sample predicted value for $CFO_{i,t+1}$, which we label $BD\ CFO_{i,t+1}$. We then estimate the following model, similar to Barth, Cram, and Nelson (2001), to assess the predictive ability of the $BD\ CFO_{i,t+1}$ variable.

$$CFO_{i,t+1} = \alpha_0 + \alpha_1\, CFO_{i,t} + \alpha_2\, \Delta REC_{i,t} + \alpha_3\, \Delta INV_{i,t} + \alpha_4\, \Delta AP_{i,t} + \alpha_5\, \Delta TAX_{i,t} + \alpha_6\, \Delta OTH_{i,t} + \alpha_7\, Dep_{i,t} + \alpha_8\, BD\ CFO_{i,t+1} + \varepsilon_{i,t} \quad (10)$$

The $\Delta REC_{i,t}$, $\Delta INV_{i,t}$, $\Delta AP_{i,t}$, $\Delta TAX_{i,t}$, and $\Delta OTH_{i,t}$ variables are from the statement of cash flows and equal changes in receivables, inventory, accounts payable, taxes payable, and other net assets and liabilities, respectively, scaled by lagged total assets for firm *i* in year *t*. The $Dep_{i,t}$ variable equals depreciation expense scaled by lagged total assets for firm *i* in year *t*. Panel D of Table 9 presents the results using Equation 10. In column 1, we exclude the $BD\ CFO_{i,t+1}$ variable and obtain an adjusted R^2 of 0.538. We include the $BD\ CFO_{i,t+1}$ variable in column 2 and find that the adjusted R^2 increases to 0.566, which corresponds to a 5.2% increase relative to the base model. The coefficient on the $BD\ CFO_{i,t+1}$ variable equals 0.358 and is significant at the 1% level. Overall, these results suggest that the MD&A contains useful information for predicting future cash flows.

7.5. Conference Calls Analysis

We next apply our big data method to earnings conference calls to compare the narrative content of the MD&A to the conference call. Research discusses the benefits of studying conference calls relative to formal disclosures such as 10-Ks and press releases (Matsumoto, et al 2011; Larcker and Zakolyukina 2012). For example, 10-K disclosures are relatively constant over time while conference calls are more variable. In addition, conference calls are more spontaneous than 10-Ks providing an opportunity for managers to convey useful information. Research suggests that conference calls provide incremental information relative to the earnings press release (Matsumoto, et al 2011). We therefore test whether managers provide information during conference calls that is useful for understanding current accruals. Our method also allows us to compare the narrative content of the MD&A to that provided during the conference call.

We obtain 55,869 quarterly earnings conference call transcripts from Factiva's FD Wire over the period from 2002 to 2013 representing 18,232 firm-years in our sample, which equates to approximately 3 calls per firm-year. We aggregate the quarterly transcripts by firm at the annual level and obtain counts of all one- and two-word phrases, removing highly frequent and highly infrequent words following the procedure outlined above. We then use the resulting dataset to predict accruals using the SVR procedure by industry and by year. We then compute the average predicted accrual generated by estimating SVR by industry and by year and call the resulting variable $\textit{BD Accruals (Conf Call)}_{i,t}$.

For comparison purposes, we re-estimate Equation 2 for the sub-sample of firm-years for which we were able to obtain conference call transcripts. We present the results in Column 1 of Table 10. The adjusted R^2 is equal to 0.159, which is slightly lower than that reported for the full sample in Table 4. In Column 2, we include the $\textit{BD Accruals}_{i,t}$ variable generated by the SVR procedure using the MD&A. The adjusted R^2 is equal to 0.180, an increase of 0.021, or approximately 13.2 percent, relative to the base model. In Column 3, we include the *BD Accruals (Conf Call)$_{i,t}$* variable generated by the SVR procedure using the conference call transcripts. The adjusted R^2 is equal to 0.179, an increase of 0.020, or approximately 12.6 percent, relative to the base model, and is lower than that obtained from the MD&A. In Column 4, we include both the *BD Accruals$_{i,t}$* variable and the *BD Accruals (Conf Call)$_{i,t}$* variable and find an adjusted R^2 of 0.188 representing an increase of 0.029, or approximately 18.2%, relative to the base model. These results suggest that the MD&A is as useful in understanding current accruals relative to the conference call. However, both the conference call and the MD&A contain incremental useful information for understanding accruals relative to each other as evidenced by the greatest increase in explanatory power when both predicted accrual variables are included in the model.

7.6. Time-series Analysis

Our method allows us to examine whether there is a systematic change in the information content of the MD&A over time. The SEC issued new MD&A guidelines in 2003, encouraging firms to reduce boilerplate disclosure and improve the narrative content of the MD&A. The Sarbanes-Oxley Act also added an additional layer of certification requiring the CEO and CFO to certify that their financial statements and footnote disclosures fairly represent their financial conditions and results of operations. Li (2010) provides a test of the change in narrative content over time and finds a marginally significant decrease, though the result is not robust to multiple specifications. His result suggests that the SEC's disclosure regulation has been ineffective.

We test the change in the narrative content of the MD&A over time by estimating Equations 2 and 3 by year and computing the incremental adjusted R^2 each year. The incremental adjusted R^2 provides an indication of the increase in the explanatory power of the accrual model when including the big data predicted accrual generated from the MD&A. We restrict our estimation to firms with at least 10 firm-years throughout the 20-year sample period, though our results are robust to including all firm-years. Figure 1 plots the incremental adjusted R^2 for each year. We observe a discernable increase in the incremental adjusted R^2 over time. To determine whether the increase in the narrative content of the MD&A is statistically significant, we estimate a time-series regression (N=20) with the incremental adjusted R^2 as the dependent variable and a trend variable equal to 1 in 1994, equal to 2 in 1995, etc. The coefficient on the trend variable is equal to 0.0012 with a t-stat of 2.90, suggesting that the narrative content of the MD&A has improved over time. Thus, in contrast to Li (2010), we find evidence that the SEC has at least partially achieved its intended objective to improve the narrative content of the MD&A.

8. Conclusion

We develop and explore the empirical properties of a technique that converts unstructured and qualitative disclosures into a prediction of a firm-level fundamental. In particular, we use word counts found in the MD&A to predict current firm-year accruals, which we call *big data accruals*, using support-vector regressions. Use of support-vector regressions is necessary given the high number of unique words found in the MD&A relative to the number of firm-year observations available. We find that *big data accruals* help to explain accruals beyond the Dechow and Dichev (2002) accrual model, as modified by McNichols (2002) and Ball and Shivakumar (2005). In fact, we find that our ability to explain accruals increases 24.1%, relative to the modified Dechow and Dichev (2002) model.

In additional cross-sectional tests, we find that big data accruals are less helpful in explaining accruals when the 10-K is less readable, suggesting that big data techniques are not able to parse managerial obfuscation. We also find that *big data accruals* are useful in predicting accruals when earnings are more difficult to predict using fundamentals, which is consistent with managers providing additional information to better explain accruals when earnings are difficult to predict.

We examine the type of accruals that we are identifying using support-vector regressions to predict accruals. We find that the accruals predicted using the MD&A (i.e., *big data accruals*) are more persistent than those accruals that are not, providing insights useful in equity valuation (Dechow et al., 2010). After separating accruals into discretionary and nondiscretionary accruals, we find similar evidence that big data technologies can identify the more persistent component.

Lastly, we find evidence consistent with *big data accruals* helping to predict cash flows realized in the next fiscal period. We further our investigation by examining whether the MD&A can directly predict future cash flows rather than the accruals that result in future cash flows. Using the Barth, Cram, and Nelson (2001) paper to model future cash flows, we find that support-vector regressions can be used to further our understanding of future cash flows using the words and phrases found in the MD&A for the current reporting period.

The prior literature uses textual analysis to develop measures for specific disclosure characteristics (e.g., tone, scripting, readability) as well as identify word groups that are assigned by the researcher to clearly identified fundamentals (e.g., firm risk). The method we use in this paper differs prior methods in two key ways. First, our method allows the computer to identify patterns in the narrative that occur in conjunction with firm fundamentals that may not be a priori known by the researcher. Second, our method does not rely on the manual classification of words, allowing the estimation process to be easily applied across languages and contexts. Our approach offers a path for developing a measure of narrative content for qualitative, unstructured disclosure.

References

Allee, K. and M. DeAngelis, 2015, The Structure of Voluntary Disclosure Narratives: Evidence from Conference Calls, Journal of Accounting Research, Forthcoming

Autor, D., 2014. Polanyi's paradox and the shape of employment growth. MIT, NBER, JPAL working paper.

Ball, R., and L. Shivakumar. 2005. Earnings quality in UK private firms: comparative loss recognition timeliness. *Journal of Accounting and Economics* 39, 83–128.

Barth, M., D. Cram, and K. Nelson. 2001. Accruals and the prediction of future cash flows. *The Accounting Review* 76(1), 27–58.

Bloomfield, R. 2002. The "incomplete revelation hypothesis" and financial reporting. *Accounting Horizons* 16, 233–243.

Bao, Y., and Datta, A., 2014, Simultaneously discovering and quantifying risk types from textual risk disclosures, Management Science 60: 1371-1391.

Bonsall, S., A. Leone, and B. Miller. 2015. A Plain English Measure of Financial Reporting Readability. Working paper.

Brown, S., and J. Tucker. 2011. Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research* 49(2), 309–346.

Campbell, J., Chen, H., Dhaliwal, D., Lu, H., and Steele, L., 2014, The disclosure content of mandatory risk factor disclosures in corporate filings, Review of Accounting Studies 19: 396-455.

Cherkassky, V., and Y. Ma. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* 17, 113–126.

Core, J. 2001, A Review of the Empirical Disclosure Literature: Discussion. Journal of Accounting and Economics 31: 441–56.

Dechow, P., and I. Dichev. 2002. The Quality of Accruals and Earnings: The Role of Accrual Estimation Errors. *The Accounting Review* 77 (supplement), 35–59.

Dechow, P., W. Ge, and C. Schrand. 2010. Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics* 50, 344–401.

The Economist. Turning the Tables. September 3, 2014.

The Economist. Why mobile data to prevent Ebola has not yet been released. November 9, 2014.

Feldman R, Govindaraj S, Livnat J, Segal B, 2010, Management's tone change, post earnings announcement drift and accruals. Review Accounting Studies 15:915–953.

Jennings, J., Lee, J., and D. Matsumoto. 2015. The effect of industry co-location on analysts' information acquisition costs. Working paper.

Kakutani, M. Watched by the Web: Surveillance Is Reborn. *The New York Times*. June 10, 2013.

Kogan, S., D. Levin, B. Routledge, J. Sagi, and N. Smith. Predicting Risk from Financial Reports with Regression. NAACL-HLT 2009, Boulder, Colo., May–June 2009.

Kothari, S. P., X. Li, and Short, J., 2009, The Effect of Disclosures by Management, Analysts, and Financial Press on the Equity Cost of Capital: A Study Using Content Analysis. Accounting Review 84: 1639–70.

Kravet, T., Muslu, V., 2011, Textual risk disclosures and investors' risk perceptions. Review Accounting Studies 18:1088–1122.

Lang, M., and L. Stice-Lawrence. 2014. Textual Analysis and International Financial Reporting: Large Sample Evidence. Working paper.

Larker and Zakolyukina, 2012, Detecting Deceptive Discussions in Conference Calls, Journal of Accounting Research 50: 495-540.

Lee, J. 2015. Scripted Earnings Conference Calls as a Signal of Future Firm Performance. Working paper.

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

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45, 221–447.

Li, F., 2010, The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach, Journal of Accounting Research 48: 1049-1102.

Li, F., Lundholm, R., and M. Minnis, 2013, A Measure of Competition Based on 10-K Filings, Journal of Accounting Research 51(2): 399-436

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

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

Loughran, T. and B. McDonald. 2015. Measuring Readability in Financial Disclosures. *Journal of Finance*, forthcoming.

Manela, A., and A. Moreira. 2014. News Implied Volatility and Disaster Concerns. Working paper.

Matsumoto, D., Pronk, M., and Roelofsen, E., 2011, What makes conference calls useful? The information content of managers presentations and analysts' discussion sessions, The Accounting Review 86: 1383-1414.

McNichols, M. 2002. Discussion of The Quality of Accruals and Earnings: The Role of Accrual Estimation Errors. *The Accounting Review* 77(supplement), 61–69.

Muslu, V., S. Radhakrishnan, K.R. Subramanyam, and D. Lim, 2014. Forward-Looking MD&A Disclosures and the Information Environment, Management Science.

Pentland, A. 2014. *Social Physics How Good Ideas Spread—The Lessons From a New Science*. The Penguin Press. New York, NY.

Petersen, K., Schmardebeck, R., and Wilks, 2015, The earnings quality and information processing effects of accounting consistency. Forthcoming The Accounting Review.

Rogers, J., and A. Van Buskirk. 2009. Shareholder litigation and changes in disclosure behavior. *Journal of Accounting and Economics* 47, 136–156.

Rogers, J., Van Buskirk, A., Zechman, S., 2011 Disclosure tone and shareholder litigation. The Accounting Rev. 86: 2155–2183.

Sloan, R. 1996. Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings? *The Accounting Review* 71(3), 289–315.

Tetlock, P., 2007, Giving content to investor sentiment: The role of media in the stock market, Journal of Finance 62, 1139-1168.

Tetlock, P., Saar-Tsechansky, M., Sofus, M., 2008, More than words: Quantifying language to measure firms' fundamentals, *The Journal of Finance* 63(3): 1437-1467.

Xie, H. 2001. The Mispricing of Abnormal Accruals. *The Accounting Review* 76(3), 357–373.

APPENDIX A
Support-Vector Regression

We apply support-vector regressions (SVR) to a linear model, which is depicted in Equation 1.

$$y_t \ = \ \omega_0 + \boldsymbol{\omega} \cdot \boldsymbol{x_t} + \ e_t \qquad\qquad t = 1 \ldots . T \tag{1}$$

The ω parameter represents a vector of K weights (i.e., regression coefficients), one for each x independent variable. In our sample, the independent variables represent the normalized word and phrase counts included in the MD&A. Since the number of independent variables exceeds the number of observations included in our sample, we cannot estimate Equation 1 using ordinary least squares (OLS) and must use alternative statistical methods, such as SVR.

Rather than minimizing the sum of the squared residuals, as is done in OLS, SVR chooses the weights by minimizing the following equation.

$$minimize \ \frac{1}{2}\|\omega\|^2 + \ C \sum_{t \in Train}^{n} g_\epsilon(e_t) \tag{2}$$

SVR works to simultaneously minimize the $\|\omega\|$ (i.e., the norm of the weights vector) and the prediction errors that are outside the "ϵ-insensitivity" margin. Since SVR allows for many independent variables, potentially leading to over-fitting the data, SVR includes the $\|\omega\|$ to "penalize" the minimization equation depicted in Equation 2 and reduces the likelihood that the model over-fits the data. In other words, as the weights get larger or more independent variables receive weights, the minimization equation requires smaller prediction errors. SVR essentially trades off the prediction error minimization with model parsimony to minimize Equation 2.

The $g_\epsilon(e) = \max\{0, |e| - \epsilon\}$ represents prediction errors that are outside the "ϵ-insensitivity" margin. In other words, the prediction error is included in the minimization problem only when the absolute value of the prediction error ($|e|$) is greater than ϵ; otherwise, the prediction error equals 0. The prediction error equals the residual from Equation 1.

Following Cherkassky and Ma (2004), we calculate "ϵ-insensitivity" margin using Equation 3.

$$\epsilon = 3\sigma\sqrt{\frac{\ln n}{n}} \tag{3}$$

The "ϵ-insensitivity" margin is calculated using the "nearest neighbor" technique, which matches each observation in the training sample to a predetermined number of "neighbors" or matched firms based on the $\boldsymbol{x}$ vector of independent variables. The average y value of the "nearest neighbors" or matched firms represents the predicted y value. The σ parameter is then calculated as the standard deviation of the difference between the y values and the predicted y values for all observations in the training sample. The n parameter represents the number of observations included in the training sample. Following Manela and Moreira (2004), we choose five "neighbors" to estimate the predicted y value.

The C parameter represents the "regularization" parameter and is also estimated in Cherkassky and Ma (2004) using the following equation.

$$C = max\left(\left|\bar{y} + 3\sigma_y\right|, \left|\bar{y} - 3\sigma_y\right|\right) \tag{4}$$

The σ_y parameter represents the standard deviation of the y variables. The $\bar{y}$ value represents the average y value in for all observations in the training sample. The C parameter essentially shifts the emphasis of the minimization equation (Equation 2). For example, as the C parameter gets larger, more emphasis is shifted toward minimizing the prediction errors rather than minimizing $||\omega||$.

The ω vector is a weighted average of regressors and is depicted using the below equation and the solution to Equation 2.

$$\omega_{SVR} = \sum_{t \in Train} (\hat{\alpha}_t^* - \hat{\alpha}_t)\, x_t \tag{5}$$

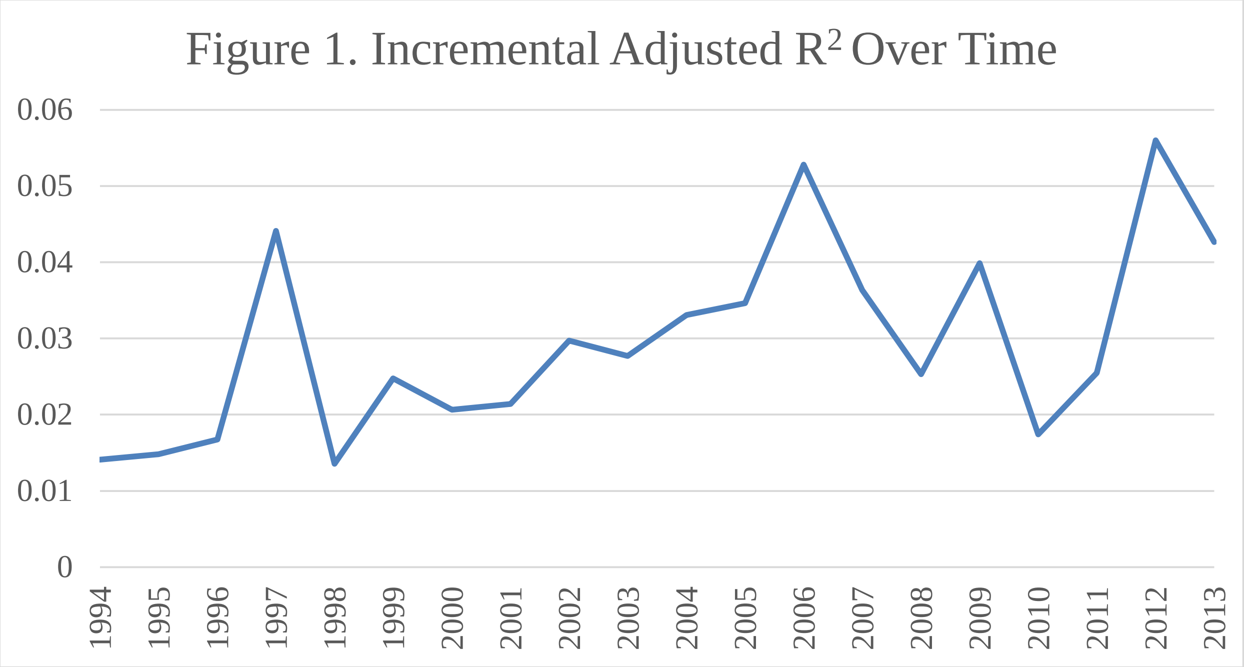
Only some of the dual weights α_t are nonzero, suggesting that SVR chooses a subset of independent variables to explain the variation in the dependent variable (i.e., y). The subset of independent variables is determined based on the influential observations (i.e., most useful in predicting the y values) in the training sample. As a result, not all observations are used when determining the weights for each variable in the x vector.

APPENDIX B
Variable Definitions

Accruals$_{i,t}$	Working capital accruals, defined as −1 multiplied by the sum of Compustat items RECCH, INVCH, APALCH, TXACH, and AOLOCH scaled by total assets at the beginning of the year.
BD Accruals$_{i,t}$	Big data accruals, defined as the mean predicted value obtained from two separate estimations (by industry and by year) of the SVR procedure described in the body of the paper. The SVR procedure includes *Accruals*$_{i,t}$ as the dependent variable in the regression model and the counts of all one- and two-word phrases in the MD&A as independent variables. The SVR procedure involves estimating coefficients on the independent variables over an estimation window prior to year *t* and then applying the estimated coefficients to values in year *t* to obtain out-of-sample predicted values for accruals in year *t*.
BD Accruals (Excl Acc Phrases)$_{i,t}$	Big data accruals, defined as the mean predicted value obtained from two separate estimations (by industry and by year) of the SVR procedure described in the body of the paper. The SVR procedure includes *Accruals*$_{i,t}$ as the dependent variable in the regression model and the counts of all one- and two-word phrases in the MD&A as independent variables, excluding any phrases containing any of the following words: asset*, liabilit*, receivable*, inventor*, payable*, and prepaid*. The SVR procedure involves estimating coefficients on the independent variables over an estimation window prior to year *t* and then applying the estimated coefficients to values in year *t* to obtain out-of-sample predicted values for accruals in year *t*.
BD CFO$_{i,t+1}$	Big data future cash flows, defined as the mean predicted value obtained from two separate estimations (by industry and by year) of the SVR procedure described in the body of the paper. The procedure includes *CFO*$_{i,t+1}$ as the dependent variable in the regression model and the counts of all one- and two-word phrases in the MD&A as independent variables. It involves estimating coefficients on the independent variables over an estimation window prior to year *t* and then applying the estimated coefficients to values in year *t* to obtain out-of-sample predicted values for future cash flows in year *t*.
BD Disc Accruals$_{i,t}$	Big data discretionary accruals, defined as the mean predicted value obtained from two separate estimations (by industry and by year) of the SVR procedure described in the body of the paper. The SVR procedure includes *Disc Accruals*$_{i,t}$ as the dependent variable in the regression model and the counts of all one- and two-word phrases in the MD&A as independent variables. The SVR procedure involves estimating coefficients on the independent variables over an

	estimation window prior to year *t* and then applying the estimated coefficients to values in year *t* to obtain out-of-sample predicted values for discretionary accruals in year *t*.
BD Non-Disc Accruals$_{i,t}$	Big data nondiscretionary accruals, defined as the mean predicted value obtained from two separate estimations (by industry and by year) of the SVR procedure described in the body of the paper. The SVR procedure includes *Non-Disc Accruals*$_{i,t}$ as the dependent variable in the regression model and the counts of all one- and two-word phrases in the MD&A as independent variables. The SVR procedure involves estimating coefficients on the independent variables over an estimation window prior to year *t* and then applying the estimated coefficients to values in year *t* to obtain out-of-sample predicted values for nondiscretionary accruals in year *t*.
CFO$_{i,t}$	Cash flow from operations, defined as Compustat item OANCF scaled by total assets at the beginning of the year.
Disc Accruals$_{i,t}$	Discretionary accruals, defined as *Accruals*$_{i,t}$ less *Non-Disc Accruals*$_{i,t}$.
Earn Vol$_{i,t}$	An indicator variable equal to 1 if earnings volatility is above the sample median and 0 otherwise, where earnings volatility is defined as the standard deviation of seasonally adjusted earnings over the preceding 16 quarters (requiring a minimum of eight prior quarters).
File Size$_{i,t}$	An indicator variable equal to 1 if the residual file size of the 10-K filing is above the sample median and equal to 0 otherwise, where the file size includes only the 10-K document and the EX-13 (i.e., annual report) and excludes all other exhibits and images. The residual file size is the residual obtained from regressing the file size variable on an indicator variable equal to 1 if the document is HTML and equal to 0 if the document is text-based.
Fog$_{i,t}$	An indicator variable equal to 1 if the Fog index of the MD&A is above the sample median and 0 otherwise.
Neg CFO$_{i,t}$	An indicator variable equal to 1 if *CFO*$_{i,t}$ is less than 0 and 0 otherwise.
Non-BD Accruals$_{i,t}$	Non-big data accruals, defined as *Accruals*$_{i,t}$ less *BD Accruals*$_{i,t}$.
Non-BD Disc Accruals$_{i,t}$	Non-big data discretionary accruals, defined as *Disc Accruals*$_{i,t}$ less *BD Disc Accruals*$_{i,t}$.
Non-BD Non-Disc Accruals$_{i,t}$	Non-big data nondiscretionary accruals, defined as *Non-Disc Accruals*$_{i,t}$ less *BD Non-Disc Accruals*$_{i,t}$.
Non-Disc Accruals$_{i,t}$	Nondiscretionary accruals, computed by estimating the modified Dechow and Dichev (2002) model (Equation 2) by industry/year in year *t−1* and then applying the estimated coefficients to observations in year *t* to obtain a predicted value for the nondiscretionary component of accruals in year *t*.
PPE$_{i,t}$	Gross property, plant, and equipment, defined as Compustat item PPEGT scaled by total assets at the beginning of the year.
Pre-Suit$_{i,t}$	An indicator variable equal to 1 if a class action lawsuit is filed against the firm in year *t+1* and 0 otherwise.

$ROA_{i,t}$	Return on assets, defined as the sum of $Accruals_{i,t}$ and $CFO_{i,t}$.
$Suit_{i,t}$	An indicator variable equal to 1 if a class action lawsuit is filed against the firm in year *t* and 0 otherwise.
$Word\ Length_{i,t}$	An indicator variable equal to 1 if the word length of the MD&A is above the sample median and 0 otherwise.
$\Delta AP_{i,t}$	Change in accounts payable, defined as Compustat item APALCH scaled by total assets at the beginning of the year.
$\Delta INV_{i,t}$	Change in inventory, defined as −1 multiplied by Compustat item INVCH scaled by total assets at the beginning of the year.
$\Delta INV_{i,t}$	Change in other net assets and liabilities, defined as −1 multiplied by Compustat item AOLOCH scaled by total assets at the beginning of the year.
$\Delta REC_{i,t}$	Change in receivables, defined as −1 multiplied by Compustat item RECCH scaled by total assets at the beginning of the year.
$\Delta Sales_{i,t}$	Change in sales, defined as the change in Compustat item REVT less the change in item RECT scaled by total assets at the beginning of the year.
$\Delta TAX_{i,t}$	Change in taxes payable, defined as Compustat item TXACH scaled by total assets at the beginning of the year.



This figure plots the yearly incremental adjusted R^2 of estimating Equation 3 relative to Equation 2 from 1994 to 2013. A time-series regression of the incremental adjusted R^2 on a trend variable equal to 1 in 1994, equal to 2 in 1995, etc. yields a trend-variable coefficient of 0.0012 with a t-statistic equal to 2.90, suggesting an increasing trend over time.

TABLE 1
Highly Predictive Words and Phrases

This table presents the words and phrases identified in the MD&A with the largest estimated coefficients from SVR regressions. Panel A presents the important words and phrases obtained from estimating the support-vector regressions by industry (four example industries are presented), and Panel B presents the important words and phrases obtained from estimating the support-vector regressions cross-sectionally by year. '_' indicates the term appears with a frequently repeated word (e.g, the, a, and, etc.). Estimated coefficients for each word and phrase are also presented. Coefficients are multiplied by word counts to produce the big data accrual estimate as a percentage of lagged total assts.

Panel A: Industry Estimation

Communications Equipment (GICS = 452010)		*Computers & Peripherals (GICS = 452020)*		*Metals and Mining (GICS = 151040)*		*Biotechnology (GICS = 352010)*	
Positive Words and Phrases	**Coeff.**	**Positive Words and Phrases**	**Coeff.**	**Positive Words and Phrases**	**Coeff.**	**Positive Words and Phrases**	**Coeff.**
accounts	0.090	increased	0.086	resulted _	0.052	income	0.148
cash _	0.075	may	0.062	resulted	0.052	through	0.102
group	0.072	used	0.061	selling	0.040	below	0.092
access	0.066	inc	0.057	agreement	0.040	paid	0.091
_ equipment	0.062	before	0.052	_ income	0.038	control	0.090
quarter ended	0.061	net income	0.051	certain	0.037	february	0.088
system	0.060	_ revenues	0.051	continued	0.037	merger	0.079
development expenses	0.058	_ exercise	0.050	_ non	0.034	_ lease	0.079
_ primarily	0.058	borrowings	0.049	_ certain	0.033	_ no	0.073
value _	0.058	during _	0.049	_ agreement	0.032	full	0.070
Negative Words and Phrases	**Coeff.**	**Negative Words and Phrases**	**Coeff.**	**Negative Words and Phrases**	**Coeff.**	**Negative Words and Phrases**	**Coeff.**
loss	-0.073	inventory	-0.080	plant	-0.091	received _	-0.132
going	-0.068	reduced	-0.070	phase	-0.055	deficit	-0.123
fees _	-0.064	capital	-0.066	plant _	-0.053	date _	-0.108
_ going	-0.059	decreased	-0.066	commercial	-0.052	date	-0.100
decreased	-0.059	information	-0.065	during _	-0.046	stage	-0.098
provided	-0.058	quarter _	-0.063	no	-0.043	deficit _	-0.098
_ decreased	-0.058	_ january	-0.059	_ loss	-0.041	_ recognized	-0.095
adverse	-0.056	internet	-0.059	_ no	-0.041	increase _	-0.092
plan	-0.056	_ technology	-0.057	_ aluminum	-0.036	estimates _	-0.087
financing	-0.055	quarter	-0.057	loss _	-0.035	_ employee	-0.081

Panel B: Cross-sectional Estimation by Year

Positive Words and Phrases	**Coeff.**	**Positive Words and Phrases**	**Coeff.**	**Negative Words and Phrases**	**Coeff.**	**Negative Words and Phrases**	**Coeff.**
titles	0.092	net profit	0.055	capital deficit	-0.117	_ no	-0.071
receivable	0.082	discounts	0.055	deficit	-0.115	delay	-0.068
private placement	0.079	expand	0.054	lack	-0.095	deficit _	-0.067
nasdaq	0.072	cash used	0.053	_ lack	-0.093	going	-0.065
placement	0.071	_ working	0.051	lack _	-0.086	b preferred	-0.065
expenses decreased	0.066	combination	0.050	accrued	-0.085	successful	-0.064
remaining	0.065	expects _	0.050	net loss	-0.081	promissory notes	-0.063
income loss	0.063	receivables _	0.050	_ default	-0.078	_ officers	-0.062
marketing expenses	0.061	proceeds _	0.049	stage	-0.078	creditors	-0.059
commissions	0.058	loan _	0.049	default	-0.078	substantial doubt	-0.058

TABLE 2
Descriptive Statistics

This table presents the descriptive statistics for the variables used in the main empirical analyses. All variables are defined in Appendix B. All continuous variables are winsorized at the 1st and 99th percentiles. The sample spans 1994 to 2013 and includes 71847 observations.

Panel A: Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
$Accruals_{i,t}$	-0.007	0.190	-1.278	-0.029	0.006	0.045	0.534
$BD\ Accruals_{i,t}$	-0.001	0.110	-0.527	-0.042	0.008	0.054	0.285
$CFO_{i,t-1}$	-0.085	0.630	-4.556	-0.058	0.061	0.136	0.552
$CFO_{i,t}$	-0.063	0.533	-3.800	-0.053	0.061	0.134	0.506
$Neg\ CFO_{i,t}$	0.331	0.471	0.000	0.000	0.000	1.000	1.000
$\Delta Sales_{i,t}$	0.096	0.406	-1.306	-0.034	0.051	0.188	2.128
$PPE_{i,t}$	0.565	0.504	0.000	0.201	0.423	0.785	2.801

TABLE 3
Correlations

This table presents the Pearson and Spearman correlations for the variables used in the main empirical analyses. Pearson (Spearman) correlations are above (below) the diagonal. Values are bolded if significant at the 5 % level or lower. All variables are defined in Appendix B. All continuous variables are winsorized at the 1st and 99th percentiles.

	Accruals$_{i,t}$	*BD Accruals*$_{i,t}$	*CFO*$_{i,t-1}$	*CFO*$_{i,t}$	*Neg CFO*$_{i,t}$	*ΔSales*$_{i,t}$	*PPE*$_{i,t}$
Accruals$_{i,t}$	**1**	**0.31**	**0.29**	**0.26**	**-0.03**	**0.10**	**-0.06**
BD Accruals$_{i,t}$	**0.18**	**1**	**0.25**	**0.33**	**-0.18**	**0.10**	**-0.05**
CFO$_{i,t-1}$	**0.09**	**0.02**	**1**	**0.66**	**-0.41**	**-0.04**	**0.05**
CFO$_{i,t}$	**-0.14**	**0.15**	**0.68**	**1**	**-0.51**	**0.05**	**-0.01**
Neg CFO$_{i,t}$	**0.08**	**-0.14**	**-0.58**	**-0.82**	**1**	**-0.09**	**-0.16**
ΔSales$_{i,t}$	**0.18**	**0.17**	**0.11**	**0.25**	**-0.17**	**1**	**0.07**
PPE$_{i,t}$	**0.00**	**-0.01**	**0.24**	**0.25**	**-0.23**	**0.07**	**1**

TABLE 4
Working Capital Accruals Prediction Model

This table includes all firm/year observations from 1994 to 2013 with sufficient data to calculate the dependent and independent variables. The dependent variable is working capital accruals ($Accruals_{i,t}$). In Panel A, a pooled model is estimated for all firms in the sample with industry and year fixed effects included as additional independent variables (unreported) and standard errors clustered by firm. In Panel B, the model is estimated separately for each GICS industry. The mean [median] (t-statistic) for each variable is presented based on the distribution of the 56 coefficients obtained from the industry-specific regressions. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Pooled Regression

	[1]	[2]
Intercept	0.078***	0.070***
	(11.199)	(11.487)
$CFO_{i,t-1}$	0.070***	0.066***
	(16.569)	(15.720)
$CFO_{i,t}$	-0.349***	-0.345***
	(-25.880)	(-26.816)
$\Delta Sales_{i,t}$	0.069***	0.060***
	(17.101)	(15.210)
$PPE_{i,t}$	-0.014***	-0.010***
	(-3.487)	(-2.749)
$Neg\ CFO_{i,t}$	0.019***	0.024***
	(7.559)	(10.385)
$CFO_{i,t}$ * $Neg\ CFO_{i,t}$	0.427***	0.402***
	(28.951)	(28.914)
$BD\ Accruals_{i,t}$		0.370***
		(21.632)
#OBS	71,847	71,847
Adjusted R^2	0.162	0.201

Panel B: Industry-specific Regression

	[1]	[2]
Intercept	0.033***	0.028***
	[0.032]	[0.027]
	(12.82)	(11.86)
$CFO_{i,t-1}$	0.101***	0.1***
	[0.076]	[0.077]
	(9.86)	(9.31)
$CFO_{i,t}$	-0.376***	-0.377***
	[-0.359]	[-0.364]
	(-16.69)	(-16.78)
$\Delta Sales_{i,t}$	0.068***	0.06***
	[0.071]	[0.066]
	(9.41)	(8.91)
$PPE_{i,t}$	-0.013***	-0.01***
	[-0.009]	[-0.006]
	(-1.65)	(-1.33)
$Neg\ CFO_{i,t}$	0.019***	0.023***
	[0.018]	[0.019]
	(4.76)	(6.02)
$CFO_{i,t}$ * $Neg\ CFO_{i,t}$	0.429***	0.412***
	[0.429]	[0.412]
	(17.04)	(16.91)
$BD\ Accruals_{i,t}$		0.286***
		[0.317]
		(12.97)
# Industries	56	56
Adjusted R^2	0.253	0.283
	[0.202]	[0.237]

TABLE 5
Earnings Persistence Model

This table includes all firm/year observations from 1994 to 2013 with sufficient data to calculate the dependent and independent variables. The dependent variable is return on assets in year t+1 (*$ROA_{i,t+1}$*). All variables are defined in Appendix B. Industry and year fixed effects are included as additional independent variables (unreported). Standard errors are clustered by firm. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]
Intercept	0.077***	0.077***	0.077***	0.075***
	(9.808)	(9.767)	(9.680)	(9.451)
$CFO_{i,t}$	0.635***	0.633***	0.633***	0.631***
	(41.577)	(41.143)	(42.266)	(41.902)
$Accruals_{i,t}$	0.472***			
	(18.092)			
BD $Accruals_{i,t}$		0.518***		
		(19.814)		
Non-BD $Accruals_{i,t}$		0.466***		
		(17.375)		
Disc $Accruals_{i,t}$			0.507***	
			(20.728)	
Non-Disc $Accruals_{i,t}$			0.506***	
			(14.428)	
BD Disc $Accruals_{i,t}$				0.531***
				(19.169)
Non-BD Disc $Accruals_{i,t}$				0.501***
				(20.037)
BD Non-Disc $Accruals_{i,t}$				0.633***
				(15.962)
Non-BD Non-Disc $Accruals_{i,t}$				0.489***
				(13.512)
#OBS	54,110	54,110	54,110	54,110
Adjusted R^2	0.495	0.495	0.496	0.496

F-tests:	**F-stat**	**P-value**
BD $Accruals_{i,t}$ - Non-BD $Accruals_{i,t}$ = 0	7.85	0.005***
BD Disc $Accruals_{i,t}$ - Non-BD Disc $Accruals_{i,t}$ = 0	1.97	0.161
BD Non-Disc $Accruals_{i,t}$ - Non-BD Non-Disc $Accruals_{i,t}$ = 0	20.00	0.000***

TABLE 6
Working Capital Accruals Prediction Model with Litigation

This table includes all firm/year observations from 1994 to 2013 with sufficient data to calculate the dependent and independent variables. The dependent variable is working capital accruals ($Accruals_{i,t}$). All variables are defined in Appendix B. Industry and year fixed effects are included as additional independent variables (unreported). Standard errors are clustered by firm. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]
Intercept	0.038
	(0.323)
$CFO_{i,t-1}$	0.047***
	(3.806)
$CFO_{i,t}$	-0.334***
	(-8.380)
$\Delta Sales_{i,t}$	0.071***
	(5.379)
$PPE_{i,t}$	-0.008
	(-0.342)
$Neg\ CFO_{i,t}$	0.012
	(1.281)
$CFO_{i,t}$ * $Neg\ CFO_{i,t}$	0.256***
	(5.432)
$Suit_{i,t}$	-0.022***
	(-4.453)
$BD\ Accruals_{i,t}$	0.317***
	(4.378)
$BD\ Accruals_{i,t}$ * $Suit_{i,t}$	-0.229***
	(-2.715)
#OBS	1,708
Adjusted R^2	0.211

TABLE 7
Working Capital Accruals Prediction Model with Readability

This table includes all firm/year observations from 1994 to 2013 with sufficient data to calculate the dependent and independent variables. The dependent variable is working capital accruals (*Accruals*$_{i,t}$). All variables are defined in Appendix B. Industry and year fixed effects are included as additional independent variables (unreported). Standard errors are clustered by firm. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	Low Readability$_{i,t}$ = *Fog*$_{i,t}$	*Low Readability*$_{i,t}$ = *Word Length*$_{i,t}$	*Low Readability*$_{i,t}$ = *File Size*$_{i,t}$
Intercept	0.071***	0.067***	0.065***
	(11.561)	(11.008)	(10.496)
CFO$_{i,t-1}$	0.066***	0.065***	0.065***
	(15.714)	(15.372)	(15.618)
CFO$_{i,t}$	-0.345***	-0.338***	-0.341***
	(-26.847)	(-26.296)	(-26.478)
ΔSales$_{i,t}$	0.060***	0.059***	0.059***
	(15.190)	(15.037)	(15.102)
PPE$_{i,t}$	-0.010***	-0.009***	-0.009***
	(-2.781)	(-2.690)	(-2.673)
Neg CFO$_{i,t}$	0.024***	0.025***	0.024***
	(10.403)	(11.232)	(10.783)
CFO$_{i,t}$ * *Neg CFO*$_{i,t}$	0.402***	0.393***	0.397***
	(28.955)	(28.159)	(28.523)
Low Readability$_{i,t}$	-0.002	0.014***	0.008***
	(-1.591)	(7.566)	(5.366)
BD Accruals$_{i,t}$	0.367***	0.417***	0.419***
	(14.728)	(18.318)	(17.602)
BD Accruals$_{i,t}$ * *Low Readability*$_{i,t}$	0.005	-0.150***	-0.133***
	(0.141)	(-4.806)	(-4.380)
#OBS	71,847	71,847	71,847
Adjusted R^2	0.201	0.204	0.203

TABLE 8
Working Capital Accruals Prediction Model with Earnings Volatility

This table includes all firm/year observations from 1994 to 2013 with sufficient data to calculate the dependent and independent variables. The dependent variable is working capital accruals ($Accruals_{i,t}$). All variables are defined in Appendix B. Industry and year fixed effects are included as additional independent variables (unreported). Standard errors are clustered by firm. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]
Intercept	0.076***
	(12.472)
$CFO_{i,t-1}$	0.063***
	(15.042)
$CFO_{i,t}$	-0.340***
	(-26.569)
$\Delta Sales_{i,t}$	0.060***
	(15.227)
$PPE_{i,t}$	-0.009***
	(-2.709)
$Neg\ CFO_{i,t}$	0.032***
	(14.775)
$CFO_{i,t}$ * $Neg\ CFO_{i,t}$	0.394***
	(28.452)
$Earn\ Vol_{i,t}$	-0.024***
	(-19.007)
$BD\ Accruals_{i,t}$	0.203***
	(16.162)
$BD\ Accruals_{i,t}$ * $Earn\ Vol_{i,t}$	0.211***
	(8.767)
#OBS	71,847
Adjusted R^2	0.206

TABLE 9
Additional Tests

This table includes all firm/year observations from 1994 to 2013 with sufficient data to calculate the dependent and independent variables. The dependent variable is working capital accruals (*$Accruals_{i,t}$*) in Panels A to C and future cash flows (*$CFO_{i,t+1}$*) in Panel D. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Working Capital Accruals Prediction Excluding Accrual-based Words

	[1]
Intercept	0.070***
	(11.476)
$CFO_{i,t-1}$	0.066***
	(15.716)
$CFO_{i,t}$	-0.345***
	(-26.809)
$\Delta Sales_{i,t}$	0.060***
	(15.205)
$PPE_{i,t}$	-0.010***
	(-2.776)
Neg $CFO_{i,t}$	0.024***
	(10.342)
*$CFO_{i,t}$ * Neg $CFO_{i,t}$*	0.402***
	(28.915)
BD Accruals (Excl Acc Phrases)$_{i,t}$	0.368***
	(21.523)
#OBS	71,847
Adjusted R^2	0.201

Panel B: Working Capital Accruals Prediction Including Future Cash Flows

	[1]	[2]
Intercept	0.074***	0.068***
	(8.756)	(8.189)
$CFO_{i,t-1}$	0.055***	0.054***
	(38.114)	(37.750)
$CFO_{i,t}$	-0.380***	-0.375***
	(-46.550)	(-46.680)
$CFO_{i,t+1}$	0.113***	0.104***
	(59.106)	(54.815)
$\Delta Sales_{i,t}$	0.060***	0.053***
	(36.416)	(32.677)
$PPE_{i,t}$	-0.020***	-0.016***
	(-13.043)	(-10.721)
Neg $CFO_{i,t}$	0.034***	0.036***
	(17.700)	(19.275)
*$CFO_{i,t}$ * Neg $CFO_{i,t}$*	0.397***	0.378***
	(47.689)	(46.187)
BD $Accruals_{i,t}$		0.310***
		(49.286)
#OBS	65,941	65,941
Adjusted R^2	0.207	0.235

Panel C: Working Capital Accruals Prediction Including *$Accruals_{i,t-1}$*

	[1]	[2]
Intercept	0.068***	0.067***
	(7.865)	(7.752)
$CFO_{i,t-1}$	0.070***	0.068***
	(51.107)	(49.911)
$CFO_{i,t}$	-0.341***	-0.342***
	(-41.301)	(-41.598)
$\Delta Sales_{i,t}$	0.066***	0.061***
	(40.111)	(37.463)
$PPE_{i,t}$	-0.009***	-0.008***
	(-5.992)	(-5.544)
Neg $CFO_{i,t}$	0.019***	0.022***
	(9.822)	(11.482)
*$CFO_{i,t}$ * Neg $CFO_{i,t}$*	0.394***	0.391***
	(46.652)	(46.574)
$Accruals_{i,t-1}$	0.202***	0.131***
	(59.571)	(30.997)
BD $Accruals_{i,t}$		0.222***
		(27.505)
#OBS	71,847	71,847
Adjusted R^2	0.202	0.210

Panel D: Future Cash Flows Prediction

	[1]	[2]
Intercept	-0.006	-0.029***
	(-0.539)	(-2.916)
$CFO_{i,t-1}$	0.614***	0.420***
	(66.226)	(32.559)
$\Delta REC_{i,t}$	0.803***	0.704***
	(23.193)	(21.371)
$\Delta INV_{i,t}$	0.625***	0.518***
	(14.570)	(12.428)
$\Delta AP_{i,t}$	-0.436***	-0.384***
	(-11.603)	(-10.534)
$\Delta TAX_{i,t}$	-0.624***	-0.749***
	(-4.918)	(-6.242)
$\Delta OTH_{i,t}$	0.570***	0.485***
	(12.721)	(11.249)
$Dep_{i,t}$	0.502***	0.436***
	(8.845)	(8.274)
BD $CFO_{i,t+1}$		0.358***
		(24.554)
#OBS	65,843	65,843
Adjusted R^2	0.538	0.566

TABLE 10
Working Capital Accruals Prediction Model with Conference Calls

This table includes all firm/year observations from 2002 to 2013 with at least one conference call during the year and with sufficient data to calculate the dependent and independent variables. The dependent variable is working capital accruals (*Accruals$_{i,t}$*). A pooled model is estimated for all firms in the sample with industry and year fixed effects included as additional independent variables (unreported) and standard errors clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]
Intercept	0.041***	0.036***	0.043***	0.039***
	(6.774)	(6.291)	(7.713)	(7.125)
CFO$_{i,t-1}$	0.096***	0.106***	0.107***	0.111***
	(7.480)	(7.542)	(7.715)	(7.664)
CFO$_{i,t}$	-0.340***	-0.350***	-0.348***	-0.354***
	(-15.164)	(-15.464)	(-15.505)	(-15.597)
ΔSales$_{i,t}$	0.072***	0.070***	0.069***	0.069***
	(12.316)	(12.007)	(11.719)	(11.644)
PPE$_{i,t}$	0.008**	0.009***	0.009***	0.010***
	(2.297)	(2.740)	(2.857)	(3.025)
Neg CFO$_{i,t}$	0.013***	0.014***	0.013***	0.013***
	(3.911)	(4.050)	(3.982)	(4.063)
*CFO$_{i,t}$ * Neg CFO$_{i,t}$*	0.263***	0.262***	0.264***	0.263***
	(10.892)	(10.929)	(11.032)	(11.020)
BD Accruals$_{i,t}$		0.173***		0.124***
		(9.337)		(7.242)
BD Accruals (Conf Call)$_{i,t}$			0.386***	0.271***
			(10.631)	(8.646)
#OBS	18,232	18,232	18,232	18,232
Adjusted R^2	0.159	0.180	0.179	0.188