Estimating the likelihood of future tax settlements using firm fundamentals and text disclosures

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

Joshua A. Lee
Marriott School of Business
Brigham Young University
Provo, UT 84602
joshlee84@byu.edu

Erin Towery*
Terry College of Business
University of Georgia
Athens, GA 30602-6269
etowery@uga.edu

January 2020

We thank Andrew Finley, Ryan Hess, Lillian Mills, Terry Shevlin, Anthony Welsch, Brady Williams, and workshop participants at Singapore Management University, National University of Singapore, and Virginia Tech, and the National Tax Association Annual Meeting for helpful comments. We are also grateful for financial support from the Terry College of Business, Marriott School of Business, and Olin Business School.

*Contact author: Erin Towery. Physical address: Terry College of Business, University of Georgia, A322 Moore-Rooker Hall, Athens, GA, 30601. Telephone: (706) 542-3620. Email address: etowery@uga.edu.

ABSTRACT

This study develops an *ex-ante* measure of the likelihood of future corporate income tax settlements using qualitative (i.e., language) and quantitative information (i.e., firm fundamentals) from the Form 10-K. We find that both qualitative and quantitative information are incrementally useful in predicting future tax settlements. We then use our new measure to examine the association between predicted future tax settlements and cash holdings. In contrast with prior research documenting a positive association between reserves for uncertain tax positions and cash holdings (Hanlon, Maydew and Saavedra 2017), we find that firms with a higher likelihood of future tax settlements hold *less* cash than firms with a lower likelihood of future tax settlements. In addition to providing an *ex-ante* measure of future cash outflows paid to tax authorities that can be used by future researchers, our findings suggest a more nuanced relation between tax uncertainty and cash holdings.

Keywords: tax settlements, machine learning, textual analysis, tax uncertainty, cash holdings

JEL Classifications: H20; H25; H26; M40; M41

I. INTRODUCTION

Settlements paid to tax authorities represent a frequent and substantial cash outflow for many U.S. firms. Among the firms in Compustat that report reserves for uncertain tax positions, 54 percent report at least one tax settlement over the period 2008 to 2016, and the average settlement equals 5.7 percent of net income. A tax settlement occurs when a tax authority overturns a tax position claimed by the taxpayer. In addition to reducing cash balances, prior research suggests potential reputational consequences of underpaying taxes (Graham, Hanlon, Shevlin, and Shroff 2014; Dyreng, Hoopes, and Wilde 2016; Chen, Powers, and Stomberg 2019; Dhaliwal, Goodman, Hoffman, and Schwab 2017). Therefore, predicting the likelihood of a future tax settlement is important to market participants. In this paper, we examine the extent to which qualitative information (i.e., language) and quantitative information (i.e., firm fundamentals) reported in the Form 10-K predict future corporate income tax settlements with tax authorities. Our objective is two-fold. First, we develop an *ex-ante* measure for the likelihood of a future corporate income tax settlement using both qualitative and quantitative information. Second, we use our measure to examine the relation between the *ex-ante* likelihood of tax settlements and cash holdings.

To our knowledge, the only *ex-ante* measure of future tax outcomes currently used in the accounting literature is the reserve for uncertain tax benefits (UTBs) reported in the financial statements, which represents the maximum amount that could potentially be paid out to tax authorities in the future.[1] However, because FASB Interpretation No. 48, *Accounting for Uncertainty in Income Taxes* (FIN 48/ASC 740-10, FASB 2006) requires firms to assume both that the tax authority will audit tax positions with certainty and that the tax authority has full

[1] We view effective tax rate measures and book-tax difference measures as capturing *ex-post* realizations of tax planning.

knowledge of the positions claimed, reserves often do not reflect future tax settlements (Robinson, Stomberg, and Towery 2016). Consistent with this conjecture, we find that UTB reserves are negatively associated with the likelihood of a future tax settlement. Further, the area under the Receiver Operating Characteristic (ROC) curve for a model predicting future tax settlements using UTB reserves is only 52.2%, which suggests that UTB reserves are ineffective at distinguishing between observations with and without a future tax settlement. We therefore develop an alternative *ex-ante* measure of future tax settlements using machine learning techniques that incorporate both the quantitative (i.e., firm fundamentals) and qualitative (i.e., language) information found in the Form 10-K.

We first use three machine learning methods (support vector machines [SVM], supervised latent Dirichlet allocation [sLDA], and random forest regression trees [RF]) to model the relation between the incidence of future tax settlements and language contained in the tax footnote and tax-related information in the management's discussion and analysis section (MD&A) of the Form 10-K.[2] Similar to Frankel, Jennings, and Lee (2016) and Frankel, Jennings, and Lee (2019), we use rolling windows that include historical data for training the models and then apply the models to current-year disclosures to obtain out-of-sample future tax settlement predictions.[3]

To assess the usefulness of our tax settlement predictions generated from the language in the Form 10-K, we examine the relation between the likelihood of a future tax settlement and our *ex-ante* tax settlement predictions using 17,117 firm-year observations between 2008 and 2016. We find that the area under the ROC curve is equal to 80.3% when using the language from the tax footnote and 77.6% when using the tax-related language from the MD&A to predict the

[2] We model the incidence of future tax settlements in our primary analyses because we are interested in modelling all settlements regardless of the magnitude. However, we also model the magnitude of future tax settlements in Section V.

[3] We discuss our estimation of the textual disclosure variables more in depth in Section III.

likelihood of future tax settlement. We find that information contained in the tax footnote and information contained in the MD&A are both highly significant when included in the same tax settlement prediction model, which suggests that each prediction captures some non-overlapping information about the likelihood of future tax settlements. We also create a measure that combines the tax footnote and MD&A predictions using factor analysis and find that it yields an area under the ROC curve of 81.0%.

Next, we investigate the relation between the likelihood of a future tax settlement and the predictions generated by quantitative firm fundamentals.[4] We rely on Wilson (2009) and Lisowsky (2010) to identify quantitative firm fundamentals that are likely associated with future tax settlements (e.g., leverage, performance, and size). Similar to our procedure for developing the qualitative tax settlement predictions, we estimate a prediction for future tax settlements using quantitative data for each year using rolling training samples that include only historical quantitative data. We then apply the parameters estimated from the rolling models to current-year data to obtain out-of-sample future tax settlement predictions. The area under the ROC curve for the tax settlement prediction model using only quantitative information is 76.0%, which suggests that our quantitative predictions have reasonable discriminatory ability, but less discriminatory ability than the qualitative predictions.

It is possible that the quantitative and qualitative predictions capture overlapping information. We find that when we include both the quantitative and qualitative tax settlement predictions in a single model, both predictions are highly significant. Therefore, we produce an aggregate tax settlement prediction that combines both the qualitative and quantitative predictions using factor analysis. We report that the area under the ROC curve increases to 82.3% for our

[4] We include UTB reserves in our prediction model for completeness, but we obtain similar results when we exclude UTB reserves from the quantitative model.

combined tax settlement prediction. Taken together, our results suggest two key inferences. First, firm fundamentals and textual disclosures can be used to produce reasonable predictions of future tax settlements. Second, textual disclosures provide incremental information that is economically significant beyond the quantitative signals obtained from the firm's financial statements.

To demonstrate one application of our measure, we examine the level of cash holdings when firms face a higher *ex-ante* likelihood of tax settlement. On one hand, firms could hold more cash when the likelihood of making a cash payment in the future is higher. If firms expect to pay additional taxes in the future, we expect firms to hold more cash today to satisfy those liabilities. Consistent with this prediction, Hanlon, Maydew, and Saavedra (2017) find that firms with higher reserves for uncertain tax positions hold more cash. On the other hand, firms that are less conservative in their tax positions (i.e., firms with a higher likelihood of a tax settlement) might also be less conservative in the amount of cash that they hold (i.e., hold less cash) when claiming more aggressive tax positions, which would suggest a negative association between the likelihood of future tax settlements and cash holdings. Similar to Hanlon et al. (2017), we continue to find that firms with higher reserves for uncertain tax positions hold more cash. However, we find that firms with a higher likelihood of future tax settlements hold *less* cash than firms with a lower likelihood of future tax settlements. We also find that the positive relation between reserves for uncertain tax positions and cash holdings documented in Hanlon et al. (2017) is attenuated when the likelihood of a future tax settlement is higher.

In addition to being concerned about future tax cash outflows, investors are also likely interested in settlements that could have a negative impact on earnings in the year of settlement. Therefore, in a supplemental analysis, we examine whether qualitative and quantitative information in the Form 10-K can be used to predict settlements that have an unfavorable impact

on earnings.[5] We consider settlements to be 'unfavorable' when the firm's reserves for uncertain tax positions are estimated to be less than the settlements paid to tax authorities. We find that both qualitative and quantitative information predict 'unfavorable' tax settlements. Thus, future researchers can use this prediction when specifically trying to identify firms that will pay tax settlements that will likely have a negative effect on earnings.

This paper contributes to the literature in the following ways. First, we provide a measure that researchers can use to approximate the *ex-ante* likelihood that a firm will pay a tax settlement to the tax authority in the future. Future tax settlements are particularly interesting because they represent uncertain tax positions claimed by the firm that result in future cash payments to tax authorities. Prior studies generally measure tax planning with effective tax rate and book-tax difference measures (Hanlon and Heitzman 2010). While these measures capture *ex-post* realizations of tax planning, they are not necessarily predictive of future tax outcomes or future cash outflows paid to tax authorities. UTB reserves represent liabilities for tax positions that are unlikely to be sustained if challenged by a tax authority. Although these reserves are an *ex-ante* measure of future tax outcomes, practitioners have argued that the uniform requirements for recording UTB reserves can result in overstated liabilities. Consistent with practitioner concerns, Robinson, Stomberg and Towery (2016) find that only 24 cents of every dollar of reserves are paid out to tax authorities. We provide evidence that UTB reserves are not positively associated with future tax settlements, which questions their usefulness as an *ex-ante* measure of tax outcomes. Our study develops an alternative *ex-ante* measure of *future* tax outcomes.

[5] If a firm is perfectly reserved, a settlement will not affect tax expense (and earnings) because the firm recognized the tax expense in a prior period. If a firm is over-reserved (under-reserved), a settlement can potentially decrease tax expense because the firm recognized too much (too little) tax expense in a prior period. Being over-reserved or under-reserved will only affect earnings to the extent that the reserves pertain to positions that will impact the effective tax rate when the position lapses or is settled.

Second, we develop our future tax settlement predictions using both qualitative and quantitative information found in the Form 10-K. Prior research primarily uses quantitative signals (i.e., firm fundamentals) to assess the determinants of tax sheltering and tax avoidance (e.g., Dyreng, Hanlon, and Maydew, 2008; Wilson, 2009; Lisowsky, 2010). We add to the literature by providing evidence that both quantitative and qualitative information in firm disclosures incrementally predict the likelihood of a future settlement with a tax authority. We therefore encourage future researchers to consider the usefulness of both quantitative and qualitative information in firm disclosures when modeling concurrent or future tax outcomes.

Third, our results suggest that the qualitative and quantitative information in the Form 10-K can also be used to predict tax settlements that also have a financial statement reporting impact (i.e., 'unfavorable' settlements). Firms that experience an 'unfavorable' settlement have understated liabilities and overstated net income. We anticipate that investors are interested in understanding when a tax settlement is more likely to be unfavorable as they consider portfolio allocation decisions.

Fourth, we use our measure to further our understanding of the relation between UTB reserves and cash holdings. Similar to Hanlon et al. (2017), we report a positive association between UTB reserves and cash holdings, which suggests that firms hold more cash as a precaution in case of future tax settlements. However, among firms with UTB reserves, we find that firms claiming more uncertain tax positions (i.e., firms with a higher *ex-ante* likelihood of future tax settlements) actually hold less cash than firms claiming less uncertain tax positions, which suggests a more nuanced relation between UTB reserves and cash holdings. These findings are important to investors in assessing whether firms maintain adequate cash holdings in the event of future tax settlements.

II. BACKGROUND & RELATED LITERATURE

Despite the budget reductions faced by the Internal Revenue Service (IRS) in recent years (Marr and Murray 2016; Nessa, Schwab, Stomberg and Towery 2019), settlements paid by corporate taxpayers to tax authorities are common.[6] Understanding the likelihood of future tax settlements is important to investors because they represent a substantial cash outflow for firms. Indeed, Bauer and Klassen (2017) find a negative market reaction when firms report unfavorable settlements in their financial statements, consistent with market participants negatively valuing settlements paid to tax authorities. The negative market reaction to the settlement announcement suggests the market finds it difficult to predict future settlements. Similarly, Chow, Klassen, and Liu (2016) show that target firms in merger and acquisition transactions are more highly valued when they disclose that they have not participated in a tax shelter.

Further, firms face increasing public pressure to pay their 'fair share' of taxes or risk losing business to competitors. Dyreng, Hoopes, and Wilde (2016) report that a sample of U.K. firms reduced their tax avoidance after facing pressure from a nonprofit activist group to disclose the location of their corporate subsidiaries. Dhaliwal, Goodman, Hoffman, and Schwab (2017) report that a hedge portfolio long (short) in low (high) tax avoidance firms generates significant positive abnormal returns during 2011, a period with numerous protests that increased scrutiny of corporate tax avoidance. Chen, Powers, and Stomberg (2018) report that the likelihood of media tax coverage is higher for firms with greater visibility and firms with effective tax rates below the top U.S. statutory rate, and the degree of negative tone in the media coverage is increasing in cash tax avoidance and firm size.[7] Given the cash outflows and the reputational costs that can result from

[6] We discuss the prevalence and magnitude of tax settlements on Page 1.
[7] During their sample period, the top U.S. statutory rate was constant at 35 percent. The Tax Cuts and Jobs Act (TCJA) reduced the top U.S. statutory tax rate to 21 percent for tax years beginning after December 31, 2017.

tax settlements, the goal of this study is to examine whether quantitative firm fundamentals and qualitative textual disclosures can predict future tax settlements.

The only *ex-ante* measure of future tax outcomes currently used in the literature is the reserve for uncertain tax benefits (UTBs). However, because ASC 740-10 requires firms to assume that the tax authority will audit tax positions with certainty and that the tax authority has full knowledge of the positions when recording reserves, the reserves often do not reflect future tax settlements (Robinson, Stomberg, and Towery 2016). Our study overcomes the limitations of UTB reserves by developing an *ex-ante* measure of future tax settlements using both the quantitative (i.e., firm fundamentals) and qualitative (i.e., language) information found in the Form 10-K.

Our study is closely related to the Wilson (2009) and Lisowsky (2010) models of tax shelter likelihood. Wilson (2009) develops a model of tax shelter participation using 59 firms accused of engaging in tax shelter activity by the U.S. government. His sample includes tax shelters identified by Graham and Tucker (2006) via Tax Court docket searches, press articles, and tax shelters identified in the Factiva database. He reports that tax shelter firms have higher book-tax differences and more aggressive financial reporting. His finding that firms with a high likelihood of tax sheltering and poor corporate governance have negative abnormal returns suggests the market negatively values tax shelter participation. However, he also reports positive abnormal returns for firms with a high likelihood of tax sheltering and strong corporate governance, which suggests tax sheltering creates wealth for well-governed firms. Lisowsky (2010) expands the Wilson (2009) model in two ways. First, he uses the Treasury (1999) white paper on tax shelters to expand the number and types of tax shelter determinants. Second, he broadens the sample of identified tax shelter transactions to 267 using confidential administrative tax return data on reportable

transactions.[8] With these innovations, he reports that the likelihood of tax shelter participation is higher for larger and more profitable firms, firms with corporate subsidiaries located in tax havens, firms with foreign-source income, firms with inconsistent book-tax treatment, litigation losses, and firms using tax shelter promoters.

Our study builds on Wilson (2009) and Lisowsky (2010) in four important ways. First, not all settlements necessarily relate to tax shelter transactions; tax law ambiguity can result in disagreements between the IRS and the taxpayer, especially with regard to setting a transfer price or determining which expenses are eligible for various tax credits. Thus, our sample of settlements is broader than the identified tax shelter samples. Second, not all of the tax shelters identified in Wilson (2009) and Lisowsky (2010) result in additional payments to the IRS because some of the accused taxpayers eventually end up winning over the IRS in litigation. By examining settlements made to tax authorities, we capture only cases where the taxpayer loses some or all of their tax savings as a result of audit. Third, we examine income tax settlements made to *all* tax authorities rather than focusing exclusively on federal tax settlements. Finally, while Wilson (2009) and Lisowsky (2010) model the determinants of tax shelter participation in the same year of the tax shelter event, we examine whether current firm fundamentals and qualitative textual disclosures can predict *future* tax settlements.

III. *EX-ANTE* MEASURE OF FUTURE CORPORATE INCOME TAX SETTLEMENTS

Prior literature suggests that both quantitative and qualitative information is informative to market participants (Donovan, Jennings, Koharki, and Lee, 2019; Frankel et al., 2019). In this section, we develop measures for the likelihood of a future tax settlement using both quantitative

[8] Beginning in 2000, firms participating in transactions deemed 'potentially abusive tax shelters' must disclose such transactions to the IRS in their corporate tax return on Form 8886. The list of transactions warranting disclosure ('reportable transactions') has grown over time as the IRS becomes aware of new tax planning strategies.

and qualitative information reported in the Form 10-K. We then develop a summary measure that can be used by future researchers to predict the likelihood of a future tax settlement.

Descriptive Statistics

Our sample includes all firm-year observations with positive uncertain tax benefit balances from 2008 to 2016. We require both quantitative Form 10-K data in Compustat and Form 10-K filing data obtained from the SEC's EDGAR database. We predict the likelihood of a future tax settlement using both quantitative data from Compustat and language contained in the tax footnote and MD&A disclosures in the Form 10-K. Our sample consists of 17,117 firm-year observations. We require additional variables in our cash holdings tests, which reduces our sample to 13,977 firm-year observations.

We report the descriptive statistics for our sample in Table 1. Our future tax settlement predictions include $SETTLE\ PRED\ (TF)_{i,t}$, $SETTLE\ PRED\ (MD\&A)_{i,t}$, $SETTLE\ PRED\ (QUAL)_{i,t}$), $SETTLE\ PRED\ (QUANT)_{i,t}$, and $SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$. We construct these variables using factor analysis; therefore, each variable has a mean of zero and a standard deviation of one. We more fully describe how each of these variables are constructed in subsequent subsections. The average and median firm in our sample is profitable, with a mean (median) $ROA_{i,t}$ of 0.018 (0.049). The average (median) firm in the sample has total debt equal to 24% (19.5%) of the total assets ($LEV_{i,t}$). Approximately, 85.5% of the firms have a Big N auditor. The average (median) firm in our sample has an effective tax rate during year *t-1* of 19.3% (29%). In addition, approximately 66.2% of the firms in our sample have net operating loss carryforwards ($NOL_{i,t}$). Table 2 provides Pearson and Spearman correlation coefficients for all variables.

[Insert Tables 1 and 2 here]

Estimating Future Tax Settlements using Qualitative Information

Machine learning methods

Following prior research (Frankel et al., 2019; Donovan et al., 2019), we use three machine learning methods to predict the likelihood of a future tax settlement using the one- and two-word phrases contained in the tax footnote and tax-related sentences in management's discussion and analysis section (MD&A) of the 10-K.[9] The three machine learning methods we use include support vector machines (SVM), supervised latent Dirichlet allocation (sLDA) (Blei and McAuliffe, 2007), and random forest regression trees (RF).[10] We use the three machine learning techniques because each likely identifies a different aspect of the Form 10-K's narrative content (Frankel et al., 2019). We briefly describe how we estimate each machine learning technique below.[11]

We define $SETTLE_{i,t+1}$ as a dichotomous variable that is equal to one if firm *i* settles with tax authorities in year *t+1* (Compustat TXTUBSETTLE), and 0 otherwise. Similar to prior literature (e.g., Frankel et al., 2016), we then apply a two-step process to predict future tax settlements using each of the three machine learning methods. First, we use rolling training samples that include historical data over the previous four years (i.e., year *t-4* to *t-1*) to estimate the relation between $SETTLE_{i,t+1}$ and all one- and two-word phrases contained in each disclosure (e.g., tax footnote and MD&A). Second, we apply the parameters estimated from the rolling

[9] We identify tax-related sentences in the MD&A by extracting all sentences from the MD&A that contain at least one of the following tax-related terms: tax, taxed, taxes, taxing, taxation, ETR, IRS, internal revenue, UTB, UTP, DTA, DTL, NOL, permanently reinvest, permanently reinvested, permanently reinvests, permanently reinvesting, indefinitely reinvest, indefinitely reinvested, indefinitely reinvests, indefinitely reinvesting, internal revenue, valuation allowance, carryforward, or carryforwards. We exclude instances of tax-related words that do not relate directly to income taxes. For example, we exclude instances of "pre-tax," "property tax," "sales tax," and "net of income taxes" from our search. The full list of exclusions is available from the authors upon request.

[10] We note that machine learning methods are not the only methods used in textual analysis. Tone (e.g., Feldman, Govindaraj, Livnat, and Segal, 2010; Loughran and McDonald, 2011; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Kothari, Li, and Short, 2009), readability (e.g., Li, 2008; Lehavy, Li, and Merkley, 2011), cosine similarity (Brown and Tucker, 2011; Lang and Stice-Lawrence, 2015; Peterson, Schmardebeck, and Wilks, 2015) are some of other methods used in textual analysis.

[11] See Frankel et al. (2019) for more information on these machine learning techniques.

training sample models to the one- and two-word phrases in year *t* to obtain an out-of-sample prediction for $SETTLE_{i,t+1}$. Figure 1 provides an example of how we define the training and prediction samples. In the figure, the training sample includes all observations from year 2010 to 2014 (i.e., years *t-4* to *t-1*), and the model parameters are then applied to 2015 (i.e., year *t*). This process of applying the relevant parameters out-of-sample helps to reduce overfitting the model to the data.

The three machine learning methods use different processes to identify patterns in the data that are useful for predicting future settlements. Due to the differences in how each machine learning method is estimated, each method likely identifies a different informational aspect of the disclosure's narrative content (Frankel et al., 2019).

SVM assigns weights to each one and two-word phrase included in the training sample by simultaneously minimizing the coefficient vector magnitude, which is the vector of weights on each phrase, and the prediction error.[12] Including the coefficient vector magnitude in the minimization function helps to reduce the likelihood of overfitting the model to the data. We then apply the weights for each phrase identified in the training sample to the language in the disclosures for year *t* to predict the likelihood of a future tax settlement for year *t+1*.

The random forest method uses binary recursive partitioning to create "trees" based on the phrases in the disclosures. To build each tree, the model first identifies the phrase count that splits the sample so that the sum of squared errors is minimized within each partition or node. The splitting process is repeated within each partition until the number of observations in each final node is lower than a pre-specified number, which we define as two, or the sum of the squared error is zero. The average value in the final node represents the prediction for the likelihood of a future

[12] SVM can be implemented using the SVM-light implementation of Support Vector Machines in C, which was created by Thorsten Joachims and can be accessed at http://svmlight.joachims.org/.

tax settlement. The final predicted value is the average predicted value generated from all trees. The random forest method (RF) works to reduce overfitting in two ways. First, each tree uses only a subset of the total number of available phrases (e.g., the square root of the total number of phrases) so that any one single phrase cannot have undue influence. Second, RF creates many regression trees (e.g., 5,000) using bootstrapped samples, where observations are selected with replacement. We apply the trees constructed using the training samples to the phrases from the disclosures in year *t* to obtain an out-of-sample future tax settlement prediction for year *t+1*.

Finally, we obtain a future tax settlement prediction using sLDA to identify 200 latent topics in each training sample that are predictive of future settlements. sLDA chooses latent topics that are associated with a dependent variable by grouping phrases based on the probability of the phrases co-occurring within documents (Blei and McAuliffe, 2007). We apply the topics and their respective weightings to the disclosures in year *t* to obtain an additional out-of-sample future tax settlement prediction for year *t+1*.

Similar to Donovan et al. (2019), we *ex-post* manually categorize the top 200 most important words and phrases generated by the RF and SVM methods and the top 10 most important words and phrases for each of the 200 topics generated by sLDA in both the tax footnote and the MD&A. We use the importance weighting supplied by the random forest method, the absolute value of the coefficient weighting supplied by the support vector machines method, and the importance weighting supplied by the sLDA method. We identify 14 categories during the coding process (Audit/Tax Authority, Credits, Cross-jurisdictional, Deferred Tax Accounts – General, Deferred Tax Accounts – NOLs/VA, General Tax, Interest & Penalties, Legislation, Permanently-reinvested Earnings, Reserves, Tax Rate, Tax Returns, Performance, and Other Financial). We

leave unclassified any words and phrases that do not fit into one of these categories and report them separately.

In Appendix A, we present the 200 most influential words and phrases for SVM and RF and the top 10 most negatively predictive and top 10 most positively predictive topics for sLDA for both the tax footnote (Appendices A.1 to A.3) and the MD&A (Appendices A.4 to A.6). We graphically summarize the importance of each category in Figure 2 (for the tax footnote) and Figure 3 (for the MD&A). For SVM and RF, we sum the relative importance among the top 200 words and phrases for each category. For sLDA, we represent each topic as a weighted average of the 14 categories, where the category weights are determined by the relative importance of the assigned categories of the top 10 words and phrases for each topic.

We note that each method identifies different information to predict future tax settlements. For example, when using the tax footnote, RF primarily identifies words and phrases that relate to audit examinations and tax authorities, while sLDA and RF tend to identify phrases that relate more to general tax information and other financial information in their predictions. This analysis provides preliminary evidence that the machine learning methods identify language that is associated with future tax settlements.

Prediction of Future Tax Settlements

Our estimate of future tax settlements using SVM (sLDA) [RF] is labeled *SETTLE PRED SVM*$_{i,t}$ (*SETTLE PRED sLDA*$_{i,t}$) [*SETTLE PRED RF*$_{i,t}$]. We label the predictions based on the tax footnote with "TF," and we label the predictions based on the tax-related sentences in the MD&A with "MD&A." For example, the SVM prediction based on the tax footnote is labeled *SETTLE PRED SVM (TF)*$_{i,t}$, and the SVM prediction based on the MD&A is labeled *SETTLE PRED SVM (MD&A)*$_{i,t}$. We also combine the SVM, sLDA, and RF predictions using factor analysis and label

the resulting variables *SETTLE PRED (TF)*$_{i,t}$ and *SETTLE PRED (MD&A)*$_{i,t}$, for the predictions based on the tax footnote and the MD&A, respectively. We note that our factor analysis in both cases yields a single factor with an eigenvalue greater than one. As previously stated, this produces a factor variable with a mean of zero and a standard deviation of one.

To understand the usefulness of the language in the tax footnote and the MD&A in predicting future tax settlements, we examine the association between actual future tax settlements and our predictions of future tax settlements (*SETTLE PRED (TF)*$_{i,t}$ and *SETTLE PRED (MD&A)*$_{i,t}$) using the following logistic regression equation:

$$SETTLE_{i,t+1} = \beta_0 + \beta_1 SETTLE\ PRED_{i,t} + \varepsilon \quad (1)$$

We first examine whether each of the machine learning predictions calculated using the tax footnote is associated with future tax settlements. We estimate Equation (1) separately replacing *SETTLE PRED*$_{i,t}$ with each of the individual tax footnote predictions (i.e., *SETTLE PRED RF (TF)*$_{i,t}$, *SETTLE PRED SVM (TF)*$_{i,t}$, and *SETTLE PRED sLDA (TF)*$_{i,t}$). We also estimate Equation (1) using the composite measure *SETTLE PRED (TF)*$_{i,t}$. We present the results in Table 3. We find that each of the predictions using the tax footnote is significant at the 1% level. Of the three predictions using the individual machine learning method, we find that *SETTLE PRED RF (TF)*$_{i,t}$ (*SETTLE PRED SVM (TF)*$_{i,t}$) produces the highest (lowest) pseudo-R^2 of 19.2% (9.9%). The composite measure (*SETTLE PRED (TF)*$_{i,t}$) produces the highest pseudo-R^2 of 21.5%. We evaluate the accuracy of our model using the area under the ROC curve, which Hosmer, Lemeshow, and Sturdivant (2013) suggest is the appropriate statistic for evaluating model fit. Our composite measure yields an area under the ROC curve of 80.3% (Column 4), which Hosmer et al. (2013) classifies as excellent discriminatory power. These results suggest the language in the tax footnote is useful for predicting future tax settlements.

[Insert Table 3 here]

Next, we examine whether machine learning methods can extract relevant language from the tax-related sentences in the MD&A to predict future tax settlements. Similar to the tax footnote predictions, we first examine the individual predictions of future tax settlements using the three machine learning methods. Similar to Table 3, we note the coefficient on each of the machine learning predictions using the tax-related sentences in the MD&A (e.g., *SETTLE PRED RF (MD&A)*$_{i,t}$, *SETTLE PRED SVM (MD&A)*$_{i,t}$, *SETTLE PRED sLDA (MD&A)*$_{i,t}$, and *SETTLE PRED (MD&A)*$_{i,t}$,) is significant in Table 4. We find that *SETTLE PRED RF (MD&A)*$_{i,t}$ (*SETTLE PRED sLDA (MD&A)*$_{i,t}$) produces the highest (lowest) pseudo-R^2 of 16.5% (8.1%). We note that the composite measure (*SETTLE PRED (MD&A)*$_{i,t}$) yields the highest pseudo-R^2 of 18.0%. The area under the ROC curve for the composite measure model is 77.6%, which Hosmer et al. (2013) classifies as acceptable discriminatory power. These results suggest that machine learning methods are able to extract language from the MD&A that is relevant to predicting future tax settlements.

[Insert Table 4 here]

Our findings suggest that both the tax footnote and the MD&A contain language that is useful in predicting future tax settlements. However, our results thus far do not examine whether the predictions from the tax footnote and the MD&A identify distinct or overlapping information. To test this, we include both *SETTLE PRED (MD&A)*$_{i,t}$ and *SETTLE PRED (TF)*$_{i,t}$ in the same regression. In Table 4 Column 5, we find a positive coefficient on both *SETTLE PRED (MD&A)*$_{i,t}$ and *SETTLE PRED (TF)*$_{i,t}$, which suggests that the tax-related sentences from the MD&A and the tax footnote identify distinct information that is useful in predicting future tax settlements. The pseudo-R^2 after including both composite variables in the model is equal to 23.1%, which yields an increase of 7.0% (27.8%) relative to the model that includes *SETTLE PRED (TF)*$_{i,t}$ (*SETTLE*

PRED (MD&A)$_{i,t}$) as the sole regressor. The area under the ROC curve also increases to 81.2% in Table 4 Column 5.

To provide a parsimonious prediction of future tax settlements using language from the Form 10-K, we next use factor analysis to combine the three machine learning predictions derived from the tax footnote with the three predictions derived from the MD&A. We label this measure *SETTLE PRED (QUAL)*$_{i,t}$. We include the regression results using *SETTLE PRED (QUAL)*$_{i,t}$ as the only independent variable in Table 4 Column 6. We find that the coefficient on *SETTLE PRED (QUAL)*$_{i,t}$ is positive and significant at the 1% level. We also find that the pseudo-R^2 is equal to 23.0%, and the area under the ROC curve for the aggregate measure model is 81.0%, which Hosmer et al. (2013) classifies as excellent discriminatory power. Because the information from the tax footnote and the MD&A appears to be most succinctly summarized by *SETTLE PRED (QUAL)*$_{i,t}$, we use this variable in future tests.

Estimating Future Tax Settlements using Uncertain Tax Benefits (UTBs)

Next, we examine how well UTB reserves predict the likelihood of a future tax settlement. We note that the uncertain tax benefits are measured in year *t* and the tax settlement variable (*SETTLE*$_{i,t+1}$) is measured in year *t+1*, which allows us to predict tax settlements with only data that is known prior to the settlement. We use the following logistic regression model to predict the likelihood of a future tax settlement as a function of UTB reserves.

$$SETTLE_{i,t+1} = \beta_0 + \beta_1 UTB_{i,t} + \varepsilon \quad (2)$$

SETTLE$_{i,t+1}$ is as previously defined. *UTB*$_{i,t}$ equals UTB reserves in year *t* divided by total assets in year *t-1*. A positive (negative) coefficient would suggest that firms with higher (lower) UTB reserves are more likely to experience a future tax settlement. We present the results for the pooled sample in Table 5 Column 1. We find a negative and significant association between UTB

reserves and the likelihood of a future tax settlements. This result highlights the concern that UTB reserves are not able to accurately capture the *ex-ante* likelihood of a future tax settlement.

Similar to how we construct the future tax settlement prediction measures using the language in the Form 10-K, we obtain a UTB reserves tax settlement prediction by estimating Equation (2) for each year using rolling training windows of the previous 4 years. We then apply the coefficient on $UTB_{i,t}$ from each estimation to current-year $UTB_{i,t}$ values and label the predictions $SETTLE\ PRED\ (UTB)_{i,t}$. In Table 5 Column 2, we find that although the $SETTLE\ PRED\ (UTB)_{i,t}$ coefficient is positive and significant, the pseudo-R^2 is equal to 0.3% and the area under the ROC curve is equal to 52.2%, which suggests that the model is unable to distinguish between observations with and without future tax settlements. These results highlight the need for future research to carefully consider whether the liability for uncertain tax benefits is a useful measure for *ex-ante* corporate tax settlements.

For completeness, we then include $SETTLE\ PRED\ (UTB)_{i,t}$ and $SETTLE\ PRED\ (QUAL)_{i,t}$ in Table 5 Column 4 to examine whether $SETTLE\ PRED\ (UTB)_{i,t}$ is incrementally useful in predicting future tax settlements. We reproduce regression results with just $SETTLE\ PRED\ (QUAL)_{i,t}$ as the regressor from Table 4 Column 6 in Table 5 Column 3 to ease comparisons. While $SETTLE\ PRED\ (UTB)_{i,t}$ is positive and significant when $SETTLE\ PRED\ (QUAL)_{i,t}$ is included in the estimation (Column 4), the incremental increase in pseudo-R^2 (increase from 0.230 to 0.231) and the area under the ROC curve (increase from 0.810 to 0.812) from Column 3 to 4 is minimal.

These results suggest that information provided by the UTB reserves is not economically meaningful despite the coefficient being statistically significant.

Estimating Future Tax Settlements using Quantitative Information

We next examine whether quantitative data (i.e., firm fundamentals) from the Form 10-K can be used to predict the likelihood of a future tax settlement. We draw on prior studies (Wilson, 2009; Lisowsky, 2010) that estimate the likelihood of tax shelter participation to identify quantitative information that is likely to be useful in predicting the likelihood of a future tax settlement. We add the firm's uncertain tax positions ($UTB_{i,t}$) for completeness.[13] We note that all of the independent variables are measured in year t and the tax settlement variable ($SETTLE_{i,t+1}$) is measured in year $t+1$, which allows us to predict tax settlements with only information that is known prior to the settlement. We use the following logistic regression model to predict the likelihood of a future tax settlement using quantitative information from the financial statements.

$$SETTLE_{i,t+1} = \beta_0 + \beta_1 UTB_{i,t} + \beta_2 BTD_{i,t} + \beta_3 LEV_{i,t} + \beta_4 ln(ASSETS_{i,t}) + \beta_5 ROA_{i,t} + \beta_6 FOR\ INC_{i,t} + \beta_7 RD_{i,t} + \beta_8 ETR_{i,t-1} + \beta_9 EQ\ EARN_{i,t} + \beta_{10} MEZZ\ FIN_{i,t} + \beta_{11} BIG\ N_{i,t} + \beta_{12} LITIG_{i,t} + \beta_{13} NOL_{i,t} + \varepsilon \quad (3)$$

$SETTLE_{i,t+1}$ is as previously defined, and all other variables are defined in Appendix B. We present the results for the pooled sample using Equation (3) in Table 6 Column 1. We discuss expected signs and realized signs for each of the coefficients on the variables included in Equation (3) below. Inconsistent with Table 5 Column 1 but consistent with expectations, the coefficient on $UTB_{i,t}$ is positive and statistically significant in Table 6 Column 1, which suggests that firms with higher uncertain tax positions are more likely to experience a future tax settlement. Consistent with firms being less likely to engage in aggressive tax positions when the interest tax shield is higher,

[13] We note that the results in Table 6 are similar if $UTB_{i,t}$ is excluded from the quantitative model (Equation 3).

we find that highly levered firms ($LEV_{i,t}$) are less likely to pay a future tax settlement. We find that firms with higher performance ($ROA_{i,t}$) are more likely to pay a future tax settlement, which is consistent with more profitable firms using more tax aggressive strategies to shield their income from taxes. We find that firms with larger auditors ($BIG\ N_{i,t}$) are more likely to pay a tax settlement in the following year, which is consistent with Big N auditors being tax planning promoters (Treasury 1999). We also find that firms involved in litigation ($LITIG_{i,t}$) are more likely to pay a settlement in the following year, which is consistent with litigious firms also being involved with aggressive tax behavior. In Table 2, we find univariate evidence that firms with NOL carryforwards ($NOL_{i,t}$) are less likely to have a future tax settlement, which is consistent with loss firms not needing to engage in aggressive tax strategies due to the benefit of the NOLs. However, we find no evidence of this relation in our multivariate tests.

We expect that more tax aggressive firms are more likely to be audited and settle with the tax authority (Mills 1998). Therefore, we expect firms with larger differences between pretax income and tax income ($BTD_{i,t}$) and lower effective tax rates ($ETR_{i,t-1}$) to be more likely to have a future tax settlement. Inconsistent with our expectations, we find a negative and significant coefficient on $BTD_{i,t}$ and an insignificant coefficient on $ETR_{i,t-1}$.

Because research and development expenses are expensed when incurred and lower taxable income, firms with more research and development expenses might be less likely to engage in more aggressive tax positions, which could result in a lower likelihood of a future tax settlement. Alternatively, firms with research and development expenses could aggressively claim R&D tax credits, which could increase the likelihood of tax settlements. We find a negative and significant coefficient on $RD_{i,t}$, which is consistent with the former explanation. We find that firms with more foreign income ($FOR\ INC_{i,t}$) are more likely to have a future tax settlement, which is consistent

with foreign operations providing greater opportunities to engage in aggressive tax planning.[14] We also find that larger firms ($ln(Assets_{i,t})$) are more likely to have a tax settlement, which could be due to their increased visibility with the tax authorities and/or their participation in continual audit programs.

We predict a positive association between the presence of equity method earnings (*EQ EARN*) and future tax settlements because the equity method enables a firm to report greater income for financial reporting purposes relative to tax reporting purposes (Lisowsky 2010). However, we fail to find a significant association between equity earnings and future tax settlements in Table 6. Finally, we expect a positive association between financial engineering products and future tax settlements because Treasury (1999) identifies financial engineering as indicative of tax sheltering. Following Lisowsky (2010), we measure financial engineering products using mezzanine financing (*MEZZ FIN*), but we fail to find a significant association with future tax settlements. The area under the ROC curve is equal to 75.7% in Table 6 Column 1, which suggests that the discriminatory ability of the model is acceptable (Hosmer et al., 2013), but lower than the discriminatory ability of models based on Form 10-K qualitative information.

[Insert Table 6 here]

Similar to how we construct the future tax settlement prediction measures using the language in the Form 10-K, we obtain quantitative tax settlement predictions by estimating Equation (3) for each year using rolling training windows of the previous 4 years. We then apply the coefficients of these models to current-year values and label the quantitative tax settlement prediction $SETTLE\ PRED\ (QUANT)_{i,t}$. In Table 6 Column 2, we find that the pseudo-R^2 with

[14] For example, with the Tax Cuts and Jobs Act, firms no longer have to pay United States income taxes on income earned abroad. However, firms could still be subject to Global Intangible Low-Taxed Income (GILTI) taxes on their foreign-sourced income.

SETTLE PRED (QUANT)$_{i,t}$ as the only independent variable is equal to 15.5%, which is identical to the pseudo-R^2 reported in Column 1. This result suggests that the rolling prediction model provides a summary measure that has explanatory power consistent with the pooled model. The area under the ROC curve of 76.0%, which is classified as acceptable discriminatory power, is lower than the area under the ROC curve of 81.0% for our aggregate qualitative measure (*SETTLE PRED (QUAL)*$_{i,t}$) presented in Table 6 Column 3.

Because the purpose of our paper is to predict the likelihood of a tax settlement using quantitative and qualitative information from the 10-K, we next include both *SETTLE PRED (QUAL)*$_{i,t}$ and *SETTLE PRED (QUANT)*$_{i,t}$ in the same regression to see if they are incrementally useful in predicting the likelihood of a future tax settlement. In Table 6 Column 4, we find that the coefficients on both *SETTLE PRED (QUAL)*$_{i,t}$ and *SETTLE PRED (QUANT)*$_{i,t}$ are positive and significant at the 1% level, which suggests that the quantitative and qualitative disclosures in the Form 10-K provide incremental information that is useful in predicting future tax settlements. The pseudo-R^2 is equal to 25.8%, which is a 66% increase from the regression that only includes *SETTLE PRED (QUANT)*$_{i,t}$ in Column 2. This result suggests that *SETTLE PRED (QUAL)*$_{i,t}$ identifies an economically significant amount of information that does not overlap with the quantitative information in the Form 10-K and highlights that qualitative disclosure is an important source of information when predicting the likelihood of a future tax settlement.

For the sake of parsimony, we produce a measure (*SETTLE PRED (QUAL + QUANT)*$_{i,t}$) that combines both the quantitative and qualitative information from the Form 10-K. *SETTLE PRED (QUAL + QUANT)*$_{i,t}$ is the first factor obtained from a factor analysis by year on *SETTLE PRED sLDA (TF)*$_{i,t}$, *SETTLE PRED RF (TF)*$_{i,t}$, *SETTLE PRED SVM (TF)*$_{i,t}$, *SETTLE PRED sLDA (MD&A)*$_{i,t}$, *SETTLE PRED RF (MD&A)*$_{i,t}$, *SETTLE PRED SVM (MD&A)*$_{i,t}$, and *SETTLE PRED*

(QUANT)$_{i,t}$. We find that the pseudo-R^2 decreases slightly after combining all the quantitative and qualitative measures into a single measure from 25.8% to 24.9%; however, we believe that the benefit from producing a single measure of future tax settlement using both quantitative and qualitative information outweighs the small decrease in the pseudo-R^2. The area under the ROC curve is 82.3%, which is similar to the area under the ROC curve when *SETTLE PRED(QUANT)*$_{i,t}$ and *SETTLE PRED(QUAL)*$_{i,t}$ are separately included in the model as independent regressors (Table 6 Column 4). Hosmer et al. (2013) classifies an output of 82.3% as excellent discriminatory power for a classification model.

IV. PREDICTED FUTURE TAX SETTLEMENTS & CASH HOLDINGS

In this section, we use our new measure to examine the relation between predicted future tax settlements and cash holdings. To do so, we build on the model developed by Hanlon et al. (2017) and estimate the following OLS regression:

$$\begin{aligned} CASH\ RATIO_{i,t} = {} & \beta_0 + \beta_1\ MODIFIED\ SETTLE\ PRED\ (QUAL+QUANT)_{i,t} + \beta_2\ UTB_{i,t} \\ & + \beta_3\ REPATRIATION\ TAX\ COST_{i,t} + \beta_4\ CASH\ ETR_{i,t} + \beta_5\ FIN\ CONSTRAINED_{i,t} + \beta_6\ NOL_{i,t} + \beta_7\ LOSS_{i,t} + \beta_8\ NET\ WORKING\ CAPITAL_{i,t} + \beta_9\ LEV_{i,t} + \beta_{10}\ CF\ VOL_{i,t} + \beta_{11}\ MTB_{i,t} + \beta_{12}\ ln(ASSETS)_{i,t} + \beta_{13}\ DIVIDEND\ PAYOUT_{i,t} + \beta_{14}\ CAPEX_{i,t} + \beta_{15}\ ACQUISITIONS_{i,t} + \beta_{16}\ AFTER\ TAX\ CF + \beta_{17}\ RD_{i,t} + INDUSTRY\ FE + YEAR\ FE + \varepsilon \end{aligned} \quad (4)$$

The dependent variable, *CASH RATIO*$_{i,t}$, equals the ratio of cash and short-term equivalents to total assets. *UTB*$_{i,t}$ is equal to UTB reserves divided by total assets. Hanlon et al. (2017) report that firms with greater UTB reserves hold more cash in order to satisfy potential future tax settlements. We include all of the control variables included in Hanlon et al. (2017). *INDUSTRY*

FE represent industry fixed effected defined as two-digit SIC codes, and *YEAR FE* represent year fixed effects. All other variables are defined in the appendix. Similar to Hanlon et al. (2017), we also cluster the standard errors by firm and year to correct for serial and cross-sectional correlation.

Our variable of interest is *MODIFIED SETTLE PRED (QUAL+QUANT)*$_{i,t}$, which is equal to the *SETTLE PRED (QUAL+QUANT)*$_{i,t}$ variable excluding the *UTB*$_{i,t}$ variable from the estimation of the quantitative prediction in Equation (3). We exclude *UTB*$_{i,t}$ from the estimation of the Form 10-K disclosure prediction because *UTB*$_{i,t}$ is a variable of interest in Equation (4). We have a two-sided prediction for the relation between *MODIFIED SETTLE PRED (QUAL+QUANT)*$_{i,t}$ and future tax settlements. On one hand, firms with a higher likelihood of a future tax settlement could be holding more cash because the expected future cash outflow associated with uncertain tax positions is higher. On the other hand, firms that are less conservative in their tax positions (i.e., firms with a higher likelihood of a tax settlement) may also be less conservative in the amount of cash that they hold (i.e., hold less cash) when claiming more aggressive tax positions, which would result in a negative association between *CASH RATIO*$_{i,t}$ and *MODIFIED SETTLE PRED (QUAL+QUANT)*$_{i,t}$.

[Insert Table 7 here]

Table 7 Panel A provides descriptive statistics for the variables included in our model, and Panel B provides Pearson and Spearman correlations for the variables included in Equation (4). We present our results from estimating Equation (4) in Table 7 Panel C. To establish a baseline, Column 1 excludes our main variable of interest (*MODIFIED SETTLE PRED (QUAL+QUANT)*$_{i,t}$). Consistent with Hanlon et al. (2017), we report a positive association between UTB reserves (*UTB*$_{i,t}$) and cash holdings. In Column 2, we estimate Equation (3) excluding *UTB*$_{i,t}$ and find a *negative* coefficient on *MODIFIED SETTLE PRED*

(QUAL+QUANT)$_{i,t}$, which suggests that firms with more uncertain tax positions maintain *lower* cash reserves. In Column 3, we estimate the full version of Equation (4) and continue to find a positive coefficient on $UTB_{i,t}$ and a negative coefficient on *MODIFIED SETTLE PRED (QUAL+QUANT)*$_{i,t}$.

We further explore this result by examining whether there is cross-sectional variation in the association between the size of the UTB and cash holdings when the likelihood of a future tax settlement varies. To do so, we create two indicator variables: *HIGH SETTLE PRED (QUAL + QUANT)*$_{i,t}$ is equal to 1 if *MODIFIED SETTLE PRED (QUAL + QUANT)*$_{i,t}$ is above its median value and zero otherwise, and *HIGH* $UTB_{i,t}$ is equal to 1 if $UTB_{i,t}$ is above its median value and zero otherwise. In Column 4, we estimate Equation (4) replacing the continuous measures of predicted future tax settlements and UTB reserves with the indicator variables. This specification confirms the results in Column 3 and facilitates a comparison of the magnitude of the coefficients. The results suggest the positive effect of high UTB reserves is greater than the negative effect of predicted future tax settlements.

In Column 5, we add the interaction between *HIGH SETTLE PRED (QUAL + QUANT)*$_{i,t}$ and *HIGH* $UTB_{i,t}$ to the regression. The main effect of UTB reserves (*HIGH* $UTB_{i,t}$) continues to be positive and significant, which suggests that larger UTB reserves are positively associated with a firm's cash holdings when the likelihood of a future tax settlement is lower. The interaction term is negative and significant (1% level), which suggests a smaller positive association between the firm's cash holdings and tax uncertainty when the likelihood of a future tax settlement is higher. We find that the sum of the coefficients on *HIGH* $UTB_{i,t}$ and the interaction between *HIGH SETTLE PRED (QUAL+QUANT)*$_{i,t}$ and *HIGH* $UTB_{i,t}$ is statistically greater than zero (p-value less than 0.001), which suggests that the positive association between tax uncertainty and cash holdings

is still positive when the likelihood of future tax settlements is higher. In other words, the effect of tax uncertainty on cash holdings is positive regardless of whether the likelihood of a future tax settlement is high or low, but the effect of tax uncertainty on cash holdings is less positive when the likelihood of tax settlement is higher. This evidence suggests that firms claiming more uncertain tax positions hold less precautionary cash when the likelihood of a future tax settlement is higher. This result is surprising given that one would expect firms with a higher likelihood of future tax settlements to be *more* likely to hold more cash. In sum, our results suggest a nuanced relation between reserves for uncertain tax positions and cash holdings.

V. SUPPLEMENTAL ANALYSES

Current Year Tax Settlements as a Predictor

We also examine whether our predictions for future tax settlements are incrementally informative after controlling for whether there is a tax settlement in the current year. We first examine whether tax settlements are serially correlated in Column 1 of Table 8. With $SETTLE_{i,t+1}$ as the dependent variable, we find a positive and significant coefficient on $SETTLE_{i,t}$. The pseudo-R^2 in Column 1 is equal to 17.9% and the area under the ROC curve is 74.1%, which Hosmer et al. (2013) classifies as acceptable discriminatory power. This result suggests that the occurrence of a tax settlement is correlated over time and that our measure predicting the likelihood of a future tax settlement could be identifying current year tax settlements. We examine this possibility by including $SETTLE_{i,t}$ as an additional control variable in our analyses.

[Insert Table 8 here]

In Column 2, we include $SETTLE_{i,t}$ and $SETTLE\ PRED\ (QUAL)_{i,t}$ as separate independent variables. We continue to find a positive and significant coefficient on both $SETTLE_{i,t}$ and $SETTLE\ PRED\ (QUAL)_{i,t}$. The coefficient on $SETTLE\ PRED\ (QUAL)_{i,t}$ suggests that the machine learning

methods capture incremental information that is predictive of future tax settlements after controlling for current year settlements. The pseudo-R^2 increase from Column 1 to 2 is approximately 29.6% of the pseudo-R^2 in Column 1, which suggests that machine learning methods extract an economically significant amount of useful information disclosed in the Form 10-K. The area under the ROC curve also increases from 74.1% in Column 1 to 81.2% in Column 2.

In Column 3, we include $SETTLE_{i,t}$ and $SETTLE\ PRED\ (QUANT)_{i,t}$ as separate independent variables. Once again, we find that both coefficients on the two variables are positive and significant at the 1% level. The pseudo-R^2 is equal to 24.5%, which represents a 36.9% increase from Column 1 to 3. These results suggest that the quantitative information found in the Form 10-K provides incrementally useful information about the likelihood of a future tax settlements. The area under the ROC curve increases from 74.1% in Column 1 to 82.0% in Column 3.

In Column 4, we include $SETTLE_{i,t}$, $SETTLE\ PRED\ (QUANT)_{i,t}$, and $SETTLE\ PRED\ (QUAL)_{i,t}$ as independent variables. Each of the coefficients is positive and significant at the 1% level. The pseudo-R^2 (area under the ROC curve) is equal to 26.3% (83.0%) and represents a 46.9% (12.0%) increase from Column 1 to 4. This result suggests that both the quantitative and qualitative information reported in the Form 10-K is useful in predicting future tax settlements after controlling for current year tax settlements. We also examine the incremental effect of including $SETTLE_{i,t}$ as an additional regressor when predicting the likelihood of a future tax settlement. We note that the percentage increase in pseudo-R^2 (area under the ROC curve) is just 1.9% (0.4%) when adding $SETTLE_{i,t}$ to a regression model with $SETTLE\ PRED(QUAL)_{i,t}$ and $SETTLE\ PRED(QUANT)_{i,t}$ as regressors (Column 4 of Table 6). Therefore, our predictions of

future tax settlements using qualitative and quantitative information from the Form 10-K are incrementally more useful than the existence of a tax settlement in the current year.

In Column 5 of Table 8, we include $SETTLE_{i,t}$ and our composite measure $SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$ in the same regression model to test the incremental usefulness of our composite measure. We continue to find a positive and significant coefficient (1% level) on $SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$. The pseudo-$R^2$ (area under the ROC curve) is equal to 25.0% (82.4%), which is slightly lower than the pseudo-R^2 (area under the ROC curve) in Column 4. Overall, these results alleviate the concern that our measure predicting the likelihood of a future tax settlement is solely identifying current year tax settlements.

Predicting the Magnitude of Future Tax Settlements

Our primary focus in this study is on predicting the likelihood of future tax settlements. In this section, we also examine whether quantitative and qualitative information in the Form 10-K is useful in predicting the *large* future tax settlements. We follow the same methods previously described to calculate the quantitative and qualitative predictions for large future tax settlement amounts. We define large settlements as those with an above median dollar value of settlements in year $t+1$ scaled by total assets in year t. We set the median value using only data with non-zero settlements in year $t+1$. The quantitative (qualitative) prediction is labeled $HIGH\ SETTLE\ PRED(QUANT)_{i,t}$ ($HIGH\ SETTLE\ PRED(QUAL)_{i,t}$). We run regressions similar to that in Equation (1) to compare the explanatory power of each prediction and test the incremental informativeness. We present our results in Table 9.

[Insert Table 9 here]

Columns 1 and 2 present the results using $HIGH\ SETTLE\ PRED(QUAL)_{i,t}$ and $HIGH\ SETTLE\ PRED(QUANT)_{i,t}$ to predict large future tax settlements, respectively. The coefficients on

both predictions are both statistically positive at the 1% level. The Pseudo R^2 (area under the ROC curve) is equal to 5.3% (0.669) when including *HIGH SETTLE PRED(QUAL)*$_{i,t}$ as the sole regressor and 11.5% (0.746) when including *HIGH SETTLE PRED(QUANT)*$_{i,t}$ as the sole regressor. Column 3 presents the results including both *HIGH SETTLE PRED(QUAL)*$_{i,t}$ and *HIGH SETTLE PRED(QUANT)*$_{i,t}$ as regressors. The coefficients on both predictions remain positive and statistically significant at the 1% level, and the pseudo R^2 (area under the ROC curve) increases to 12.1% (0.749).

In Column 4, we present the regression for the prediction combining the quantitative and qualitative information from the Form 10-K (*HIGH SETTLE PRED (QUAL + QUANT)*$_{i,t}$). We continue to find a positive and significant coefficient (1% level) on *HIGH SETTLE PRED (QUAL + QUANT)*$_{i,t}$. We note that the pseudo R^2 and area under the ROC curve are similar to values reported in Column 3. Because the model with the combined measure (Column 4) has acceptable discriminatory power, future researchers can use this measure to identify the likelihood of large settlements.

Predicting Unfavorable Future Tax Settlements

In our last set of analyses, we examine whether we can predict "unfavorable" future tax settlements. We define an unfavorable tax settlement to be a settlement for which the firm is under-reserved (i.e., the total settlement is greater than the liabilities for uncertain tax positions). If the amount of the firm's tax settlement is less than or equal to the reserve for uncertain tax positions, then a settlement has no effect on earnings because the firm has already recognized the tax expense in a prior period. However, if a firm is under-reserved (i.e., the amount of the tax settlement is greater than the amount of reserves for uncertain tax positions), then a settlement can increase tax expense and decrease net income because the firm recognized too little tax expense in a prior

period.[15] To determine whether a tax settlement is unfavorable, we rely on information reported in the tax footnote related to reserves for uncertain tax positions. Specifically, we sum increases to reserves related to prior year positions and decreases to reserves related to prior year positions. If a firm is over-reserved (under-reserved), the firm will report a net decrease (increase) in reserves related to prior year positions. Therefore, if in the year of a settlement, prior year increases are greater than prior year decreases, we consider the settlement to be unfavorable.[16]

We follow the same method previously described in Section III to calculate the quantitative and qualitative predictions for the likelihood of an unfavorable tax settlement. Rather than predicting the likelihood of tax settlement in year *t+1*, we predict the likelihood of an unfavorable tax settlement in *t+1*. The quantitative (qualitative) prediction is labeled *UNFAV SETTLE PRED(QUANT)*$_{i,t}$ (*UNFAV SETTLE PRED(QUAL)*$_{i,t}$). We run regressions similar to that in Equation (1) to compare the explanatory power of each prediction and to test the incremental informativeness. We present the result from predicting unfavorable tax settlements in Table 10.

[Insert Table 10 here]

The four columns in Table 10 mirror the columns in Table 9. We find that the coefficients on *UNFAV SETTLE PRED(QUANT)*$_{i,t}$ and *UNFAV SETTLE PRED(QUAL)*$_{i,t}$ are positive and significant at the 1% level when they are separately included as regressors in Column 1 and 2. We note that the area under the ROC curve is equal to 72.2% (70.6%) when *UNFAV SETTLE PRED(QUAL)*$_{i,t}$ (*UNFAV SETTLE PRED(QUANT)*$_{i,t}$) is the sole regressor. We continue to find

[15] Not all reserves for uncertain tax positions will affect the ETR if recognized. We still view a settlement as unfavorable if the firm is under-reserved.

[16] Importantly, prior year increases and decreases can also reflect changes in judgement. To the extent that prior year changes are unrelated to settlements, we acknowledge that it is possible that we classify favorable settlements as unfavorable settlements, and vice versa. We also considered using a measure developed by Finley (2019) that relies on interest and penalties to identify unfavorable settlements, but we chose to use prior year increases and decreases to preserve our sample size given that interest and penalties are missing for a substantial number of firm-years (e.g., Finley (2019) loses 33.6% of observations because of missing interest and penalties).

that the coefficients on the quantitative and qualitative predictions are positive and significant at the 1% level when both are included simultaneously in the regression model (see Column 3). The area under the ROC curve in Column 3 increases to 75.4%, which suggests that the model has acceptable discriminatory power (Hosmer et al. 2013). Lastly, we find that the coefficient on the combined quantitative and qualitative prediction (*UNFAV SETTLE PRED (QUAL + QUANT)*$_{i,t}$) is positive and significant at the 1% level in Column 4. Similar to prior tables, we note that the area under the ROC curve decreases from 75.4% in Column 3 to 74.7% in Column 4.

Collectively, the results in this section suggest that we are not only able to identify information that is useful in assessing the likelihood of a future tax settlement, but we are also able to predict the likelihood of an unfavorable future tax settlement.

VI. CONCLUSION

In this study, we examine whether quantitative information and qualitative information reported in the Form 10-K can predict the likelihood of a future tax settlement. Following prior research (Donovan et al., 2019; Frankel et al., 2019), we use three machine learning methods to identify language in the Form 10-K that is useful in predicting the likelihood of a future tax settlement. We also follow prior research to identify quantitative characteristics that are likely to be associated with the likelihood of a future tax settlement.

We find that both quantitative information and qualitative information in the Form 10-K are separately useful in predicting the likelihood of a future tax settlement. When including both the qualitative and quantitative predictions in the same regression model, we find that both predictions continue to predict future tax settlements, which suggests that each prediction identifies non-overlapping information. When adding our qualitative prediction to a model including the quantitative prediction, we note an economically significant increase in our ability to predict future

tax settlements, which suggests that the language in the Form 10-K is particularly useful in evaluating the likelihood of future tax settlements.

Next, we examine the relation between predicted future tax settlements and cash holdings. We find that among firms with reserves for uncertain tax positions, firms with a higher likelihood of future tax settlements hold *less* cash than firms with a lower likelihood of future tax settlements. We also find that we are able to predict the amount of a future tax settlement as well as "unfavorable" tax settlements. Overall, our results suggest that the quantitative and qualitative information in the Form 10-K are useful in predicting tax settlement outcomes.

Our study contributes to the tax literature by providing an *ex-ante* measure of the likelihood that a firm will settle with tax authorities in the future. The only other ex-ante measure of tax outcomes used in prior research is the reserve for uncertain tax positions, which our results suggest poorly predicts future tax settlements. This result highlights the need for future research to carefully consider whether the reserve for uncertain tax positions represents an accurate *ex-ante* measure for corporate tax outcomes. Prior studies have used effective tax rate and book-tax difference measures (Hanlon and Heitzman 2010) to measure tax planning. These measures capture *ex-post* realizations from tax planning, but do not necessarily capture future tax outcomes. To the best of our knowledge, we are the first to use both quantitative and qualitative information from the Form 10-K to create an *ex-ante* measure of future tax settlements. We provide evidence that the language in the tax footnote and MD&A is particularly useful in predicting the likelihood of a tax settlement. Our composite measure, which combines both the language and firm fundamentals from the Form 10-K, is a succinct measure that can be used by future researchers to predict future tax settlements. Finally, our findings on the effect of predicted future tax settlements on cash holdings suggest a nuanced relation between tax uncertainty and cash holdings.

REFERENCES

Bauer, A., Klassen, K., 2017. Assessing the market reaction to unfavorable tax settlements: Using textual analysis to categorize ambiguous tabulated disclosures. Working paper.

Blei, D., McAuliffe, J., 2007. Supervised topic models. Advances in Neural Information Processing Systems, 20.

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

Chen, S., Powers, K., Stomberg, B., 2019. Media coverage of corporate taxes. The Accounting Review, Forthcoming.

Chow, T., Klassen, K.J., Liu, Y., 2016. Targets' tax shelter participation and takeover premiums. Contemporary Accounting Research, 33(4), 1440-1472.

Dhaliwal, D.S., Goodman, T.H., Hoffman, P.J., Schwab, C.M., 2017. The incidence, valuation and management of tax-related reputational costs: Evidence from a period of protest. Working paper.

Donovan, J., Jennings, J.N., Koharki, K., Lee, J.A., 2019. Determining Credit Risk Using Qualitative Disclosure. Working paper.

Dyreng, S.D., Hoopes, J.L., Wilde, J.H., 2016. Public pressure and corporate tax behavior. Journal of Accounting Research, 54(1), 147-186.

Dyreng, S., Hanlon, M., Maydew, E., 2008. Long-run corporate tax avoidance. The Accounting Review, 83, 61-82.

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

Financial Accounting Standards Board (FASB). 2006. Accounting for Uncertainty in Income Taxes: An Interpretation of FASB Statement No. 109. Interpretation No. 48 (FIN 48). (June). Norwalk, CT: FASB.

Finley, A., 2019. The impact of large tax settlement favorability on firms' subsequent tax avoidance. Review of Accounting Studies, 24, 156-187.

Frankel, R., Jennings, J., Lee, J., 2016. Using unstructured and qualitative disclosures to explain accruals. Journal of Accounting and Economics, 62(2-3), 209-227.

Frankel, R., Jennings, J., Lee, J., 2019. Assessing the relative explanatory power of narrative content measures using conference calls, earnings predictions and analyst revisions. Working paper.

Graham, J.R., Hanlon, M., Shevlin, T., Shroff, N., 2014. Incentives for tax planning and avoidance: Evidence from the field. The Accounting Review, 89(3), 991-1023.

Graham, J.R., Tucker, A.L., 2006. Tax shelters and corporate debt policy. Journal of Financial Economics, 81(3), 563-594.

Hanlon, M., Heitzman, S., 2010. A review of tax research. Journal of Accounting and Economics, 50(2-3), 127-178.

Hanlon, M., Maydew, E., Saavedra, D., 2017. The taxman cometh: Does tax uncertainty affect corporate cash holdings? Review of Accounting Studies, 22(3), 1198-1228.

Hosmer, D., Lemeshow, S., Sturdivant, R., 2013. Applied Logistic Regression. Wiley Series in Probability and Statistics, 3rd ed. New York: John Wiley & Sons Inc.

Houlder, V., Pickard, J., Lucas, L., Jopson, B., December 6, 2012. Starbucks to pay £20m UK corporate tax. Financial Times.

Kothari, S.P., Li, X., Short, J., 2009. The effect of disclosures by management, analysts, and financial press on the equity cost of capital: A study using content analysis. The Accounting Review, 84, 1639-70.

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

Lehavy, R., Li, F., Merkley, K., 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. The Accounting Review, 86(3), 1087-1115.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45, 221-247.

Lisowsky, P., 2010. Seeking shelter: Empirically modeling tax shelters using financial statement information. The Accounting Review, 85(5), 1693-1720.

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

Marr, C., Murray, C., April 4, 2016. IRS funding cuts compromise taxpayer service and weaken enforcement. Center on Budget and Policy Priorities (online at http://www.cbpp.org/research/federal-tax/irs-funding-cuts-compromise-taxpayer-service-and-weaken-enforcement).

Nessa, M., Schwab, C., Stomberg, B., Towery, E., 2019. How do IRS resources affect the corporate audit process? The Accounting Review, Forthcoming.

Peterson, K., Schmardebeck, R., Wilks, T.J., 2015. The earnings quality and information processing effects of accounting consistency. The Accounting Review, 90(6), 2483-2514.

Robinson, L.A., Stomberg, B., Towery, E., 2016. One size does not fit all: How the uniform rules of FIN 48 affect the relevance of income tax accounting. The Accounting Review, 91(4), 1195-1217.

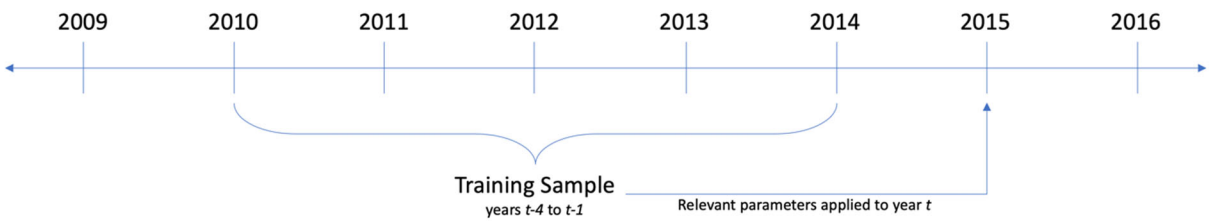
Tetlock, P., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. Journal of Finance, 63(3), 1437-1467.

Thomas, J., Zhang, F., 2014. Valuation of tax expense. Review of Accounting Studies, 19(4), 1436-1467.

U.S. Department of the Treasury., 1999. The Problem of Corporate Tax Shelters: Discussion, Analysis and Legislative Proposals. July. Washington, D.C.: Government Printing Office.

Wilson, R.J., 2009. An examination of corporate tax shelter participants. The Accounting Review, 84(3), 969-999.

Figure 1
Example of training sample and prediction year



In this example, the disclosures in years 2010 to 2014 represent the training sample for the future tax settlement predictions. The relevant parameters from each prediction method are then applied to 2015.

Figure 2
Tax Footnote Category Importance

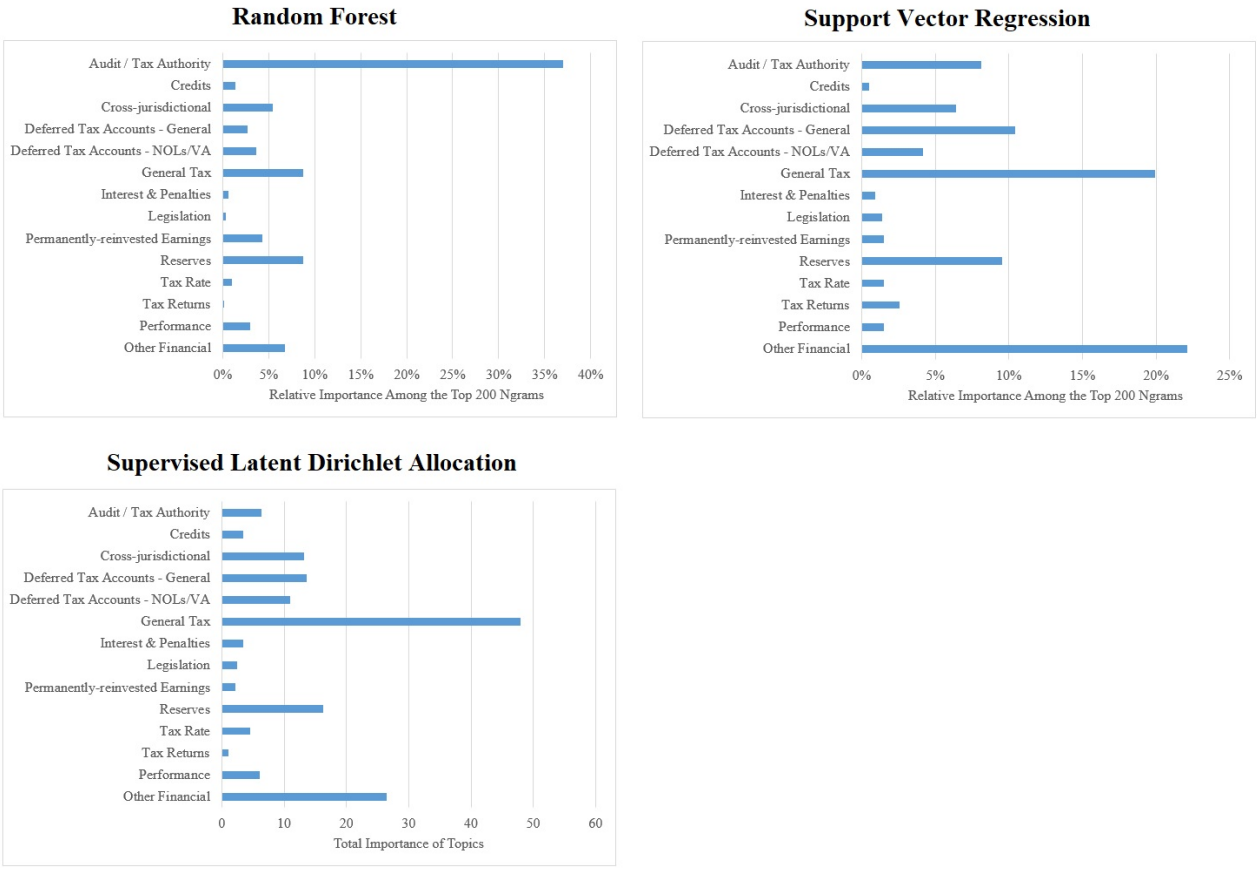
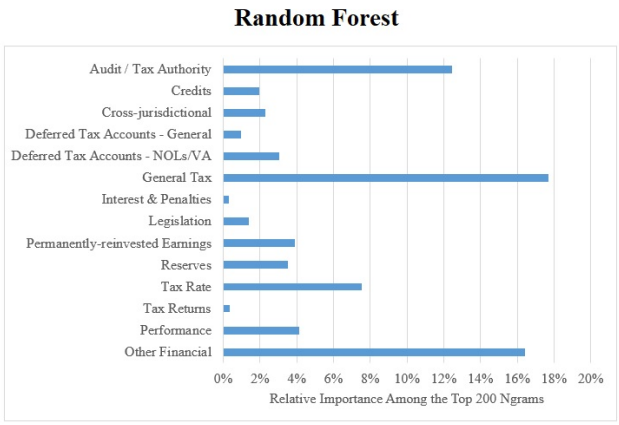
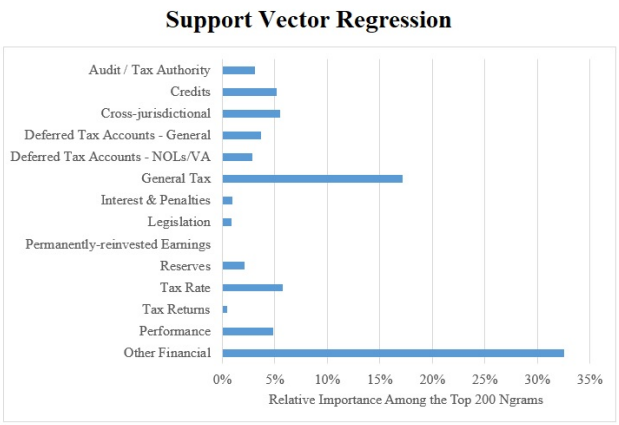
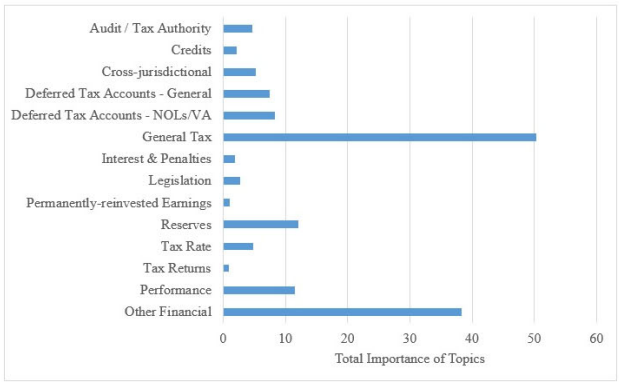


Figure 3
MD&A Category Importance







Appendix A.1 – Top 200 Important Words and Phrases
Tax Footnote – Random Forest

CATEGORY	OVERALL IMPORTANCE	RELATIVE IMPORTANCE AMONG THE TOP 200
AUDIT/TAX AUTHORITY		
settlement	9.52%	26.10%
audit	1.36%	3.74%
resolution	1.00%	2.74%
audit settlement	0.53%	1.46%
settlement balance	0.30%	0.82%
irs irs	0.16%	0.44%
taxing authority	0.12%	0.32%
subject	0.09%	0.24%
examination	0.08%	0.23%
authority	0.08%	0.21%
settlement lapse	0.07%	0.20%
under audit	0.07%	0.20%
settled	0.07%	0.18%
irs	0.06%	0.18%
	13.51%	37.04%
CREDITS		
development	0.17%	0.48%
foreign tax-credit-carryforward	0.15%	0.40%
development credit	0.09%	0.25%
research	0.08%	0.21%
	0.49%	1.34%
CROSS-JURISDICTIONAL		
jurisdiction	0.36%	1.00%
local	0.24%	0.67%
foreign	0.19%	0.52%
major jurisdiction	0.12%	0.33%
united kingdom	0.11%	0.31%
international	0.11%	0.29%
state-tax	0.10%	0.27%
certain jurisdiction	0.09%	0.25%
various state	0.08%	0.22%
brazil	0.08%	0.22%
state jurisdiction	0.08%	0.21%
state-income-tax	0.07%	0.20%
kingdom	0.07%	0.20%
canada	0.07%	0.20%
switzerland	0.06%	0.18%
state local	0.06%	0.18%
jurisdiction remain	0.06%	0.18%
	1.97%	5.40%
DEFERRED TAX ACCOUNTS - GENERAL		
asset	0.10%	0.26%
deferred-tax-asset insurance	0.09%	0.25%
deferred-tax-asset	0.09%	0.24%
expire	0.09%	0.23%
carryforward	0.08%	0.22%
net deferred-tax-liability	0.08%	0.21%
net deferred-tax-asset	0.08%	0.21%
expire between	0.07%	0.20%
employee benefit	0.07%	0.18%
deferred	0.07%	0.18%
expire over	0.07%	0.18%
deferred-tax-expense	0.07%	0.18%
expiration	0.06%	0.18%
	1.00%	2.73%
DEFERRED TAX ACCOUNTS - NOLS/VA		
ownership change	0.80%	2.18%
realization	0.10%	0.28%
nol	0.08%	0.23%
less valuation-allowance	0.08%	0.22%
nole	0.07%	0.19%
valuation-allowance related	0.07%	0.19%
nol-carryforward	0.07%	0.18%
valuation-allowance	0.06%	0.18%
	2.33%	3.64%
GENERAL TAX		
tax position	0.18%	0.49%
federal	0.15%	0.42%
current-asset	0.14%	0.38%
lease	0.11%	0.31%
table exclude	0.10%	0.28%
income-tax-expense benefit	0.10%	0.27%
minimum	0.10%	0.27%
asc	0.10%	0.27%
due	0.09%	0.26%
excess	0.09%	0.24%

CATEGORY	OVERALL IMPORTANCE	RELATIVE IMPORTANCE AMONG THE TOP 200
GENERAL TAX (CONTINUED)		
purpose	0.09%	0.24%
taxable-income	0.09%	0.24%
income-taxes the	0.08%	0.22%
income-tax	0.08%	0.22%
tax	0.08%	0.22%
domestic	0.08%	0.22%
federal-income-tax	0.08%	0.21%
stock-based-compensation	0.07%	0.21%
federal-income-tax-return	0.07%	0.20%
position	0.07%	0.20%
significant component	0.07%	0.20%
income-tax income	0.07%	0.20%
non deductible	0.07%	0.20%
reversal	0.07%	0.20%
liability	0.07%	0.20%
income-tax net	0.07%	0.19%
federal state	0.07%	0.19%
issue	0.07%	0.19%
component	0.07%	0.19%
off-set	0.07%	0.19%
federal-tax-rate	0.07%	0.18%
plan	0.07%	0.18%
income before-income-tax	0.07%	0.18%
payment	0.07%	0.18%
income-tax matter	0.06%	0.17%
attributable	0.06%	0.17%
income-tax-provision benefit	0.06%	0.17%
reconciliation	0.06%	0.17%
	3.18%	8.71%
INTEREST AND PENALTIES		
penalty	0.12%	0.34%
interest	0.10%	0.27%
	0.22%	0.61%
LEGISLATION		
internal-revenue-code	0.14%	0.38%
	0.14%	0.38%
PERMANENTLY-REINVESTED EARNINGS		
practicable	0.97%	2.66%
cumulative	0.11%	0.29%
undistributed-earnings	0.08%	0.22%
unremitted	0.07%	0.20%
permanently reinvested	0.07%	0.20%
repatriation	0.07%	0.19%
indefinitely	0.07%	0.18%
permanently	0.06%	0.18%
foreign earnings	0.06%	0.17%
	1.57%	4.30%
RESERVES		
statute	1.52%	4.17%
lapse	0.72%	1.96%
certain	0.23%	0.64%
benefit	0.17%	0.46%
limitation	0.11%	0.31%
tax benefit	0.10%	0.29%
taken	0.07%	0.20%
reserve	0.07%	0.18%
exclude	0.07%	0.18%
position may	0.06%	0.18%
being recognized	0.06%	0.18%
	3.19%	8.74%
TAX RATE		
effective-tax-rate	0.09%	0.24%
statutory tax-rate	0.07%	0.20%
rate	0.07%	0.20%
statutory income-tax-rate	0.07%	0.19%
consolidated effective-tax-rate	0.07%	0.19%
	0.37%	1.02%
TAX RETURNS		
tax-return	0.06%	0.17%
	0.06%	0.17%
PERFORMANCE		
earnings	0.44%	1.20%
income loss	0.16%	0.45%
income	0.09%	0.26%

CATEGORY	OVERALL IMPORTANCE	RELATIVE IMPORTANCE AMONG THE TOP 200
PERFORMANCE (CONTINUED)		
loss before-income-tax	0.09%	0.24%
expense	0.09%	0.24%
net-income	0.08%	0.22%
operation	0.08%	0.22%
historical	0.07%	0.19%
	1.10%	3.02%
OTHER FINANCIAL		
ownership	0.37%	1.00%
merchandise	0.12%	0.33%
recorded	0.11%	0.32%
accrued-liability	0.11%	0.31%
year-ended	0.11%	0.30%
investment	0.10%	0.27%
prior-year	0.10%	0.26%
subsidiary	0.10%	0.26%
employee	0.09%	0.25%
adjustment	0.09%	0.24%
fiscal	0.09%	0.23%
compensation	0.08%	0.23%
amount computed	0.08%	0.22%
book	0.08%	0.22%
depreciation	0.08%	0.21%
amount	0.08%	0.21%
pension	0.08%	0.21%
year	0.08%	0.21%
financial-statement	0.07%	0.20%
foreign-currency-translation	0.07%	0.20%
cash	0.07%	0.18%
gain	0.07%	0.18%
recorded thereon	0.06%	0.18%
consolidated-financial-statement	0.06%	0.18%
consolidated	0.06%	0.18%
cost	0.06%	0.17%
	2.46%	6.75%
UNCATEGORIZED		
possible	2.29%	6.27%
increase	0.14%	0.38%
net	0.14%	0.37%
various	0.13%	0.35%
note	0.13%	0.35%
numerous	0.11%	0.30%
during	0.11%	0.29%
computed	0.11%	0.29%
decreased	0.10%	0.29%
connection	0.10%	0.28%
upon	0.10%	0.26%
table	0.10%	0.26%
relate	0.09%	0.25%
addition	0.09%	0.25%
resulted	0.09%	0.24%
related	0.09%	0.24%
totaled	0.09%	0.24%
additional	0.09%	0.24%
expected	0.09%	0.23%
remain	0.08%	0.23%
prior	0.08%	0.23%
following table	0.08%	0.22%
whether	0.08%	0.22%
may	0.08%	0.22%
both	0.08%	0.22%
applying	0.08%	0.21%
dollar	0.08%	0.21%
completed	0.08%	0.21%
content	0.08%	0.21%
decrease	0.07%	0.20%
ending	0.07%	0.20%
change	0.07%	0.19%
remaining	0.07%	0.19%
received	0.07%	0.19%
result	0.07%	0.19%
represented	0.07%	0.18%
begin	0.07%	0.18%
increase decrease	0.07%	0.18%
provided	0.07%	0.18%
associated	0.06%	0.17%
based	0.06%	0.17%
between	0.06%	0.17%
each	0.06%	0.17%
	5.89%	16.14%

Appendix A.2 – Top 200 Important Words and Phrases
Tax Footnote – Support Vector Machines

CATEGORY	COEFFICIENT	RELATIVE IMPORTANCE AMONG THE TOP 200
AUDIT/TAX AUTHORITY		
settled position	-1.99	1.33%
state-tax audit	-0.92	0.62%
irs commenced	-0.91	0.61%
irs concluded	-0.83	0.56%
agreement resulted	-0.67	0.45%
examination process	0.62	0.41%
under review	0.62	0.41%
settlement reduction	0.62	0.42%
settlement resulted	0.63	0.42%
being audited	0.68	0.45%
authority decrease	0.72	0.48%
authority balance	1.17	0.78%
authority reduction	1.73	1.16%
		8.10%
CREDITS		
expense credit	-0.79	0.53%
		0.53%
CROSS-JURISDICTIONAL		
indiana	-0.72	0.48%
canadian tax	-0.71	0.48%
state-tax refund	-0.65	0.44%
united-states australia	-0.63	0.42%
commonwealth	-0.60	0.40%
international restructuring	-0.59	0.39%
multi	-0.57	0.38%
city tax	0.60	0.40%
brazil india	0.63	0.42%
foreign outside	0.65	0.44%
income apportionment	0.66	0.44%
various taxing	0.70	0.47%
united-states foreign	0.87	0.58%
canada australia	0.97	0.65%
		6.39%
DEFERRED TAX ACCOUNTS - GENERAL		
asset before	-1.52	1.02%
deferred-tax-asset goodwill	-0.76	0.51%
deferred-tax effect	-0.67	0.45%
unrecognized deferred-tax-asset	-0.64	0.43%
deferred credit	-0.57	0.38%
deferred-tax-expense	-0.57	0.38%
cumulative translation	-0.56	0.38%
deferred-tax-liability established	0.59	0.39%
debt expense	0.61	0.41%
kingdom deferred-tax-asset	0.62	0.41%
instrument	0.62	0.41%
deferred-tax-liability accumulated	0.63	0.42%
deferred-tax-rate	0.68	0.45%
deferred-tax-expense benefit	0.70	0.47%
deferred income	0.72	0.48%
deferred-tax-asset decreased	0.75	0.50%
asset-impairment	0.75	0.50%
accounting adjustment	0.80	0.54%
noncurrent-liability	0.84	0.56%
deferred-tax-liability inventory	0.91	0.61%
long-term deferred-revenue	1.08	0.73%
		10.42%
DEFERRED TAX ACCOUNTS - NOLS/VA		
such nol-carryforward	-0.78	0.52%
allowed	-0.77	0.52%
valuation-allowance increase	-0.60	0.40%
nol-carryforward generated	-0.58	0.39%
tax valuation-allowance	0.63	0.42%
gross nol-carryforward	0.66	0.44%
associated valuation-allowance	0.69	0.46%
loss expire	0.78	0.52%
nol-carryforward total	0.78	0.52%
		4.19%
GENERAL TAX		
management judgment	-1.75	1.17%
income-tax receivable	-0.93	0.62%
income-tax pre-tax	-0.88	0.59%
period expense	-0.85	0.57%
income-taxesincome-tax-expense benefit	-0.77	0.51%
united-states tax	-0.77	0.51%
federal-income-tax liability	-0.75	0.50%
income-taxesthe following	-0.75	0.50%
deductible until	-0.74	0.50%

CATEGORY	COEFFICIENT	RELATIVE IMPORTANCE AMONG THE TOP 200
GENERAL TAX (CONTINUED)		
federal valuation-allowance	-0.72	0.48%
federal effect	-0.67	0.45%
prepaid tax	-0.66	0.44%
tax claim	-0.65	0.43%
income-taxesfor	-0.64	0.43%
current accrued	-0.63	0.42%
net benefit	-0.63	0.42%
depletion	-0.63	0.42%
stock-based-compensation deduction	-0.59	0.39%
deductible non	-0.58	0.39%
recognize tax	-0.56	0.38%
current expense	-0.56	0.38%
federal purpose	-0.56	0.38%
partnership tax	0.59	0.39%
net accrued	0.59	0.40%
payable federal	0.61	0.41%
income-tax-benefit during	0.66	0.44%
character	0.67	0.45%
certain income-tax	0.69	0.46%
net noncurrent	0.72	0.48%
recognized during	0.74	0.49%
income-tax-benefit recognized	0.77	0.52%
recognized accrued	0.83	0.56%
significant interest	0.85	0.57%
expense tax	0.85	0.57%
deductible portion	0.87	0.58%
income-taxesthe reconciliation	0.88	0.59%
foreign-currency	0.89	0.60%
recognized reduce	0.99	0.66%
income-tax-expense attributable	1.36	0.91%
		19.95%
INTEREST AND PENALTIES		
penalty balance	-0.58	0.39%
interest net	0.83	0.56%
		0.94%
LEGISLATION		
internal-revenue-code limit	-0.89	0.60%
state law	-0.66	0.44%
legal settlement	-0.57	0.38%
		1.42%
PERMANENTLY-REINVESTED EARNINGS		
cumulative foreign	-0.77	0.51%
invested earnings	-0.57	0.38%
reinvest indefinitely	0.92	0.61%
		1.50%
RESERVES		
released during	-0.91	0.61%
expect change	-0.81	0.54%
reduce utb	-0.77	0.52%
limitation utb	-0.76	0.50%
limitation balance	-0.69	0.46%
uncertain-tax-position relate	-0.68	0.45%
limitation related	-0.64	0.43%
remaining utb	-0.62	0.42%
gross unrecognized	-0.59	0.39%
uncertain-tax-position due	0.59	0.39%
remaining liability	0.62	0.41%
exclude interest	0.62	0.41%
decrease current	0.62	0.42%
uncertain state-tax	0.63	0.42%
income-tax reserve	0.68	0.46%
prior-year increase	0.73	0.48%
utb totaled	1.02	0.68%
material increase	1.05	0.70%
limitation gross	1.27	0.85%
		9.55%
TAX RATE		
therefore subject	-1.07	0.71%
effective-tax-rate benefited	-0.56	0.38%
lower effective-tax-rate	0.66	0.44%
		1.53%
TAX RETURNS		
separate federal-income-tax-return	-1.18	0.79%
prior-year tax-return	-0.72	0.48%
filed amended	-0.58	0.39%
return adjustment	1.33	0.89%
		2.55%

CATEGORY	COEFFICIENT	RELATIVE IMPORTANCE AMONG THE TOP 200
PERFORMANCE		
loss benefit	-1.04	0.70%
earnings before-tax	0.60	0.40%
continuing-operations united-states	0.63	0.42%
		1.52%
OTHER FINANCIAL		
amount resulting	-1.12	0.75%
certain investment	-1.09	0.73%
total consolidated	-0.89	0.59%
until paid	-0.74	0.50%
payment made	-0.73	0.49%
officer	-0.69	0.46%
discrete benefit	-0.68	0.46%
closure	-0.68	0.45%
lifo	-0.65	0.44%
cash paid	-0.65	0.43%
subsequent year	-0.65	0.43%
receipt	-0.64	0.43%
compensation cost	-0.63	0.42%
subsequent period	-0.61	0.41%
paid received	-0.60	0.40%
year including	-0.59	0.40%
fasb guidance	-0.59	0.39%
equity adjustment	-0.58	0.39%
pension obligation	-0.58	0.38%
book purpose	-0.57	0.38%
foreign-exchange loss	-0.57	0.38%
year still	0.59	0.39%
insurance policy	0.59	0.40%
stock exercise	0.60	0.40%
stock appreciation	0.61	0.40%
unrealized capital	0.61	0.41%
book expense	0.62	0.41%
reduce goodwill	0.62	0.42%
restructured	0.64	0.43%
investment security	0.65	0.43%
acquisition completed	0.65	0.44%
before interest	0.66	0.44%
subsidiary except	0.66	0.44%
restructuring-cost	0.66	0.44%
indirect effect	0.68	0.45%
account deduction	0.69	0.46%
comparable amount	0.70	0.47%
year remaining	0.70	0.47%
participated	0.75	0.50%
divestiture	0.80	0.53%
spinoff	0.83	0.55%
continuing-operation before-tax	0.85	0.57%
recorded interest	0.92	0.62%
todecember	1.05	0.70%
benefit before	1.12	0.75%
net expense	1.32	0.88%
		22.13%
UNCATEGORIZED		
ended dollar	-1.14	0.76%
along	-1.02	0.68%
both fiscal	-0.71	0.48%
removed	-0.70	0.47%
change cannot	-0.62	0.42%
announced	-0.61	0.41%
adjusting	-0.59	0.40%
unlikely	-0.59	0.40%
making	-0.59	0.39%
none	-0.57	0.38%
giving	-0.57	0.38%
during both	-0.56	0.38%
remaining decrease	0.59	0.39%
below fiscal-year-ended	0.64	0.43%
addition due	0.66	0.44%
engaged	0.68	0.46%
conducting	0.72	0.48%
increase during	0.75	0.50%
consent	0.75	0.50%
terminal	0.80	0.53%
		9.26%

Appendix A.3 – Top 10 Most Negative and Most Positive Topics
Tax Footnote – Supervised Latent Dirichlet Allocation

MOST NEGATIVE TOPICS

Topic	Coefficient	Audit / Tax Authority	Credits	Cross-jurisdictional	Deferred Tax Accounts - General	Deferred Tax Accounts - NOLs/VA	General Tax	Interest & Penalties	Legislation	Permanently-reinvested Earnings	Reserves	Tax Rate	Tax Returns	Performance	Other Financial
income-tax, tax, current, provision, related, balance, subject, rate, liability, statutory	-118.42	0.06	0.00	0.00	0.00	0.00	0.68	0.00	0.05	0.00	0.00	0.06	0.00	0.00	0.00
less, less valuation-allowance, balance, current, follow, deferred-tax, income-tax-expense, deferred-tax-asset, reconciliation, tax	-9.08	0.00	0.00	0.00	0.14	0.14	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
financial-statement, income-tax, recognition, recognized, expected, measurement, position, amount, fasb, penalty	-8.39	0.00	0.00	0.00	0.00	0.00	0.49	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.37
income-tax-return, jurisdiction, file, various, file income-tax-return, various state, state, federal jurisdiction, subsidiary file, federal	-8.34	0.00	0.00	0.31	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.44	0.00	0.05
valuation-allowance, more, cumulative, future, deferred-tax-asset, result, whether, portion, realized, required	-7.89	0.00	0.00	0.00	0.09	0.32	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
nol-carryforward, state, state nol-carryforward, federal, federal nol-carryforward, expire, nol-carryforward expire, utilized, intangible-asset, anticipate	-7.88	0.00	0.00	0.16	0.07	0.63	0.12	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
open, year, tax year, remain, remain open, examination, open tax, tax, penalty related, component	-7.15	0.05	0.00	0.00	0.00	0.00	0.23	0.01	0.00	0.00	0.40	0.00	0.00	0.00	0.19
difference, permanent, expense, utb, permanent difference, current, tax, benefit, follow, federal	-6.39	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.15	0.30	0.00	0.12	0.00
income-tax-expense, income-tax-expense benefit, benefit, total income-tax-expense, total, current income-tax-expense, deferred-income-tax-expense, before-income-tax-expense, deferred-income-tax-expense benefit, benefit federal	-6.17	0.00	0.00	0.00	0.04	0.00	0.78	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.02
year-ended, follow year-ended, follow, following year-ended, equipment, during, reconciliation, statutory, tax year-ended, below year-ended	-6.16	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.91

MOST POSITIVE TOPICS

Topic	Coefficient	Audit / Tax Authority	Credits	Cross-jurisdictional	Deferred Tax Accounts - General	Deferred Tax Accounts - NOLs/VA	General Tax	Interest & Penalties	Legislation	Permanently-reinvested Earnings	Reserves	Tax Rate	Tax Returns	Performance	Other Financial
examination, under, state, year, under examination, utb, federal-income-tax-return, federal, tax year, various state	20.96	0.41	0.00	0.13	0.00	0.00	0.17	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.08
tax, year, reduction, result, tax liability, addition, provided, certain, multiple, benefit	12.44	0.00	0.00	0.03	0.00	0.00	0.56	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.10
audit, irs, settlement, issue, completed, settled, under audit, irs irs, audit settlement, irs audit	12.38	0.91	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
accrued, accrued interest, interest, penalty, utb, penalty related, end, beginning, recognize, recognize accrued	11.77	0.00	0.00	0.00	0.00	0.00	0.07	0.74	0.00	0.00	0.11	0.00	0.00	0.00	0.00
between, expire, related, expire between, utb, deferred-tax-asset, income-tax-expense, total, liability, difference between	8.59	0.00	0.00	0.00	0.38	0.00	0.08	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.03
income-tax, determination, upon, unrecognized, income-tax liability, calculation, associated, amount, component, state-income-tax	7.67	0.00	0.00	0.06	0.00	0.00	0.42	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.07
foreign, foreign jurisdiction, certain, jurisdiction, certain foreign, foreign nol-carryforward, foreign tax, foreign deferred-tax-asset, federal state, before-income-tax	7.51	0.00	0.00	0.82	0.01	0.05	0.01	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.01
related, increase, decrease, increase related, decrease related, balance, prior-year tax, penalty related, month, subject	7.27	0.01	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.00	0.18	0.00	0.00	0.00	0.01
permanently, permanently reinvested, earnings, reinvested, subsidiary, considered, undistributed-earnings, tax, jurisdiction, provided	7.20	0.00	0.00	0.03	0.00	0.00	0.04	0.00	0.00	0.56	0.00	0.00	0.00	0.14	0.11
entity, above, impact, related, future, certain, legal, position, described, off-set	6.89	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.00	0.20	0.00	0.00	0.00	0.25

Appendix A.4 – Top 200 Important Words and Phrases
MD&A – Random Forest

CATEGORY	OVERALL IMPORTANCE	RELATIVE IMPORTANCE AMONG THE TOP 200
AUDIT/TAX AUTHORITY		
settlement	1.27%	4.96%
resolution	0.75%	2.91%
tax audit	0.36%	1.40%
audit	0.12%	0.46%
audit settlement	0.11%	0.43%
taxing authority	0.10%	0.39%
tax-authority regarding	0.09%	0.34%
irs	0.08%	0.30%
examination	0.07%	0.28%
irs audit	0.06%	0.24%
favorable resolution	0.06%	0.24%
authority	0.06%	0.24%
foreign tax-authority	0.06%	0.24%
	3.19%	12.44%
CREDITS		
research	0.14%	0.54%
general	0.13%	0.49%
development	0.10%	0.38%
tax-credit	0.08%	0.31%
foreign tax-credit	0.07%	0.27%
	0.51%	1.99%
CROSS-JURISDICTIONAL		
state	0.12%	0.48%
foreign	0.12%	0.48%
jurisdiction	0.12%	0.47%
various jurisdiction	0.08%	0.32%
local	0.08%	0.31%
country	0.06%	0.25%
	0.59%	2.31%
DEFERRED TAX ACCOUNTS - GENERAL		
reduce deferred-tax-asset	0.09%	0.35%
deferred-tax-liability	0.09%	0.34%
deferred-income-tax asset	0.07%	0.28%
	0.25%	0.97%
DEFERRED TAX ACCOUNTS - NOLS/VA		
full valuation-allowance	0.25%	0.96%
valuation-allowance	0.12%	0.48%
nol-carryforward	0.10%	0.40%
deferred-tax-asset valuation-allowance	0.09%	0.34%
value	0.09%	0.34%
valuation-allowance against	0.07%	0.28%
available evidence	0.07%	0.26%
	0.79%	3.06%
GENERAL TAX		
tax	0.75%	2.93%
favorable	0.46%	1.81%
matter	0.31%	1.21%
taxing	0.16%	0.64%
income-tax-expense	0.14%	0.54%
asc	0.12%	0.48%
stock-based-compensation	0.11%	0.44%
federal	0.11%	0.42%
purpose	0.11%	0.41%
income-tax	0.10%	0.41%
tax planning	0.10%	0.41%
basis	0.10%	0.38%
obligation	0.10%	0.38%
income-tax-benefit	0.09%	0.36%
due	0.09%	0.36%
provision	0.09%	0.34%
adjusted	0.08%	0.32%
discrete tax	0.08%	0.31%
minimum	0.08%	0.29%
favorably	0.08%	0.29%
issue	0.07%	0.29%
taxable-income	0.07%	0.28%
income-tax position	0.07%	0.28%
unfavorably	0.07%	0.28%
recurring	0.07%	0.28%
recognized	0.07%	0.27%
tax jurisdiction	0.07%	0.27%
content income-tax	0.07%	0.27%
tax-expense	0.07%	0.26%
attributable	0.07%	0.26%
taxable	0.07%	0.26%
taxing jurisdiction	0.07%	0.26%
payment	0.07%	0.25%
strategy	0.07%	0.25%

CATEGORY	OVERALL IMPORTANCE	RELATIVE IMPORTANCE AMONG THE TOP 200
GENERAL TAX (CONTINUED)		
tax position	0.06%	0.25%
material	0.06%	0.25%
income-tax payment	0.06%	0.25%
refund	0.06%	0.25%
liability related	0.06%	0.24%
	4.55%	17.72%
INTEREST AND PENALTIES		
accrued	0.08%	0.30%
	0.08%	0.30%
LEGISLATION		
law change	0.11%	0.41%
regulation	0.10%	0.39%
internal-revenue-code	0.09%	0.34%
tax law	0.06%	0.24%
	0.36%	1.38%
PERMANENTLY-REINVESTED EARNINGS		
reinvested	0.52%	2.03%
repatriation	0.18%	0.72%
repatriated	0.09%	0.34%
foreign earnings	0.07%	0.29%
anticipated	0.07%	0.28%
permanently reinvest	0.07%	0.25%
	1.00%	3.91%
RESERVES		
impacted	0.10%	0.39%
utb including	0.10%	0.38%
reserve	0.10%	0.38%
utb	0.09%	0.36%
benefit	0.09%	0.36%
tax benefit	0.09%	0.34%
certain	0.08%	0.32%
uncertain-tax-position	0.07%	0.27%
impact	0.07%	0.26%
taken	0.07%	0.25%
applicable	0.06%	0.25%
	0.91%	3.54%
TAX RATE		
effective-tax-rate	1.01%	3.95%
effective	0.30%	1.15%
effective income-tax-rate	0.27%	1.06%
income-tax-rate	0.09%	0.35%
rate	0.09%	0.35%
tax-rate	0.09%	0.34%
statutory rate	0.08%	0.33%
	1.93%	7.52%
TAX RETURNS		
filing	0.09%	0.34%
	0.09%	0.34%
PERFORMANCE		
earnings	0.49%	1.92%
revenue	0.14%	0.56%
income	0.08%	0.32%
operating	0.08%	0.31%
net-income	0.07%	0.27%
operation	0.07%	0.26%
loss	0.06%	0.25%
adjusted-ebitda	0.06%	0.25%
	1.06%	4.13%
OTHER FINANCIAL		
pension	0.45%	1.74%
liability method	0.29%	1.14%
restructuring	0.29%	1.12%
divestiture	0.25%	0.96%
year-ended	0.21%	0.82%
sale	0.17%	0.66%
accounting	0.14%	0.55%
revenue recognition	0.14%	0.53%
period	0.12%	0.48%
fiscal	0.11%	0.41%
amount	0.10%	0.37%
equity	0.09%	0.37%
year ended	0.09%	0.36%
cash	0.09%	0.36%
include amount	0.09%	0.35%
fiscal-year	0.09%	0.35%

CATEGORY	OVERALL IMPORTANCE	RELATIVE IMPORTANCE AMONG THE TOP 200
OTHER FINANCIAL (CONTINUED)		
prior-year	0.09%	0.33%
financing	0.08%	0.31%
operating-expense	0.08%	0.31%
pay	0.08%	0.30%
financial reporting	0.08%	0.30%
consolidated	0.08%	0.29%
pension-plan	0.08%	0.29%
labor cost	0.07%	0.28%
postretirement	0.07%	0.28%
marketing	0.07%	0.28%
year	0.07%	0.28%
share-repurchase	0.07%	0.28%
employee	0.07%	0.27%
operating-profit	0.07%	0.27%
benefit related	0.07%	0.26%
prepaid	0.07%	0.26%
manufacturing	0.07%	0.26%
retirement plan	0.07%	0.26%
successful	0.06%	0.25%
improvement	0.06%	0.25%
subsidiary	0.06%	0.24%
	4.21%	16.41%
UNCATEGORIZED		
lower	0.64%	2.48%
note	0.61%	2.36%
higher	0.34%	1.32%
information	0.30%	1.17%
difference	0.18%	0.69%
various	0.16%	0.63%
item	0.14%	0.54%
increase	0.13%	0.52%
table	0.12%	0.47%
change	0.12%	0.45%
content	0.11%	0.43%
additional	0.11%	0.43%
outside	0.11%	0.42%
decreased	0.11%	0.42%
portion	0.11%	0.41%
additional information	0.10%	0.41%
decrease	0.10%	0.40%
exist	0.10%	0.38%
driven	0.10%	0.37%
less	0.10%	0.37%
reduction	0.09%	0.37%
include	0.09%	0.36%
increased	0.09%	0.36%
common	0.09%	0.34%
different	0.09%	0.34%
discussed	0.09%	0.34%
compared	0.08%	0.32%
expected	0.08%	0.32%
total	0.08%	0.31%
reduce	0.08%	0.31%
under	0.08%	0.31%
against	0.08%	0.30%
lower average	0.07%	0.29%
during	0.07%	0.29%
most	0.07%	0.29%
prior	0.07%	0.28%
resulted	0.07%	0.28%
believe	0.07%	0.28%
combined	0.07%	0.27%
more	0.07%	0.27%
included	0.07%	0.27%
incurred	0.07%	0.27%
both	0.07%	0.26%
further information	0.07%	0.26%
annual	0.07%	0.26%
addition	0.07%	0.26%
potential	0.07%	0.26%
further	0.06%	0.25%
future	0.06%	0.25%
related	0.06%	0.25%
continued	0.06%	0.24%
determining	0.06%	0.24%
	6.15%	23.96%

Appendix A.5 – Top 200 Important Words and Phrases
MD&A – Support Vector Machines

CATEGORY	COEFF ICIENT	RELATIVE IMPORTANCE AMONG THE TOP 200
AUDIT/TAX AUTHORITY		
examination including	-0.50	0.45%
state-tax audit	-0.44	0.40%
disallowed	0.43	0.39%
irs tax	0.44	0.39%
state-income-tax settlement	0.44	0.40%
under audit	0.57	0.51%
longer subject	0.61	0.55%
		3.09%
CREDITS		
minimum tax-credit	-0.94	0.85%
increased research	-0.84	0.75%
related income-tax-credit	-0.60	0.54%
credit provision	-0.47	0.42%
retroactive extension	0.42	0.38%
higher research	0.49	0.45%
tax-credit state	0.62	0.56%
federal-income-tax-credit	0.64	0.58%
tax-credit resulting	0.71	0.64%
		5.17%
CROSS-JURISDICTIONAL		
lower state-income-tax	-0.91	0.82%
state jurisdiction	-0.64	0.58%
additional foreign	-0.59	0.53%
state-income-tax-provision	0.48	0.43%
jurisdiction due	0.49	0.44%
expense foreign	0.53	0.47%
higher foreign	0.53	0.48%
international statutory	0.55	0.49%
canadian subsidiary	0.65	0.58%
lower state-tax-rate	0.74	0.67%
		5.49%
DEFERRED TAX ACCOUNTS - GENERAL		
content deferred-income-tax	-2.00	1.80%
method deferred-tax	-0.65	0.58%
recorded deferred-tax-liability	-0.57	0.52%
deferred-tax-asset include	-0.45	0.41%
temporary-difference reverse	0.44	0.40%
		3.71%
DEFERRED TAX ACCOUNTS - NOLS/VA		
nol position	-0.51	0.46%
including nol-carryforward	-0.50	0.45%
net realizable	-0.43	0.38%
valuation-allowance relate	0.42	0.38%
asset valuation-allowance	0.43	0.38%
sharing	0.43	0.39%
percentage-point increase	0.46	0.41%
		2.86%
GENERAL TAX		
income-tax derivative	-0.98	0.88%
refund received	-0.88	0.80%
judgment assumption	-0.80	0.72%
received deduction	-0.78	0.70%
total obligation	-0.74	0.67%
index income-tax	-0.55	0.49%
content income-tax-benefit	-0.54	0.49%
tax-expense included	-0.50	0.45%
tax refund	-0.50	0.45%
income-tax include	-0.49	0.44%
reduced tax	-0.49	0.44%
amount paid	-0.46	0.42%
significant estimate	-0.44	0.40%
require judgment	-0.44	0.39%
increased domestic	-0.43	0.39%
off-set future	-0.43	0.39%
excess income-tax-benefit	-0.42	0.38%
certain item	0.42	0.37%
tax relating	0.45	0.41%
income-tax-provision related	0.46	0.42%
higher current	0.47	0.42%
higher income-tax-benefit	0.48	0.43%
tax gain	0.51	0.46%
substantial tax	0.53	0.47%
material tax	0.54	0.48%
higher income-tax	0.57	0.51%
refund related	0.60	0.54%
acceleration	0.63	0.56%

CATEGORY	COEFF ICIENT	RELATIVE IMPORTANCE AMONG THE TOP 200
GENERAL TAX (CONTINUED)		
income-tax-benefit recorded	0.64	0.57%
income income-tax-provision	0.68	0.62%
lower provision	0.69	0.63%
estimated federal-income-tax	0.70	0.63%
income-tax income	0.84	0.76%
		17.17%
INTEREST AND PENALTIES		
interest-expense increased	-0.54	0.48%
including penalty	-0.53	0.48%
		0.97%
LEGISLATION		
principle	0.43	0.39%
extended retroactively	0.54	0.48%
		0.88%
PERMANENTLY-REINVESTED EARNINGS		
N/A		
RESERVES		
uncertain-tax-position off-set	-0.50	0.45%
prior-year resulting	-0.45	0.40%
including interest	0.43	0.39%
reserve resulting	0.45	0.41%
limitation lapse	0.48	0.43%
		2.08%
TAX RATE		
blended state-tax-rate	-0.85	0.77%
statutory federal	-0.80	0.72%
low tax-rate	-0.68	0.61%
overall rate	-0.58	0.52%
effective-tax-rate resulting	-0.54	0.48%
different income-tax-rate	-0.49	0.44%
state-tax-rate	-0.43	0.39%
rate decrease	-0.42	0.38%
effective income-tax	0.43	0.39%
underlying effective-tax-rate	0.54	0.48%
estimated tax-rate	0.63	0.57%
		5.76%
TAX RETURNS		
united-states federal-income-tax-return	-0.50	0.45%
		0.45%
PERFORMANCE		
income level	-0.70	0.63%
earnings-per-diluted-share	-0.51	0.46%
higher profitability	-0.43	0.38%
profit loss	-0.43	0.38%
historical earnings	0.45	0.41%
worldwide income	0.51	0.46%
higher pre-tax	0.54	0.49%
incomea	0.59	0.53%
profit sharing	0.60	0.54%
profit before-tax	0.63	0.57%
		4.86%
OTHER FINANCIAL		
united-states gaap	-1.25	1.13%
inventory allowance	-1.02	0.92%
gain recognized	-0.75	0.67%
payable long-term	-0.74	0.67%
adjustment relating	-0.65	0.58%
before benefit	-0.63	0.57%
content asset	-0.61	0.55%
foreign-currency hedging	-0.59	0.53%
negative impact	-0.58	0.53%
stock-based-compensation litigation	-0.56	0.50%
interest received	-0.55	0.50%
continued result	-0.53	0.48%
recognized upon	-0.52	0.47%
included elsewhere	-0.52	0.47%
period during	-0.51	0.46%
cash proceed	-0.49	0.44%
content expense	-0.47	0.42%
hedging	-0.47	0.42%
financing cash	-0.46	0.42%
insurance reserve	-0.45	0.40%
increased expense	-0.45	0.40%

CATEGORY	COEFF ICIENT	RELATIVE IMPORTANCE AMONG THE TOP 200
OTHER FINANCIAL (CONTINUED)		
marketable-security held	-0.44	0.40%
future cash	-0.44	0.40%
accrued payroll	-0.44	0.39%
net liability	-0.43	0.39%
continuing-operation before-tax	-0.43	0.39%
revenue increased	-0.43	0.39%
long-term-debt obligation	-0.43	0.38%
capitalized interest	-0.43	0.38%
subordinated debenture	-0.42	0.38%
common-stock under	-0.42	0.38%
lender	-0.42	0.38%
payment award	0.42	0.38%
beneficial impact	0.42	0.38%
service business	0.42	0.38%
repaid	0.43	0.38%
limited liability	0.43	0.38%
benefit arising	0.43	0.39%
liability using	0.43	0.39%
interest reserve	0.43	0.39%
lower interest-rate	0.43	0.39%
lived intangible-asset	0.44	0.40%
bank credit-facility	0.45	0.40%
before interest-expense	0.45	0.41%
volume contract	0.45	0.41%
beneficial	0.45	0.41%
debt during	0.45	0.41%
related item	0.45	0.41%
fair-value method	0.46	0.42%
benefit cost	0.47	0.42%
content contractual	0.47	0.43%
recent acquisition	0.47	0.43%
continued provision	0.50	0.45%
inflow related	0.51	0.46%
revenue decreased	0.53	0.47%
booked	0.54	0.48%
transfer fund	0.58	0.52%
utility expense	0.58	0.53%
prepaid expense	0.67	0.60%
accrued compensation	0.67	0.60%
expense compared	0.69	0.62%
impairment-charge recorded	0.78	0.70%
continuing-operation related	0.80	0.72%
liquidation	1.13	1.02%
realizable value	1.73	1.56%
		32.52%
UNCATEGORIZED		
portrayal	-0.93	0.84%
revision	-0.80	0.72%
variation	-0.58	0.53%
remaining increase	-0.57	0.51%
ending	-0.55	0.50%
content net	-0.55	0.50%
flat	-0.52	0.47%
determinable	-0.49	0.44%
overall increase	-0.49	0.44%
over time	-0.47	0.42%
item such	-0.46	0.41%
finalization	-0.45	0.41%
various non	-0.44	0.40%
efficiency	-0.44	0.40%
detailed	-0.44	0.40%
shop	-0.44	0.40%
neta	-0.43	0.38%
complementary	-0.42	0.38%
caused	0.43	0.39%
handling	0.43	0.39%
farm	0.44	0.40%
eliminated	0.45	0.41%
usual	0.46	0.41%
repeated	0.46	0.41%
unit award	0.46	0.41%
larger proportion	0.46	0.41%
consistent	0.47	0.42%
increasing	0.47	0.43%
satisfaction	0.48	0.43%
lower due	0.50	0.45%
priority	0.51	0.46%
may need	0.54	0.48%
content significant	0.63	0.56%
		15.01%

Appendix A.6 – Top 10 Most Negative and Most Positive Topics
MD&A – Supervised Latent Dirichlet Allocation

MOST NEGATIVE TOPICS

Topic	Coeff icient	Audit / Tax Authority	Credits	Cross-jurisdictional	Deferred Tax Accounts - General	Deferred Tax Accounts - NOLs/VA	General Tax	Interest & Penalties	Legislation	Permanently-reinvested Earnings	Reserves	Tax Rate	Tax Returns	Performance	Other Financial
tax, income-tax, filed, india, appeal, year, department, based, subject, indian	-6.62	0.12	0.00	0.12	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.06
tax, liability, additional, tax liability, jurisdiction, record, additional tax, tax jurisdiction, determine, extent	-6.50	0.00	0.00	0.09	0.00	0.00	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
benefit, tax benefit, benefit related, net tax, tax, benefit associated, benefit resulting, related, tax-expense benefit, recognized tax	-6.02	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.08
income-tax-expense, compared, recorded income-tax-expense, income-tax-expense increased, income-tax-expense income-tax-expense, income-tax income-tax-expense, income-tax-expense related, table, income-tax-expense decreased, content income-tax-expense	-4.18	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
during, period, recorded, recognized, resulting, addition, current, recorded during, off-set, content	-4.04	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24
limitation, statute, uncertain-tax-position, expiration, liability, due, reduction, applicable, significant, tax-authority	-3.86	0.02	0.00	0.00	0.17	0.00	0.11	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.00
jurisdiction, united-states, international, lower, foreign, tax jurisdiction, foreign jurisdiction, subject, taxed, country	-3.20	0.06	0.00	0.58	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fee, service, expense, management, including, based, salary, public, provide, commission	-3.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.69
difference, between, difference between, permanent, book, basis, tax basis, permanent difference, expected, based	-3.10	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.22
year, ended, year ended, change, net, operation, result, table, income-tax, comparison	-2.95	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.55

MOST POSITIVE TOPICS

Topic	Coeff icient	Audit / Tax Authority	Credits	Cross-jurisdictional	Deferred Tax Accounts - General	Deferred Tax Accounts - NOLs/VA	General Tax	Interest & Penalties	Legislation	Permanently-reinvested Earnings	Reserves	Tax Rate	Tax Returns	Performance	Other Financial
tax, income-tax, result, liability, certain, additional, based, required, addition, more	165.03	0.00	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00
charge, adjusted, pension, restructuring, postretirement, tax charge, gain, include, plan, divestiture	6.23	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.70
audit, tax, jurisdiction, various, tax audit, settlement, multiple, favorable, various tax, included	4.94	0.48	0.00	0.19	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
effective-tax-rate, impact, impacted, tax impact, reduced, negatively, percentage-point, negatively impacted, positively, positively impacted	4.36	0.00	0.00	0.00	0.00	0.02	0.04	0.00	0.00	0.00	0.23	0.64	0.00	0.00	0.05
higher, lower, due, off-set, increased, compared, decreased, lower effective-tax-rate, driven, higher effective-tax-rate	2.98	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00
matter, tax, resolution, tax matter, related, favorable, well, legal, various, favorable resolution	2.70	0.15	0.00	0.00	0.00	0.00	0.66	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
tax, income-tax, cost, increased, include, provision, liability, payroll, resulting, related	2.64	0.00	0.00	0.00	0.00	0.00	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19
net, interest, total, income, average, revenue, interest income, capital, asset, non	2.63	0.00	0.00	0.00	0.16	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.18	0.08
tax, global, earnings, lower, effective-tax-rate, operation, including, taxed, tax-rate, non	2.43	0.00	0.00	0.15	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.14	0.00	0.21	0.00
change, mix, effective-tax-rate, tax, geographic, factor, period, recurring, such, non recurring	2.28	0.00	0.00	0.09	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.07

Appendix B

Variable Definitions

Variable	Definition
$ACQUISITIONS_{i,t}$	Acquisition expenses calculated as ACQ in year *t* divided by AT in year *t-1*.
$AFTER\ TAX\ CF_{i,t}$	After-tax cash flows calculated as (OIBDP - XINT - TXT - DVC) in *t* divided by AT in *t-1*.
$ASSETS_{i,t}$	Total assets (AT).
$BIG\ N_{i,t}$	Big N auditor equal to an indicator variable equal to 1 if AU is equal to 1, 4, 5, 6, 7, or 8, and equal to 0 otherwise.
$BTD_{i,t}$	Book tax difference calculated as (PI - [(TXFED+TXFO)/0.35])/AT.
$CAPEX_{i,t}$	Capital expenditures calculated as CAPX in year *t* divided by AT in year *t-1*.
$CASH\ ETR_{i,t}$	Five-year cash effective tax rate calculated as the sum of TXPD from *t-4* to *t* divided by the sum of pretax income adjusted for special items (PI - SPI) from *t-4* to *t*. The measure is winsorized at 0 and 1.
$CASH\ RATIO_{i,t}$	The ratio of cash and short-term equivalents to total assets calculated as CHE/AT.
$CF\ VOL_{i,t}$	Cash flow volatility calculated as the standard deviation of OIBDP from *t-4* to *t* divided by average noncash assets (AT - CHE) from *t-4* to *t*.
$DIVIDEND\ PAYOUT_{i,t}$	An indicator variable equal to 1 if the firm pays a common dividend (DVC) in year *t* and equal to 0 otherwise.
$EQ\ EARN_{i,t}$	Equity in earnings equal to an indicator variable equal to 1 if ESUB is greater than 0 and equal to 0 otherwise.
$ETR_{i,t-1}$	Lagged effective tax rate calculated as year *t-1* TXT divided by PI in year *t-1*.
$FIN\ CONSTRAINED_{i,t}$	An indicator variable equal to 1 if the firm's Whited and Wu (2006) financial constraints index is above the sample median, and equal to 0 otherwise.
$FOR\ INC_{i,t}$	Foreign income calculated as PIFO divided by AT in year *t-1*.
$HIGH\ SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$	An indicator variable equal to 1 if $SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$ is above the sample median, and equal to 0 otherwise.
$HIGH\ UTB_{i,t}$	An indicator variable equal to 1 if $UTB_{i,t}$ is above the sample median, and equal to 0 otherwise.
$LEV_{i,t}$	Leverage calculated as (DLTT+DLC)/AT.
$LITIG_{i,t}$	Litigation equal to an indicator variable equal to 1 if SETP or SETA are negative, and equal to 0 otherwise.
$LOSS_{i,t}$	An indicator variable equal to 1 if the firm has negative pretax income (PI) in year *t* and equal to 0 otherwise.
$MEZZ\ FIN_{i,t}$	Mezzanine financing calculated as DCPSTK/AT.
$MTB_{i,t}$	Market to book ratio calculated as (CSHO × PRCC_F + [AT - CEQ])/AT
$NET\ WORKING\ CAPITAL_{i,t}$	Net working capital calculated as (WCAP-CHE) in year *t* divided by AT in year *t-1*.
$NOL_{i,t}$	Net operating loss carryforwards equal to an indicator variable equal to 1 if TLCF is greater than 0, and equal to 0 otherwise.
$RD_{i,t}$	R&D calculated as XRD dividec by year *t-1* AT.
$REPATRIATION\ TAX\ COST_{i,t}$	Five-year reptriation tax costs calculated as the sum of (PIFO × 35% - TXFO) from *t-4* to *t* divided by AT in year *t-1*. The measure is winsorized at zero when it is negative.
$ROA_{i,t}$	Return on assets calculated as PI/AT.
$SETTLE\ AMT_{i,t+1}$	Future settlement amount calculated as 100 multiplied by TXTUBSETTLE in year *t+1* divided by AT in year *t*.

SETTLE AMT PRED (QUAL)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *SETTLE AMT PRED sLDA (TF)$_{i,t}$*, *SETTLE AMT PRED RF (TF)$_{i,t}$*, *SETTLE AMT PRED SVM (TF)$_{i,t}$*, *SETTLE AMT PRED sLDA (MD&A)$_{i,t}$*, *SETTLE AMT PRED RF (MD&A)$_{i,t}$*, and *SETTLE AMT PRED SVM (MD&A)$_{i,t}$*. The estimation of the *SETTLE AMT PRED* variables mirrors the estimation of the *SETTLE PRED* variables, replacing the prediction variable with *SETTLE AMT$_{i,t+1}$*.
SETTLE AMT PRED (QUAL + QUANT)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *SETTLE AMT PRED sLDA (TF)$_{i,t}$*, *SETTLE AMT PRED RF (TF)$_{i,t}$*, *SETTLE AMT PRED SVM (TF)$_{i,t}$*, *SETTLE AMT PRED sLDA (MD&A)$_{i,t}$*, *SETTLE AMT PRED RF (MD&A)$_{i,t}$*, *SETTLE AMT PRED SVM (MD&A)$_{i,t}$*, and *SETTLE AMT PRED (QUANT)$_{i,t}$*. The estimation of the *SETTLE AMT PRED* variables mirrors the estimation of the *SETTLE PRED* variables, replacing the prediction variable with *SETTLE AMT$_{i,t+1}$*.
SETTLE AMT PRED (QUANT)$_{i,t}$	The predicted value of a yearly OLS regression with *SETTLE AMT$_{i,t}$* as the dependent variable and the following independent variables: *BTD$_{i,t}$*, *LEV$_{i,t}$*, *ln(ASSETS)$_{i,t}$*, *ROA$_{i,t}$*, *FOR INC$_{i,t}$*, *RD$_{i,t}$*, *ETR$_{i,t-1}$*, *EQ EARN$_{i,t}$*, *MEZZ FIN$_{i,t}$*, *BIG N$_{i,t}$*, *LITIG$_{i,t}$*, and *NOL$_{i,t}$*. The yearly OLS regressions are estimated using the previous four years of data and predictions are obtained using the coefficients applied to current-year values.
SETTLE$_{i,t+1}$	An indicator variable equal to 1 if settlements for firm *i* in year *t+1* (TXTUBSETTLE) are greater than 0, and equal to 0 otherwise.
SETTLE PRED (QUAL)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *SETTLE PRED sLDA (TF)$_{i,t}$*, *SETTLE PRED RF (TF)$_{i,t}$*, and *SETTLE PRED SVM (TF)$_{i,t}$*, *SETTLE PRED sLDA (MD&A)$_{i,t}$*, *SETTLE PRED RF (MD&A)$_{i,t}$*, and *SETTLE PRED SVM (MD&A)$_{i,t}$*.
SETTLE PRED (QUAL + QUANT)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *SETTLE PRED sLDA (TF)$_{i,t}$*, *SETTLE PRED RF (TF)$_{i,t}$*, *SETTLE PRED SVM (TF)$_{i,t}$*, *SETTLE PRED sLDA (MD&A)$_{i,t}$*, *SETTLE PRED RF (MD&A)$_{i,t}$*, *SETTLE PRED SVM (MD&A)$_{i,t}$*, and *SETTLE PRED (QUANT)$_{i,t}$*.
SETTLE PRED (MD&A)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *SETTLE PRED sLDA (MD&A)$_{i,t}$*, *SETTLE PRED RF (MD&A)$_{i,t}$*, and *SETTLE PRED SVM (MD&A)$_{i,t}$*.
SETTLE PRED (QUANT)$_{i,t}$	The predicted value of a yearly OLS regression with *SETTLE$_{i,t}$* as the dependent variable and the following independent variables: *BTD$_{i,t}$*, *LEV$_{i,t}$*, *ln(ASSETS)$_{i,t}$*, *ROA$_{i,t}$*, *FOR INC$_{i,t}$*, *RD$_{i,t}$*, *ETR$_{i,t-1}$*, *EQ EARN$_{i,t}$*, *MEZZ FIN$_{i,t}$*, *BIG N$_{i,t}$*, *LITIG$_{i,t}$*, and *NOL$_{i,t}$*. The yearly OLS regressions are estimated using the previous four years of data and predictions are obtained using the coefficients applied to current-year values.
SETTLE PRED (TF)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *SETTLE PRED sLDA (TF)$_{i,t}$*, *SETTLE PRED RF (TF)$_{i,t}$*, and *SETTLE PRED SVM (TF)$_{i,t}$*.
SETTLE PRED RF (MD&A)$_{i,t}$	The out-of-sample Random Forest prediction of *SETTLE$_{i,t+1}$* using the counts of one- and two-word phrases of the management's discussion and analysis (MD&A) for firm *i* in fiscal year *t*. The training data includes all MD&As from year *t-4* to year *t-1*.
SETTLE PRED RF (TF)$_{i,t}$	The out-of-sample Random Forest prediction of *SETTLE$_{i,t+1}$* using the counts of one- and two-word phrases of the tax footnote for firm *i* in fiscal year *t*. The training data includes all tax footnotes from year *t-4* to year *t-1*.
SETTLE PRED sLDA (MD&A)$_{i,t}$	The out-of-sample supervised Latent Dirichlet Allocation prediction of *SETTLE$_{i,t+1}$* using the counts of one- and two-word phrases of the management's discussion and analysis (MD&A) for firm *i* in fiscal year *t*. The training data includes all MD&As from year *t-4* to year *t-1*.
SETTLE PRED sLDA (TF)$_{i,t}$	The out-of-sample supervised Latent Dirichlet Allocation prediction of *SETTLE$_{i,t+1}$* using the counts of one- and two-word phrases of the tax footnote for firm *i* in fiscal year *t*. The training data includes all tax footnotes from year *t-4* to year *t-1*.

SETTLE PRED SVM (MD&A)$_{i,t}$	The out-of-sample Support Vector Machines prediction of *SETTLE*$_{i,t+1}$ using the counts of one- and two-word phrases of the management's discussion and analysis (MD&A) for firm *i* in fiscal year *t*. The training data includes all MD&As from year *t-4* to year *t-1*.
SETTLE PRED SVM (TF)$_{i,t}$	The out-of-sample Support Vector Machines prediction of *SETTLE*$_{i,t+1}$ using the counts of one- and two-word phrases of the tax footnote for firm *i* in fiscal year *t*. The training data includes all tax footnotes from year *t-4* to year *t-1*.
UNFAV SETTLE$_{i,t+1}$	An indicator variable equal to 1 if both (TXTUBPOSPINC - TXTUBPOSPDEC) in year t+1 is greater than 0 and *SETTLE*$_{i,t+1}$ is equal to 1, and equal to 0 otherwise..
UNFAV SETTLE PRED (QUAL)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *UNFAV SETTLE PRED sLDA (TF)*$_{i,t}$, *UNFAV SETTLE PRED RF (TF)*$_{i,t}$, *UNFAV SETTLE PRED SVM (TF)*$_{i,t}$, *UNFAV SETTLE PRED sLDA (MD&A)*$_{i,t}$, *UNFAV SETTLE PRED RF (MD&A)*$_{i,t}$, and *UNFAV SETTLE PRED SVM (MD&A)*$_{i,t}$. The estimation of the *UNFAV SETTLE PRED* variables mirrors the estimation of the *SETTLE PRED* variables, replacing the prediction variable with *UNFAV SETTLE*$_{i,t+1}$.
UNFAV SETTLE PRED (QUAL + QUANT)$_{i,t}$	The factor obtained from estimating a factor analysis by year on *UNFAV SETTLE PRED sLDA (TF)*$_{i,t}$, *UNFAV SETTLE PRED RF (TF)*$_{i,t}$, *UNFAV SETTLE PRED SVM (TF)*$_{i,t}$, *UNFAV SETTLE PRED sLDA (MD&A)*$_{i,t}$, *UNFAV SETTLE PRED RF (MD&A)*$_{i,t}$, *UNFAV SETTLE PRED SVM (MD&A)*$_{i,t}$, and *UNFAV SETTLE PRED (QUANT)*$_{i,t}$. The estimation of the *UNFAV SETTLE PRED* variables mirrors the estimation of the *SETTLE PRED* variables, replacing the prediction variable with *UNFAV SETTLE*$_{i,t+1}$.
UNFAV SETTLE PRED (QUANT)$_{i,t}$	The predicted value of a yearly OLS regression with *UNFAV SETTLE*$_{i,t}$ as the dependent variable and the following independent variables: *BTD*$_{i,t}$, *LEV*$_{i,t}$, *ln(ASSETS)*$_{i,t}$, *ROA*$_{i,t}$, *FOR INC*$_{i,t}$, *RD*$_{i,t}$, *ETR*$_{i,t-1}$, *EQ EARN*$_{i,t}$, *MEZZ FIN*$_{i,t}$, *BIG N*$_{i,t}$, *LITIG*$_{i,t}$, and *NOL*$_{i,t}$. The yearly OLS regressions are estimated using the previous four years of data and predictions are obtained using the coefficients applied to current-year values.
UTB$_{i,t}$	Unrecognized tax benefits calculated as TXTUBEND in year *t* divided by AT in year *t-1*. Following Hanlon, et al. (2017), we eliminate observations in the top 1% of this variable.

TABLE 1

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 2008 to 2016 and includes 17,117 observations.

Variable	Mean	Std. Dev.	Q1	Median	Q3
$SETTLE_{i,t+1}$	0.387	0.487	0.000	0.000	1.000
$SETTLE\ PRED\ (TF)_{i,t}$	0.000	1.000	-0.873	-0.287	0.977
$SETTLE\ PRED\ (MD\&A)_{i,t}$	0.000	1.000	-0.860	-0.275	0.949
$SETTLE\ PRED\ (QUAL)_{i,t}$	0.000	1.000	-0.855	-0.335	0.969
$SETTLE\ PRED\ (UTB)_{i,t}$	0.395	0.031	0.379	0.400	0.420
$SETTLE\ PRED\ (QUAL + UTB)_{i,t}$	0.000	1.000	-0.853	-0.332	0.969
$SETTLE\ PRED\ (QUANT)_{i,t}$	0.406	0.214	0.238	0.397	0.563
$SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$	0.000	1.000	-0.843	-0.277	0.917
$UTB_{i,t}$	0.015	0.024	0.003	0.007	0.017
$BTD_{i,t}$	-0.031	0.162	-0.030	0.009	0.035
$LEV_{i,t}$	0.240	0.235	0.031	0.195	0.364
$ASSETS_{i,t}$	8,551	26,092	345	1,267	4,676
$ROA_{i,t}$	0.018	0.189	-0.011	0.049	0.104
$FOR\ INC_{i,t}$	0.018	0.044	0.000	0.002	0.029
$RD_{i,t}$	0.051	0.106	0.000	0.003	0.057
$ETR_{i,t-1}$	0.193	0.537	0.057	0.290	0.369
$EQ\ EARN_{i,t}$	0.188	0.391	0.000	0.000	0.000
$MEZZ\ FIN_{i,t}$	0.022	0.088	0.000	0.000	0.000
$BIG\ N_{i,t}$	0.855	0.352	1.000	1.000	1.000
$LITIG_{i,t}$	0.107	0.309	0.000	0.000	0.000
$NOL_{i,t}$	0.662	0.473	0.000	1.000	1.000

TABLE 2
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. Correlations significant at the 5% level or lower are bolded. All variables are defined in Appendix B. All continuous variables are winsorized at the 1st and 99th percentiles.

		I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.	XIII.	XIV.	XV.	XVI.	XVII.	XVIII.
I.	$SETTLE_{i,t+1}$	**1.00**	**0.52**	**0.48**	**0.54**	**0.44**	**0.55**	**0.15**	**0.06**	**0.40**	**0.18**	**0.15**	**-0.19**	**0.06**	**0.13**	**-0.08**	**0.19**	**0.06**	**-0.05**
II.	$SETTLE\ PRED\ (TF)_{i,t}$	**0.51**	**1.00**	**0.74**	**0.94**	**0.55**	**0.93**	**0.23**	**0.08**	**0.51**	**0.26**	**0.21**	**-0.27**	**0.10**	**0.17**	**-0.11**	**0.24**	**0.08**	**-0.07**
III.	$SETTLE\ PRED\ (MD\&A)_{i,t}$	**0.47**	**0.71**	**1.00**	**0.92**	**0.52**	**0.91**	**0.21**	**0.07**	**0.48**	**0.24**	**0.22**	**-0.24**	**0.11**	**0.16**	**-0.09**	**0.23**	**0.08**	**-0.06**
IV.	$SETTLE\ PRED\ (QUAL)_{i,t}$	**0.52**	**0.93**	**0.91**	**1.00**	**0.58**	**0.99**	**0.23**	**0.08**	**0.53**	**0.27**	**0.23**	**-0.27**	**0.11**	**0.18**	**-0.11**	**0.25**	**0.08**	**-0.07**
V.	$SETTLE\ PRED\ (QUANT)_{i,t}$	**0.44**	**0.57**	**0.53**	**0.59**	**1.00**	**0.69**	**0.36**	**0.15**	**0.92**	**0.42**	**0.32**	**-0.45**	**0.14**	**0.31**	**-0.17**	**0.44**	**0.15**	**-0.10**
VI.	$SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$	**0.54**	**0.92**	**0.89**	**0.98**	**0.72**	**1.00**	**0.27**	**0.10**	**0.64**	**0.31**	**0.26**	**-0.32**	**0.12**	**0.21**	**-0.12**	**0.30**	**0.10**	**-0.08**
VII.	$BTD_{i,t}$	**0.13**	**0.22**	**0.20**	**0.22**	**0.32**	**0.26**	**1.00**	**-0.04**	**0.34**	**0.93**	**0.30**	**-0.45**	**0.07**	**0.12**	**-0.20**	**0.14**	**0.02**	**-0.10**
VIII.	$LEV_{i,t}$	**0.11**	**0.16**	**0.14**	**0.17**	**0.26**	**0.20**	**-0.02**	**1.00**	**0.28**	**-0.08**	**-0.07**	**-0.19**	0.00	**0.12**	**0.20**	**0.11**	**0.02**	**0.05**
IX.	$ln(ASSETS)_{i,t}$	**0.40**	**0.53**	**0.49**	**0.55**	**0.92**	**0.67**	**0.25**	**0.41**	**1.00**	**0.32**	**0.19**	**-0.37**	**0.10**	**0.29**	**-0.10**	**0.43**	**0.10**	**-0.04**
X.	$ROA_{i,t}$	**0.19**	**0.29**	**0.27**	**0.30**	**0.41**	**0.34**	**0.74**	**-0.12**	**0.22**	**1.00**	**0.35**	**-0.43**	**0.11**	**0.10**	**-0.21**	**0.15**	0.01	**-0.16**
XI.	$FOR\ INC_{i,t}$	**0.21**	**0.29**	**0.28**	**0.31**	**0.34**	**0.33**	**0.36**	**-0.03**	**0.25**	**0.40**	**1.00**	**-0.08**	**0.02**	**0.07**	**-0.07**	**0.07**	**0.04**	-0.01
XII.	$RD_{i,t}$	**-0.11**	**-0.19**	**-0.16**	**-0.19**	**-0.39**	**-0.25**	**-0.13**	**-0.31**	**-0.33**	**-0.14**	**0.11**	**1.00**	**-0.13**	**-0.17**	**0.17**	**-0.08**	**-0.03**	**0.13**
XIII.	$ETR_{i,t-1}$	**0.12**	**0.17**	**0.16**	**0.18**	**0.23**	**0.21**	**0.13**	**0.04**	**0.15**	**0.30**	-0.01	**-0.31**	**1.00**	**0.02**	**-0.05**	**0.03**	0.00	**-0.07**
XIV.	$EQ\ EARN_{i,t}$	**0.13**	**0.18**	**0.16**	**0.18**	**0.31**	**0.22**	**0.12**	**0.17**	**0.30**	**0.08**	**0.10**	**-0.16**	**0.03**	**1.00**	**-0.04**	**0.10**	**0.04**	**-0.02**
XV.	$MEZZ\ FIN_{i,t}$	**-0.02**	**-0.03**	**-0.03**	**-0.03**	**-0.02**	**-0.03**	**-0.08**	**0.23**	**0.07**	**-0.14**	**-0.05**	**0.05**	**-0.08**	**0.03**	**1.00**	**-0.03**	0.01	**0.05**
XVI.	$BIG\ N_{i,t}$	**0.19**	**0.25**	**0.24**	**0.27**	**0.45**	**0.33**	**0.09**	**0.14**	**0.42**	**0.13**	**0.11**	**-0.08**	**0.07**	**0.10**	**0.02**	**1.00**	**0.04**	**-0.02**
XVII.	$LITIG_{i,t}$	**0.06**	**0.08**	**0.07**	**0.09**	**0.15**	**0.10**	-0.01	**0.03**	**0.10**	-0.01	**0.05**	**0.03**	-0.01	**0.04**	**0.03**	**0.04**	**1.00**	0.00
XVIII.	$NOL_{i,t}$	**-0.05**	**-0.07**	**-0.06**	**-0.07**	**-0.10**	**-0.08**	**-0.09**	**0.04**	**-0.03**	**-0.18**	**0.02**	**0.14**	**-0.15**	**-0.02**	**0.05**	**-0.02**	**0.00**	**1.00**

TABLE 3

Future Tax Settlement Prediction with the Tax Footnote

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* settles with the tax authority in year *t+1*, and equal to 0 otherwise ($SETTLE_{i,t+1}$). Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]
Intercept	-1.362***	-1.158***	-0.287***	-0.575***
	(-57.453)	(-49.484)	(-16.312)	(-30.835)
SETTLE PRED RF (TF)$_{i,t}$	2.273***			
	(61.720)			
SETTLE PRED SVM (TF)$_{i,t}$		1.551***		
		(46.038)		
SETTLE PRED sLDA (TF)$_{i,t}$			2.008***	
			(51.522)	
SETTLE PRED (TF)$_{i,t}$				1.229***
				(61.788)
#OBS	17,117	17,117	17,117	17,117
Pseudo R^2	0.192	0.099	0.152	0.215
Area Under ROC	0.747	0.683	0.760	0.803

TABLE 4

Future Tax Settlement Prediction with the MD&A

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* settles with the tax authority in year *t+1*, and equal to 0 otherwise ($SETTLE_{i,t+1}$). Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]	[5]	[6]
Intercept	-1.240***	-1.094***	-0.284***	-0.550***	-0.577***	-0.573***
	(-55.104)	(-47.795)	(-16.714)	(-30.483)	(-30.617)	(-30.460)
SETTLE PRED RF (MD&A)$_{i,t}$	2.119***					
	(57.842)					
SETTLE PRED SVM (MD&A)$_{i,t}$		1.440***				
		(43.114)				
SETTLE PRED sLDA (MD&A)$_{i,t}$			1.594***			
			(39.865)			
SETTLE PRED (MD&A)$_{i,t}$				1.104***	0.497***	
				(57.622)	(19.221)	
SETTLE PRED (TF)$_{i,t}$					0.875***	
					(33.093)	
SETTLE PRED (QUAL)$_{i,t}$						1.281***
						(63.327)
#OBS	17,117	17,117	17,117	17,117	17,117	17,117
Pseudo R^2	0.165	0.086	0.081	0.180	0.231	0.230
Area Under ROC	0.727	0.670	0.697	0.776	0.812	0.810

TABLE 5

Future Tax Settlement Prediction - UTB vs. Qualitative Disclosure

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* settles with the tax authority in year *t+1*, and equal to 0 otherwise ($SETTLE_{i,t+1}$). Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]
Intercept	-0.410***	-2.045***	-0.573***	-1.929***
	(-22.165)	(-9.912)	(-30.460)	(-7.627)
$UTB_{i,t}$	-3.390***			
	(-5.017)			
SETTLE PRED (UTB)$_{i,t}$		4.005***		3.418***
		(7.717)		(5.381)
SETTLE PRED (QUAL)$_{i,t}$			1.281***	1.277***
			(63.327)	(63.206)
#OBS	17,117	17,117	17,117	17,117
Pseudo R^2	0.001	0.003	0.230	0.231
Area Under ROC	0.442	0.522	0.810	0.812

TABLE 6

Future Tax Settlement Prediction - Quantitative vs. Qualitative

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* settles with the tax authority in year *t+1*, and equal to 0 otherwise ($SETTLE_{i,t+1}$). Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]	[5]
Intercept	-4.135***	-2.558***	-0.573***	-1.702***	-0.590***
	(-41.833)	(-57.226)	(-30.460)	(-34.130)	(-30.809)
$UTB_{i,t}$	8.836***				
	(9.957)				
$BTD_{i,t}$	-3.016***				
	(-9.516)				
$LEV_{i,t}$	-0.539***				
	(-6.363)				
$ln(ASSETS)_{i,t}$	0.451***				
	(35.873)				
$ROA_{i,t}$	2.818***				
	(10.756)				
$FOR\ INC_{i,t}$	2.902***				
	(6.557)				
$RD_{i,t}$	-3.745***				
	(-11.838)				
$ETR_{i,t-1}$	0.049				
	(1.440)				
$EQ\ EARN_{i,t}$	0.024				
	(0.545)				
$MEZZ\ FIN_{i,t}$	-0.907***				
	(-3.231)				
$BIG\ N_{i,t}$	0.353***				
	(5.510)				
$LITIG_{i,t}$	0.174***				
	(3.155)				
$NOL_{i,t}$	-0.041				
	(-1.079)				
$SETTLE\ PRED\ (QUANT)_{i,t}$		4.919***		2.691***	
		(52.837)		(25.093)	
$SETTLE\ PRED\ (QUAL)_{i,t}$			1.281***	1.023***	
			(63.327)	(45.799)	
$SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$					1.367***
					(64.392)
#OBS	17,117	17,117	17,117	17,117	17,117
Pseudo R^2	0.155	0.155	0.230	0.258	0.249
Area Under ROC	0.757	0.760	0.810	0.827	0.823

TABLE 7

Current Cash Holdings and Future Tax Settlement Prediction

This table includes 13,977 firm-year observations from 2008 to 2016 with sufficient data necessary to estimate the cash ratio analysis. Panel A provides descriptive statistics, Panel B provides Pearson (Spearman) correlations above (below) the diagonal with values bolded if significant at the 5% level or lower, and Panel C provides the multivariate regression analysis where the dependent variable is equal to the ratio of cash and short-term investments to total assets for firm *i* in year *t* ($CASH\ RATIO_{i,t}$). Standard errors are clustered by firm. All continuous variables are winsorized at the 1% and 99% levels. All variables are defined in Appendix B. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Descriptive Statistics

Variable	Mean	Std. Dev.	Q1	Median	Q3
$CASH\ RATIO_{i,t}$	0.189	0.187	0.048	0.127	0.270
$MODIFIED\ SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$	0.000	1.000	-0.848	-0.274	0.921
$UTB_{i,t}$	0.013	0.017	0.003	0.007	0.016
$REPATRIATION\ TAX\ COST_{i,t}$	0.012	0.031	0.000	0.000	0.009
$CASH\ ETR_{i,t}$	0.230	0.199	0.075	0.225	0.319
$FIN\ CONSTRAINED_{i,t}$	0.523	0.499	0.000	1.000	1.000
$NOL_{i,t}$	0.666	0.472	0.000	1.000	1.000
$LOSS_{i,t}$	0.259	0.438	0.000	0.000	1.000
$NET\ WORKING\ CAPITAL_{i,t}$	0.059	0.171	-0.032	0.055	0.160
$LEV_{i,t}$	0.225	0.215	0.025	0.190	0.338
$CF\ VOL_{i,t}$	0.089	0.205	0.021	0.041	0.084
$MTB_{i,t}$	1.907	1.234	1.166	1.528	2.169
$ASSETS_{i,t}$	5,536	16,063	345	1,137	3,819
$DIVIDEND\ PAYOUT_{i,t}$	0.420	0.494	0.000	0.000	1.000
$CAPEX_{i,t}$	0.046	0.049	0.016	0.030	0.057
$ACQUISITIONS_{i,t}$	0.040	0.103	0.000	0.000	0.025
$AFTER\ TAX\ CF_{i,t}$	0.070	0.125	0.041	0.080	0.122
$RD_{i,t}$	0.045	0.084	0.000	0.006	0.059

TABLE 7

Panel B: Correlations

		I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.	XIII.	XIV.	XV.	XVI.	XVII.	XVIII.
I.	*CASH RATIO$_{i,t}$*	**1.00**	**-0.28**	**0.29**	**0.15**	**-0.13**	**0.30**	**0.03**	**0.14**	**-0.27**	**-0.39**	**0.48**	**0.37**	**-0.35**	**-0.23**	**-0.14**	**-0.12**	**-0.22**	**0.54**
II.	*MODIFIED SETTLE PRED (QUAL + QUANT)$_{i,t}$*	**-0.22**	**1.00**	**-0.04**	**0.14**	**0.18**	**-0.50**	**-0.07**	**-0.30**	0.01	**0.11**	**-0.24**	**-0.05**	**0.65**	**0.38**	**-0.02**	-0.01	**0.23**	**-0.29**
III.	*UTB$_{i,t}$*	**0.31**	**0.06**	**1.00**	**0.24**	**-0.12**	**0.11**	**0.06**	**0.09**	**-0.13**	**-0.10**	**0.17**	**0.14**	**-0.09**	**-0.12**	**-0.09**	**0.03**	**-0.08**	**0.38**
IV.	*REPATRIATION TAX COST$_{i,t}$*	**0.10**	**0.27**	**0.19**	**1.00**	**-0.11**	**-0.08**	**0.02**	**-0.14**	0.01	**-0.07**	-0.01	**0.13**	**0.15**	**0.06**	**-0.02**	0.00	**0.16**	**0.11**
V.	*CASH ETR$_{i,t}$*	**-0.12**	**0.30**	**-0.13**	**-0.03**	**1.00**	**-0.09**	**-0.14**	**-0.13**	**0.14**	**-0.03**	**-0.12**	**-0.09**	**0.07**	**0.14**	**-0.03**	**-0.02**	**0.14**	**-0.23**
VI.	*FIN CONSTRAINED$_{i,t}$*	**0.28**	**-0.51**	**0.05**	**-0.17**	**-0.18**	**1.00**	**0.04**	**0.27**	**0.02**	**-0.19**	**0.20**	**0.07**	**-0.63**	**-0.54**	**-0.04**	-0.01	**-0.15**	**0.29**
VII.	*NOL$_{i,t}$*	**0.03**	**-0.07**	**0.06**	**0.06**	**-0.19**	**0.04**	**1.00**	**0.11**	**-0.08**	**0.07**	**0.04**	-0.01	0.02	**-0.11**	**-0.04**	**0.04**	**-0.10**	**0.11**
VIII.	*LOSS$_{i,t}$*	**0.08**	**-0.32**	**0.06**	**-0.21**	**-0.27**	**0.27**	**0.11**	**1.00**	**-0.13**	**0.06**	**0.22**	**-0.10**	**-0.29**	**-0.30**	**-0.07**	**-0.03**	**-0.48**	**0.24**
IX.	*NET WORKING CAPITAL$_{i,t}$*	**-0.18**	0.00	**-0.10**	**0.08**	**0.18**	**0.03**	**-0.06**	**-0.12**	**1.00**	**-0.15**	**-0.13**	**-0.22**	**-0.11**	**0.05**	**-0.06**	-0.01	**0.15**	**-0.18**
X.	*LEV$_{i,t}$*	**-0.53**	**0.20**	**-0.13**	**-0.03**	**-0.03**	**-0.26**	**0.07**	**0.02**	**-0.14**	**1.00**	**-0.14**	**-0.09**	**0.32**	**0.03**	**0.05**	**0.11**	**-0.02**	**-0.24**
XI.	*CF VOL$_{i,t}$*	**0.49**	**-0.40**	**0.18**	**-0.10**	**-0.15**	**0.38**	0.01	**0.30**	0.01	**-0.38**	**1.00**	**0.18**	**-0.30**	**-0.16**	**-0.06**	**-0.06**	**-0.40**	**0.43**
XII.	*MTB$_{i,t}$*	**0.31**	**0.04**	**0.18**	**0.15**	**-0.02**	**-0.02**	0.00	**-0.23**	**-0.19**	**-0.12**	**0.02**	**1.00**	**-0.08**	0.00	**0.07**	**-0.02**	**0.03**	**0.34**
XIII.	*ASSETS$_{i,t}$*	**-0.33**	**0.67**	0.00	**0.26**	**0.13**	**-0.67**	**0.03**	**-0.28**	**-0.15**	**0.43**	**-0.49**	**0.02**	**1.00**	**0.36**	**0.07**	**0.06**	**0.25**	**-0.30**
XIV.	*DIVIDEND PAYOUT$_{i,t}$*	**-0.20**	**0.39**	**-0.07**	**0.15**	**0.24**	**-0.54**	**-0.11**	**-0.30**	**0.04**	**0.08**	**-0.27**	**0.09**	**0.35**	**1.00**	0.01	**-0.03**	**0.08**	**-0.26**
XV.	*CAPEX$_{i,t}$*	**-0.16**	**0.07**	**-0.10**	-0.01	**0.02**	**-0.12**	**-0.04**	**-0.13**	0.01	**0.07**	**-0.04**	**0.09**	**0.11**	**0.07**	**1.00**	0.00	**0.19**	**-0.07**
XVI.	*ACQUISITIONS$_{i,t}$*	**-0.14**	**0.13**	0.01	**0.12**	**0.04**	**-0.12**	**0.06**	**-0.15**	0.00	**0.12**	**-0.23**	**0.05**	**0.22**	**0.06**	**-0.04**	**1.00**	**0.12**	0.01
XVII.	*AFTER TAX CF$_{i,t}$*	**-0.04**	**0.23**	0.02	**0.19**	**0.19**	**-0.12**	**-0.10**	**-0.53**	**0.09**	**-0.02**	**-0.15**	**0.33**	**0.18**	**0.05**	**0.33**	**0.18**	**1.00**	**-0.35**
XVIII.	*RD$_{i,t}$*	**0.49**	**-0.22**	**0.37**	**0.17**	**-0.25**	**0.29**	**0.12**	**0.15**	-0.01	**-0.31**	**0.33**	**0.27**	**-0.26**	**-0.22**	**-0.17**	**0.01**	**-0.07**	**1.00**

TABLE 7

Panel C: Multivariate Regression Analysis

	[1]	[2]	[3]	[4]	[5]
Intercept	0.206***	0.187***	0.183***	0.200***	0.194***
	(8.512)	(7.349)	(7.363)	(8.637)	(8.280)
$UTB_{i,t}$	0.376**		0.407***		
	(2.566)		(2.789)		
MODIFIED SETTLE PRED (QUAL + QUANT)$_{i,t}$		-0.012***	-0.012***		
		(-5.179)	(-5.439)		
HIGH UTB$_{i,t}$				0.025***	0.038***
				(6.446)	(6.304)
HIGH MODIFIED SETTLE PRED (QUAL + QUANT)$_{i,t}$				-0.015***	-0.004
				(-4.283)	(-0.990)
HIGH UTB$_{i,t}$ × *HIGH MODIFIED SETTLE PRED (QUAL + QUANT)*$_{i,t}$					-0.025***
					(-3.898)
REPATRIATION TAX COST$_{i,t}$	0.517***	0.573***	0.532***	0.521***	0.531***
	(5.976)	(6.672)	(6.151)	(6.100)	(6.174)
CASH ETR$_{i,t}$	-0.002	0.002	0.003	0.001	0.002
	(-0.214)	(0.224)	(0.308)	(0.131)	(0.215)
FIN CONSTRAINED$_{i,t}$	0.012***	0.011***	0.011***	0.011**	0.011***
	(2.777)	(2.656)	(2.593)	(2.575)	(2.656)
$NOL_{i,t}$	-0.006	-0.007	-0.007*	-0.007	-0.007*
	(-1.575)	(-1.615)	(-1.683)	(-1.640)	(-1.693)
$LOSS_{i,t}$	-0.006	-0.007	-0.007	-0.007	-0.007
	(-1.263)	(-1.392)	(-1.514)	(-1.548)	(-1.586)
NET WORKING CAPITAL$_{i,t}$	-0.245***	-0.249***	-0.246***	-0.244***	-0.245***
	(-14.660)	(-14.935)	(-14.795)	(-14.728)	(-14.796)
$LEV_{i,t}$	-0.202***	-0.205***	-0.204***	-0.201***	-0.201***
	(-16.564)	(-16.738)	(-16.873)	(-16.613)	(-16.651)
CF VOL$_{i,t}$	0.227***	0.228***	0.226***	0.227***	0.227***
	(13.054)	(12.955)	(13.115)	(13.273)	(13.394)
$MTB_{i,t}$	0.023***	0.024***	0.024***	0.023***	0.023***
	(10.834)	(11.022)	(11.015)	(10.611)	(10.607)

$ln(ASSETS)_{i,t}$	-0.011***	-0.007***	-0.007***	-0.010***	-0.009***
	(-5.895)	(-3.350)	(-3.334)	(-5.129)	(-4.920)
$DIVIDEND\ PAYOUT_{i,t}$	-0.023***	-0.021***	-0.020***	-0.021***	-0.021***
	(-5.164)	(-4.627)	(-4.577)	(-4.772)	(-4.780)
$CAPEX_{i,t}$	-0.352***	-0.374***	-0.365***	-0.355***	-0.357***
	(-8.805)	(-9.331)	(-9.154)	(-8.837)	(-8.862)
$ACQUISITIONS_{i,t}$	-0.144***	-0.146***	-0.148***	-0.147***	-0.147***
	(-13.319)	(-13.369)	(-13.526)	(-13.559)	(-13.564)
$AFTER\ TAX\ CF_{i,t}$	0.022	0.025	0.022	0.022	0.022
	(0.943)	(1.054)	(0.952)	(0.951)	(0.945)
$RD_{i,t}$	0.366***	0.379***	0.354***	0.358***	0.346***
	(8.128)	(8.763)	(7.901)	(8.224)	(7.902)
#OBS	13,977	13,977	13,977	13,977	13,977
Adjusted R^2	0.577	0.578	0.579	0.580	0.581

TABLE 8
Future Tax Settlement Prediction - Including Current Settlements

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* settles with the tax authority in year *t+1*, and equal to 0 otherwise ($SETTLE_{i,t+1}$). Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]	[5]
Intercept	-1.359***	-0.736***	-2.757***	-2.031***	-0.709***
	(-56.721)	(-24.649)	(-57.121)	(-34.274)	(-24.364)
$SETTLE_{i,t}$	2.170***	0.436***	1.713***	0.684***	0.322***
	(60.036)	(7.089)	(44.449)	(10.778)	(5.477)
$SETTLE\ PRED\ (QUAL)_{i,t}$		1.096***		0.717***	
		(33.472)		(20.032)	
$SETTLE\ PRED\ (QUANT)_{i,t}$			3.719***	2.865***	
			(37.049)	(26.243)	
$SETTLE\ PRED\ (QUAL + QUANT)_{i,t}$					1.231***
					(37.920)
#OBS	17,117	17,117	17,117	17,117	17,117
Pseudo R^2	0.179	0.232	0.245	0.263	0.250
Area Under ROC	0.741	0.812	0.820	0.830	0.824

TABLE 9

High Future Tax Settlement Amount Prediction

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* has above median dollar amount of settlements divided by total assets in year *t+1*, and equal to 0 otherwise (HIGH $SETTLE_{i,t+1}$). The median value is based on firms with non-zero tax settlements in year *t+1*. Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]
Intercept	-2.777***	-1.604***	-2.141***	-1.617***
	(-52.579)	(-72.707)	(-37.253)	(-72.627)
HIGH SETTLE PRED (QUANT)$_{i,t}$	6.242***		2.585***	
	(29.367)		(10.369)	
HIGH SETTLE PRED (QUAL)$_{i,t}$		0.794***	0.694***	
		(42.752)	(33.289)	
HIGH SETTLE PRED (QUAL + QUANT)$_{i,t}$				0.826***
				(43.372)
#OBS	17,117	17,114	17,114	17,114
Pseudo R^2	0.053	0.115	0.121	0.121
Area Under ROC	0.669	0.746	0.749	0.750

TABLE 10
Future Unfavorable Tax Settlement Prediction

This table includes all firm-year observations from 2008 to 2016 with sufficient data to calculate the dependent and independent variables. The dependent variable is an indicator variable equal to 1 if firm *i* settles unfavorably with the tax authority in year *t+1*, and equal to 0 otherwise (*UNFAV SETTLE*$_{i,t+1}$). Standard errors are clustered by firm. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. T-statistics are reported in parentheses. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]
Intercept	-1.530***	-2.972***	-2.673***	-1.564***
	(-72.898)	(-56.768)	(-49.116)	(-72.744)
UNFAV SETTLE PRED (QUAL)$_{i,t}$	0.660***		0.494***	
	(37.091)		(25.910)	
UNFAV SETTLE PRED (QUANT)$_{i,t}$		3.472***	2.600***	
		(35.219)	(24.433)	
UNFAV SETTLE PRED (QUAL + QUANT)$_{i,t}$				0.753***
				(40.138)
#OBS	17,117	17,117	17,117	17,117
Pseudo R^2	0.085	0.081	0.121	0.103
Area Under ROC	0.722	0.706	0.754	0.747