Review of Accounting Studies (2026) 31:374–412
https://doi.org/10.1007/s11142-025-09914-3

Using narrative disclosures to predict tax outcomes

Olga Bogachek[1] · **Antonio De Vito**[2] · **Paul Demeré**[3] · **Francesco Grossetti**[3,4]

Accepted: 5 August 2025 / Published online: 27 October 2025

Abstract

We examine whether narrative discussion in financial disclosures can help corporate stakeholders better predict tax outcomes. To measure qualitative discussion, we use topic modeling analysis to create measures of the thematic content of 10-K disclosures. We find that qualitative discussion in financial disclosures can substantially improve prediction of tax outcomes in in- and out-of-sample tests. We also find that prediction-relevant discussion is distributed throughout the 10-K, supporting that disclosures should be analyzed holistically rather than examining only limited pieces of larger disclosures. Further, we find that analysts often do not use this information effectively, resulting in predictable and economically meaningful forecast errors. These findings illustrate the wealth of qualitative information in 10-K disclosures for stakeholders concerned about tax outcomes and offer a practical approach to examining qualitative disclosures and using them to predict tax outcomes.

Keywords Tax outcomes · Tax planning · Textual analysis · Topic modeling · Forecasting · Qualitative disclosure · Machine learning

JEL Classification G39 · H25 · H26 · M21 · M41

✉ Paul Demeré
paul.demere@unibocconi.it

Olga Bogachek
olga.bogachek@unibz.it

Antonio De Vito
a.devito@unibo.it

Francesco Grossetti
francesco.grossetti@unibocconi.it

1 Faculty of Economics and Management, Free University of Bolzano, Bolzano, Italy

2 Department of Management, University of Bologna, Bologna, Italy

3 Department of Accounting, Bocconi University, Milan, Italy

4 Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy

1 Introduction

Corporate income taxes are an important financial statement item for firms around the world. They are a critical input into valuation models (Graham et al. 2012; Rajgopal 2022) and play an important role in corporate decision-making (Graham et al. 2012; Jacob 2022) and executive career outcomes and compensation (Armstrong et al. 2012; Brown et al. 2016a; Chyz and Gaertner 2018). Corporate taxes also affect the behavior of corporate stakeholders, such as analysts and lenders (Hasan et al. 2014; Bratten et al. 2017; Platikanova 2017; Isin 2018; He et al. 2020), and are important to governments' funding, budgeting, and policy (Green et al. 2022). Thus, understanding and predicting tax outcomes is important for the decision making of corporate stakeholders.

However, understanding firms' past tax outcomes and current tax circumstances, much less predicting future tax outcomes, can be very challenging (Rajgopal 2022). Tax returns themselves are typically protected from disclosure, leaving financial statements as the primary source of public tax information (Graham et al. 2012; Demeré 2023). Meanwhile, GAAP accounting for income taxes is often noisy, with tax accounts being susceptible to manipulation for earnings management objectives (Dhaliwal et al. 2004; Krull 2004; Frank and Rego 2006; Cazier et al. 2015) and subject to one-size-fits-all rules that reduce the relevance of tax numbers (Robinson et al. 2016). Tax disclosures can also be opaque, particularly when firms engage in substantial tax planning that might draw scrutiny from tax authorities or incur reputational costs (Ayers et al. 2015; Chen et al. 2018; Inger et al. 2018; Balakrishnan et al. 2019; Nguyen 2021; Chychyla et al. 2022; Richter et al. 2024). In these circumstances, nontax disclosures may better illuminate firms' tax outcomes, and considerable research has examined *quantitative* nontax factors that can explain tax outcomes (Wilde and Wilson 2018; Guenther et al. 2023). However, quantitative disclosures are only a small portion of firms' financial statement disclosures. We examine the extent to which the narrative content of firms' *qualitative* textual disclosures can also help corporate stakeholders better predict tax outcomes.

Accounting research examining the narrative content of firms' qualitative disclosures is relatively new (Bochkay et al. 2023) but already shows that they have value relevance and can be used to detect financial misreporting (Campbell et al. 2014; Brown et al. 2020). However, it is unclear how much the narrative content of firms' qualitative disclosures can aid stakeholders in anticipating tax outcomes. On one hand, the length and potential richness of qualitative disclosures may contain value-relevant information that can aid stakeholders in understanding firms' tax functions and predicting future tax outcomes. On the other hand, firms may strategically limit the informativeness of qualitative discussions with tax implications to minimize the information available to adversarial tax authorities or competitors (Graham et al. 2012; Kubick et al. 2015; Donohoe et al. 2022). Research indicates that firms often use vague, boilerplate language in their disclosures, particularly when faced with proprietary costs and judicial and regulatory review (Hope et al. 2016; Cazier et al. 2021). Given the significant presence of these costs in the tax setting, it is unsurprising that research finds that tax-specific language in financial statements is particularly likely to be highly standardized (Bernard et al. 2023) and that tax executives often use vague language in tax footnotes to limit the information available to tax authorities (Richter et al. 2024).

We begin our study by downloading all 10-K filings from the period of 2006 through 2019. We then use a variant of Latent Dirichlet Allocation (LDA) (Blei et al. 2003; Chen et al. 2015) to estimate the narrative topics discussed in each item of each 10-K. LDA is an established method of analyzing text (Dyer et al. 2017; Huang et al. 2018; Brown et al. 2020). Our machine-learning approach identifies general discussion topics and produces a measure of the proportion of each 10-K or 10-K item that is composed of each topic.

We then examine whether our qualitative discussion topic measures can improve the in-sample prediction of a key tax outcome: the cash effective tax rate (ETR) (Dyreng et al. 2008; Brown et al. 2016a). We find that general discussion topics have strong predictive power for cash ETRs for the future three years incremental to numerous previously used quantitative measures of cash ETR determinants. In particular, discussions about (1) governance that could constrain tax planning, (2) risks that can increase the nontax costs of tax planning, (3) complex transactions, and (4) complex organizational structures seem to be particularly informative. Economically, a standard deviation increase in the proportion of a 10-K devoted to a single one of these topics is associated with cash ETR deviations of up to 5.1% of the mean cash ETR, equivalent to approximately $6.6 million in cash taxes paid.

We also run out-of-sample analyses to ensure that stakeholders can use qualitative discussion topic measures to predict tax outcomes in a practical manner. Using a commonly used machine learning method known as extreme gradient boosting (XGBoost) (Chen et al. 2022; Campbell et al. 2023; Geertsema and Lu 2023; Guenther et al. 2023), we find that incorporating qualitative discussion topics significantly improves the performance of cash ETR prediction models by substantially reducing the out-of-sample root mean square error (RMSE). These results are incremental to numerous established quantitative cash ETR determinants and lagged cash ETRs. The reduction in RMSE ranges from 57% to 62%, which is economically meaningful.

We next explore whether the prediction-relevant discussion is concentrated within a particular 10-K item or is spread throughout the 10-K. Research has examined tax keyword searches within narrow portions of the 10-K, such as risk factor disclosures (Item 1A) or tax footnotes (Balakrishnan et al. 2019; Campbell et al. 2019; Luo et al. 2024). We find that narrative discussion that is relevant to tax outcome prediction can be found throughout the 10-K. These findings provide novel evidence that qualitative discussions throughout the 10-K can provide unique and value-relevant information to stakeholders and suggest that research and stakeholders focusing on narrow subsets of 10-K disclosure likely miss considerable information.

While our evidence supports the idea that qualitative discussion topics contain useful information for predicting tax outcomes, it is unclear whether corporate stakeholders are aware of and using this information. As such, we also examine whether analysts' tax forecast accuracy is associated with qualitative discussion topics. We find that analysts react to some of the information in narrative discussions, but also that a number of topics (e.g., executive pay, operating trends, financial reporting of intangibles) are significantly associated with reduced analyst forecast accuracy, indicating that analysts are not appropriately using this information in their forecasts.

Economically, a standard deviation increase in the proportion of a 10-K devoted to one of these individual topics is associated with analyst forecast errors equivalent to 4.1% to 38.4% of the mean forecast error.

As a final analysis, we find that our results are largely consistent when examining alternative tax outcome variables, such as GAAP ETRs and settlements with tax authorities. As such, our research can benefit corporate stakeholders who want to use tax metrics other than the cash ETR in their decision-making.

Our study provides two substantial contributions to tax accounting research. First, by showing that narrative topics in the 10-K can be used to predict tax outcomes, we document an important source of *qualitative* information that can be used to evaluate and predict tax outcomes in addition to quantitative measures examined in prior research (Wilde and Wilson 2018; Guenther et al. 2023). Second, while some research suggests that investors and analysts can use qualitative tax-specific disclosures (Luo et al. 2024; Burd et al. 2023), we show that analysts do not appropriately use the information in broad qualitative disclosures, leading to substantial and predictable tax forecast errors.

We also contribute to research on qualitative disclosures. In particular, our evidence that prediction-relevant discussion is found throughout the 10-K and that many 10-K items contain unique prediction-relevant discussion illustrates the wealth of qualitative information in 10-Ks. These results should also encourage future research on qualitative disclosures to examine disclosures holistically using the approach we describe here, rather than focusing only on single pieces of the 10-K, such as risk factor disclosures, management discussion and analysis, or individual footnotes. Our study also offers an example of how narrative disclosures can be used for prediction that can guide future research that uses qualitative information to predict important firm outcomes beyond taxes.

Finally, our research can practically benefit corporate stakeholders interested in tax outcomes. Notably, we describe a practical method for deriving prediction-relevant information from 10-K disclosures that can be used to improve tax forecasts. This method involves (1) using the processes described here to calculate topic measures for 10-Ks and (2) using our evidence on which topics are most relevant to predicting tax outcomes to evaluate when a specific firm's narratives indicate that cash ETRs or other tax outcomes are likely to be higher or lower than a prediction model based solely on quantitative information would indicate. Our evidence that narrative disclosure can predict analyst forecast errors supports the notion that even sophisticated stakeholders may benefit from applying our methodology.[1] This practical approach

[1] Informal discussions with analysts indicate that they often read narrative disclosures to refine their forecasts, consistent with research showing that analysts use the 10-K and that their forecast accuracy is affected by 10-K readability (Bozanic and Thevenot 2015; Brown et al. 2016b). Our discussions indicate that analysts will often look beyond the tax footnote (e.g., examining discussions about mergers, regulatory exposure, geographic footprints) to inform their tax forecasts. However, reading narrative disclosures entails effort and time, especially when analyzing multiple firms. Further, analysts acknowledged that their interpretations require judgment and can be imprecise and raised concerns that boilerplate language and spin from management make discerning novel information challenging. Our method for summarizing the narrative content of 10-K disclosures offers a structured alternative that reduces effort while enhancing precision. Our evidence that 10-K narrative topics can also substantially improve predictions of tax outcomes indicates that our methodology can help stakeholders see beyond boilerplate and spin.

can also benefit regulators and tax authorities, as we document that narrative discussion in 10-Ks can be helpful in evaluating firms' tax planning and can even provide guidance in identifying firms with unsustainable tax positions.

2 Background and theory

2.1 Income tax outcomes

Corporate income tax outcomes are very important to the decisions of a variety of corporate stakeholders (Kim et al. 2011; Hasan et al. 2014; Brown et al. 2016a; Goh et al. 2016; Bratten et al. 2017; Graham et al. 2017; Chyz and Gaertner 2018; Doellman et al. 2020; He et al. 2020; Green et al. 2022). While there may be several specific outcomes that stakeholders care about, ETRs are considered one of the most important (Graham et al. 2014; Brown et al. 2016a; Demeré et al. 2024). As such, considerable research examines their determinants, primarily with the objective of understanding the causal factors that affect them (Hanlon and Heitzman 2010; Lietz 2013). To provide context to the many documented factors, Wilde and Wilson (2018) develop a conceptual framework that organizes tax outcome determinants into three major categories: agency factors, implementation factors, and outcome factors.

Agency factors are those that affect the alignment of different stakeholders' interests. For example, managers may have different tax planning preferences than shareholders, which can create agency costs, such as managers engaging in insufficient tax planning because they do not want to bear the risk of tax authority scrutiny or managers engaging in excessive tax avoidance to build reserves they can extract rents from (Chen and Chu 2005; Desai and Dharmapala 2006).[2] Research indicates that executive compensation incentives (Robinson et al. 2010; Rego and Wilson 2012; Chi et al. 2017; Chyz and Gaertner 2018; Caglio et al. 2022), individual executive characteristics (Boone et al. 2013; Chyz 2013; Olsen and Stekelberg 2016; Koester et al. 2017), and corporate governance controls (Brown 2011; Brown and Drake 2014; Armstrong et al. 2015; Khan et al. 2017; Gallemore et al. 2019) can all affect tax outcomes by affecting the alignment of managers' tax planning behavior with the interests of shareholders and other stakeholders.

While agency factors address a single important cost, implementation factors encompass a variety of firm attributes and operating environment features that shape firms' economic fundamentals and their impact on tax outcomes. Firm attributes are economic characteristics of firms that can affect tax outcomes by providing opportunities for tax planning, introducing tax planning frictions, and changing firms' cost-benefit trade-offs for particular tax strategies. These firm attributes include such factors as firm size, multinational operations, business strategies, entity structures, and financial constraints (Zimmerman 1983; Rego 2003; Dyreng and Lindsey 2009; Higgins et al. 2015; Edwards et al. 2016; Demeré et al. 2020; Campbell et al. 2021).

[2] The primary focus in prior tax accounting research is on the interest alignment of managers and shareholders, although manager-employee, shareholder-lender, and other dual-interest alignments can also affect tax outcomes.

Firms' operating environments can also significantly affect tax outcomes. Relevant factors include competition (Kubick et al. 2015), macroeconomic conditions (Kim et al. 2022), corruption (Al-Hadi et al. 2022), litigation risk (Arena et al. 2021), and foreign exchange risk (Deng 2020). Financial reporting pressures are another important operating environment characteristic that can substantially affect tax outcomes when firms (a) engage in additional tax avoidance to reduce tax expense and thus increase after-tax earnings (Lisowsky 2010; Klassen and Laplante 2012), (b) manipulate tax accruals to manage earnings (Krull 2004; Frank and Rego 2006; Cazier et al. 2015), or (c) pay additional taxes on manipulated earnings (Erickson et al. 2004; Morrow and Ricketts 2014).

Unlike agency and implementation factors, which affect tax outcomes ex ante, outcome factors influence tax planning by affecting the anticipated ex post costs of reporting a particular tax outcome. A significant outcome factor is regulation and oversight by regulators. Firms that take aggressive tax positions or use tax accruals to manipulate earnings may face tax authority audits or financial regulator investigations and may bear the costs of, for example, having tax positions overturned, being required to restate financial statements, or paying interest, penalties, and fines. As such, firms consider the expected value of these costs when evaluating tax positions and tax reporting (Hoopes et al. 2012; Kubick et al. 2016, 2017). Political, market, and reputational costs are other outcome factors that can substantially affect tax outcomes. Managers claim that reputational costs are one of the most important considerations in tax planning (Graham et al. 2014), and empirical evidence suggests that firms can face reputational damage and customer backlash when their tax outcomes indicate tax avoidance in a manner perceived to be aggressive or unfair (Hanlon and Slemrod 2009; Mills et al. 2013; Dyreng et al. 2016; Austin and Wilson 2017; Dhaliwal et al. 2022). Like customers, employees often react ex post to the tax planning choices of the tax department and upper managers, with evidence suggesting that they often view aggressive tax avoidance negatively (Chyz et al. 2013; Wilde 2017; Lee et al. 2021).

2.2 Text and narrative disclosures

Considerable research in accounting examines descriptive features of qualitative disclosures, such as length, tone/sentiment, readability, similarity, and embedded quantitative content (Li 2008; Brown and Tucker 2011; Rogers et al. 2011; Loughran and McDonald 2014, 2016; Lang and Stice-Lawrence 2015; Hutchens 2017; Crowley and Wong 2022; Bochkay et al. 2023). While these features of qualitative disclosures can richly describe *how* firms speak in disclosures, they often miss out on the content of *what* firms actually disclose (Loughran and McDonald 2016). Some recent accounting research goes further to examine *what* firms are saying in their qualitative disclosures (Brown et al. 2020). This research typically takes one of two approaches: (a) using word counts to evaluate text subject to predetermined word dictionaries (Kothari et al. 2009; Loughran and McDonald 2011) or machine learning techniques (Basu et al. 2022) or (b) using natural language processing to determine latent topics of discussion (Dyer et al. 2017; Hoberg and Lewis 2017; Huang et al. 2018; Brown et al. 2020).

Research examining the narrative topics in qualitative disclosures finds that these topics respond to major changes in financial accounting and reporting standards (Dyer et al. 2017) and regulator oversight (Lowry et al. 2020). Discussions of certain topics can also illuminate firm risks and financial misreporting (Bao and Datta 2014; Campbell et al. 2014; Brown et al. 2020; Donovan et al. 2021). However, research on narrative topics in accounting is still in its early stages (Bochkay et al. 2023), and researchers know little about how topical discussion in firm disclosures can be used by corporate stakeholders.

2.3 Theoretical development

Given the importance of corporate income tax outcomes to corporate stakeholders (Brown et al. 2016a; Bratten et al. 2017; Chyz and Gaertner 2018; He et al. 2020), particularly for firm valuation (Graham et al. 2012; Inger 2014; Goh et al. 2016; Rajgopal 2022), *predicting* tax outcomes is very important to the decisions of many stakeholders. To predict tax outcomes, stakeholders can use quantitative information on past tax outcomes (e.g., ETRs; Dyreng et al. 2008; Guenther et al. 2023; Demeré et al. 2024) or models of tax outcome determinants (Hanlon and Heitzman 2010; Wilde and Wilson 2018). However, these quantitative metrics only represent a portion of the content in corporate disclosures, as qualitative discussions can contain significant incremental information (Brown et al. 2020; Burd et al. 2023).

Some research has begun to touch on the value of qualitative tax information. For example, Campbell et al. (2019) find that the presence of certain tax keywords in risk factor disclosures (i.e., 10-K Item 1A) can predict cash ETRs. Concurrent studies also examine tax footnotes and find that the presence of certain keywords and textual characteristics can affect stakeholder processing of tax information (Hutchens 2017; Luo et al. 2024). Stakeholders also appear to pay attention to qualitative discussions of quantitative disclosures with tax-related XBRL tags (Burd et al. 2023). Further, in a concurrent working paper, Jennings et al. (2020) show that tax-related language in the management discussion and analysis (10-K Item 7) and tax footnote can improve predictions of the likelihood of a firm experiencing a tax settlement.[3] In total, these studies suggest that stakeholders use the tax-specific discussion in certain limited pieces of disclosure to understand tax outcomes.

However, by focusing only on the tax-specific discussion within narrow portions of the 10-K, these studies do not consider most of the qualitative disclosure in that filing. For example, the average firm has an average of less than four tax mentions in Item 1A of the 10-K that are the focus of Campbell et al. (2019). Similarly, the average firm has an average of less than 25 sentences discussing tax numbers per 10-K, the focus of Burd et al. (2023). The qualitative disclosure outside tax footnotes and limited tax sentences is extensive and could contain substantial information about items that, although not explicitly related to a tax keyword, might influence tax outcomes. Specifically, if this qualitative disclosure can provide information about tax outcome

[3] In contrast to Jennings et al. (2020), we examine (a) tax outcomes besides the likelihood of having a tax settlement, (b) a much wider array of qualitative disclosures that have incremental information content beyond what is found in tax-specific discussion in Item 7 and the tax footnote, and (c) whether and how stakeholders use qualitative information.

determinants or how they might map into tax outcomes (e.g., Wilde and Wilson 2018), then it could be very informative to the prediction of tax outcomes. For example, a discussion of internal controls may reveal corporate governance strategies and culture that could constrain tax aggressiveness, while a discussion of firms' legal structure and use of partnerships may indicate the presence of of special purpose entities or structures that can facilitate future tax savings through cross-border income shifting.

But there are also reasons to question the amount of useful information for predicting tax outcomes in the 10-K. First, because these topics are not necessarily tax-specific and can only inform about tax outcomes indirectly, they may not have substantial information content for tax outcomes. Second, firms may strategically manage their qualitative discussions to limit the informativeness for tax outcomes. Because tax benefits and planning can offer competitive advantages (Kubick et al. 2015; Donohoe et al. 2022), disclosures that provide tax information may help competitors adopt similar tax strategies and cut into a firm's competitive tax advantage or may help competitors identify when a firm might face increased tax costs and thus be less able to respond to competitive actions. Additionally, disclosures that provide information about tax outcomes may be used by adversarial tax authorities to select firms or tax positions for costly tax audits (Mills 1998; Graham et al. 2012; Bozanic et al. 2017). These forces provide strong incentives for firms to strategically manage their qualitative disclosures to limit their informativeness for taxes, consistent with evidence that firms will use vague boilerplate language in disclosures when faced with proprietary costs and judicial and regulatory review (Hope et al. 2016; Cazier et al. 2021) and that firms will make tax disclosure formatting more opaque when facing greater risk of tax authority scrutiny (Chychyla et al. 2022; Flagmeier et al. 2023). Third, tax disclosures tend to be highly standardized and subject to one-size-fits-all rules that can limit informativeness (Robinson et al. 2016; Bernard et al. 2023). While this might mean that narrative discussion that is not tax-specific will be even more informative than tax-specific disclosures, the standardization tendencies of financial accounting standards may also limit the amount of tax-relevant information that can be gleaned from narratives. Fourth, even if the narrative does contain information regarding tax outcomes, it may not be incremental to that already contained in key quantitative tax metrics, such as past tax outcomes. As such, it is important to document the extent to which narrative disclosures can help stakeholders better predict tax outcomes.

3 Research design

3.1 Textual data

Our textual data comes from U.S. public firms' annual report (Form 10-K) filings from 2006 through 2019.[4] We begin our sample in 2006, as this is the first full year

[4] We download 10-K reports as Stage One 10-X Parse Data from the Software Repository for Accounting and Finance of the University of Notre Dame (Loughran and McDonald 2016). This repository provides 10-Ks after removal of extraneous text, such as HTML tags, non-ASCII characters, and tables, which helps ensure consistent machine-readable content for textual analysis. A detailed description of the preprocessing steps is provided at https://sraf.nd.edu/data/stage-one-10-x-parse-data/.

after the Securities and Exchange Commission (SEC) mandated risk factor disclosures (i.e., Item 1A), which guarantees a consistent number of 10-K items throughout the sample period. We then individually extract all 20 items in the annual reports and estimate a series of topic models at the item level. A description of our pre-processing steps is reported in Online Appendix OA-A.I. After generating our topic and textual measures, we merge the 10-K sample with Compustat.

3.2 Identification of thematic content

To measure qualitative disclosure content, we use a topic modeling approach based on a variant of LDA (Blei et al., 2003; Chen et al., 2015). LDA automatically discovers the topics contained within a collection of documents in an intuitive manner without requiring many assumptions (Dyer et al. 2017; Huang et al. 2018; Brown et al. 2020).[5] It treats a document as a mixture of a finite number of unobserved (i.e., latent) groups, which are called *topics*. An LDA topic is a collection of words or groups of words that share a theme. For example, a potential 10-K topic might include words like "price," "stock," and "security," reflecting a discussion of the stock market, while another topic could focus on an industry theme by gathering words like "integrate," "circuit," and "semiconductor." To better capture topical context and facilitate interpretability, we use bigrams as the basic components of our model (Wang et al. 2007).[6] Each related bigram collectively contributes to the thematic content of a topic.

The final output of the LDA model is represented by a document-topic-weights matrix (DTW), where each entry indicates the estimated proportion of a specific topic within each document. For example, an LDA model with 10 topics will produce a 10-topic DTW with each topic within a document being assigned a value between 0 and 1 and the topic values for a given document summing to 1. This reflects the proportion of content related to each topic within each document. For example, a proportion of 0.16 for topic A indicates that 16% of the document pertains to topic A. The DTW is what we use when constructing our topic measures, so that each topic variable will have a value between 0 and 1 for each observation, with the value representing the proportion of the document's text that discusses this topic.[7]

[5] We provide a detailed discussion of the method in Online Appendix OA-A.II.

[6] A bigram is formed by two consecutive words (or tokens) in the document. After removing stopwords and punctuation, a sentence like "The company contracts with numerous third parties to offer their digital content to customers" is decomposed into such bigrams as "company_contracts," "contracts_numerous," "numerous_third," "third_parties," "parties_offer," "offer_digital," "digital_content," and "content_customers."

[7] LDA also generates a topic-word-weights matrix (TWW), which represents the estimated probability of each word or bigram appearing in a particular topic. This matrix is instrumental in identifying the words or bigrams that most strongly characterize each topic, as those with the highest probabilities are considered the most indicative of the topic's content. By analyzing the TWW matrix, we can pinpoint key bigrams and use them to assign an intuitive label to each topic, as detailed in Online Appendix OA-B Table B1. This labeling approach follows Brown et al. (2020), who note that interpreting and labeling LDA topics requires human judgment.

We estimate a topic model at the 10-K item level, which allows us to determine the proportion of each topic in each 10-K item.[8] Since the primary input parameter required by LDA is the number of topics, we estimate a series of models ranging from 20 to 200 topics in increments of 20.[9] The choice of the final number of topics is usually guided by a combination of the researchers' domain knowledge and formal statistical tools. While we use our own domain knowledge and the Blei et al. (2003) perplexity measure, we primarily rely on the OpTop measure of Lewis and Grossetti (2022) because it removes the subjectivity that accompanies interpreting perplexity charts and bases our inferences on a formal statistical test.[10] We conclude that an 80-topic model gives the best balance of statistical power and topic interpretability.[11] However, because the 80 topic variables sum to 1 due to the multinomial nature of LDA, we have to omit some topics to avoid perfect multicollinearity. Following prior research (e.g., Brown et al. 2020), we omit topics that appear to be industry-specific and thus more likely to be collinear with industry fixed effects, leaving us with 54 topic variables.

To understand the variation in our topic variables, we plot the distribution of topic proportions (i.e., the percentage of the 10-K composed of discussion about a topic) across 10-Ks in Fig. 1. For ease of presentation in this figure, we aggregate our topic variables into three categories following the framework of Wilde and Wilson (2018), with the aggregation provided in Column (2) of Online Appendix Table B1. Each panel represents the distribution of a given topic aggregate (i.e., the proportion of qualitative discussion in a 10-K that is about a topic aggregate) across our sample of 10-Ks. These panels illustrate that there is substantial variability in the amount of emphasis that 10-Ks give to each topic.

[8] Estimating at the 10-K item level greatly increases the sample size to more than 2 million documents. This large collection of documents poses computational challenges, as the classic LDA developed by Blei et al. (2003) cannot handle such dimensionality. As such, we use a cache-efficient LDA implementation called WarpLDA (Chen et al. 2015) to facilitate estimation. The method proposed by Chen et al. (2015) replaces collapsed Gibbs sampling with Monte-Carlo expectation maximization while maintaining the same working principle as Blei et al. (2003).

[9] We estimate our LDA model using optimized hyperparameter values of α (the DTW Dirichlet prior) equal to 0.1 and β (the TWW Dirichlet prior) equal to 0.001. We determined these values through a grid-search approach (Kuhn and Frick 2024), which prevents arbitrary tuning and ensures that our topic model captures meaningful data structures. This approach also follows established practices in machine learning and statistics research (e.g., Hastie et al. 2009).

[10] Lewis and Grossetti (2022) show how the perplexity measure, used by Dyer et al. (2017) and Huang et al. (2018), often overestimates the optimal number of LDA topics.

[11] OpTop is a fully parametric chi-square statistic test of whether the observed distribution of words in a document is statistically similar to the distribution predicted by a K-topic LDA model (Lewis and Grossetti 2022). By evaluating models with different numbers of topics (in our case, a sequence $k = 20, \ldots, 200$ in increments of 20), the chi-square statistic identifies as optimal the range of K-topic models for which we cannot reject the hypothesis that the observed and estimated word distributions are identical, indicating a good fit. Within the range of optimal K-topic models, 80 topics was a round number that, when we reviewed the LDA output, produced topics that could be clearly understood and labeled and thus that have high human interpretability and could be easily used by stakeholders. In addition, following Dyer et al. (2017), we execute a word intrusion task by extracting nine words suggested by the model as pertaining to a given topic and one intruder. We conducted this task manually, and the results give us confidence about the model's ability to correctly cluster topics.

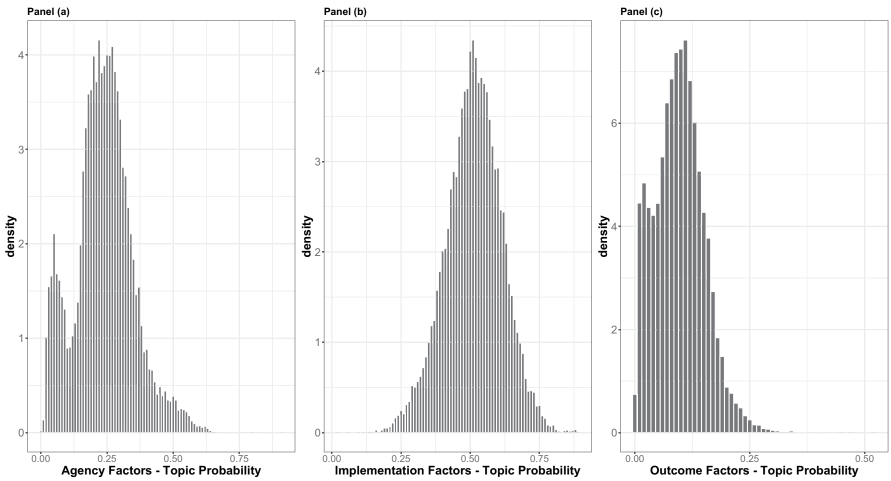

Fig. 1 Topic Distribution Histograms. Panels (a), (b), and (c) show the distribution of topic probabilities (i.e., the proportion of a 10-K containing discussion related to a topic) for three topic aggregates (i.e., agency factors, implementation factors, and outcome factors, respectively). For ease of presentation, we aggregate our 54 topic variables into three categories, following the Wilde and Wilson (2018) framework, as mapped in Online Appendix Table B1

3.3 In-sample empirical design

To evaluate the ability of narrative discussion topics in 10-Ks to predict future corporate income tax outcomes, we begin by estimating the following in-sample OLS regressions for firm i at time t:

$$Cash\ ETR_{i,t+\tau} = \gamma_0 + \sum_{l=1}^{L} \boldsymbol{\gamma}_l \text{Controls}_{itl} + \varepsilon_{it}, \tag{1}$$

$$Cash\ ETR_{i,t+\tau} = \beta_0 + \sum_{k=1}^{K} \boldsymbol{\beta}_k \text{Topic}_{itk} + \sum_{l=1}^{L} \boldsymbol{\theta}_l \text{Controls}_{itl} + \varepsilon_{it}, \tag{2}$$

where $Cash\ ETR_{i,t+\tau}$ is the cash ETR (i.e., cash taxes paid divided by pretax income) measured for firm i at year $t+\tau$, where τ is alternately one, two, or three fiscal years ahead. We focus on the $Cash\ ETR$, given evidence that ETRs are one of the most important tax outcomes (Graham et al. 2014; Flagmeier et al. 2023) and that cash ETRs are increasingly important to evaluating tax outcomes (Brown et al. 2016a).[12] Cash ETRs are also important metrics for understanding firms' tax planning (Dyreng et al. 2008) and closely relate to governmental tax revenues.

[12] Focusing on predicting *cash* tax outcomes also ties closely to the goal of financial reporting outlined in SFAC 8 of helping stakeholders make decisions informed by "their assessment of the amount, timing,

Our primary focus is on the additional explanatory power provided by including our 54 topic variables (the **Topic** vector) described in Section 3.2. We measure this additional explanatory power in two ways. First, we use a joint F-test of the topic variables in Eq. 2. Second, we compare the difference in R^2 between Eqs. 1 and 2, which we formally test using a cluster-robust Vuong test (Vuong 1989; Wooldridge 2010).[13] Together, these analyses can identify whether there is a statistically significant improvement in explanatory power for models predicting cash ETRs when topic variables are included.[14]

We also include a vector of control variables (**Controls**) in all models. We begin by controlling for the $Cash\ ETR$ measured at time *t*. This control is important because prior tax outcomes strongly predict future tax realizations, and we want to test whether our qualitative topic variables can predict tax outcomes beyond major quantitative variables (Guenther et al. 2023).[15] We also control for quantitative factors used in research predicting tax sheltering behavior (Wilson 2009; Lisowsky 2010). Specifically, we control for firm size (*Firm Size*), leverage (*Leverage*), profitability (*Profitability*), research and development expense (*R&D*), foreign income (*Foreign activity*), net operating loss carryforwards (*NOL*), equity earnings (*Equity Earnings*), mezzanine financing (*Mezzanine Finance*), litigation risk (*Litigation Risk*), and use of a Big Four auditor (*Big4*). We further control for quantitative variables that represent economic factors commonly included in tax avoidance models, including the book-to-market ratio (*Book-to-Market*), property and equipment (*Capital Intensity*), intangible assets (*Intangible Intensity*), cash holdings (*Cash*), sales growth (*Sales Growth*), and business segmentation (*Business Segments*) (McGuire et al. 2014; Demeré et al. 2020; Al-Hadi et al. 2022). To ensure that our topics are not simply

and uncertainty of (the prospects for) *future net cash inflows* to the entity" (FASB 2021 page 2) (emphasis added). In Section 4.6, we report the results of prediction tests for additional tax outcome variables.

[13] To quantify the magnitude of the explanatory power improvement provided by including our topic variables, we use the difference in R^2 between Eqs. 1 and 2. Additionally, to provide an easily comprehensible economic magnitude estimate, we use the magnitude of statistically significant coefficients on our topic variables as described in Section 4.2.

[14] Some research has also used Shapley values to evaluate the relative importance of explanatory variables (Guenther et al. 2023). However, recent evidence from computer science and machine learning research indicates that Shapley values suffer from a number of limitations, including excessive model dependence and an inability to properly weight model features with dependence on other features (Kumar et al. 2020, 2021) as well as theoretical limitations that prevent useful inference in machine-learning analyses (Huang and Marques-Silva 2024). Despite these limitations, in untabulated analyses, we evaluate the Shapley values of our topic measures. Depending on the future time horizon examined, Shapley values indicate that our topic variables explain 19.8% to 28.7% of the variation in cash ETRs. However, if we hold aside variation from fixed effects and the lagged tax outcome variable, which will provide the most explanatory power (Guenther et al. 2023), we find that our topic variables explain 61.5% to 64.0% of the remaining variation in cash ETRs. To the extent that Shapley values can be used for inference, these results indicate that our topic variables provide substantial explanatory power beyond common quantitative tax outcome predictors.

[15] Controlling for the time *t* dependent variable also removes past variation from the dependent variable, making the regression similar to a first-differences design. However, while a first-differences design implicitly assumes perfect serial correlation of the dependent variable (i.e., a random-walk expectation), controlling for the lag of the dependent variable controls more flexibly for imperfect serial correlation of the dependent variable.

capturing overall 10-K readability or the extent of qualitative disclosures, we also control for the Fog Index (*Fog Index*) (Li 2008) and 10-K length (*10-K Length*) (Loughran and McDonald 2014). We winsorize continuous variables at the first and 99th percentiles to address outliers and report detailed variable definitions in Appendix A. Finally, we include industry and year fixed effects and cluster standard errors at the firm level.[16]

3.4 Out-of-sample empirical design

While in-sample analyses are an important baseline, they implicitly assume that prediction estimates remain stable over time (Lev et al. 2010). As such, we next estimate out-of-sample models using a rolling-window approach (Brown et al. 2020) to evaluate the ability of 10-K discussion topics to predict tax outcomes using extreme gradient boosting (XGBoost) (Chen and Guestrin 2016; Chen et al. 2022; Geertsema and Lu 2023; Guenther et al. 2023). XGBoost is a tree-based ensemble learning algorithm (Friedman 2001) that builds multiple decision trees sequentially, improving predictions at each step by identifying mistakes made by previous trees. Unlike traditional regression models, which assume linear relationships between variables, XGBoost captures nonlinear patterns and complex interactions between predictors, making it particularly well suited for financial forecasting.

We train XGBoost to predict future cash ETRs using the same set of independent variables as specified in Eqs. 1 and 2.[17] Following Brown et al. (2020), we train the model using five-year rolling windows to avoid a look-ahead bias. In total, we have nine annual prediction cross-sections for cash ETRs.[18] For each fiscal year, we randomly select 90% of observations as our training sample and use the remaining 10% as our testing sample. To ensure consistency, we resample with replacement 25

[16] We use industry fixed effects rather than firm fixed effects because (a) the latter can introduce endogeneity in the form of Nickell bias (Nickell 1981) and often do so in finance and tax accounting settings (Grieser and Hadlock 2019; Demeré 2023), (b) firm fixed effects can incorporate future information into predictor variables in the same way they introduce endogeneity bias, creating a look-ahead bias that is dangerous to prediction inferences, and (c) we want to maintain conformity between our in- and out-of-sample analyses, as our out-of-sample analyses are estimated on a year-by-year basis that does not accommodate firm fixed effects. Nevertheless, in untabulated analyses, we find that our in-sample results (a) are largely unchanged when replacing industry fixed effects with industry-by-year fixed effects and (b) continue to show that our topic variables improve model predictive power when replacing industry fixed effects with firm fixed effects and predicting Cash ETRs in $t + 1$. However, the statistical significance of prediction improvements in firm-fixed-effect models diminishes over longer horizons as firm fixed effects become more likely to overcontrol away persistent information and introduce biases.

[17] We provide additional discussion of the estimation in Online Appendix OA-A.III.

[18] For example, the first out-of-sample prediction we run uses topic and control variable data from 2006, 2007, 2008, and 2009 to predict cash ETRs in years 2007, 2008, 2009, and 2010, respectively. The coefficient weights from these models are then averaged and applied to topic and control variable data from 2010 to predict cash ETRs in the 2011 prediction year. This rolling window approach prevents look-ahead bias by ensuring that 2011 cash ETRs are predicted only using data from years prior to (and thus known by) 2011. To prevent overfitting, we include a regularization term using the L1 metric (also known as *Lasso*) (Tibshirani 1996).

Table 1 Sample selection

This table reports the sample selection process

Firm-years from SEC EDGAR	107,268
Excluding firm-years with missing data in Compustat	−30,970
Excluding financial and utility firms and firm-years with negative total assets, cash holdings, or equity	−8,839
Excluding firm-years with missing tax information or control variables, and negative pre-tax income	−49,810
Final sample	**17,649**

times and report the bootstrapped results (Efron 1992).[19] The metric of interest in this analysis is the RMSE from prediction models that contain versus omit our 54 topic measures. To the extent that narrative discussion topics can help predict tax outcomes, then the RMSE for the model including our topic measures will be *smaller* than the RMSE for the model that omits our topic measures.

3.5 Sample selection

To construct our sample, we begin with the intersection of our narrative topic measures and Compustat. We then exclude (1) observations with negative total assets, cash holdings, or equity and (2) utilities and financial firms (Standard Industrial Classification (SIC) codes 4900–4999 and 6000–6999) because these firms face fundamentally different tax rules and economic and regulatory pressures. We then exclude observations with negative pre-tax income (Compustat item *pi* < 0) because ETRs are difficult to interpret for firms with negative pre-tax income (Dyreng et al. 2016). We also remove observations in which the calculated cash ETR exceeds one (Dyreng et al. 2008). Finally, we remove observations that are missing necessary data to calculate our variables of interest and control variables. As shown in Table 1, we obtain a final sample of 17,649 firm-year observations for 3,241 public U.S. firms. The annual number of firm-year observations varies from 1,072 in 2019 to 1,401 in 2013. At the 10-K item level, there are an average of approximately 6.8 items reported in each 10-K in our sample, resulting in 120,239 firm-year-item observations, which we use to evaluate the location of prediction-relevant information in Section 4.4.

[19] Machine learning models come with a series of parameters that need to be tuned to optimize prediction performance. To optimize our XGBoost models, we tune the following parameters: *eta*, *max_depth*, *min_child_weight*, and *lambda*. *eta* controls the learning rate used to prevent overfitting. A lower value for *eta* means that the model is more robust to overfitting but also slower to fit. *max_depth* is the depth at which a tree can grow (Breiman 2001; Breiman et al. 2017). Ideally, a tree with more branches and leaves should be able to find more complex patterns but can also lead to overfitting problems. *min_child_weight* controls how many times a given leaf will be further partitioned, given the number of instances. The larger this value, the more conservative the algorithm will be. *lambda* controls the amount of L1 regularization in the model to prevent overfitting (Tibshirani 1996). Following Chen et al. (2022), we initially set these parameters to their default values as given in the R package xgboost (Chen et al. 2025) and then tune them to minimize the RMSE. The default values are *eta* = 0.3, *max_depth* = 6, *min_child_weight* = 1, and *alpha* = 0. The optimized values for cash ETR are *eta* = 0.04, *max_depth* = 100, *min_child_weight* = 10, and *alpha* = 7.

Table 2 Descriptive statistics

Variable	N	Mean	SD	p25	p50	p75
Cash ETR	17,649	0.2255	0.1794	0.0864	0.2137	0.3195
Street Tax Expense	91,741	0.0269	0.0265	0.0130	0.0229	0.0350
Street ETR	91,741	0.2843	0.1235	0.2161	0.3000	0.3647
Est Tax Expense	91,741	0.0270	0.0241	0.0136	0.0237	0.0356
Est ETR	91,741	0.2871	0.1241	0.2214	0.3071	0.3689
AccTax	91,741	−0.0062	0.0120	−0.0062	−0.0026	−0.0011
AccETR	91,741	−0.0812	0.3209	−0.0453	−0.0166	−0.0068
GAAP ETR	18,553	0.2631	0.1661	0.1616	0.2831	0.3590
ΔMVA	17,351	0.0135	0.0129	0.0035	0.0109	0.0198
Log(Cash Taxes Paid)	16,099	2.7947	2.3734	1.3985	2.9907	4.3526
ISA Cash ETR	17,649	−0.0072	0.1756	−0.1331	−0.0188	0.0795
UTB	12,926	0.0097	0.0163	0.0008	0.0043	0.0116
Tax Settlements	15,143	0.0003	0.0008	0.0000	0.0000	0.0001
10-K Length	17,649	10.6689	0.4744	10.4367	10.7199	10.9661
Fog Index	17,649	24.5715	2.0152	23.1483	24.4277	25.7233
Leverage	17,649	0.2085	0.1888	0.0276	0.1808	0.3325
Book-to-Market	17,649	0.5360	0.4680	0.2599	0.4292	0.6788
Firm size	17,649	7.0724	2.0095	5.7971	7.0798	8.3834
R&D	17,649	0.0297	0.0613	0.0000	0.0000	0.0261
Capital intensity	17,649	0.2252	0.2262	0.0553	0.1441	0.3217
Profitability	17,649	0.1659	0.1009	0.1012	0.1450	0.2045
Intangible intensity	17,649	0.2086	0.2122	0.0185	0.1418	0.3501
Cash	17,649	0.1915	0.2293	0.0413	0.1133	0.2611
Sales Growth	17,649	0.1333	0.3203	0.0095	0.0780	0.1800
NOL	17,649	0.7162	0.4508	0	1	1
Foreign activity	17,649	0.9273	0.2596	1	1	1
Equity earnings	17,649	0.2495	0.4327	0	0	0
Mezzanine finance	17,649	0.0105	0.0424	0	0	0
Big4	17,649	0.7636	0.4249	1	1	1
Litigation risk	17,649	0.0854	0.2795	0	0	0
Business segments	17,649	1.7321	0.8110	1.3863	1.3863	2.3026

This table reports summary statistics for the main variables in the empirical models. We report the number of observations (N), mean (Mean), standard deviation (SD), and the 25th, 50th, and 75th percentile of the distribution (p25, p50, p75, respectively). All continuous variables are winsorized at the first and 99th percentiles. Appendix A provides the variable definitions

4 Empirical results

4.1 Descriptive statistics

Table 2 presents descriptive statistics for the variables used in our analysis.[20] The average firm in our sample has a cash ETR of approximately 0.23. The composition and characteristics of our sample resembles those reported by Campbell et al. (2019).

[20] For clarity and brevity, we omit our topic variables from Table 2 and report them in Online Appendix Table B1.

4.2 In-sample prediction results

Table 3 presents the estimation of Eq. 2 using $Cash\ ETR$ as the dependent variable. For brevity, we only report joint F-tests of the topic variables in Eq. 2 and the differences in R^2 between Eqs. 1 and 2, along with a cluster-robust Vuong (1989) test.[21] In the first row with cash ETRs one period into the future, we find that our topic variables are jointly significant additions to Eq. 2, and the Vuong test indicates that there is a statistically significant improvement in explanatory power between Eqs. 1 and 2. Economically, the addition of topic variables increases the prediction model's power by 4.7%. Similarly, when examining cash ETRs two and three periods into the future, we find that our topic variables are statistically significant model additions and significantly improve the explanatory power of the predictive model. For these future periods, however, the increase in explanatory power is even more economically significant, with a 7.9% (9.7%) increase in the model's explanatory power when predicting cash ETRs two (three) years into the future. Together, these results indicate that narrative topics provide important information about factors that affect tax planning and can substantially improve predictions of cash ETR outcomes.

To provide some additional context about what information in qualitative narratives appears most important in predicting cash ETRs, we summarize the statistically significant coefficients on our topic variables in Fig. 2. A light gray *plus* indicates that a greater emphasis on a given topic significantly predicts a higher cash ETR for a given future period, while a dark gray *minus* indicates that a greater emphasis on a given topic predicts a lower cash ETR for a given future period. In general, we find that topics indicating (1) governance that can constrain aggressive tax planning (e.g., internal controls and shareholder governance), (2) risks that could increase the costs of aggressive tax planning (e.g., customer relations, negative events, foreign exchange risk, and financial reporting), and (3) the presence of complex transactions that may increase the costs of tax planning (e.g., M&A and intangibles) tend to predict higher cash ETRs. We also find that topics that indicate the presence of complex organizational structures that can facilitate tax planning (Demeré et al. 2020) (e.g., partnership issues, REITs, firm operations, and reorganizations) tend to predict lower cash ETRs. As such, these are the narrative topics that stakeholders interested in forecasting cash tax outcomes may find most valuable. Economically, a standard deviation increase in the proportion of a 10-K containing discussion of each topic that positively (negatively) predicts cash ETRs is associated with future cash ETRs that are 0.38% to 1.13% higher (0.36% to 1.15% lower), depending on the topic.[22] Given that the mean cash ETR in our sample is 22.55%, the information in narrative topics appears to predict economically significant variation in cash ETRs. In monetary terms, one standard deviation increase in the proportion of a 10-K devoted to any

[21] This display is consistent with other studies focused on relative information content and predictive ability (Ayers et al. 2009; Guenther et al. 2023). Because we are interested in the incremental predictive ability of our topic variables rather than a causal effect, we include all control variables in every reported model. We report all coefficient estimates in Online Appendix Table B5.

[22] We calculate these economic significance estimates by multiplying coefficients from Online Appendix Table B5 by the standard deviations in Online Appendix Table B1.

Table 3 OLS in-sample prediction analyses

Dep.Var	Joint F-Test	Adjusted R^2				Vuong Test	Fixed Effects	Controls	Obs.
		Topics	*No Topics*	*Diff.*	%				
$Cash\ ETR_{t+1}$	F = 4.40; Prob >F = 0.0000	0.2227	0.2127	0.0100	4.7%	[8.21]***	Industry, Year	Yes	17,649
$Cash\ ETR_{t+2}$	F = 3.94; Prob >F = 0.0000	0.1740	0.1613	0.0127	7.9%	[8.18]***	Industry, Year	Yes	14,500
$Cash\ ETR_{t+3}$	F = 3.03; Prob >F = 0.0000	0.1448	0.1320	0.0128	9.7%	[7.34]***	Industry, Year	Yes	12,445

This table presents the results of OLS regressions predicting cash ETR values for various time horizons ($t+1$, $t+2$, and $t+3$) using models that either include or exclude textual topic variables (i.e., Eqs. 2 and 1). The table reports (1) the results of an F-test for the joint significance of the topic variables, (2) the adjusted R^2 for models with and without topics and their difference, (3) the percentage improvement when topics are included, and (4) the t-statistic of a cluster-robust Vuong (1989) test. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels (two tailed), respectively. Industry and year fixed effects and all control variables are included in every model. Appendix A provides the variable definitions

	Topics	Cash ETR $t+1$	Cash ETR $t+2$	Cash ETR $t+3$
2	Executive pay			-
3	Stock option compensation	+	+	+
9	Internal controls		+	+
10	Stock issues and payouts	-	-	
12	Shareholder governance and trading	+	+	+
17	Valuation issues			-
18	Firm operations and reorganizations	-	-	-
22	M&A activity and reorganizations	+	+	
25	Operating trends	-		
27	Partnership issues	-	-	-
29	Basic filing descriptions		+	
31	Financial measurement			-
32	Financial reporting	+	+	+
34	Financial reporting of intangibles	+		
35	Financial reporting and REITs	-	-	-
39	Financial statement notes	+	+	+
43	Foreign exchange risk	+	+	+
45	Negative event impacts	+	+	+
51	Customer and shareholder relations	+	+	

Fig. 2 Individual Topics and Cash ETR Predictions. This figure shows which topics are significantly associated with future ($t + 1$, $t + 2$, and $t + 3$) cash ETRs in our in-sample prediction analyses. For brevity, we only include topics with statistical significance ($p < 0.10$) for one or more future periods. Light gray cells with a plus sign (+) represent a positive effect (i.e., greater emphasis on these topics predicts *higher* future cash ETRs), while dark gray cells with a minus sign (-) indicate a negative effect (i.e., greater emphasis on these topics predicts *lower* future cash ETRs). Empty cells reflect cases where there is no statistically significant association between that topic variable and the cash ETR of that future period

single topic is associated with a change of up to approximately $6.6 million in cash taxes paid.[23]

4.3 Out-of-sample prediction results

We also run out-of-sample tests to ensure that qualitative discussion topics can be used to predict tax outcomes in a practicable manner. In doing so, we follow studies that rely on machine learning methods to predict a given outcome of interest (Jones

[23] We compute the monetary effect by multiplying the mean cash taxes paid in our sample ($131 million) by the maximum cash ETR deviation associated with a standard deviation increase in a topic variable (1.15%) and dividing by the mean cash ETR in our sample (22.55%).

Table 4 Out-of-sample predictions of future cash ETR via extreme gradient boosting

Training Window	Prediction Year	Baseline RMSE	With Topics RMSE	Δ RMSE %	Obs. (train)	Obs. (test)
		(1)	(2)	(3)	(4)	(5)
2006 - 2010	2011	0.034631	0.013146	62.04%	6,053	1,400
2007 - 2011	2012	0.034541	0.014374	58.38%	6,194	1,389
2008 - 2012	2013	0.034427	0.014702	57.29%	6,337	1,401
2009 - 2013	2014	0.034694	0.014524	58.14%	6,640	1,332
2010 - 2014	2015	0.034898	0.013700	60.74%	6,822	1,274
2011 - 2015	2016	0.035761	0.013953	60.98%	6,796	1,235
2012 - 2016	2017	0.038778	0.015217	60.76%	6,631	1,262
2013 - 2017	2018	0.038149	0.015219	60.11%	6,504	1,231
2014 - 2018	2019	0.038102	0.016210	57.46%	6,334	1,072

This table reports the root mean square error (RMSE) goodness of fit metric for each training window and its prediction year. Out-of-sample predictions for cash ETR are computed via extreme gradient boosting (XGBoost) using L1 regularization. Column (1) reports the RMSE for the baseline model specification that includes control variables (including industry and year fixed effects) only. Column (2) reports the RMSE when disclosure topics are included. Column (3) reports the percentage RMSE decline of the topic-inclusive model compared to the baseline. Columns (4) and (5) report the number of observations in the training and testing groups, respectively

2017; Chen et al. 2022; Geertsema and Lu 2023; Guenther et al. 2023) by implementing a widely accepted and reliable method called XGBoost (Friedman 2001; Chen and Guestrin 2016). XGBoost and similar machine-learning algorithms are increasingly used for generating robust out-of-sample predictions without overfitting the data (Chen et al. 2022; Uddin et al. 2022; Campbell et al. 2023; Geertsema and Lu 2023; Guenther et al. 2023). Using XGBoost allows us to predict with greater accuracy than simple OLS while mapping our analyses to a prediction tool common in practice and increasingly used in research.

Table 4 reports the empirical results when predicting cash ETRs. To evaluate the models' performance, we report the RMSE, a typical goodness of fit metric used with machine learning models. A smaller RMSE indicates that the model is better at predicting tax outcomes out-of-sample. Columns (1) and (2) report the RMSE for a *Baseline* model with control variables only and a model *With Topics* that includes our topic variables, respectively. Column (3) reports the relative percentage decrease in RMSE by adding topic variables to our *Baseline* model. We find that the inclusion of topic variables substantially improves the model's ability to predict cash ETRs. Depending on the prediction year, adding topic variables results in a 57% to 62% reduction in RMSE. Economically, these RMSE reductions translate to reduced forecast errors of approximately \$11.46 million to \$13.69 million in cash taxes paid, on average.[24] These results support that qualitative discussion topics provide important information that stakeholders can practicably use to improve their cash ETR forecasts.

[24] We compute the monetary effect by multiplying the mean cash taxes paid in our sample (\$131 million) by the difference between the baseline RMSEs and With Topics RMSEs and dividing by the mean cash ETR in our sample (22.55%).

4.4 10-K item-by-item analysis

Our topic modeling approach also allows us to identify the exact 10-K items where these discussion topics occur (e.g., Item 1A versus Item 7). Exploring the value of qualitative disclosure across 10-K items is a novel idea in accounting research but offers the opportunity to derive unique insights. For example, examining and comparing individual 10-K items allows us to see whether specific topics of interest are concentrated in a single item or are spread throughout the 10-K. Understanding where the prediction-relevant discussion occurs would provide useful guidance to researchers and corporate stakeholders on where to focus their attention. Additionally, by evaluating the distribution of prediction-relevant discussions across 10-K items, we can evaluate the importance of studying the information content of an annual report as a whole versus limiting the focus to specific individual 10-K items, as is common in prior research (e.g., Balakrishnan et al. 2019, Campbell et al. 2019, and Luo et al. 2024).

As such, we next modify Eq. 2 by estimating it at a firm-year-item level rather than a firm-year level. Thus, each firm-year will have the same number of observations as it has 10-K items, with the values of the topic variables being the only thing that differs for observations within a firm-year. We then use our out-of-sample testing methodology with XGBoost to calculate the RMSE for a model with all firm-year-item observations included over our training windows and report the results of this baseline model in Column (1) of Table 5.[25]

We then examine how the RMSE Changes as specific 10-K items are iteratively dropped from this analysis in subsequent columns of Table 5.[26] Our results indicate that there is substantial variability in how much predictive value is lost when omitting a given item. At the high end, omitting the narrative discussion in risk factor disclosures (Item 1A) results in a 7.74% increase in RMSE, while at the low end, the other information in Item 9B does not appear relevant to cash ETR prediction. Importantly, however, the predictive value of narrative disclosure is not concentrated in a single 10-K item but is distributed among several. Discussions in the business description (Item 1), risk factor disclosures (Item 1A), management discussion and analysis (MD&A, Item 7), and financial statement exhibits (Item 15) are the most important, and all contain relatively similar amounts of prediction-relevant information. Discussions in the financial statements (Item 8) and controls and procedures disclosures (Item 9A) are somewhat less important, but still omitting either of these items results in an RMSE increase equivalent to approximately two-thirds of the RMSE increase from omitting risk factor disclosures. Beyond these six items, the predictive value

[25] The RMSE from the baseline model in Column (1) of Table 5 is smaller than the RMSE reported for the *With Topics* model in Column (2) of Table 4, indicating that the prediction model is better (i.e., that there is unique information content) in knowing where qualitative information is and how it is distributed across different items.

[26] By focusing on the firm-year-item level, we ensure a constant sample of firm-year observations where no firm-year is dropped, only specific items from the 10-K. In Online Appendix Table B4, we repeat this analysis for the five most populated 10-K items after reducing our sample to only firm-years that report all five of these items. Inferences from this analysis are similar to those reported here, i.e., that prediction-relevant qualitative discussion is spread across multiple 10-K items.

Table 5 Cash ETR and 10-K item-by-item analyses

Training	Prediction		Excluding									
Window	Year	Baseline	It.1	It.1A	It.1B	It.2	It.3	It.4	It.5	It.6	It.7	It.7A
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
2006 - 2010	2011	0.006866	0.007157	0.007318	0.006871	0.006897	0.006616	0.006869	0.006773	0.006836	0.007347	0.006812
2007 - 2011	2012	0.006486	0.006852	0.006804	0.006531	0.006513	0.006475	0.006597	0.006650	0.006588	0.006874	0.006724
2008 - 2012	2013	0.006250	0.006648	0.006714	0.006307	0.006300	0.006376	0.006361	0.006478	0.006333	0.006663	0.006436
2009 - 2013	2014	0.006531	0.006823	0.006879	0.006541	0.006751	0.006632	0.006531	0.006765	0.006565	0.006849	0.006726
2010 - 2014	2015	0.006459	0.007074	0.006954	0.006453	0.006623	0.006641	0.006603	0.006727	0.006565	0.007044	0.006640
2011 - 2015	2016	0.007237	0.008386	0.008486	0.007272	0.007683	0.007384	0.007267	0.008014	0.007845	0.008340	0.007561
2012 - 2016	2017	0.005873	0.006420	0.006505	0.005880	0.006170	0.006340	0.005831	0.006287	0.005813	0.006462	0.006454
2013 - 2017	2018	0.006778	0.007111	0.007088	0.006807	0.006772	0.006830	0.006861	0.006927	0.006782	0.007112	0.006908
2014 - 2018	2019	0.006772	0.007215	0.007089	0.006761	0.006770	0.006774	0.006783	0.006611	0.006188	0.007051	0.006887
Average		0.006584	0.007076	0.007093	0.006603	0.006720	0.006674	0.006634	0.006804	0.006613	0.007082	0.006794
Delta			**−7.49%**	**−7.74%**	**−0.29%**	**−2.07%**	**−1.38%**	**−0.76%**	**−3.34%**	**−0.44%**	**−7.58%**	**−3.20%**
Observations		120,239	103,591	103,435	120,174	117,802	116,592	118,309	112,522	118,362	103,710	112,984
Training	Prediction		Excluding									
Window	Year	Baseline	It.8	It.9	It.9A	It.9B	It.10	It.11	It.12	It.13	It.14	It.15
		(1)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
2006 - 2010	2011	0.006866	0.006897	0.006869	0.006893	0.006938	0.006746	0.006896	0.006860	0.006796	0.006893	0.007127
2007 - 2011	2012	0.006486	0.006731	0.006545	0.006707	0.006448	0.006539	0.006529	0.006508	0.006550	0.006504	0.006819
2008 - 2012	2013	0.006250	0.006507	0.006241	0.006627	0.006256	0.006380	0.006395	0.006297	0.006333	0.006326	0.006704
2009 - 2013	2014	0.006531	0.006813	0.006569	0.007011	0.006526	0.006494	0.006562	0.006522	0.006494	0.006568	0.006825
2010 - 2014	2015	0.006459	0.006761	0.006430	0.006809	0.006444	0.006621	0.006626	0.006515	0.006483	0.006517	0.007035
2011 - 2015	2016	0.007237	0.008216	0.007237	0.007853	0.007156	0.007849	0.007801	0.007818	0.007759	0.007178	0.008414
2012 - 2016	2017	0.005873	0.006424	0.005830	0.006523	0.005878	0.005837	0.005855	0.005815	0.005832	0.005797	0.006473
2013 - 2017	2018	0.006778	0.006997	0.006847	0.006946	0.006854	0.006861	0.006798	0.006777	0.006794	0.006797	0.007097
2014 - 2018	2019	0.006772	0.006938	0.006739	0.006935	0.006753	0.006822	0.006796	0.006803	0.006752	0.006739	0.007075
Average		0.006584	0.006921	0.006590	0.006923	0.006584	0.006683	0.006695	0.006657	0.006644	0.006591	0.007063

Table 5 (continued)

Delta		**−5.12%**	**−0.10%**	**−5.15%**	**0.00%**	**−1.51%**	**−1.70%**	**−1.12%**	**−0.92%**	**−0.12%**	**−7.29%**
Observations	120,239	109,495	120,147	109,408	119,572	117,548	118,959	118,802	119,498	119,946	103,685

This table reports the root mean square error (RMSE) goodness of fit metric for each training window and its prediction year. Out-of-sample predictions for cash ETR are computed via extreme gradient boosting (XGBoost) using L1 regularization. Column (1) reports the RMSE for the baseline model specification that includes control variables (including industry and year fixed effects) and the full text of the annual report (i.e., topic variables for all firm-year-item observations). Columns (2) through (21) iteratively exclude the firm-year-item observations for a particular item. The (1) average RMSE across all years, (2) % difference (Delta) with the baseline model, and (3) number of observations that are included in one or more of the training windows or prediction years are reported at the bottom of the narrative discussions in other items is less but not insubstantial. For example, narrative discussions in property listings (Item 2), market disclosures (Item 5), and disclosures of market risks (Item 7A) together have more prediction-relevant information than the risk factor disclosures.

In total, these results suggest that prediction-relevant qualitative discussion is not limited to only one or two 10-K items, like risk factor disclosures or income tax footnotes (Campbell et al. 2019; Luo et al. 2024), but rather exists across many 10-K items. This evidence supports our focus on qualitative disclosure across the entire 10-K and indicates that studies focused on qualitative disclosures in a single 10-K item are likely to miss substantial incremental information. Further, these results support that our primary findings are not attributable to any single 10-K disclosure and thus do not have substantial overlap with the findings from studies examining tax keyword disclosures within a single 10-K item (Balakrishnan et al. 2019; Campbell et al. 2019; Luo et al. 2024).

4.5 Analyst tax forecasts

Prior analyses show that narrative discussion topics contain useful information that can predict cash tax outcomes and that should thus be valuable to a variety of corporate stakeholders. However, it is unclear whether stakeholders can effectively use this information, as processing and evaluating qualitative disclosure may be difficult and costly, especially without clear guidance on how to efficiently do so. As such, we next use financial analysts as a stand-in for sophisticated corporate stakeholders (Campbell et al. 2015; Blankespoor et al. 2020) and examine whether analysts use the information in narrative disclosures when forecasting tax outcomes and whether they do so effectively.

Given their sophisticated tools and training (Hope 2003; Dhaliwal et al. 2012), analysts may very well understand and incorporate the useful information in narrative topics into their forecasts, which should increase forecast accuracy. However, research also shows that analysts can miss important information in their forecasts, particularly when the information is more complex (Chang et al. 2016; Engelberg et al. 2020) or when they face an overload of information (Hirshleifer et al. 2019; Impink et al. 2022). Findings are mixed within tax accounting research generally, as analysts can produce better tax forecasts than managers in complex environments (Bratten et al. 2017) but can also struggle to forecast taxes and miss important cues in quantitative tax metrics (Plumlee 2003; Weber 2009; Kim et al. 2020; Guenther et al. 2023). However, research on limited tax-specific *qualitative* disclosures indicates that analysts and investors use these disclosures (Luo et al. 2024; Burd et al. 2023). Given mixed evidence, we have no expectation regarding whether analysts can effectively use the information in narrative topics to forecast tax outcomes.

Following Bratten et al. (2017), we focus on the accuracy of analysts' implied tax expense and ETR forecasts and report the results in Table 6.[27] As in Table 3, we

[27] Analyst forecast data come from I/B/E/S, which does not contain explicit analyst tax forecasts. Following Bratten et al. (2017), we impute the tax expense forecast as the difference between the pre-tax and after-tax earnings forecasts in I/B/E/S and then divide this imputed tax expense forecast by forecasted pre-

Table 6 Analyst analyses

Dep.Var	Joint F-Test	Adjusted R^2				Vuong Test	Fixed Effects	Controls	Obs.
		Topics	*No Topics*	*Diff.*	%				
Panel A. Actual Analyst Tax Outcomes									
$Street\ Tax\ Expense_{t+1}$	F = 2.80; Prob >F = 0.0000	0.3126	0.2939	0.0187	6.4%	[8.72]***	Analyst, Industry, Year	Yes	91,741
$Street\ ETR_{t+1}$	F = 2.09; Prob >F = 0.0000	0.3596	0.3481	0.0115	3.3%	[15.13]***	Analyst, Industry, Year	Yes	91,741
Panel B. Analysts' Tax Outcome Estimates									
$Est\ Tax\ Expense_{t+1}$	F = 3.17; Prob >F = 0.0000	0.3828	0.3642	0.0186	5.1%	[9.98]***	Analyst, Industry, Year	Yes	91,741
$Est\ ETR_{t+1}$	F = 2.66; Prob >F = 0.0000	0.3958	0.3854	0.0104	2.7%	[15.04]***	Analyst, Industry, Year	Yes	91,741
Panel C. Analysts' Tax Outcome Accuracy									
$AccTax_{t+1}$	F = 2.80; Prob >F = 0.0000	0.2293	0.2115	0.0178	8.4%	[10.88]***	Analyst, Industry, Year	Yes	91,741
$AccETR_{t+1}$	F = 2.02; Prob >F = 0.0000	0.1130	0.1065	0.0065	6.1%	[7.33]***	Analyst, Industry, Year	Yes	91,741

This table presents the results of OLS regressions predicting (1) actual outcomes for street tax expense and ETRs, (2) analyst implied forecasts of tax expense and ETRs, and (3) analyst forecast accuracy using models that either include or exclude textual variables (topics). The table reports (1) the results of an F-test for the joint significance of the topic variables, (2) the adjusted R^2 for models with and without topics and their difference, (3) the % improvement when topics are included, and (4) the t-statistic of a cluster-robust Vuong (1989) test. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels (two tailed), respectively. Industry and year fixed effects and all control variables are included in every model. Appendix A provides the variable definitions focus on joint F-tests of the topic variables and the differences in R^2 between models with and without the topic variables, along with a cluster-robust Vuong (1989) test. Because analysts do not forecast a cash ETR but rather a street ETR, we begin by testing whether our topic variables significantly improve the explanatory power of models forecasting actual street ETR outcomes.[28] In Table 6 Panel A, we find that our topic variables are jointly significant additions to models predicting future street tax expense and ETRs and significantly improve the explanatory power of these models.

In Table 6 Panel B, we examine whether our topic variables improve models predicting analyst forecasts of tax expense and ETRs to see whether analysts use the information in narrative discussions. Here we find that our topic variables are jointly significant additions to these models and significantly improve their explanatory power. These results suggest that analysts use the information in narrative disclosures, consistent with research indicating that stakeholders react to tax-specific qualitative disclosures (Luo et al. 2024; Burd et al. 2023).

Our next analysis examines whether analysts use the information in narrative disclosures effectively. To do so, we examine whether our topic variables can predict future analyst forecast accuracy. Finding that narrative topics predict analyst forecast accuracy would indicate that analysts do not appropriately use and weight the available information in their forecasts, while an inability to predict analyst forecast accuracy would be consistent with analysts understanding and appropriately using the information in narrative disclosures. We report the results of these analyses in Panel C of Table 6. We find that our topic variables are jointly significant additions to the models predicting analyst forecast accuracy for tax expense and ETRs and that these variables significantly improve the explanatory power of these models.

In sum, the results in Table 6 indicate that narrative topics provide important information about street tax expense and ETRs. Further, analysts can use this information to some degree in their forecasts. However, they struggle to appropriately use this information, resulting in predictable forecast errors.

To further support that our topic variables contain useful information about the street tax outcomes that analysts care about, we also run out-of-sample analyses, like

tax earnings to create the ETR forecast. Approximately 70% of analyst forecasts have the data necessary to impute the tax expense forecast and calculate forecast accuracy. We use all one-period-ahead forecasts in these analyses. Analysts forecasting before the 10-K is released are effectively choosing not to use the narrative disclosures in the 10-K and thus should be included in these analyses; however, we find similar results when requiring that the analyst forecasts are made only after release of the 10-K. Following Bratten et al. (2017), we run these analyses at the analyst-firm-year level by averaging all forecasts of an analyst within a firm-year but find similar results if we use only the most recent forecast. Because the average firm is covered by multiple analysts, we have more observations than in prior analyses. We cluster standard errors at the firm level to address serial correlation and cross-correlation between analysts following the same firm. In untabulated analyses, we find that our results strengthen when we cluster standard errors at the analyst level.

[28] Street tax expense and ETRs resemble GAAP tax expense and ETRs (we examine GAAP ETRs in Section 4.6) but with adjustments for (a) tax-specific items and (b) the tax effects of items excluded from street earnings (Beardsley et al. 2021). Beardsley et al. (2021) document that approximately 80% of firms in 2016 had material differences between their street and GAAP ETRs. They also find that approximately 35% of firm-years with differences exclude at least one tax-specific item, such as uncertain tax benefit reserves (UTBs) and valuation allowances, while over 90% of firm-years with differences adjust for the tax effects of items excluded from street earnings, such as stock compensation and restructuring charges.

those reported in Table 4, predicting future street tax expense and ETRs. We report the results of these out-of-sample analyses in Table 7. Results in Panel A show RMSE reductions of 43% to 45%, depending on the prediction year, when topic variables are included in the street tax expense prediction model. Similarly, results in Panel B indicate a 44% to 51% reduction in RMSE for street ETR prediction models when topic variables are included. Together, these results support that our topic variables substantially improve the predictive ability of models for the street tax outcomes that analysts focus on.

We additionally run out-of-sample analyses predicting analyst forecast accuracy for tax expense and ETRs to support that analysts do not appropriately use the information in narrative disclosures in their forecasts. We report these results in Table 8. Panel A reports RMSE reductions of 44% to 46%, depending on the prediction year, when topic variables are included in the tax expense accuracy prediction model. Similarly, Panel B reports RMSE reductions of 48% to 53% from including topic variables in the ETR accuracy prediction model. In total, the results from Table 8

Table 7 Out-of-sample predictions of future street tax expense and future street ETR via extreme gradient boosting

Training Window	Prediction Year	Baseline RMSE	With Topics RMSE	Δ RMSE %	Obs. (train)	Obs. (test)
		(1)	(2)	(3)	(4)	(5)
Panel A. Street Tax Expense (*Street Tax Expense*)						
2006 - 2010	2011	0.050103	0.027592	44.93%	24,867	7,825
2007 - 2011	2012	0.050341	0.027761	44.85%	28,371	7,591
2008 - 2012	2013	0.050891	0.028362	44.27%	31,678	7,932
2009 - 2013	2014	0.051030	0.028406	44.33%	35,478	7,687
2010 - 2014	2015	0.050912	0.028310	44.39%	37,958	7,749
2011 - 2015	2016	0.051109	0.028385	44.46%	38,784	7,340
2012 - 2016	2017	0.051797	0.029271	43.49%	38,299	7,802
2013 - 2017	2018	0.051607	0.028841	44.12%	38,510	7,573
2014 - 2018	2019	0.051912	0.029268	43.62%	38,151	6,117
Panel B. Street ETR (*Street ETR*)						
2006 - 2010	2011	0.025681	0.014494	43.56%	24,867	7,825
2007 - 2011	2012	0.024693	0.012339	50.03%	28,371	7,591
2008 - 2012	2013	0.025055	0.013006	48.09%	31,678	7,932
2009 - 2013	2014	0.025378	0.012806	49.54%	35,478	7,687
2010 - 2014	2015	0.025909	0.013215	49.00%	37,958	7,749
2011 - 2015	2016	0.027373	0.013812	49.54%	38,784	7,340
2012 - 2016	2017	0.033137	0.016265	50.92%	38,299	7,802
2013 - 2017	2018	0.033164	0.016551	50.09%	38,510	7,573
2014 - 2018	2019	0.034090	0.017810	47.75%	38,151	6,117

This table reports the Root Mean Square Error (RMSE) goodness of fit metric for each training window and its prediction year. Out-sample predictions for Street Tax Expense (Panel A) and Street ETR (Panel B) are computed via Extreme Gradient Boosting (XGBoost) using L1 regularization. Column (1) reports the RMSE for the baseline model specification that includes control variables (including industry and year fixed effects) only. Column (2) reports the RMSE when disclosure topics are included. Column (3) reports the percentage RMSE decline of the topic-inclusive model compared to the baseline. Columns (4) and (5) report the number of observations in the training and testing groups, respectively

Table 8 Out-of-sample predictions of future tax expense accuracy and future ETR accuracy via extreme gradient boosting

Training Window	Prediction Year	Baseline RMSE (1)	With Topics RMSE (2)	Δ RMSE % (3)	Obs. (train) (4)	Obs. (test) (5)
Panel A. Tax Expense Accuracy (*AccTax*)						
2006 - 2010	2011	0.054606	0.030491	44.16%	24,867	7,825
2007 - 2011	2012	0.054363	0.029400	45.92%	28,371	7,591
2008 - 2012	2013	0.054316	0.029280	46.09%	31,678	7,932
2009 - 2013	2014	0.054321	0.029552	45.60%	35,478	7,687
2010 - 2014	2015	0.054420	0.029916	45.03%	37,958	7,749
2011 - 2015	2016	0.054394	0.029815	45.19%	38,784	7,340
2012 - 2016	2017	0.054347	0.029927	44.93%	38,299	7,802
2013 - 2017	2018	0.054385	0.029885	45.05%	38,510	7,573
2014 - 2018	2019	0.054485	0.030350	44.30%	38,151	6,117
Panel B. ETR Accuracy (*AccETR*)						
2006 - 2010	2011	0.071079	0.034351	51.67%	24,867	7,825
2007 - 2011	2012	0.064402	0.029958	53.48%	28,371	7,591
2008 - 2012	2013	0.069326	0.034075	50.85%	31,678	7,932
2009 - 2013	2014	0.071254	0.036431	48.87%	35,478	7,687
2010 - 2014	2015	0.068844	0.033024	52.03%	37,958	7,749
2011 - 2015	2016	0.073501	0.038328	47.85%	38,784	7,340
2012 - 2016	2017	0.075768	0.037277	50.80%	38,299	7,802
2013 - 2017	2018	0.075408	0.036552	51.53%	38,510	7,573
2014 - 2018	2019	0.088843	0.044818	49.55%	38,151	6,117

This table reports the Root Mean Square Error (RMSE) goodness of fit metric for each training window and its prediction year. Out-sample predictions for Tax Expense Accuracy (Panel A) and ETR Accuracy (Panel B) are computed via Extreme Gradient Boosting (XGBoost) using L1 regularization. Column (1) reports the RMSE for the baseline model specification that includes control variables (including industry and year fixed effects) only. Column (2) reports the RMSE when disclosure topics are included. Column (3) reports the percentage RMSE decline of the topic-inclusive model compared to the baseline. Columns (4) and (5) report the number of observations in the training and testing groups, respectively

indicate that analysts could improve their tax forecasts by better using the information in narrative disclosures.

We also provide further analysis to understand *which* topics analysts might benefit the most from paying additional attention to in Fig. 3.[29] As shown, 21 of our topic variables predict analyst forecast errors (i.e., negative accuracy) for tax expense or ETR forecasts.[30] Economically, a standard-deviation increase in the proportion of

[29] For brevity, we only report the topics that are significantly *negatively* associated with analyst forecast accuracy in this figure, as these are the topics that are most likely to deserve further attention from analysts. A blank in the accuracy column indicates that there is no significant association with analyst forecast accuracy, while a topic being omitted may either (a) not be associated with or (b) be positively associated with analyst forecast accuracy. We report all coefficient estimates underlying this figure in Online Appendix Table B6.

[30] In the *Estimate* and *Street* columns, we also report whether the topic has a significant association with the analyst forecast estimate and the actual street outcome, respectively. In general, analyst forecast errors can arise in this setting in three ways: (a) analysts do not use information that can predict actual outcomes, which should produce a significant association with the actual outcome and an insignificant association with the forecast; (b) analysts use information that they shouldn't because it is irrelevant, which should

	Topics	Tax Expense			ETR		
		Estimate	Street	Accuracy	Estimate	Street	Accuracy
2	Executive pay		*	-			-
9	Internal controls						-
10	Stock issues and payouts	*	*	-	*	*	
15	Consolidations		*	-		*	
16	Credit agreements			-			
21	Loan portfolio						-
22	M&A activity and reorganizations			-			
25	Operating trends	*		-			-
26	Partnerships			-		*	
27	Partnership issues			-			-
28	Revenues	*	*				-
29	Basic filing descriptions						-
33	Financial reporting estimates			-			
34	Financial reporting of intangibles			-			-
38	Financial statement filing	*	*				-
44	Legal actions						-
48	Clean energy						-
50	Consumer retail	*	*	-			-
52	Customer support			-			
53	Product development						-
54	Postretirement benefits						-

Fig. 3 Analyst Forecast Estimates and Accuracy for Tax Expense and ETR. This figure summarizes the associations between specific narrative topics and analyst forecasts, actual outcomes, and analyst forecast accuracy for both street tax expense and street ETRs. $*$ indicates a statistically significant association ($p < 0.10$) between the topic and the analyst forecast or actual outcome. Negative signs indicate a statistically significant ($p < 0.10$) negative association between the topic and analyst forecast accuracy. For brevity, we only report topics that are negatively associated with forecast accuracy

a 10-K discussing a single one of these topics is associated with analyst forecast errors for tax expense (ETRs) that are approximately 4.1% to 20.4% (7.2% to 38.4%) greater than the mean forecast error. While this is only a subset of our topic variables, that there are 21 narrative discussion topics that analysts could use to significantly improve their forecast accuracy is consistent with analysts not appropriately using much of the information in narrative disclosures.

4.6 Alternative tax outcomes

In our final analysis, we examine whether our primary results are similar when using alternative tax outcome variables. While cash ETRs are increasingly important and may map better to tax planning and governmental tax revenues (Dyreng et al. 2008;

produce an insignificant association with the actual outcome and a significant association with the forecast; or (c) analysts misweight information in their forecasts. For topics with a significant association with actual outcomes and analyst forecasts, a negative association with analyst forecast accuracy arises when analysts use prediction-relevant information but severely misweight it. For topics with an insignificant association with actual outcomes and analyst forecasts, a negative association with analyst forecast accuracy could arise if (a) there is an insignificant positive association with actual outcomes and an insignificant negative association with forecasts (or vice versa), but the diverging signs open a wide enough gap between actual and forecasted tax outcomes that the difference is statistically significant or (b) actual and forecasted outcomes contain common sources of noise that differencing eliminates, increasing test power in accuracy analyses.

Brown et al. 2016a), GAAP ETRs are also important metrics (Graham et al. 2014; Flagmeier et al. 2023). As such, in the first three rows of Online Appendix Table B2, we use GAAP ETRs as our tax outcome in Eq. 2. Consistent with our cash ETR analyses, we find that adding our topic variables results in statistically significant improvements in the explanatory power of the predictive model, consistent with narrative discussion also containing prediction-relevant information about future GAAP ETRs. In Online Appendix Table B3, we confirm using out-of-sample analyses that adding topic variables to our prediction model reduces the RMSE by 48% to 77%.

Next, to evaluate whether the increased prediction value for ETRs from topic variables is attributable to tax-specific information in the ETR numerator or is solely driven by the pre-tax income denominator, we examine two non-ETR measures. First, we use an alternative measure of tax avoidance (ΔMVA) developed by Henry and Sansing (2018) that uses the market value of assets as a scalar rather than pre-tax income. Second, we use the natural log of cash taxes paid. As shown in Online Appendix Tables B2 and B3, we find that our topic variables significantly improve the explanatory power of the prediction models for these tax outcomes, with 33% to 58% reductions in RMSE for our out-of-sample models. Taken together, these results support that narrative discussion topics, even when not directly related to taxes, nevertheless help predict tax outcomes.

Most stakeholders likely want to predict the future value of firms' future cash tax burdens. In predicting the tax outcomes of multiple firms, stakeholders can also directly compare firms. Nevertheless, some stakeholders may also want to directly predict a benchmarked cash ETR.[31] As such, we follow Balakrishnan et al. (2019) and evaluate industry- and size-adjusted cash ETRs ($ISA\ Cash\ ETR$) as tax outcomes. As shown in Online Appendix Tables B2 and B3, these results are generally consistent with our primary cash ETR results.

We also examine uncertain tax benefit reserves (UTBs), as these reserves can speak to the outcomes of more aggressive tax planning (Lisowsky et al. 2013) and as such are closely monitored by tax authorities (Bozanic et al. 2017).[32] The results in Online Appendix Table B2 indicate that our topic variables improve the explanatory power of models predicting UTBs by a statistically significant amount. However, results here and in Table B3 are not very economically significant, with R^2 improvements of less than 0.4% and RMSE reductions of 1% to 19%. This lower benefit of narrative disclosures in predicting UTBs is unsurprising, however, given that (a) UTBs are particularly prone to one-size-fits-all rules that can result in UTBs that are not well aligned with economic fundamentals (Robinson et al. 2016), (b) UTBs are subject to manipulation to both manage earnings and avoid attention from tax authorities (Cazier et al. 2015; Towery 2017), and, (c) given the attention paid to UTBs by tax authorities, firms may be particularly cautious about using boilerplate

[31] Given that industry- and size-adjusted ETRs can function as good measures of aggressive tax avoidance (Balakrishnan et al. 2019), stakeholders who want to predict how aggressive a firm will be in tax planning (e.g., tax authorities) may find value in this measure.

[32] In our analyses with the *UTB* and $Tax\ Settlements$, we only have seven training windows, rather than nine for other variables, because data for these variables were not available until Financial Interpretation No. 48 was implemented.

language when discussing items of relevance to UTBs (Cazier et al. 2021; Bernard et al. 2023; Richter et al. 2024).

Finally, we examine how narrative discussion can predict the amount of future settlements with tax authorities.[33] Results in Online Appendix Tables B2 and B3 indicate that our topic variables significantly improve prediction models for tax settlements, with out-of-sample RMSE reductions of 56% to 57%. These findings should be of interest to tax authorities looking to identify firms with uncertain and aggressive tax positions and corporate stakeholders who are concerned about the costs of tax authority scrutiny.

5 Conclusion

Understanding and predicting corporate income tax outcomes is important to a variety of corporate stakeholders (Hasan et al. 2014; Goh et al. 2016; Bratten et al. 2017; Chyz and Gaertner 2018; Rajgopal 2022). However, limited disclosures (Graham et al. 2012; Rajgopal 2022), earnings management through tax accounts (Dhaliwal et al. 2004), one-size-fits-all rules (Robinson et al. 2016), and strategic opacity to reduce proprietary costs and adverse tax authority scrutiny (Hope et al. 2016; Cazier et al. 2021; Chychyla et al. 2022) can make understanding and predicting tax outcomes difficult. Given these challenges, we examine whether the narrative discussion content in financial disclosures can improve forecasts of tax outcomes.

To examine narrative discussion content, we use LDA topic analysis to estimate measures of the topical content of Form 10-K qualitative disclosures. Using both in- and out-of-sample analyses, we find that our topic variables provide significant increases in predictive power for models forecasting cash ETRs and other tax outcomes. These model improvements are incremental to many traditional quantitative tax planning determinants, showing that narrative discussion can help stakeholders better predict tax outcomes. We also take advantage of the richness of 10-K data by examining narrative discussion content by 10-K item. Our results suggest that prediction-relevant discussion is distributed throughout the 10-K, indicating that corporate disclosures should be analyzed in their entirety rather than examining only certain subsets of larger disclosures. We further find that analysts seem to struggle to use this qualitative information effectively in their tax forecasts, resulting in substantial and predictable tax forecast errors.

Our study documents an important but underexplored source of information about corporate tax outcomes. Our findings also illustrate the richness of qualitative disclosure found throughout the 10-K and should encourage researchers and corporate stakeholders to avoid focusing only on narrow pieces of qualitative disclosure in favor of a more holistic approach. Further, our study indicates that analysts struggle

[33] These results are conditional on firms having a tax settlement, which means that these results can be interpreted as predicting the magnitude of the tax settlement, given that a settlement occurs. We choose to focus on settlement magnitude rather than whether a firm has a settlement in a given year because predicting a binary variable requires a completely different approach to out-of-sample testing than we use elsewhere in the paper to predict continuous outcomes. For research predicting the binary occurrence of a tax settlement, see Jennings et al. (2020).

to appropriately use the information in broad qualitative disclosures while providing a practical method that can be applied by analysts and other corporate stakeholders to derive prediction-relevant information from qualitative disclosures. In total, we believe that our results will be of practical interest to researchers, corporate stakeholders, analysts, regulators, and tax authorities who want to better understand and predict corporate tax outcomes and evaluate the quality of firms' tax positions.

Appendix A: Variable definitions

Variable	Definition
Cash ETR	Cash taxes paid (TXPD) divided by pre-tax income less special items (PI - SPI). The variable is bounded between 0 and 1. (Source: Compustat)
Street Tax Expense	Following Bratten et al. (2017), the average of analysts' implied actual street tax expense, divided by the number of shares in *t-1* (CSHO) and scaled by price in *t-1* (PRCC_F) (Source: Compustat and I/B/E/S)
Street ETR	Following Bratten et al. (2017), the average analysts' implied actual street effective tax rate (Source: I/B/E/S)
Est Tax Expense	Following Bratten et al. (2017), the average of analysts' implied forecasts of tax expense, divided by the number of shares in *t-1* (CSHO) and scaled by price in *t-1* (PRCC_F) (Source: Compustat and I/B/E/S)
Est ETR	Following Bratten et al. (2017), the average of analysts' implied forecasts of the firm's effective tax rate (Source: I/B/E/S)
AccTax	Following Bratten et al. (2017), the average absolute value of analysts' implied forecasts of tax expense less the implied actual tax expense, divided by the number of shares in *t-1* (CSHO), scaled by price in *t-1* (PRCC_F), multiplied by -1 (Source: Compustat and I/B/E/S)
AccETR	Following Bratten et al. (2017), the average absolute value of analysts' implied forecasts of the firm's effective tax rate less the implied actual GAAP effective tax rate, multiplied by -1. (Source: I/B/E/S)
GAAP ETR	Income taxes (TXT) divided by pre-tax income less special items (PI - SPI). The variable is bounded between 0 and 1. (Source: Compustat)
ΔMVA	Following Henry and Sansing (2018), cash taxes paid (TXPD) minus the change in a firm's tax refund receivable (TXR) and minus the statutory corporate tax rate multiplied by pretax-income (PI), scaled by the prior year's market value of assets (MVA). MVA is computed as $(AT + ((PRCC\_F \times CSHO) - SEQ))$. (Source: Compustat)
Log(Cash Taxes Paid)	Natural logarithm of the firm's cash taxes paid (TXPD). (Source: Compustat)
ISA Cash ETR	Following Balakrishnan et al. (2019), the firm's *Cash ETR* less the firm's average industry size *Cash ETR*, where *Cash ETR* is cash taxes paid (TXPD) divided by pre-tax income less special items (PI - SPI). The variable is bounded between 0 and 1. (Source: Compustat)
UTB	The reserve for unrecognized tax benefits (TXTUBEND) scaled by the prior year's total assets (AT). (Source: Compustat)
Tax Settlements	The amount of tax settlements (TXTUBSETTLE) scaled by the prior year's total assets (AT). We replace missing values with 0. (Source: Compustat)
10-K Length	Natural logarithm of the firm's 10-K number of tokens plus 1. (Source: SEC EDGAR - Form 10-K)

Variable	Definition
Fog Index	Fog index, a statistic that combines the number of words per sentence and the number of syllables per word to create a measure of readability (Li 2008) (Source: SEC EDGAR - Form 10-K)
Firm Size	Natural logarithm of the firm's total assets (AT). (Source: Compustat)
Leverage	Total debt (DLTT+ DLC) relative to total assets (AT). (Source: Compustat)
Capital Intensity	Ratio of property, plant, and equipment (PPEGT) relative to the prior year's total assets (AT). (Source: Compustat)
Intangible Intensity	Intangible assets (INTAN) relative to total assets (AT). (Source: Compustat)
R&D	R&D expenses (XRD) relative to total sales (SALE). We replace missing values with 0. (Source: Compustat)
Profitability	Earnings before interest, taxes, depreciation, and amortization relative to the prior year's total assets (AT). (Source: Compustat)
Sales Growth	Natural logarithm of the growth rate of sales (SALES) from year $t - 1$ to t. (Source: Compustat)
Book-to-Market	Total common equity (CEQ) divided by common shares outstanding (CSHO) multiplied by the stock price at the fiscal year-end (PRCCF). (Source: Compustat)
Cash	Cash and short-term equivalents (CHE) divided by the prior year's total assets (AT). (Source: Compustat)
NOL	Indicator variable that equals one if the lagged tax loss carryforward (TLCF) is positive and zero otherwise. (Source: Compustat)
Foreign Activity	Indicator variable that equals one if foreign income taxes (TXFO) are non-missing and zero otherwise. (Source: Compustat)
Equity Earnings	Indicator variable that equals one if equity method earnings (ESUB) are present and zero otherwise. (Source: Compustat)
Mezzanine Finance	Convertible debt and preferred stock (DCPSTK) scaled by total assets (AT). (Source: Compustat)
Big4	Indicator variable that equals one if the firm is audited by PwC, E&Y, KPMG, or Deloitte and zero otherwise. (Source: Compustat)
Litigation Risk	Indicator variable that equals one if pre-tax or after-tax litigation/insurance settlement (SETP or SETA) is negative and zero otherwise. (Source: Compustat)
Business Segments	Natural logarithm of the firm's number of business segments plus 1. (Source: Compustat)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11142-025-09914-3.

Acknowledgements We appreciate helpful feedback from Henry Friedman (editor), two anonymous reviewers, Harald Amberger, Elisa Casi-Eberhard, Richard Crowley, Carlo D'Augusta, Marco Greco, Martin Jacob, Marco Maria Mattei, Arthur Stenzel, Jaron Wilde, and workshop participants at Bocconi University, EDHEC Business School, ESSEC Business School, the Free University of Bolzano, IE Business School, the Norwegian School of Economics, the University of Paderborn, the University of Padova, WHU Otto Besheim School of Management, the 2021 SIDREA Conference on Financial Reporting, the 2022 Accounting Summer Camp, the 2022 Berlin-Vallendar Conference on Tax Research, the 2022 European Accounting Association Annual Congress, and the 2023 EIASM Conference on Current Research in Taxation. An earlier version of this paper was titled "Risk Disclosure and (In)-Effective Tax Planning."

Funding Open access funding provided by Università Commerciale Luigi Bocconi within the CRUI-CARE Agreement.

References

Al-Hadi, A., G. Taylor, and G. Richardson. 2022. Are corruption and corporate tax avoidance in the United States related? *Review of Accounting Studies* 27 (1): 344–389.

Arena, M. P., B. Wang, and R. Yang. 2021. Securities litigation and corporate tax avoidance. *Journal of Corporate Finance* 66:101546.

Armstrong, C. S., J. L. Blouin, A. D. Jagolinzer, and D. F. Larcker. 2015. Corporate governance, incentives, and tax avoidance. *Journal of Accounting and Economics* 60 (1): 1–17.

Armstrong, C. S., J. L. Blouin, and D. F. Larcker. 2012. The incentives for tax planning. *Journal of Accounting and Economics* 53 (1–2): 391–411.

Austin, C. R., and R. J. Wilson. 2017. An examination of reputational costs and tax avoidance: Evidence from firms with valuable consumer brands. *The Journal of the American Taxation Association* 39 (1): 67–93.

Ayers, B.C., J. Jiang, and S.K. Laplante. 2009. Taxable income as a performance measure: The effects of tax planning and earnings quality. *Contemporary Accounting Research* 26 (1): 15–54.

Ayers, B. C., C. M. Schwab, and S. Utke. 2015. Noncompliance with mandatory disclosure requirements: The magnitude and determinants of undisclosed permanently reinvested earnings. *The Accounting Review* 90 (1): 59–93.

Balakrishnan, K., J. L. Blouin, and W. R. Guay. 2019. Tax aggressiveness and corporate transparency. *The Accounting Review* 94 (1): 45–69.

Bao, Y., and A. Datta. 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science* 60 (6): 1371–1391.

Basu, S., X. Ma, and H. Briscoe-Tran. 2022. Measuring multidimensional investment opportunity sets with 10-K text. *The Accounting Review* 97 (1): 51–73.

Beardsley, E.L., M.A. Mayberry, and S.T. McGuire. 2021. Street versus GAAP: Which effective tax rate is more informative? *Contemporary Accounting Research* 38 (2): 1310–1340.

Bernard, D., E. Blankespoor, T. de Kok, and S. Toynbee. 2023. Confused readers: A modular measure of business complexity. *Working paper, University of Washington. Available at SSRN 4480309*.

Blankespoor, E., E. deHaan, and I. Marinovic. 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics* 70 (2–3): 101344.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *Journal Of Machine Learning Research* 3: 993–1022.

Bochkay, K., S.V. Brown, A.J. Leone, and J.W. Tucker. 2023. Textual analysis in accounting: What's next? *Contemporary Accounting Research* 40 (2): 765–805.

Boone, J.P., I.K. Khurana, and K. Raman. 2013. Religiosity and tax avoidance. *The Journal of the American Taxation Association* 35 (1): 53–84.

Bozanic, Z., J. L. Hoopes, J. R. Thornock, and B. M. Williams. 2017. IRS attention. *Journal of Accounting Research* 55 (1): 79–114.

Bozanic, Z., and M. Thevenot. 2015. Qualitative disclosure and changes in sell-side financial analysts' information environment. *Contemporary Accounting Research* 32 (4): 1595–1616.

Bratten, B., C. A. Gleason, S. A. Larocque, and L. F. Mills. 2017. Forecasting taxes: New evidence from analysts. *The Accounting Review* 92 (3): 1–29.

Breiman, L. 2001. *Random forests. Machine learning* 45 (1): 5–32.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone. 2017. *Classification and regression trees*. Chapman and Hall/CRC.

Brown, J. L. 2011. The spread of aggressive corporate tax reporting: A detailed examination of the corporate-owned life insurance shelter. *The Accounting Review* 86 (1): 23–57.
Brown, J.L., and K.D. Drake. 2014. Network ties among low-tax firms. *The Accounting Review* 89 (2): 483–510.
Brown, J.L., K.D. Drake, and M.A. Martin. 2016a. Compensation in the post-fin 48 period: The case of contracting on tax performance and uncertainty. *Contemporary Accounting Research* 33 (1): 121–151.
Brown, L. D., A. C. Call, M. B. Clement, and N. Y. Sharp. 2016b. The activities of buy-side analysts and the determinants of their stock recommendations. *Journal of Accounting and Economics* 62 (1): 139–156.
Brown, N.C., R.M. Crowley, and W.B. Elliott. 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research* 58 (1): 237–291.
Brown, S. V., and J. W. Tucker. 2011. Large-sample evidence on firms' year-over-year MD &A modifications. *Journal of Accounting Research* 49 (2): 309–346.
Burd, C., E. Casi-Eberhard, F. Grossetti, and P. Lisowsky. 2023. Does the story matter? Putting financial statement numbers into context using XBRL data. *Working paper, North Carolina State University. TRR Working Paper 266*.
Caglio, A., C. Imperatore, and C. Valle Ruiz. 2022. Institutional investors' tax preferences and the design of CEOs' compensation packages. *Working paper, Bocconi University. Available at SSRN 4188982*.
Campbell, J., H. Ham, Z. Lu, and K. Wood. 2023. Relevance of analysts' earnings forecasts in the era of machine learning? What we learn from estimating 3,000+ models. *Working paper, University of Georgia*.
Campbell, J. L., M. Cecchini, A. M. Cianci, A. C. Ehinger, and E. M. Werner. 2019. Tax-related mandatory risk factor disclosures, future profitability, and stock returns. *Review of Accounting Studies* 24 (1): 264–308.
Campbell, J. L., H. Chen, D. S. Dhaliwal, H.-M. Lu, and L. B. Steele. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies* 19 (1): 396–455.
Campbell, J. L., J. F. Downes, and W. C. Schwartz. 2015. Do sophisticated investors use the information provided by the fair value of cash flow hedges? *Review of Accounting Studies* 20:934–975.
Campbell, J.L., N.C. Goldman, and B. Li. 2021. Do financing constraints lead to incremental tax planning? Evidence from the Pension Protection Act of 2006. *Contemporary Accounting Research* 38 (3): 1961–1999.
Cazier, R., S. Rego, X. Tian, and R. Wilson. 2015. The impact of increased disclosure requirements and the standardization of accounting practices on earnings management through the reserve for income taxes. *Review of Accounting Studies* 20:436–469.
Cazier, R. A., J. L. McMullin, and J. S. Treu. 2021. Are lengthy and boilerplate risk factor disclosures inadequate? An examination of judicial and regulatory assessments of risk factor language. *The Accounting Review* 96 (4): 131–155.
Chang, H. S., M. Donohoe, and T. Sougiannis. 2016. Do analysts understand the economic and reporting complexities of derivatives? *Journal of Accounting and Economics* 61 (2–3): 584–604.
Chen, C.-W., B.F. Hepfer, P.J. Quinn, and R.J. Wilson. 2018. The effect of tax-motivated income shifting on information asymmetry. *Review Of Accounting Studies* 23: 958–1004.
Chen, J., K. Li, J. Zhu, and W. Chen. 2015. Warplda: A cache efficient o(1) algorithm for latent dirichlet allocation. *Working paper, Tsinghua University*. arXiv:1510.08628.
Chen, K.-P., and C.C. Chu. 2005. Internal control versus external manipulation: A model of corporate income tax evasion. *Rand Journal of Economics* 36 (1): 151–164.
Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 1, pp. 785–794.
Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan. 2025. *xgboost: Extreme Gradient Boosting*. R package version 1.7.11.1.
Chen, X., Y. H. Cho, Y. Dou, and B. Lev. 2022. Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research* 60 (2): 467–515.
Chi, S., S. X. Huang, and J. M. Sanchez. 2017. CEO inside debt incentives and corporate tax sheltering. *Journal of Accounting Research* 55 (4): 837–876.
Chychyla, R., D. Falsetta, and S. Ramnath. 2022. Strategic choice of presentation format: The case of ETR reconciliations. *The Accounting Review* 97 (1): 177–211.

Chyz, J. A. 2013. Personally tax aggressive executives and corporate tax sheltering. *Journal of Accounting and Economics* 56 (2–3): 311–328.

Chyz, J. A., and F. B. Gaertner. 2018. Can paying "too much" or "too little" tax contribute to forced CEO turnover? *The Accounting Review* 93 (1): 103–130.

Chyz, J. A., W. S. C. Leung, O. Z. Li, and O. M. Rui. 2013. Labor unions and tax aggressiveness. *Journal of Financial Economics* 108 (3): 675–698.

Crowley, R. M., and M. Wong. 2022. Understanding sentiment through context. *Working paper, Singapore Management University. Available at SSRN 4316229*.

Demeré, P. 2023. Is tax return information useful to equity investors? *Review of Accounting Studies* 28 (3): 1413–1465.

Demeré, P., M. P. Donohoe, and P. Lisowsky. 2020. The economic effects of special purpose entities on corporate tax avoidance. *Contemporary Accounting Research* 37 (3): 1562–1597.

Demeré, P., L. Y. Li, P. Lisowsky, and R. W. Snyder. 2024. Smoothing GAAP ETRs through tax accruals and the quality of financial reporting. *The Journal of the American Taxation Association, Forthcoming*.

Deng, Z. 2020. Foreign exchange risk, hedging, and tax-motivated outbound income shifting. *Journal of Accounting Research* 58 (4): 953–987.

Desai, M. A., and D. Dharmapala. 2006. Corporate tax avoidance and high-powered incentives. *Journal of Financial Economics* 79 (1): 145–179.

Dhaliwal, D.S., C.A. Gleason, and L.F. Mills. 2004. Last-chance earnings management: Using the tax expense to meet analysts' forecasts. *Contemporary Accounting Research* 21 (2): 431–459.

Dhaliwal, D.S., T.H. Goodman, P. Hoffman, and C.M. Schwab. 2022. The incidence, valuation, and management of tax-related reputational costs: Evidence from a period of protest. *The Journal of the American Taxation Association* 44 (1): 49–73.

Dhaliwal, D. S., S. Radhakrishnan, A. Tsang, and Y. G. Yang. 2012. Nonfinancial disclosure and analyst forecast accuracy: International evidence on corporate social responsibility disclosure. *The Accounting Review* 87 (3): 723–759.

Dickey, J.M. 1983. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal Of The American Statistical Association* 78 (383): 628–637.

Doellman, T., F. Huseynov, T. Nasser, and S. Sardarli. 2020. Corporate tax avoidance and mutual fund ownership. *Accounting and Business Research* 50 (6): 608–635.

Donohoe, M. P., H. Jang, and P. Lisowsky. 2022. Competitive externalities of tax cuts. *Journal of Accounting Research* 60 (1): 201–259.

Donovan, J., J. Jennings, K. Koharki, and J. Lee. 2021. Measuring credit risk using qualitative disclosure. *Review of Accounting Studies* 26 (2): 815–863.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from latent dirichlet allocation. *Journal Of Accounting And Economics* 64 (2–3): 221–245.

Dyreng, S. D., M. Hanlon, and E. L. Maydew. 2008. Long-run corporate tax avoidance. *The Accounting Review* 83 (1): 61–82.

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

Dyreng, S.D., and B.P. Lindsey. 2009. Using financial accounting data to examine the effect of foreign operations located in tax havens and other countries on U.S. multinational firms' tax rates. *Journal of Accounting Research* 47 (5): 1283–1316.

Edwards, A., C. Schwab, and T. Shevlin. 2016. Financial constraints and cash tax savings. *The Accounting Review* 91 (3): 859–881.

Efron, B. 1992. *Bootstrap methods: another look at the jackknife*. Springer.

Engelberg, J., R. D. McLean, and J. Pontiff. 2020. Analysts and anomalies. *Journal of Accounting and Economics* 69 (1): 101249.

Erickson, M., M. Hanlon, and E. L. Maydew. 2004. How much will firms pay for earnings that do not exist? Evidence of taxes paid on allegedly fraudulent earnings. *The Accounting Review* 79 (2): 387–408.

FASB. 2021. *Statement of Financial Accounting Concepts No. 8: Conceptual Framework for Financial Reporting*. Financial Accounting Standards Board.

Flagmeier, V., J. Müller, and C. Sureth-Sloane. 2023. When do firms highlight their effective tax rate? *Accounting and Business Research* 53 (1): 1–37.

Frank, M.M., and S.O. Rego. 2006. Do managers use the valuation allowance account to manage earnings around certain earnings targets? *Journal of The American Taxation Association* 28 (1): 43–65.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals Of Statistics* 29 (5): 1189–1232.

Gallemore, J., B. Gipper, and E. Maydew. 2019. Banks as tax planning intermediaries. *Journal of Accounting Research* 57 (1): 169–209.
Geertsema, P., and H. Lu. 2023. Relative valuation with machine learning. *Journal of Accounting Research* 61 (1): 329–376.
Goh, B. W., J. Lee, C. Y. Lim, and T. Shevlin. 2016. The effect of corporate tax avoidance on the cost of equity. *The Accounting Review* 91 (6): 1647–1670.
Graham, J. R., M. Hanlon, T. Shevlin, and N. Shroff. 2014. Incentives for tax planning and avoidance: Evidence from the field. *The Accounting Review* 89 (3): 991–1023.
Graham, J.R., M. Hanlon, T. Shevlin, and N. Shroff. 2017. Tax rates and corporate decision-making. *The Review Of Financial Studies* 30 (9): 3128–3175.
Graham, J. R., J. S. Raedy, and D. A. Shackelford. 2012. Research in accounting for income taxes. *Journal of Accounting and Economics* 53 (1–2): 412–434.
Green, D. H., E. Henry, S. M. Parsons, and G. A. Plesko. 2022. Incorporating financial statement information to improve forecasts of corporate taxable income. *The Accounting Review* 97 (7): 169–192.
Grieser, W. D., and C. J. Hadlock. 2019. Panel-data estimation in finance: Testable assumptions and parameter (in) consistency. *Journal of Financial and Quantitative Analysis* 54 (1): 1–29.
Griffiths, T.L., and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101 (suppl 1): 5228–5235.
Guenther, D.A., K. Peterson, J. Searcy, and B.M. Williams. 2023. How useful are tax disclosures in predicting effective tax rates? A machine learning approach. *The Accounting Review* 98 (5): 1–26.
Hanlon, M., and S. Heitzman. 2010. A review of tax research. *Journal of Accounting and Economics* 50 (2–3): 127–178.
Hanlon, M., and J. Slemrod. 2009. What does tax aggressiveness signal? Evidence from stock price reactions to news about tax shelter involvement. *Journal of Public Economics* 93 (1–2): 126–141.
Hasan, I., C.K.S. Hoi, Q. Wu, and H. Zhang. 2014. Beauty is in the eye of the beholder: The effect of corporate tax avoidance on the cost of bank loans. *Journal of Financial Economics* 113 (1): 109–130.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
He, G., H. M. Ren, and R. Taffler. 2020. The impact of corporate tax avoidance on analyst coverage and forecasts. *Review of Quantitative Finance and Accounting* 54 (2): 447–477.
Henry, E., and R. Sansing. 2018. Corporate tax avoidance: Data truncation and loss firms. *Review of Accounting Studies* 23:1042–1070.
Higgins, D., T.C. Omer, and J.D. Phillips. 2015. The influence of a firm's business strategy on its tax aggressiveness. *Contemporary Accounting Research* 32 (2): 674–702.
Hirshleifer, D., Y. Levi, B. Lourie, and S. H. Teoh. 2019. Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics* 133 (1): 83–98.
Hoberg, G., and C. Lewis. 2017. Do fraudulent firms produce abnormal disclosure? *Journal Of Corporate Finance* 43: 58–85.
Hoopes, J. L., D. Mescall, and J. A. Pittman. 2012. Do irs audits deter corporate tax avoidance? *The Accounting Review* 87 (5): 1603–1639.
Hope, O.-K. 2003. Disclosure practices, enforcement of accounting standards, and analysts' forecast accuracy: An international study. *Journal Of Accounting Research* 41 (2): 235–272.
Hope, O.-K., D. Hu, and H. Lu. 2016. The benefits of specific risk-factor disclosures. *Review of Accounting Studies* 21 (4): 1005–1045.
Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng. 2018. Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science* 64 (6): 2833–2855.
Huang, X., and J. Marques-Silva. 2024. On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*. https://doi.org/10.1016/j.ijar.2023.109112.
Hutchens, M. 2017. Can disclosure characteristics improve analyst forecast accuracy? *Working paper, University of Illinois at Urbana-Champaign. Available at SSRN 3042836.*
Impink, J., M. Paananen, and A. Renders. 2022. Regulation-induced disclosures: Evidence of information overload? *Abacus* 58 (3): 432–478.
Inger, K. K. 2014. Relative valuation of alternative methods of tax avoidance. *The Journal of the American Taxation Association* 36 (1): 27–55.
Inger, K.K., M.D. Meckfessel, M. Zhou, and W. Fan. 2018. An examination of the impact of tax avoidance on the readability of tax footnotes. *The Journal of the American Taxation Association* 40 (1): 1–29.
Isin, A. A. 2018. Tax avoidance and cost of debt: The case for loan-specific risk mitigation and public debt financing. *Journal of Corporate Finance* 49:344–378.

Jacob, M. 2022. Real effects of corporate taxation: A review. *European Accounting Review* 31 (1): 269–296.
Jennings, J. N., J. A. Lee, and E. Towery. 2020. Estimating the likelihood of future tax settlements using firm fundamentals and text disclosures. *Working paper, Washington University in St. Louis. Available at SSRN 3615882*.
Jones, S. 2017. Corporate bankruptcy prediction: A high dimensional analysis. *Review of Accounting Studies* 22: 1366–1422.
Jordan, M. I. 1998. *Learning in graphical models*, vol. 89. The MIT Press.
Khan, M., S. Srinivasan, and L. Tan. 2017. Institutional ownership and corporate tax avoidance: New evidence. *The Accounting Review* 92 (2): 101–122.
Kim, J., S. McGuire, S. Savoy, and R. Wilson. 2022. Expected economic growth and investment in corporate tax planning. *Review Of Accounting Studies* 27: 745–778.
Kim, J.-B., Y. Li, and L. Zhang. 2011. Corporate tax avoidance and stock price crash risk: Firm-level analysis. *Journal of Financial Economics* 100 (3): 639–662.
Kim, S., A. P. Schmidt, and K. Wentland. 2020. Analysts, taxes, and the information environment. *The Journal of the American Taxation Association* 42 (1): 103–131.
Klassen, K.J., and S.K. Laplante. 2012. The effect of foreign reinvestment and financial reporting incentives on cross-jurisdictional income shifting. *Contemporary Accounting Research* 29 (3): 928–955.
Koester, A., T. Shevlin, and D. Wangerin. 2017. The role of managerial ability in corporate tax avoidance. *Management Science* 63 (10): 3285–3310.
Kothari, S. P., X. Li, and J. E. Short. 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review* 84 (5): 1639–1670.
Krull, L.K. 2004. Permanently reinvested foreign earnings, taxes, and earnings management. *The Accounting Review* 79 (3): 745–767.
Kubick, T.R., G.B. Lockhart, L.F. Mills, and J.R. Robinson. 2017. IRS and corporate taxpayer effects of geographic proximity. *Journal Of Accounting And Economics* 63 (2–3): 428–453.
Kubick, T. R., D. P. Lynch, M. A. Mayberry, and T. C. Omer. 2015. Product market power and tax avoidance: Market leaders, mimicking strategies, and stock returns. *The Accounting Review* 90 (2): 675–702.
Kubick, T. R., D. P. Lynch, M. A. Mayberry, and T. C. Omer. 2016. The effects of regulatory scrutiny on tax avoidance: An examination of SEC comment letters. *The Accounting Review* 91 (6): 1751–1780.
Kuhn, M., and H. Frick. 2024. Dials: Tools for creating tuning parameter values. 2024.
Kumar, I., C. Scheidegger, S. Venkatasubramanian, and S. Friedler. 2021. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems* 34:26598–26608.
Kumar, I. E., S. Venkatasubramanian, C. Scheidegger, and S. Friedler. 2020. Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, vol. 119, pp. 5491–5500. PMLR.
Lang, M., and L. Stice-Lawrence. 2015. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics* 60 (2–3): 110–135.
Lee, Y., S. Ng, T. Shevlin, and A. Venkat. 2021. The effects of tax avoidance news on employee perceptions of managers and firms: Evidence from glassdoor.com ratings. *The Accounting Review* 96 (3): 343–372.
Lev, B., S. Li, and T. Sougiannis. 2010. The usefulness of accounting estimates for predicting cash flows and earnings. *Review of Accounting Studies* 15:779–807.
Lewis, C. M., and F. Grossetti. 2022. A statistical approach for optimal topic model identification. *Journal of Machine Learning Research* 23 (58): 1–20.
Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2–3): 221–247.
Lietz, G. M. 2013. Determinants and consequences of corporate tax avoidance. *Working paper, ISM Hamburg. Available at SSRN 2363868*.
Lisowsky, P. 2010. Seeking shelter: Empirically modeling tax shelters using financial statement information. *The Accounting Review* 85 (5): 1693–1720.
Lisowsky, P., L. Robinson, and A. Schmidt. 2013. Do publicly disclosed tax reserves tell us about privately disclosed tax shelter activity? *Journal of Accounting Research* 51 (3): 583–629.
Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66 (1): 35–65.

Loughran, T., and B. McDonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance* 69 (4): 1643–1671.
Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54 (4): 1187–1230.
Lowry, M., R. Michaely, and E. Volkova. 2020. Information revealed through the regulatory process: Interactions between the SEC and companies ahead of their IPO. *The Review Of Financial Studies* 33 (12): 5510–5554.
Luo, L., M. S. Ma, T. C. Omer, and H. Xie. 2024. Tax avoidance and firm value: Does qualitative disclosure in the tax footnote matter? *Review of Accounting Studies* 29 (3): 2927–2970.
McGuire, S. T., D. Wang, and R. J. Wilson. 2014. Dual class ownership and tax avoidance. *The Accounting Review* 89 (4): 1487–1516.
Mills, L. F. 1998. Book-tax differences and internal revenue service adjustments. *Journal of Accounting Research* 36 (2): 343–356.
Mills, L.F., S.E. Nutter, and C.M. Schwab. 2013. The effect of political sensitivity and bargaining power on taxes: Evidence from federal contractors. *The Accounting Review* 88 (3): 977–1005.
Morrow, M., and R. C. Ricketts. 2014. Financial reporting versus tax incentives and repatriation under the 2004 tax holiday. *The Journal of the American Taxation Association* 36 (1): 63–87.
Nguyen, J. H. 2021. Tax avoidance and financial statement readability. *European Accounting Review* 30 (5): 1043–1066.
Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica*. https://doi.org/10.2307/1911408.
Olsen, K. J., and J. Stekelberg. 2016. CEO narcissism and corporate tax sheltering. *The Journal of the American Taxation Association* 38 (1): 1–22.
Platikanova, P. 2017. Debt maturity and tax avoidance. *European Accounting Review* 26 (1): 97–124.
Plumlee, M.A. 2003. The effect of information complexity on analysts' use of that information. *The Accounting Review* 78 (1): 275–296.
Rajgopal, S. 2022. Why investors need better corporate tax disclosures - Part I. *Forbes*.
Rego, S.O. 2003. Tax-avoidance activities of US multinational corporations. *Contemporary Accounting Research* 20 (4): 805–833.
Rego, S.O., and R. Wilson. 2012. Equity risk incentives and corporate tax aggressiveness. *Journal of Accounting Research* 50 (3): 775–810.
Richter, S., J. K. Seidman, R. K. Sinha, and B. Stomberg. 2024. How tax executives craft income tax disclosures in response to tax-based proprietary costs. *Working paper, University of Illinois at Urbana-Champaign. Available at SSRN 4834552*.
Robinson, J. R., S. A. Sikes, and C. D. Weaver. 2010. Performance measurement of corporate tax departments. *The Accounting Review* 85 (3): 1035–1064.
Robinson, L. A., B. Stomberg, and E. M. Towery. 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.
Rogers, J. L., A. Van Buskirk, and S. L. Zechman. 2011. Disclosure tone and shareholder litigation. *The Accounting Review* 86 (6): 2155–2183.
Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–288.
Towery, E. M. 2017. Unintended consequences of linking tax return disclosures to financial reporting for income taxes: Evidence from schedule UTP. *The Accounting Review* 92 (5): 201–226.
Uddin, A., X. Tao, C.-C. Chou, and D. Yu. 2022. Are missing values important for earnings forecasts? *A machine learning perspective. Quantitative Finance* 22 (6): 1113–1132.
Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pp. 307–333.
Wang, X., A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pp. 697–702. IEEE.
Weber, D.P. 2009. Do analysts and investors fully appreciate the implications of book-tax differences for future earnings? *Contemporary Accounting Research* 26 (4): 1175–1206.
Wilde, J.H. 2017. The deterrent effect of employee whistleblowing on firms' financial misreporting and tax aggressiveness. *The Accounting Review* 92 (5): 247–280.

Wilde, J.H., and R.J. Wilson. 2018. Perspectives on corporate tax planning: Observations from the past decade. *The Journal of the American Taxation Association* 40 (2): 63–81.
Wilson, R. J. 2009. An examination of corporate tax shelter participants. *The Accounting Review* 84 (3): 969–999.
Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. MIT press.
Zimmerman, J. L. 1983. Taxes and firm size. *Journal of Accounting and Economics* 5:119–149.