What drives partisan tax policy?
The effective tax code

Elliott Ash*

Abstract

This paper contributes to recent work in political economy and public finance that focuses on how details of the tax code, rather than tax rates, are used to implement redistributive fiscal policies. I use tools from natural language processing to construct a high-dimensional representation of tax code changes from the text of 1.6 million statutes enacted by state legislatures since 1963. A data-driven approach is taken to recover the *effective tax code* – the language in tax law that has the largest impact on revenues, holding major tax rates constant. These language features provide a robust out-of-sample predictor for tax collections. I then show that the effective tax code drives partisan tax policy: relative to Republicans, Democrats use revenue-increasing language for income taxes but use revenue-decreasing language for sales taxes (consistent with a more redistributive fiscal policy) despite making no changes on average to statutory tax rates. These results are consistent with the view that due to their relative salience, changing tax rates is politically more difficult than changing the tax code.

1 Introduction

Standard models in the political economy of tax policy feature tax rates, public goods, and expenditures as the key tools for implementing a redistributive fiscal policy (Persson and Tabellini, 2002). A redistribution-oriented government can implement a progressive tax on income and redistribute the proceeds as public goods or lump-sum transfers. A model of what components of income are taxable, or how those components are legally specified, is not needed for this approach.

Recent work in public finance has shown that the legal definition of the tax base has important revenue and redistributive consequences (Kopczuk, 2005; Gordon and Kopczuk, 2014). The base involves a complex set of policy choices that affect the allocation of the tax burden. For example, income tax credits for dependent children will favor families with children. Sales tax exemptions for groceries will favor individuals who spend a relatively large proportion of their income on groceries.

An attractive setting for the empirical study of tax policy is the U.S. states. With panel data on fifty different state governments, one can analyze the political determinants of redistribution. Previous work on state politics has documented that political control of state government has an impact on tax revenues (Reed, 2006; Warren, 2009). But how those revenue changes are implemented – changes in tax rates, versus changes in the tax base – presents an open question.

The difficulty in measuring the relative importance of tax rates and the tax base is that the definition of the base must be embodied in the language of the tax code. The wording of legislation can have large impacts: Legislators must specify which people count as dependents, for example, and which items count as groceries. Because statutory language is ambiguous, tax base provisions may have multiple interpretations. Legal experts, including judges tasked with enforcing the code, often disagree on the tax consequences of these provisions (Weisbach, 1999, 2002). For the empirical researcher, this means that many provisions cannot be reliably coded as data across states. The researcher interested in testing for the revenue consequences of particular provisions across state tax codes would have to make many subjective decisions.

This paper aims to provide a data-driven approach to this problem using tools from natural language processing applied to the text of state tax legislation. These tools are used to construct a high-dimensional representation of tax law from the text of 1.6 million statutes enacted by state legislatures since 1963. A supervised topic model based on word embeddings technology shrinks the feature set while extracting

the language that is most informative about tax law. Exogenous variation in the tax law comes from diffusion of legal language within regional judicial districts. This variation is used to estimate the impact of tax law text features on revenue. This method uncovers the *effective tax code* – the set of text features in the tax code that have a measurable causal impact on revenue collections. This data-driven method provides a more objective representation of the tax code than would be possible with subjective coding of complex, potentially ambiguous, provisions. More generally, these methods open up to empirical analysis a set of previously unobserved policy instruments.

The advantage of a state-level analysis (relative to the federal government) is that one can examine how variation in political party control is related to changes in the tax law. In this paper, I measure the effect of a change in political party of state government on tax rates and the tax code. Consistent with the previous literature, I document effects of political control on tax revenue. But I find no effects of political control on average to the major tax rates. Income tax revenues increase due to Democrat control, while sales tax revenues decrease.

The new contribution is in demonstrating the role of the effective tax code in the implementation of partisan tax policy. Relative to Republican-controlled state government, Democrat-controlled governments use revenue-increasing language on income taxes. On sales taxes, they use revenue-decreasing language. Because income taxes are relatively progressive, and sales taxes are relatively regressive, this pattern is consistent with more redistributive fiscal policy choices by Democrats. The results suggest that in U.S. state governments, political parties implement fiscal policy primarily through the legal definition of the tax base, rather than through changes to the major tax rate structures.

The data include state government financial accounts linked to the text of state tax laws for a 48-year time period (1963 through 2010). The panels include information for individual income taxes and sales taxes. These two state taxes together account for about 70 percent of state government tax collections and about 4 percent of U.S. GDP (as of 2014).

The first challenge is to represent the features of the tax base as analyzable data. For example, the New York state tax agency web site lists eighty major exemptions to the sales tax, and that excludes many relatively minor exemptions, deductions, and credits for the sales tax in other tax code sections. Trying to measure the effects of each of these individual rules on sales tax revenue in New York would be a difficult task – and this is just one tax source, one state, and at one point in time. Analyzing all fifty

3

states at once requires new techniques from natural language processing (Gentzkow and Shapiro, 2010; Gentzkow et al., 2015; Jelveh et al., 2015; Gentzkow et al., 2017), which are used to represent the tax base using measurable features of tax legislation. Section 5 describes the application of these methods to represent tax law changes as a frequency distribution over a vocabulary of 25,000 phrases. The goal is not to estimate precisely the revenue impact of any particular phrase, but rather to construct a ranking of the phrases that can be used to explore how political parties differ in the language they insert into the tax code.

Tax code language is chosen endogenously in response to variables that are correlated with tax revenue, so standard panel data methods comparing changes in revenue to changes in tax laws may render inconsistent estimates. Determining which phrases have a causal effect on tax revenues may require exogenous variation in these phrases. The solution to this problem is an instrumental-variables setup related to Bartik's (1991) identification of labor demand shocks. Instruments for phrase frequencies in an individual state are constructed from the lagged phrase frequencies in states in the same federal judicial circuit. This approach is motivated by previous work demonstrating a shared legal community within circuits in which legal ideas and legal language diffuse through cultural channels that are orthogonal to the economic variables that otherwise underlie tax revenues (Carp, 1972; Bird and Smythe, 2008; Hinkle, 2015). These lagged features constitute a high-dimensional set of sparse instruments, requiring the application of recently introduced methods in sufficient dimension reduction (Belloni et al., 2012; Lin et al., 2015; Chernozhukov et al., 2017).

The 2SLS regressions provide estimates of the predicted impact of phrases on tax revenue. The most predictive phrases are then aggregated in a partial least squares regression model, which can predict tax-revenue changes out of sample. The model provides good predictions with both the actual phase frequencies and the instrumented phrase frequencies, demonstrating that the textual features of legislation are predictive of and causally related to tax revenue. Analysis of the set of revenue-relevant phrases suggests the importance of language defining tax expenditures: deductions, exemptions, and credits.

The next step is to investigate the role of the tax code in the political economy of state fiscal policy. The empirical strategy is to use panel data regressions estimating the effect of Democrat control of state government, controlling for governor votes and legislative seat shares as forcing variables. When new political parties take control of state government, there are changes in revenues collected (Reed, 2006; Warren, 2009),

4

but no average change in major tax rates on average.

The main results section looks at the effect of political control on the predicted revenue impact of the effective tax code. For income taxes, Democrats choose revenue-increasing language. For sales taxes, Democrats choose revenue-decreasing language. Moving from full Republican control of government to full Democrat control of government is associated with tax code changes that are predicted to raise an additional \$2 billion of annual income tax revenue in the average state, with a corresponding decrease of \$1.7 billion in annual sales tax revenue.

Income taxes are relatively progressive, while sales taxes are regressive. The use of revenue-increasing language by Democrats on progressive taxes but revenue-decreasing language on regressive taxes is consistent with Democrats implementing a more redistributive fiscal policy through the tax code. Tax code provisions defining the base – rather than the tax rate – are the key policy tool in the political economy of fiscal policy in the U.S. states. This is consistent with the view that major tax rates are politically more difficult to change than the tax code, perhaps because rate changes would be more salient for voters (Finkelstein, 2009; Chetty et al., 2009; Cabral and Hoxby, 2012).

These results are relevant to a broad literature in political economy, reviewed in Section 2. Thereafter Section 3 presents a model to guide analysis of the data. Section 4 describes the tax data, while Section 5 details the legislative text data and methods for text processing. Section 6 provides methods and results for recovering the effective tax code using the Bartik language instruments. Section 7 uses changes in political control to estimate the effect of political control on tax policy. Section 8 relates the phrase effects on revenue to the political effects on phrases to analyze the role of the tax code in redistributive fiscal policy. Section 9 concludes.

2   Related Literature

The standard models in public finance assume that tax collections are a function of rates and audit probabilities (Mirrlees, 1971; Atkinson and Stiglitz, 1976; Feldstein, 1999; Chetty, 2009). In that case there is no scope for legal avoidance or gaming, and a deterrence model like Allingham and Sandmo (1972) or Logue (2007) will suffice to explain the interaction between tax agency and taxpayer. Good empirical evidence that increased audit rates reduce evasion include Kleven et al. (2011) and Pomeranz (2011).[1]

---

[1]In a Minnesota experiment, Slemrod et al. (2001) show that high-income individuals actually report less income when threatened with a high probability of audit. This low-ball report can be

In the standard models, tax legislation is important because it encodes policies that have socioeconomic impacts, but the wording of those statutes doesn't have independent interest because the policies are well-defined. On the other hand, there is a competing view among tax law scholars that the tax code is not a complete description of policy: There is ambiguity and indeterminacy in the language that makes a complete formal description impossible.[2] Graetz (1995), for example, notes that despite the use of accounting methods to evaluate tax reforms, there are still "massive empirical uncertainties" precluding good predictions about the revenue consequences.

More recent work has recognized that this simple model of the tax system is too limited (Andreoni et al., 1998; Slemrod and Yitzhaki, 2002). In reality, the tax code is an incomplete set of written rules, and taxpayers face administrative and legal uncertainty in their dealings with the tax authority.[3] Honest mistakes do occur, so harsh rule-based penalties are often inefficient. But discretionary standards that require adjudication are more easily gamed.[4]

The importance of interpretation and language in the operation of tax law rules is well-known in legal scholarship on tax law. Livingston (1995) and Heen (1996) discuss the importance of text, as well as the limits of plain-meaning textual analysis, in tax law. In tax law especially, judges are encouraged to interpret the intentions of legislators and not to interpret the text literally. Shaviro (2004) discusses the dual nature of legal language in tax and fiscal policy – both for furthering political goals and for describing policy. This results in indeterminate and confusing language.

Efforts in economics to extend the standard model demonstrate the pros and cons of more complex tax rules. Kopczuk (2001) uses a model of heterogeneous avoidance

---

understood as an introductory offer in a bargaining exchange between taxpayer and tax agency, on the assumption that legal ambiguity about liability creates scope for allocating a surplus. Cai and Liu (2009) report that tax avoidance among Chinese firms is higher in more competitive industries.

[2]"Between these extremes was a continuous range of transactions, and the policymaker had to decide which were taxable and which were not. This type of problem is quite general in the tax law The tax law distinguishes between debt and equity, selling and holding,an independent contractor and employees. There are hundreds of these types of distinctions" (Weisbach, 1999). Vasconcellos (2007) discusses the problems judges often face of uncertainty in tax law, and how they have to appeal to policy interests or fairness.

[3]These points are consistent with Givati's (2009) observation that tax litigation filings and IRS internal tax appeals are persistently high; if tax law was predictable, taxpayers would not invest in these costly challenges.

[4]Likhovski (2004) examines the history of tax-shelter adjudication beginning with Learned Hand's *Gregory v. Helving*. Solan and Dean (2007) identify the importance of the rule of lenity, a statutory-construction heuristic normally associated with criminal cases which advises strict construal of penal provisions against the government. Because conservative judges construe tax provisions this way, corporations can avoid taxes by structuring tax shelters that are arguably within the text of the statute but are unrelated to the policy interest motivating the provision.

ability among taxpayers to show that avoidance can be optimal if mainly performed by low earners, or if administrative costs are sufficiently high. Kleven and Kopczuk (2011) show that increased complexity in eligibility requirements for social benefits can reduce takeup, but that optimal programs must have complex eligibility rules to prevent false award grants.[5] A well-known example of complex tax targeting is the set of multiple partially overlapping definitions of child in the federal tax code, resulting in uncertainty for taxpayers about eligibility for credits (Holtzblatt and McCubbin, 2003).[6]

Other work has analyzed the political incentives for complex tax legislation. Surrey (1957) provides an early anecdotal account of the role of lobbyists in writing special tax provisions, while Graetz (2007) provides a more recent account to the same effect. Holcombe (1998) proposes that complex tax rules facilitate inefficient rent-seeking by giving legislators numerous hidden opportunities to give interest groups special tax treatment. A more innocuous view is that policymakers exploit the complexity of legislation to reduce the perceived tax burden (Krishna and Slemrod, 2003). Hettich and Winer (2005) argue that complex tax structures emerge as a byproduct of electoral competition; political parties attempt to propose and implement policies that discriminate as carefully as possible among heterogeneous voters, a process held in check only by administrative costs. [7]

An important strand of this literature has focused on the definition of the tax base: The set of transactions or components of income that are included as targets of tax collections. In Weisbach (2002), the tax base is difficult to define and can only be measured by indirect proxy. Tax shelters arise from efforts to exploit the limitations of these proxies. Kopczuk (2005) examines the relation between the tax base and the income elasticity with respect to taxes, showing that the direct effect of tax rates on

---

[5]In practice, eligibility provisions can have undesirable consequences. In analogous work on the student financial aid system, Dynarski and Scott-Clayton (2006) show that a radically simplified process could reproduce the same distribution of aid with far lower administrative costs and less invasive collection of private information.

[6]Paul (1997) shows that the number of tax law reporter volumes published in a state is correlated with state income tax revenue, suggesting some relationship between revenue and complexity. Slemrod (2005) measures tax complexity by the number of lines in tax forms and the number of pages in tax instruction booklets. He reports small correlations of higher tax complexity with older income tax systems, higher legislator salaries, lower voter turnout, higher average tax rates, and higher education levels. Katz and Bommarito II (2014) provide measurements of the complexity of the titles of the U.S. Code using measures constructed from the text and its citations. Bommarito et al. (2011) provide a descriptive survey of the population of U.S. Tax Court decisions.

[7]Yet another idea is that the drafters of tax laws have an incentive to make those laws more complex so they can earn rents after they leave government explaining the laws to clients (Weisbach, 2002). Schizer (2005) observes that private tax lawyers outmatch their government counterparts in sheer numbers, access to information, and sheer expertise.

taxable income is zero, but that there are large effects when deductions are available. This shows that previous models examining income elasticity left out an important institutional component: the tax base. Follow-up work by Gordon and Kopczuk (2014) shows that the choice of the tax base matters for the incidence of the tax burden.

Another related literature examines tax expenditures – deductions and exemptions to taxes that are designed to implement social policies (Howard, 1999). Well-known examples are the deduction for property taxes and mortgage interest, and the exclusion of imputed rental income, which favor homeowners (Poterba and Sinai, 2008). According to Slemrod (2004), revenue losses due to corporate income tax shelters are growing and account for at least half of the corporate tax gap.[8] Desai (2005) describes how the legal distinction between financial reporting of corporate income (for stock value) and tax reporting of income (for tax liabilities) has led to a large gap between the two and under-collection of corporate income taxes.[9]Zucman et al. (2015) estimates that a full 8 percent of the world's wealth is held in tax havens. On the positive side, Chetty and Hendren (2013) show that higher tax expenditures at the state and local level are related to better socioeconomic mobility across generations. Methodologically, an active issue in public finance is how to measure tax expenditures (Burman and Christopher Geissler, 2008); the text-based methods developed in this paper may be helpful in this area.

While there is less work on the tax base at the state level, Shaviro (1992) notes how every state has different definitions for taxable income. This is part of a large literature examining state tax systems. For example, Rork (2003) finds that states tend to follow the rate changes in neighboring states for excise taxes, but but not for personal income taxes or general sales taxes. Chernick (2005) shows that deductibility of state and local taxes is an important factor increasing progressivity.

The most relevant segment of this literature is that examining the effect of political party control on state fiscal policy. Besley and Case (2003) provide a review of this literature and present some evidence that Democrat control of the lower legislative chamber (but not upper chamber) is associated with higher total taxes. Reed (2006) and Warren (2009) use data from state legislatures from 1960 through 2000 and show

---

[8]The IRS estimates that the federal tax gap, based on audits, is 17%. Alm and Borders (2014) review the small set of papers and reports on state-level tax gaps. They find tax gaps similar to the federal level, ranging from 10% in Idaho to 20% in Montana.

[9]See also GAO (2003) and Plesko (2007). Ordower (2010) reviews the history of tax avoidance and the transformation of corporate tax departments from compliance centers to profit centers. This is an old issue; Griswold (1944) blamed the low tax collections in the 1940s on "uncertainty, confusion, discrimination, and inconsistency" in tax rules.

that Democrat control of both legislatures is associated with higher tax collections, but they do not look at rates nor attempt to break things out by revenue source. Leigh (2008) analyzes the effect of governor control in an RD setup using data for 1941 through 2002. He finds that the party of the governor has no effect on rates or collections for personal income or corporate income.[10] I couldn't find any papers on political control and sales taxes.

The literature in behavioral public finance on tax salience provides evidence relevant to the government's tax policy choices. Chetty et al. (2009) show that consumer demand reacts less strongly to sales taxes that are excluded from the posted purchase price. Goldin and Homonoff (2013) show that low-income individuals respond just as strongly to less salient cigarette taxes. Finkelstein (2009) shows that toll agencies increase tax rates significantly in response to the implementation of automated toll collections that are less salient to the taxpayer. Finally, the survey data reported in Cabral and Hoxby (2012) suggest that the reason homeowners disfavor property taxes is that they pay a salient lump sum once a year, rather than having the payments withheld (as is the case in payroll taxes for example). Other works in this area include Gamage and Shanske (2011) and Goldin (2015).

3 Political economy of tax policy

This section presents a model of the political economy of tax policy. The government can affect tax revenue through the tax rate, tax code, and unobserved policies. The tax code affects revenues through changing the tax base, broadly defined. The goal of the model is to isolate sources of variation in the tax code and tax revenues, in order to clarify the role of the tax code in setting fiscal policy.

3.1 Tax policy

A state government is setting policy for an income stream $Y > 0$, say personal income. Tax policy has three elements. The first is the tax rate τ, where I assume a linear marginal rate. The second is the written tax code, modeled as a vector of text features $\mathbf{x} \in \mathbb{R}^p$, where $p > 0$ is the number of text features in the vocabulary. The third element is other (unobserved) policy measures that affect tax collections, denoted by $\mathbf{u} \in \mathbb{R}^o$,

---

[10]See also Besley and Case (1995), who find that Democrat governors increase sales taxes, income taxes, and corporate taxes when they face a binding term limit. Nelson (2000) analyzes how rates relate to electoral competitiveness.

where o is the dimensionality of the unobserved policy space. This includes all policies besides the rate and the written tax code, including for example the appointment of a lax tax regulator.

Therefore tax policy is a vector $(\tau, \mathbf{x}, \mathbf{u})$. Total government revenue $G(\cdot)$ is determined by

$$G(\tau, \mathbf{x}, \mathbf{u}) = \tau B(\mathbf{x}, \mathbf{u}) Y(\tau, \mathbf{x}, \mathbf{u}),$$

where $B(\mathbf{x}, \mathbf{u}) \in (0, 1]$ is the tax base (the proportion of income that is taxable). We take "tax base" to be broadly defined, as the aggregate result of all tax policies besides the tax rate.

Define $g = \log \frac{G}{Y}$ as the government revenue as a share of income, known in the previous literature as "tax burden" (Reed, 2006). Let $b = \log B$. Then government revenue g is given by

$$g(\tau, \mathbf{x}, \mathbf{u}) = \log \tau + b(\mathbf{x}, \mathbf{u}).$$

The goal of the analysis is to understand the effect of changing text feature i on government revenue through its effect on the tax base. Holding rates and other policies fixed, the effect on log revenue of changing text feature i is

$$\frac{\partial g}{\partial x_i} = \frac{\partial b}{\partial x_i}.$$

The goal of the empirical analysis to provide estimates for this quantity. We want to identify the set of tax code features for which $\frac{\partial g}{\partial x_i} > 0$ or $\frac{\partial g}{\partial x_i} < 0$. This set of features is the effective tax code.

Extracting these features is a challenge empirically due to the presence of the unobserved policies. Assuming a linear specification for $b(\cdot)$ with data indexed by state s and year t gives:

$$g_{st} = \log(\tau_{st}) + \mathbf{x}'_{st}\beta + \mathbf{u}'_{st}\pi + \epsilon_{st}. \tag{1}$$

The basic empirical goal is to identify the set of tax code features i for which

$$\beta_i \neq 0.$$

Each coefficient gives the average effect of increasing tax code feature i on the tax base holding other policies constant.

Cross-sectional OLS could be used to estimate (1) while excluding $\mathbf{u}_{st}$. OLS would procure consistent estimates for β under the assumption that $\mathbf{x}$ is uncorrelated with

the unobserved policies $\mathbf{u}$. However, states may have different unobserved policies that are correlated with both the tax code and revenue. Cross-sectional estimates of β are therefore likely inconsistent.

Panel data improve the situation through fixed effects estimation. If state-level changes in $\mathbf{x}$ are uncorrelated with state-level changes in $\mathbf{u}$, including state fixed effects for state and year panel OLS will procure consistent estimates for β. However, if the changes are correlated, then the OLS estimates would still be biased. Again, this type of correlation is likely. The changes in $\mathbf{x}$ are likely correlated with changes in $\mathbf{u}$ because tax code reforms are chosen jointly and endogenously with other non-written policy reforms. If there is a change in the ruling political party in the state, for example, the new leaders will change the statutes $\mathbf{x}$ as well as other non-legislative policies $\mathbf{u}$. Therefore looking at the average effect of the change in text over time would procure biased estimates.

To estimate β, one needs variation in $\mathbf{x}$ that is uncorrelated with changes in $\mathbf{u}$. Obtaining this variation through instrumental variables is the goal of the empirical strategy described in Section 6.

3.2   Tax Politics

This section discusses a change in political power. In a standard model of ideological political parties without commitment, a new party will come in and change tax policy in line with their ideological preferences. In the case of U.S. politics, for example, one would expect Democrats to increase overall tax collections (Reed, 2006). They could do so through changes to the tax rate τ, as emphasized in standard political economy models, or through the base $b(\mathbf{x}, \mathbf{u})$ by changing the tax code $\mathbf{x}$. It is an open empirical question whether the tax rate or the tax code is the more important component of state fiscal policy.

Consider a model with two ideological political parties, Democrat and Republican. Let $D = 1$ for Democrat control and $D = 0$ for Republican control. The policy components can be understood as functions of the ruling party: $\tau(D)$, $\mathbf{x}(D)$, and $\mathbf{u}(D)$. The empirical work is designed to understand better the relative importance of these components in how political parties implement fiscal policy.

The effect of Democratic control on revenue can be decomposed as

$$\frac{\partial g}{\partial D} \;=\; \frac{\partial \log \tau}{\partial D} + \sum_{i=1}^{p} \frac{\partial g}{\partial x_i}\frac{\partial x_i}{\partial D} + \sum_{j=1}^{o} \frac{\partial g}{\partial u_j}\frac{\partial u_j}{\partial D}$$

$$\rho_g = \rho_\tau + \sum_{i=1}^{p} \beta_i \delta_i + U$$

where I have defined $\rho_g = \frac{\partial g}{\partial D}$, $\rho_\tau = \frac{\partial \log \tau}{\partial D}$, $\beta_i = \frac{\partial g}{\partial x_i}$, $\delta_i = \frac{\partial x_i}{\partial D}$, and $U = \sum_{j=1}^{o} \frac{\partial g}{\partial u_j} \frac{\partial u_j}{\partial D}$. The goal of this paper is to provide evidence on these quantities. Appendix A.4 uses the coefficients estimated in the empirical section to compute this decomposition and in particular measure U.

I observe g, τ, $\mathbf{x}$, and D. I do not observe $\mathbf{u}$. I have panel variation in D, as described in Section 7. The effect of Democratic control on revenue, ρ_g, and on the tax rate, ρ_τ, can be obtained from estimating

$$g_{st} = \rho_g D_{st} + \epsilon_{st}$$

$$\log \tau_{st} = \rho_\tau D_{st} + \epsilon_{st}$$

Although $\mathbf{u}$ is unobserved, it is uncorrelated with treatment under the identification assumptions described below. Therefore these quantities can be estimated consistently.

Similarly, one can estimate the average effect of Democratic control on each text feature i, δ_i, by estimating

$$x_{st}^{i} = \delta_i D_{st} + \epsilon_{st}^{i}, \forall i.$$

Again, with variation over time in D_{st}, δ_i is consistently estimated in spite of $\mathbf{u}$ being omitted from the regression. These estimates identify the set of tax code features for which $\frac{\partial x_i}{\partial D} > 0$ or $\frac{\partial x_i}{\partial D} < 0$. Then one can compare these features to those in the effective tax code – those that have a causal effect on revenue ($\frac{\partial g}{\partial x_i} \neq 0$). This will provide insight into whether and how political parties use the tax code (rather than tax rates) to implement fiscal policy.

4 Data on tax policy and state politics

This section takes account of the data sources for tax revenues and political control of state government. Subsection 4.1 accounts for the tax policy data. Subsection 4.2 accounts for the data on state politics. These data are used to analyze the role of the tax code in implementing redistributive policies.

4.1 Tax policy data

There are three sources of tax data by state: actual tax revenues, statutory tax rates, and the value of targeted income flows. The data consists of a 48-year panel (1963-2010) for all fifty states and for two taxes: personal income tax and sales tax.[11] This section discusses the sources for this data.

The data on taxes collected by state governments comes from the State Government Finances census. This data have been used in many previous papers analyzing the public finances of state government (e.g. Serrato and Zidar, 2014; Fajgelbaum et al., 2015). The census has separate categories for the taxes; here personal income tax and general sales and gross receipts taxes are used. The other two major sources of state tax revenue are corporate income taxes and excise taxes (selective sales tax), which are much smaller pieces of revenue but pose interesting topics for future work. Few state governments collect significant revenue from property taxes, which primarily fund local government.

The state tax rate data are obtained from the World Tax Database and Tax Foundation. The data include information on rates and brackets. My regressions condition on the rate structure non-parametrically by including fixed effects for sets of years where the revenue source had the same rates and brackets, excluding automatic bracket changes due to inflation. This is preferable due to non-linearity in the tax rate structure.

The data on the value of the income flows are constructed from Bureau of Economic Analysis (BEA) data. Personal income tax is the most straightforward; the BEA provides data on total personal income in each state. The income flow for sales tax is measured as sectoral GDP for retail trade (SIC 44-45).[12]

The tax data is defined for income source r (personal income or sales), state s (all fifty states), and year t (every odd-numbered year between 1963 and 2010). The main outcome measure for the regressions below is the tax burden, used in previous work on state public finance (Chernick, 2005; Reed, 2006; Leigh, 2008). The tax burden is the revenue collected divided by the value of the income flow. Define g_{st}^r, the log tax burden for source r in state s at time t, after being residualized on the source-state-rate fixed

---

[11]I have corresponding data sets for corporate income tax, estate/inheritance tax, and excise taxes (alcohol, cigarettes, and fuel). These taxes are not the focus of this paper because the political and redistributive character of these taxes is less clear. The tax code text works to predict revenues for corporate taxes and estate taxes. These data and analysis are available upon request.

[12]For robustness I used two alternative specifications for income flows. For income tax, I used federal personal income tax collections; if the rate is staying the same, the proportion of state tax collections to federal tax collections should be constant unless there are changes in the state law on the tax base. For sales tax, the alternative specification is total state GDP.

Table 1: Summary Statistics on Tax Data

Base	Variable	Mean	Median	Std. Dev.
Personal Income	Income Value ($B)	145.53	88.60	178.69
	Tax Rate	0.05	0.06	0.04
	Revenue ($B)	6.3	2.9	10.1
Sales	Income Value ($B)	173.58	19.89	535.96
	Tax Rate	0.04	0.04	0.02
	Revenue ($B)	6.68	4.02	8.26

Observation is a state-year. Dollar amounts deflated to 2007 dollars.

effects and source-year fixed effects.

Table 1 reports summary statistics on tax variables in the sample. Each of the tax bases is responsible for large amounts of revenue for state governments. As noted in Fajgelbaum et al. (2015), in recent years these state taxes have accounted for almost four percent of U.S. GDP.

4.2   State Politics Data

This section accounts for the data on state politics. The empirical goal is to determine how the revenue impacts of the effective tax code relate to the preference of the two political parties to use that language. This data has been used in many previous papers analyzing the politics of state fiscal policy (e.g. Besley and Case, 2003; Reed, 2006; Leigh, 2008).

The data include party control for both houses of the state legislatures as well as the governorship, for the years 1963 through 2010. More specifically, it has the number of Democrat and Republican seats in each legislature, and the number of Democrat and Republican votes cast in the previous governor election. These measures allow me to measure the effects of party control on policy and on legislation using panel data.

Table 2 shows summary statistics for the political variables in the dataset. Democrats had a small advantage in both legislatures and governorships during this time period. There were many changes in control, however. There was some change in the partisan makeup of state governments, whether in the legislature or governorship, in 72.8% of state-bienniums. This is the variation used in the political analysis.

Table 2: Summary Statistics on State Politics Data

Variable	Mean	Std. Dev
Democrat Governor	.5875	.4923
Democrat Lower Chamber	.6627	.4728
Democrat Upper Chamber	.6307	.4826
Previous Democrat Governor Vote Margin (%)	7.216	23.943
Lower Chamber Democrat Margin (%)	11.106	19.98
Upper Chamber Democrat Margin (%)	11.406	20.99
Tied Parties in Lower House	.0320	.1761
Tied Parties in Upper House	.0459	.2094
Log Financial Administration Expenditures	10.20	1.265

Summary statistics on state political variables.

5 Tax Legislation Data

This section describes the approach for extracting and constructing statistical representations of tax legislation. Text is becoming an important data source for empirical work in economics and political science (Gentzkow and Shapiro, 2010; Quinn et al., 2010; Jensen et al., 2012; Hansen et al., 2014; Gentzkow et al., 2015; Ash et al., 2015). This paper builds on this previous work.

Subsection 5.1 describes the source and scope of the raw legislation text. Subsection 5.2 describes the methods for tokenizing the text for analysis. Subsection 5.3 discusses how to extract tax legislation and represent it in the regression analysis.

5.1 Raw Text Data

The data on legislation consists of the full text of U.S. state session laws through 2010. The data go back to inception for most states. The "session laws" consist of the collection of statutes enacted by a legislature during a legislative session – published every year or every two years. All of the data are constructed biennially to account for this issue. The sample is all fifty states, and the 24 bienniums starting in 1963 and ending in 2010.

There is a large literature in political science examining the process of drafting and enacting legislation (Tollison, 1988; Jansa et al., 2015). State legislators can draft their own statutes, and most of them are trained to do so from attorney experience. They

also delegate the task of drafting legislation to aides. Given the difficulty of crafting bills from scratch, legislators often borrow language from other legislatures or from interest groups. For example, Hertel-Fernandez and Kashin (2015) use text analysis to measure the influence of the conservative lobbying group ALEC on state legislatures. There are also non-partisan professional organizations such as the National Council of State Legislators, and the American Law Institute, which provide model legislation. These organizations provide information about which states have adopted particular provisions. Legislators pay attention to what other states are doing to make their state appear more competitive (Berry and Baybeck, 2005).

Legislation is the ideal source of legal text for examining the legal underpinnings of tax policy. Unlike common-law subjects like criminal law and tort law, tax does not have a substantial judge-made component. Shaviro (1990) recounts the cyclical back-and-forth in tax legislation, where the base is narrowed and broadened over time.

There are some important caveats for interpreting this data. These statutes may amend or repeal previous statutory provisions, or create new provisions. These documents give the "flow," rather than the "stock," of legislation. Sometimes the laws include bills that failed or were vetoed. A team of research assistants reviewed samples and found that these practices do not change significantly within state over the time period.

Figure 1 shows an example page of a scanned statute, with the corresponding OCR. As can be seen, the OCR is quite high-quality. The scans for the period 1963-2010 are mostly high-quality. Inaccuracies from OCR are assumed to introduce random noise.

5.2 Processing Text Features

The first step is to merge and process all of this raw text. A script serves to append pages, remove headers, footers, tables of contents, indexes, and other non-statute material. Then it segments the text into individual bills, acts, and resolutions using text markers for the start of new statutes. These include indicators for new Chapters, Articles, or Titles, such as a line with "CHAPTER" followed by a Roman numeral. Some states have their own standard indicators, such as "P.A" followed by a number to reflect a new "Public Act." The script also uses common text for the beginning of a statute preamble (e.g., "An act to...") and for enacting clauses (e.g., "Be it enacted that..."). Research assistants checked samples of the statute segmenter for each state-year to make sure it worked well. This results in 1.56 million statutes for the years 1963 through 2010.

Figure 1: Scanned Session Laws and Resulting OCR

Scanned image and resulting OCR text for an example statute in the text data. This example is from the Texas Legislature for the 1889 session.

The next step is to process the text for analysis. Because the tax code is such a complex object, it is necessary to break down most of the grammatical content of language and represent it as a frequency distribution over phrases. As there are improvements in storage and computer processing power, more refined representations of language may be useful in future research.[13]

The basic methods on tokenizing text and representing documents as frequency distributions over tokens has become relatively standardized in the literature on political text analysis (Gentzkow and Shapiro, 2010; Quinn et al., 2010; Jensen et al., 2012; Gentzkow et al., 2014; Ash et al., 2015; Gentzkow et al., 2015; Jelveh et al., 2015; Gentzkow et al., 2017). A script removes upper-case, splits text into sentences, and removes punctuation. It then splits sentences into words and stems word endings using the Snowball stemmer (Porter, 2001). This stemmer is less aggressive than the better-known Porter stemmer. For example, "corporate" and "corporation" would both become "corpor." The Porter stemmer would reduce both words to "corp," which would confuse these corporation-related terms with unrelated terms like "corpus."

Most previous social science papers using text analysis represent documents as frequency distributions over stemmed words or n-grams. The disadvantage with a "bag of words" approach is that important information about word order is left out. The segments "corporate tax on sales" and "sales tax on corporations" are treated as equivalent under a bag-of-stemmed-words representation, even though they clearly concern taxes on different bases. The disadvantage of a "bag of n-grams" approach is that some phrases are counted independently even when they are clearly subordinate to a longer noun phrase. For example, the segments "corporate income tax" and "personal income tax" would both include "income tax" and "tax" as independent grams, even though the full three-word segments should be represented as singular concepts.

This paper improves on these approaches by parsing grammatical content of sentences and representing documents as frequency distributions over informative noun phrases and verb phrases. For example, "personal income tax" becomes "person_incom_tax." To do this, the script first tags each token by part of speech (nouns, verbs, adjectives, etc.) using the algorithm described in Collins (2002). Then it links up phrases based on the part-of-speech patterns, using a set of tag patterns based on Denny et al. (2015) but significantly extended for the purposes of legal language.[14] I consulted legal con-

---

[13]For example, Levy and Goldberg (2014) use grammatically parsed sentences rather than word order to train Word2vec embeddings.

[14]These include AN, NN, VN, VV, NV, VP, NNN, AAN, ANN, NAN, NPN, VAN, VNN, AVN, VVN, VPN,ANV,NVV,VDN, VVV, NNV, VVP,VAV,VVN, NCN,VCV, ACA, PAN, NCVN, ANNN,

cept dictionaries to develop the list. For example, "beyond a reasonable doubt" is preposition-determinant-adjective-noun (PDAN).

To be tokenized, phrases have to co-occur together frequently relative to how often they occur apart.[15] As an example, the sentence "`Eligible individuals must pay personal income tax on foreign business earnings`" becomes "`elig_individu must_pay person_income_tax foreign_busi_earn`".

Once the distribution of phrases is computed, infrequent phrases are excluded. Words and phrases are included if they occur in at least (roughly) 500 legislative sessions, or five states per year on average. This results in a baseline vocabulary of 25,217 tokens.

5.3 Extracting Tax Code Text Features

The next step is to construct measures of phrase frequencies for each of the three tax sources: corporate income, personal income, and sales. The approach is to weight the statutes by their similarity to these sources using Word2Vec, a natural language tool for representing words as vectors introduced in Mikolov et al. (2013). This section describes this procedure.

There is no straight-forward way to identify the tax statutes for each source. Some statutes can have an impact on the tax sources without mentioning them explicitly, while other statutes may mention the taxes but have little relation to them. This means that searching for particular keywords would result in both false positives and false negatives. With such a large database of statutes (1.56 million), meanwhile, manual classification is also infeasible.

The approach is to use a supervised topic classifier to weight statutes by their similarity to tax sources. I use Word2Vec, a popular word embeddings model which provides an off-the-shelf technique for mapping the relations between words and phrases (Mikolov et al., 2013). This tool has proven performance on web search, language translation, and speech recognition. It can be trained relatively quickly on a large corpus, and thereafter can quickly compute similarity statistics between words and documents.

NNNN, NPNN, AANN, ANNN, ANPN, NNPN, NPAN, ACAN, NCNN, NNCN, ANCN, NCAN, PDAN, PNPN, VDNN, VDAN, VVDN for Adjective, Noun, Verb, Preposition, Determinant. Verb particles are coded as "V" to ensure verb phrases such as "go along" are connected.

[15]They have to meet a point-wise mutual information threshold (Church and Hanks, 1990). This is given by $\Pr(w_1, w_2)/(\Pr(w_1)\Pr(w_2))$: the probability that the words co-occur, divided by the product of the probability (frequency) that the words occur individually.

Table 3: Most Similar Phrases to Revenue Source Labels

Personal Income Tax		
personal income tax	for that taxable year	corporate income
corporate tax	state income tax	individual income tax
income tax	taxpayer	net income tax
income tax return	individual taxpayer	individual income tax return

Sales Tax		
sales tax	local sales	use tax revenue
use tax	state sales	additional sales
sales and use tax	county sales	amount of sales
local sales tax	sales or use tax	sales tax revenue

The model is described in detail in the appendix. The important point is that Word2Vec provides a function for mapping phrases to vectors in $[-1, 1]^{300}$ using information from surrounding phrases. For a given word, Word2Vec looks at the sequence of nearby words and learns which other words/phrases in the vocabulary would fit into the same context. It is best-known for recognizing analogies. After being trained on the state session laws corpus, for example, the model knows that

$$vec["\text{corporate income tax}"] - vec["\text{corporation}"] + vec["\text{person}"]$$

$$\approx vec["\text{personal income tax}"].$$

While Word2Vec is not the only solution to the problem of identifying tax legislation, it does provide a quick and effective solution that provides intuitive rankings and can be used feasibly on such a large corpus. The tool provides relations between similar phrases that can be used to isolate tax code changes and better interpret results.

Classifying the statutes starts with text labels for the revenue sources, indexed by $r \in \{\texttt{person\_incom\_tax}, \texttt{sale\_tax}\}$. Represent by $\vec{r}$ the word vector for income label r. Table 3 gives examples of the types of phrases that are most related to the labels, as scored by the trained model.

Next the statutes k are scored by their relation to the three tax sources r. Let P_k be the set of words and phrases in k. The average cosine similarity between the phrases

in k (with corresponding vector $\vec{i}$) and tax source r (with corresponding vector $\vec{r}$) is

$$S(k,r) = \frac{1}{|P_k|} \sum_{i \in P_k} \frac{\vec{i} \cdot \vec{r}}{||\vec{i}|| \cdot ||\vec{r}||}$$

where $|P_k|$ is the number of phrases in statute k. The metric inside the summation, the cosine similarity between the phrases, is the standard metric in the NLP literature on word vectors (Levy et al., 2015).[16] It will weight highly the statutes that have words Table 3, and other words that appear in similar contexts.

Next the statute similarities $S(k,r) \in [0,1]$ are used as weights to construct phrase frequencies for each state, year, and source. Let K_{st} be the set of statutes enacted by the government of state s at period t. Let f_k^i equal the frequency of phrase i in statute k. The weighted term frequency of phrase i for source r in state s at time t is

$$\sum_{k \in K_{st}} S(k,r) f_k^i.$$

One could use this expression as the measure of text features, but in that case the effects may be driven by the volume of legislation enacted, rather than the phrases chosen. The focus is on the allocation, rather than the volume, of language, so proportional (relative) frequencies are constructed. The proportional frequency for phrase i divides the term frequency for i by the summed frequency over all phrases:

$$\dot{x}_{st}^{ir} = \frac{\sum_{k \in K_{st}^p} S(k,r) f_k^i}{\sum_{i=1}^p \sum_{k \in K_{st}^i} S(k,r) f_k^i} \tag{2}$$

The numerator is the term frequency of i in state s during year t, weighted by the similarity to tax source r of the statutes where it appeared. The denominator is the total phrase frequency in a state-year for a given source. Therefore $\dot{x}_{rst}^{ir}$ is the proportional frequency for phrase i.

As mentioned, the session laws give the flow rather than the stock of legislation. Therefore $\dot{x}_{rst}^{ir}$ can be seen as giving the within-state-source change in tax legislation. To control for nationwide legislative trends by source, each $\dot{x}_{rst}^{ir}$ is de-meaned by the

---

[16]Cosine similarity has also been used in recent political science work showing text reuse across states (e.g. Hinkle, 2015; Jansa et al., 2015).

average for each source-year. Formally, define

$$x_{rst}^{ir} = \dot{x}_{rst}^{ir} - \frac{1}{n_{jt}} \sum_j \dot{x}_{rjt}^{ir}$$

where the second term is the source-year average for the n_{jt} states who imposed tax r at biennium t. Finally, each text feature variable is standardized by dividing by the within-source standard deviation.

Let n_r be the number of state-year observations for revenue source r. Define the $n_r \times p$ matrix

$$X^r = \begin{matrix} x_{11}^{1r} & ... & x_{11}^{pr} \\ \vdots & \vdots & \vdots \\ x_{st}^{1r} & \vdots & x_{st}^{pr} \\ \vdots & \vdots & \vdots \end{matrix}$$

as the matrix of residualized proportional phrase frequencies. The corresponding column vectors are given by $\mathbf{x}^{ir} = (x_{11}^{ir}, ..., x_{st}^{ir})$, and corresponding row vectors are $\mathbf{x}_{st}^r = (x_{st}^{1r}, x_{st}^{2r}, ..., x_{st}^{Pr})$.

6 Constructing the effective tax code

This section describes the method for constructing the effective tax code by measuring the effect on tax revenues of text features in tax legislation. The goal is not to estimate precisely the effect on revenue of any particular phrase. One cannot measure the tax code perfectly, and phrases are correlated with each other, so the coefficient for any particular phrase cannot be treated as precisely estimated. Instead the objective is to construct a ranking of phrases that can be used to explore how political parties use the tax code in their implementation of state fiscal policy.

The approach is analogous to Gentzkow and Shapiro (2010), who use political floor speech to score language by its association with Democrat or Republican congressmen. They then use that measure to study political bias in newspaper articles. In this paper, phrases are scored by their effect on tax revenue, for use in studying the role of the effective tax code in the political economy of fiscal policy.

Subsection 6.1 outlines the approach for high-dimensional estimation in an OLS framework. Subsection 6.2 constructs Bartik-type instruments for legislative text using variation from statutes enacted in neighboring states. Subsection 6.3 describes the approach for regularized 2SLS estimation using these instruments.

6.1 Ordinary Least Squares

This section presents the basic econometric framework for measuring the average effect of a phrase on tax revenue collected. The estimation strategy is described first using an ordinary least squares framework, to describe the basic structure of the data.

The data is indexed by st, for state s and biennium t. Let P be the set of phrases in the vocabulary $\{1, 2, ..., p\}$. Let R be the set of revenue sources (corporate income tax, personal income tax, sales tax). The goal is to estimate the effect β_{ir} for each phrase $i \in P$ on government revenue g_{st}^r for each source $r \in R$. A linear model of the effect of the proportional frequency x_{st}^{ir} on legislation related to r enacted in state s at biennium t for phrase i on the tax burden g_{st}^r from source r, holding all other phrases constant, is

$$g_{st}^r = \beta_{ir} x_{st}^{ir} + \epsilon_{st}^r. \tag{3}$$

Recall that g_{st}^r has been residualized on a state fixed effect and a year fixed effect, while x_{st}^{ir} is the flow of legislation and has been residualized on a year fixed effect. This means that this regression controls for time-invariant state-level factors, as well as time-varying nationwide factors. A positive β_{ir} means that when phrase i appears more in statutes related to source r, there is a higher revenue for that source. A negative β_{ir} means that when phrase i appears more in statutes related to source r, there is a lower measured revenue for that source. For statistical inference one could cluster standard errors by state (Bertrand et al., 2004).

Consistent estimation of (3) using OLS relies on the assumption that there are no state-level time-varying factors affecting both the phrase frequencies x_{st}^{ir} and revenue g_{st}^r. Tax legislation is chosen endogenously in response to other economic factors affecting tax revenues; Chang (2014) documents this type of endogeneity in the context of state R&D tax credits. These other factors may include other phrases j, which are correlated with phrase i as well as government revenues. One could try to include other phrases in the regression, but there would be a problem of multi-collinearity if one tried to include all $p = 25,000$ phrases. For these reasons, OLS will likely provide inconsistent estimates for many of the phrases.

6.2 Instrumental Variables

Because of these identification issues, to estimate β_{ir} we need exogenous variation in x_{st}^{ir} that is uncorrelated with other policies that affect tax revenues. The approach to solving the identification problem is to construct a set of Bartik-style instruments for

phrase frequencies. Exogenous variation comes from diffusion of text from other states in the same regional judicial district.

Bartik (1991) constructs instruments for labor demand using nationwide industry-specific shocks, which are exogenous from the perspective of any individual locality. If one interacts this shock with the sectoral composition of a locality, one obtains exogenous cross-sectional variation in labor demand. Another related instrument is that used for state tax rates in Fajgelbaum et al. (2015), who used tax rates in neighboring states as instruments in 2SLS estimates for labor supply elasticity with respect to top tax rates.

This paper uses regional variation over time in phrase frequencies from enacted legislation by state governments. The basic motivation stems from previous work documenting diffusion of policies from state to state (Berry and Berry, 1990, 1992; Case et al., 1993; Berry and Berry, 1994; Mooney and Lee, 1995). This diffusion includes not just discrete policies but the actual wording of statutes; Jansa et al. (2015) document that state legislatures frequently borrow the text of legislation from other states. The goal is to find variation in statute text that is more or less randomly assigned conditional on the fixed effects. [17]

Cross-sectional variation is needed so that a year fixed effect can be included in the regressions to control for national trends. Because the focus is on legal language, a channel for preferential diffusion of legal language – as opposed to policies generally – is desirable. A good fit for these needs is to use lagged regional variation in language within the federal appellate court circuits, which comprise a set of eleven judicial districts in the federal court system. Figure 2 illustrates the groupings of states into circuits which has been in place since 1982. For the earlier years in the sample (1963-1981), Alabama, Florida, and Georgia were part of the Fifth Circuit (rather than the Eleventh).

These districts were founded and are administered by the federal government (rather than state governments) with a focus on federal law. The state governments have little direct influence on the circuits or the decision-making of their judges, yet circuit judges are asked to interpret and apply state law in numerous cases every year (Hoover, 1982). Previous empirical work has shown that policies diffuse between state governments in the same circuit even more than they do between neighboring states or states in the same political party (Bird and Smythe, 2008), supporting the idea that the circuit

---

[17]Balla (2001) shows that the text of insurance legislation preferentially diffuses in states whose commissioners are members of the same insurance regulation professional association. Chernick (2005) documents that the regressivity of taxes are actually negatively related to those of neighbors, showing that diffusion of language is not necessarily accompanied by diffusions in substantive policy.

Figure 2: Federal Circuit Court Map
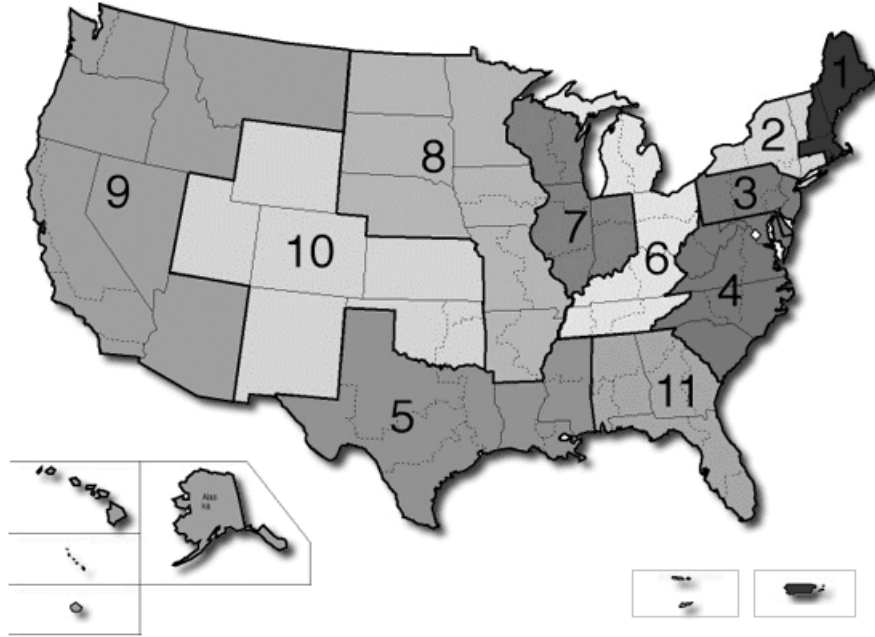
represents a regional legal community (see also Carp, 1972). Hinkle (2015) in particular shows that the actual text of statutes preferentially diffuses to states in the same federal circuit.

This is useful empirically because the timing of legislative choices in one state in a circuit is likely unrelated to non-legislative factors affecting tax collections in other states in the circuit. While the groupings are more-or-less contiguous, they are not based on historically or politically important relationships. Assignment is more or less arbitrary; for example, Washington and Utah are grouped together yet their state governments share little in common politically.

The text instruments are constructed as follows. For each source r, state s, time t, and phrase i, construct the leave-one-out average frequency for other states in the same federal circuit for the previous year,

$$z_{st}^{ir} = \frac{1}{|J(s,t)| - 1} \sum_{j \neq s, j \in J(s,t)} x_{jt-1}^{ir}$$

where j indexes the other states, $J(s,t)$ is the set of states in s's circuit at t, and $|J(s,t)|$ is the number of states in $J(s,t)$. This gives the lagged leave-one-out average phrase

frequency for phrase i on legislation for source r in the circuit.

Define

$$
Z_r = \begin{matrix}
z_{11}^{1r} & \cdots & z_{11}^{pr} \\
\vdots & \vdots & \vdots \\
z_{st}^{1r} & \vdots & z_{rt}^{pr} \\
\vdots & \vdots & \vdots
\end{matrix}
$$

the $n_r \times q$ matrix of Bartik phrase instruments for revenue source r. Let $\mathbf{z}_{rst}$ denote a row vector from this matrix and consider the following two-stage least-squares framework. The first stage for each phrase i is

$$
x_{st}^{ir} = \mathbf{z}_{rst}' \gamma_i + \eta_{st}^{ir}, \forall i, r \tag{4}
$$

where $\gamma_i \in \mathbb{R}^q$ is a row of the $p \times q$ matrix of first-stage coefficients Γ. The second stage equation for the effect of x_{st}^{ir} on revenue is the same as the OLS equation from above:

$$
g_{st}^r = \beta_{ir} x_{st}^{ir} + \epsilon_{st}^r. \tag{5}
$$

The empirical goal is to obtain consistent estimates of β_{ir} from Equation (5).

The key identifying assumption for this IV setup is that

$$
\text{Cov}(z_{st}^{ir}, \epsilon_{st}) = 0, \forall i \in P, r \in R.
$$

This requires that the instrument only affect g_{st}^r through its effect on x_{st}^{ir}. That is, a state legislature's choices of tax law phrases will have an impact on the phrases chosen by other state legislatures in the circuit, but will not otherwise affect tax revenue collections as a share of income (conditional on the fixed effects). This is justified by the same arguments that that are used for traditional Bartik instruments. With the inclusion of state-source and source-year fixed effects, this specification compares well to other recent work using related methods (e.g. Bertrand et al., 2013; Acemoglu et al., 2014). In the data, the instruments are not significantly related to current period observables, including tax revenues and state GDP. The 2SLS results reported below are not sensitive to the inclusion of a variety of sets of covariates that one would expect to be correlated with tax collections, including a state's own GDP and/or the average GDP for the rest of the circuit.

6.3 High-Dimensional IV Estimation

Even if the instruments are valid, there are too many of them. The 2SLS estimator is consistent only for small numbers of instruments relative to the sample size (Chao and Swanson, 2005; Hansen et al., 2008). In this dataset there are 25,000 instruments but just 3,500 observations. This subsection describes the use of regularization methods for dealing with high dimensionality.

A set of recent econometrics papers have made progress in solving the many-weak-instruments problem using regularization methods such as Lasso (Least Absolute Shrinkage). Lasso and related methods (such as Ridge regression and elastic net) can improve the performance of IV under the assumption of a sparse first stage, that is, when a relatively small number of instruments suffice to approximate the effect of all the instruments on the endogenous regressors. This active research area includes Caner (2009), Gautier and Tsybakov (2011), Okui (2011), and Carrasco (2012).

The main approach in this paper is based on Belloni et al. (2012), who use post-Lasso to obtain optimal instruments under sparsity. That paper provides conditions under which regularized IV is consistent and asymptotically normal under heteroskedastisticy and non-normality. Another related paper is Lin et al. (2015), who use Lasso (and more general regularization methods) in the case of a large number of instruments as well as a large number of endogenous regressors. They prove consistency for a regularized 2SLS estimator under sparse effects of the instruments and the endogenous regressors.[18]

In this case, the sparsity assumption means that there are a set of factors, traditions, cultures, or ideas that are active within the federal judicial circuits and driving changes in the tax code. Regularization provides a data-driven method for recovering proxies for these factors from the lagged leave-one-out average phrase frequencies.

Lasso is implemented as follows. There are $p = 25000$ phrases and $q = 25000$ instruments. Estimating the 625 million elements of Γ is computationally expensive. To

---

[18]An alternative approach to dimension reduction is the factor IV method using principal components analysis (PCA) to reduce the matrix of instruments (Bai and Ng, 2008). This method is widely used in the time series forecasting literature in empirical macroeconomics. Bai and Ng (2010) show that when there are underlying factors driving both the endogenous regressors and the instruments, then the principal components of the matrix of instruments will themselves provide the optimal instruments. For robustness, all of the regressions below were alternatively implemented using factor IV in the first stage (as detailed in Appendix A.2). The main results were similar under factor IV, but the out-of-sample prediction (Subsection 6.5) was worse, so the sparse-instruments specification is reported in the main text.

ease the computational burden, I first run each of the 625 million univariate regressions

$$x_{st}^i = \gamma_{ij} z_{st}^j + \eta_{st}^{ij}$$

and exclude from the first stage any elements of $\mathbf{z}$ for which $\hat{\gamma}_{ij}$ has a t-statistic below 3.

The first stage regression for phrase i solves

$$\hat{\gamma}_i = \arg\min_{\gamma_i \in \mathbb{R}^q} \{\frac{1}{2n}||x_{st}^i - Z\gamma_i||_2^2 + \lambda \sum_{j=1}^J (||\gamma_{ij}||_1)\} \tag{6}$$

where the last term is the L1 (Lasso) penalty. The penalty parameter λ is chosen following the methods in Belloni et al. (2012) and Lin et al. (2015).[19]

The regularized first stage forces sparsity; most elements of Γ go to zero. Lasso provides its own regularized estimates for $\hat{\Gamma}$, but following Belloni et al. (2012), the preferred approach is to use post-Lasso.[20] First-stage estimates are obtained by running OLS using only the non-zero phrases from Lasso, with standard errors clustered by state. An advantage of using post-Lasso is that it provides a first-stage F-statistic for evaluating instrument relevance. This is discussed further in Subsection 6.4.[21]

The rest of the IV method is standard. The estimated $\hat{\Gamma}$ is used to predict

$$\hat{X}_r = \begin{matrix} \hat{x}_{11}^{1r} & ... & \hat{x}_{11}^{pr} \\ \vdots & \vdots & \vdots \\ \hat{x}_{st}^{1r} & \vdots & \hat{x}_{st}^{pr} \\ \vdots & \vdots & \vdots \end{matrix},$$

the $n_r \times p$ matrix of instrumented (and fixed-effect-transformed) phrase frequencies for each revenue source. This matrix includes only the exogenous variation in phrase changes due to the instruments. Then the average partial effect of phrase i on tax

---

[19]An alternative specification included an L2 penalty in addition to the L1 penalty. This L1/L2 specification is the elastic net model, which has better performance than Lasso under high multi-collinearity (Zou and Hastie, 2005). The elastic net estimator also satisfies the assumptions of the more general regularization framework in Lin et al. (2015). Zou and Hastie (2005) show that one of the limiting cases for elastic net is Lasso, while the other is equivalent to choosing regressors via soft thresholding. Caner and Zhang (2014) study the elastic net in a GMM framework.

[20]The Lasso and post-Lasso second-stage results were similar in this sample.

[21]First stage regressions were implemented in Python using scikit-learn (for Lasso and elastic net) and statsmodels (for OLS). I followed the advice of Dubé et al. (2012) in setting numerical tolerance levels.

Figure 3: Distribution of First-Stage F-Statistic
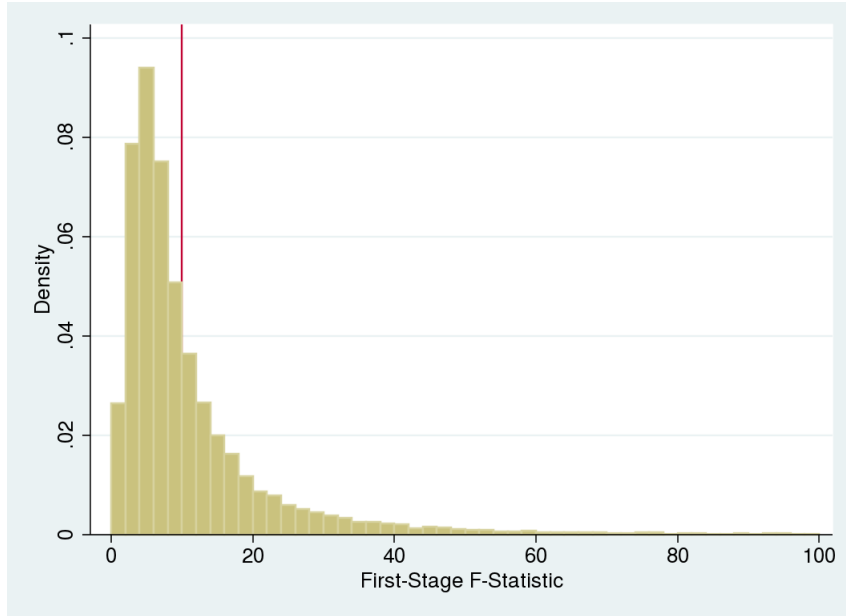
Distribution of first-stage F-statistics for main IV specification. Vertical line at $F = 10$. The mean is 14.1 and the median is 7.4. Out of a vocabulary of 25000, 8,923 phrases have an F-stat greater than 10.

revenues can be estimated using

$$g_{st}^r = \beta_{ir}\hat{x}_{st}^{ir} + \epsilon_{st}^r. \tag{7}$$

This equation uses the instrumented phrase frequency $\hat{x}_{st}^{ir}$. Holding other phrases constant, this will procure the average effect on tax revenues for source r of using phrase i once more in statutes related to r.

6.4 First Stage Statistics

This section reports statistics on the first stage regressions. The main goal is to show that the post-Lasso obtains a sufficiently high first-stage F-statistic, and therefore instrument relevance, for a large set of phrases.

Figure 3 shows the distribution of the first-stage F-statistics. A set of 8,923 phrases have a strong first stage. In the main analysis, phrases with a weak first stage are excluded. This set of phrases is still large enough for prediction and analysis, as demonstrated below. For comparison, Gentzkow and Shapiro (2010) use a vocabulary of 1,000

Figure 4: Instrument Phrases Have a Stronger Effect on Own Endogenous Phrase

(a)
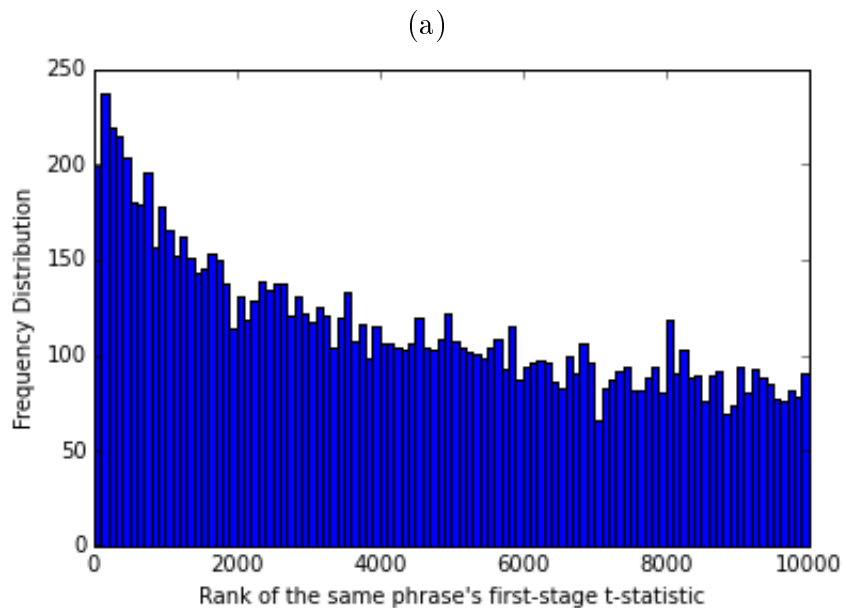
Frequency distribution over ranking of same phrase in first stage t-statistics.

phrases to predict politician ideology.

Figure 4 is designed to assess the common-sense idea of whether the instrument phrases are affecting their own phrase in other states, to substantiate the diffusion process. The figure shows that when ranking the instruments j by the t-statistic of γ_{ij} for any given endogenous regressor i, the t-statistic for one's own phrase tends to rank highly among the set of phrases. This supports the idea that language diffusion is occurring through preference for phrases in the same judicial circuit.

To further assesses the usefulness of the Bartik instrument, alternative specifications were run that intuitively should have a weaker first stage. First, a ten-year lag was used rather than a two-year lag, which results in a 20% smaller mean F-statistic and 23% smaller median F-statistic. Second, a set of instruments were constructed from non-tax statutes (rather than tax statutes), which results in a 10% decrease in the mean F-statistic and a 12% decrease in the the median F-statistic. These alternative specifications are weaker, as intuition would suggest.

6.5 Out-of-sample prediction of revenue with the effective tax code

With thousands of regressors, reporting the individual 2SLS estimates is not very informative. Many of them are significant just due to statistical noise. Therefore this section takes a machine-learning approach to see whether a regression model trained on the textual features of tax code changes can predict out-of-sample changes in tax revenue. The prediction is run conditional on a constant rate structure, and uses the exogenous variation in the tax code derived from the instruments.

The method for out-of-sample prediction is partial least squares regression (PLS). PLS is a dimension-reduction technique similar to principal component analysis (PCA), where high-dimensional data is projected down to a lower-dimensional space while retaining as much information as possible. The key difference from PCA is that PLS is a supervised technique: Components are constructed to maximize the predictiveness for an outcome variable (Chun and Keleş, 2010). Previous examples of PLS in social-science text analysis include Jensen et al. (2012) and Jelveh et al. (2015).

The outcome variable is g^r_{st}, which has been residualized on a source-year fixed effect and a source-state-rate fixed effect and then standardized. PLS is then used to predict $\hat{g}^r_{st}$. As the explanatory data, the actual phrase frequencies X_r and the instrumented phrase frequencies $\hat{X}_r$ are alternatively used. The former should predict better, but the latter only uses causal variation in the effective tax code. If the instrumented tax code changes predict changes in tax revenues, that uncovers an aggregate causal effect of the tax code on tax revenues.

Chun and Keleş (2010) show that PLS can be inconsistent with a large number of non-predictive noise variables. To avoid this problem, phrases with a weak t-statistic for β_{ir} (below three) are excluded. In the set of instrumented phrases, any phrases with a first-stage F-statistic below 10 are also excluded. The training data included a random sample of 70% of the observations, while the test data included the remaining 30% of observations. The best highest predictions were obtained for between 25 and 50 PLS components.[22]

Figure 5 illustrates the predictiveness of the PLS model for the three tax sources. In these graphs, the horizontal axis is the true tax-revenue change for each test observation. The vertical axis is the PLS-predicted tax-revenue change based on the phrase frequencies for that test observation. The red line gives the best linear fit for these

---

[22]The regressions used the Python implementation of PLS from the scikit-learn package.

Figure 5: Out-of-Sample Tax Revenue Predictions

(a) Corporate Income Tax
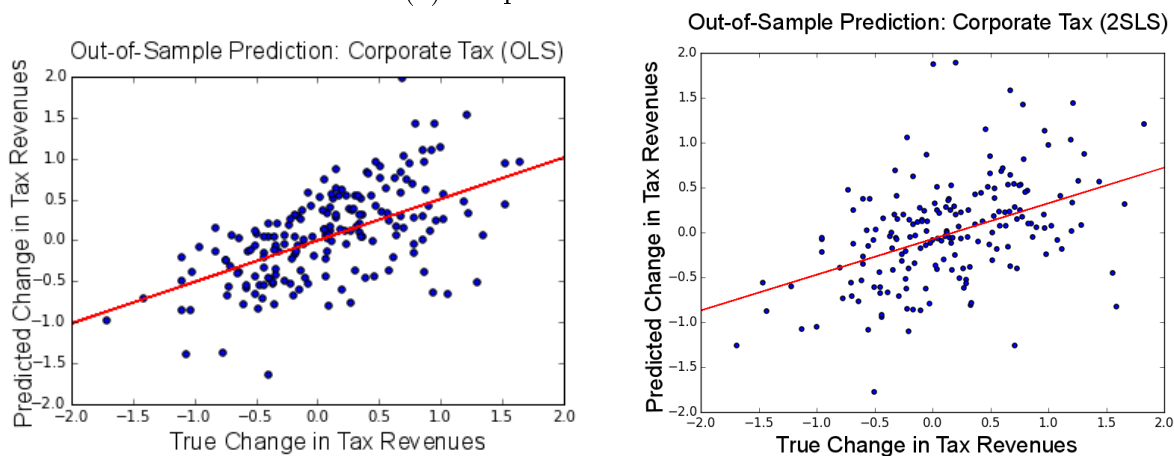
(b) Personal Income Tax
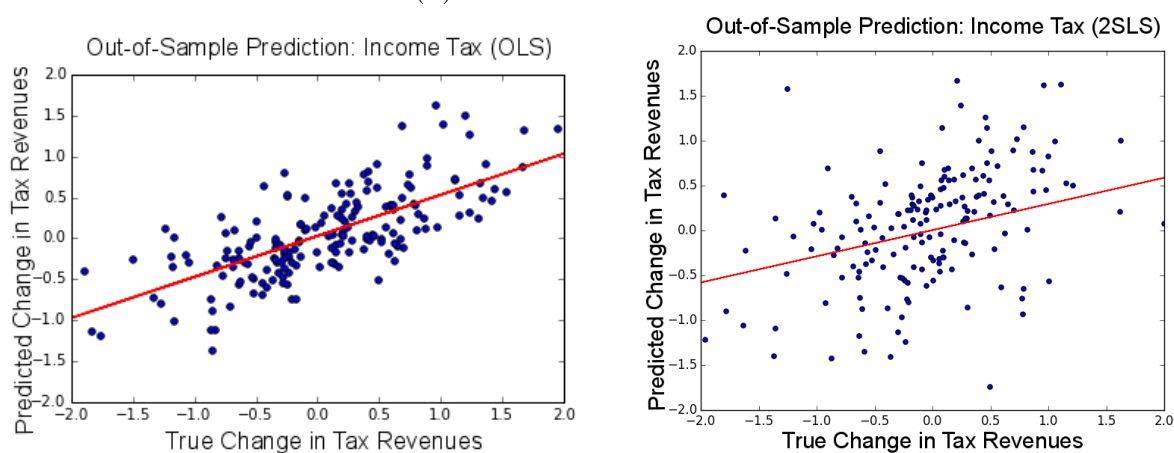
(c) Sales Tax
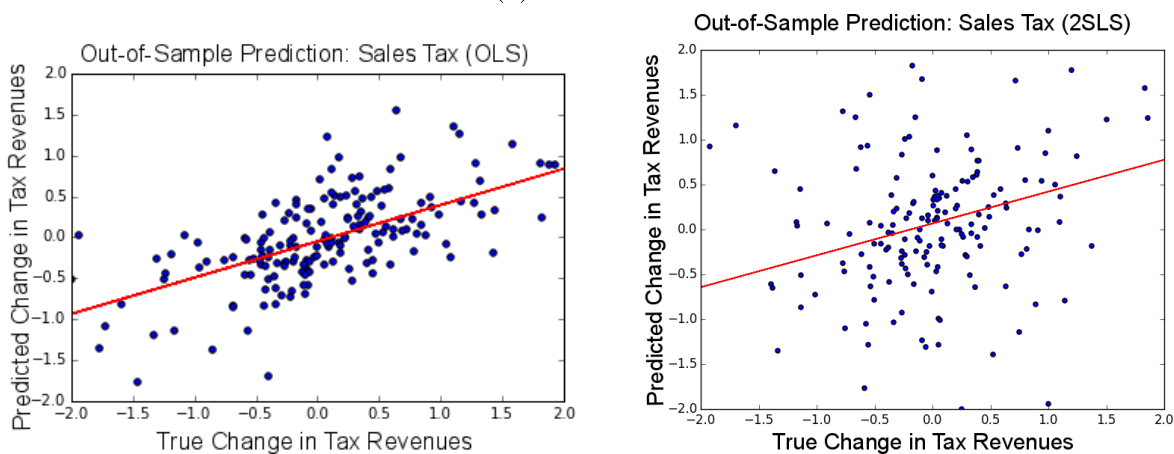
PLS model trained with most predictive phrases ($p < .01$) and 25 PLS components. Horizontal axis is the true tax-revenue change for that test observation; the vertical axis is the PLS-predicted tax-revenue change based on the phrase frequencies for that test observation. The red line gives the best linear fit. In the left column, the actual phrases are used; in the right column, the instrumented phrases are used.

observations. In the left column, the actual phrases are used; in the right column, the instrumented phrases are used.

The PLS model has good out-of-sample predictiveness. With the actual phrases, the correlation between truth and prediction is very high for all three income sources: 0.88, 0.89, and 0.84, respectively. Using the instrumented phrases results in a worse prediction, as expected (.65, .53, and .41, respectively). But there is still a clear correlation between truth and prediction. Taking the square of the correlation coefficient gives the R^2. With the actual phrase frequencies, we can say that roughly 80% of the variance in tax revenues (remaining after partialling out the source-year and source-state-year fixed effects) is explained by the text features of the tax code. As a comparison, Gentzkow and Shapiro (2010) report an *in-sample* correlation of 0.61 for their measure of political ideology (they do not report an out-of-sample correlation). The in-sample correlation for the PLS model used here is over 0.9 for all the measures.

These statistics demonstrate the out-of-sample predictiveness of tax code features, holding major tax rates constant. The PLS model is learning information about the tax base from tax code changes and using it to predict revenue changes. This validates the use of this measure in the subsequent analysis.

6.6 Analysis of phrases that affect tax revenues

The next step is to analyze the set of predictive phrases. Because the particular phrases chosen by the algorithm do not play a key role in the empirical analysis, this section can be seen as a set of descriptive statistics. These statistics are useful because they show how the phrases in the tax code relate to changes in the tax base.

The 2SLS framework discussed so far procures a set of statistics for ranking phrases by their predicted effect on tax revenues. First, the F-statistic for the first-stage regression can be used to filter out phrases for which there isn't sufficient exogenous variation in the phrase from the instruments. Second, the t-statistic for the second-stage regression summarizes the impact of the phrase on tax revenue, accounting for both the covariance and the noise in the data.

The simplest approach would be to rank all of the phrases by their t-statistic and then to look at the top and bottom phrases for each revenue source. This turns out not to be very informative, since the phrases chosen are from a variety of topics, some of which are not related to the tax base. To get more interpretability, I construct phrase topics and rank the phrases within topic by their revenue effect.

Topics are constructed by using k-means clustering to partition the Word2Vec space

into clusters of related words and phrases (Yu et al., 2013; Guo et al., 2014). Given a set of word vectors $\{\vec{q}_1, \vec{q}_2, ..., \vec{q}_P\}$, the algorithm chooses clusters $Q = \{Q_1, Q_2, ...Q_k\}$, to minimize the within-cluster sum of squares. Formally, the model solves

$$\arg\min_Q \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||^2$$

where μ_i is the mean of the points (the centroid) for cluster Q_i. Once initialized, the algorithm re-assigns samples to clusters and recomputes centroids until convergence to a threshold. The only parameter needed is the desired number of clusters. After experimenting with between $k = 5$ and $k = 250$ topics, I settled on $k = 25$, which is small enough to allow reports for all topics but still produced reasonable results in terms of interpretability.

Within topic, the F-statistics and t-statistics are collected for each phrase by revenue source. Phrases with low F-statistics and low t-statistics are filtered out, and the remainder are ranked by the t-statistic. The full ranking of phrases is available in an appendix. In Table 4, I report a selection of topics for personal income tax and sales tax, respectively, which are relatively useful for interpretation. Words in bold are discussed in the text. The numbers on the topics are arbitrary and were determined randomly by the algorithm.

The top half of Table 4 looks at phrases related to the income tax. First consider Topic 3 (panel a), which includes phrases related to pensions and dependents. The phrase "such dependent" refers to exemptions and credits for children and other dependents.[23] The phrase "such service" is found in income tax statutes giving deductions for certain service expenses.[24] Interestingly, the fact that using "such" increases revenue may reflect the effect of higher clarity in the tax code, as the word "such" serves to clarify the targets of deductions and exemptions.[25]

---

[23]E.g. 1994 Kansas H.B. 2929: "Income earned on an individual development account shall be exempt from state income taxation under the Kansas income tax act... There shall be no limit on the amount of earned income of a dependent child, who is a recipient of aid to families with dependent children, deposited in an individual development account of **such dependent** child that was created or organized to pay for educational expenses of **such dependent** child."

[24]E.g. 1995 Idaho H.B. 132: "In the case of an individual, there shall be allowed as a deduction from gross income either (1) or (2) at the option of the taxpayer: Itemized expenditures of not to exceed one thousand dollars ($1,000) per cared for member incurred in providing personal care services to or for an immediate member of the taxpayer's family; **such services** may be provided either in the taxpayer's home or the family member's home."

[25]In Appendix A.5, I show that using the 2SLS rankings to suggest replacements to increase revenue often results in adding "such" or "said" before phrases.

Table 4: Phrases with a Significant 2SLS Effect on Tax Revenues

Phrase	T-statistic	Phrase	T-statistic

Personal Income Tax

Phrase	T-statistic	Phrase	T-statistic
Topic 3		*Topic 7*	
such dependent	**5.89**	**buildings and structures**	**14.92**
retirement purposes	5.34	construct operate and maintain	14.51
such service	**5.34**	adjacent land	12.62
in excess of year	4.54	street and road	10.52
pay period	-4.31	sewage disposal plant	10.13
bi-weekly	-4.31	curb gutter	9.66
pension board	3.71	aforesaid purposes	9.07
Topic 19		*Topic 22*	
dependent children	**7.09**	school activity	-7.14
daycare service	**-5.30**	high school graduate	5.88
self-support	4.57	school graduate	5.56
legal settlement	4.44	educational purposes	4.57
center	4.00	**adult education**	**4.13**
medical condition	**-4.00**	academic	3.99
admission	3.92	**vocation**	**3.96**

Sales Tax

Phrase	T-statistic	Phrase	T-statistic
Topic 8		*Topic 12*	
not-for	5.60	retail store	-8.70
internal combustion engine	**4.73**	fell	8.11
certain motor vehicles	-4.60	**fuel dealer**	**6.84**
snow	-4.20	such distributor	6.80
such vehicle	4.00	wrapper	6.45
antique	-3.91	director of agriculture	5.59
movement of traffic	3.62	frog	4.79
Topic 14		*Topic 19*	
retail install sale	-8.70	aid to families	8.11
on the real property	-7.49	**cost of health**	**6.29**
such dwelling	6.20	retard service	4.95
certificate of sale	-4.93	state plan	4.84
other rights	-4.57	educate or train	-4.69
valuable consideration	**3.88**	psychiatrist	-4.57
execute and deliver	3.87	first aid	-4.37

Topic 7 (panel b) relates to construction projects and expenses. These phrases can affect income tax through deductions and credits for various home-related expenses. For example, the phrase "building or structure" can be used to define homes for the purposes of homeowners' exemptions.[26]

Next, Topic 19 (panel c) again has phrases related to dependents, but with an emphasis on health care. The phrase "dependent children" occurs frequently in income tax statutes in determining credits for parents of children. For example, some statutes provide for medical expense deductions for dependent children.[27] Similarly, "medical condition" is relevant to income tax for determining what types of health expenses are deductible, or for determining targeted benefits.[28] Third, "daycare service" is another relevant deductible expense in state income taxes, as part of deductible childcare expenses.[29]

Topic 22 (panel d) is related to education and training. "Adult education" is relevant to income tax in light of the deductions for adult educational expenditures provided in many states.[30] Meanwhile, the word "vocation" is often found in income tax statutes as part of the definition of income-generating activities that are taxable.[31]

---

[26] E.g. 1997 California AB 2797: "For the purposes of this section, the term 'premises' means a house or a dwelling unit used to provide living accommodations in a **building or structure** and the land incidental thereto, but does not include land only, unless the dwelling unit is a mobile home. The credit is not allowed for any taxable year for the rental of land upon which a mobile home is located if the mobile home has been granted a homeowners' exemption under Section 218 in that year."

[27] E.g. 2001 Idaho HB 121: "'Eligible medical expense' means an expense paid by the taxpayer for medical care described in section 213(d) of the Internal Revenue Code, and long-term care expenses of the account holder and the spouse, dependents and **dependent children** of the account holder."

[28] E.g. 1995 South Carolina SB 753: "There is allowed as a deduction in computing South Carolina taxable income of an individual the following: Two thousand dollars for each adopted special needs child... For purposes of this item, a special needs child is a person who is:unlikely to be adopted without assistance as determined by the South Carolina Department of Social Services because of conditions such as ethnic minority status, age, sibling group membership, **medical condition**, or physical, mental, or emotional handicaps."

[29] E.g. 1995 New Mexico HB 11: "Any resident who files an individual New Mexico income tax return and who is not a dependent of another taxpayer may claim a credit for child daycare expenses incurred and paid to a caregiver in New Mexico during the taxable year by such resident...The caregiver shall furnish the resident with a signed statement of compensation paid by the resident to the caregiver for **daycare services**. Such statements shall specify the dates and the total number of days for which payment has been made."

[30] E.g. 2006 Kentucky HB 1: "An employer who assists an individual to complete his or her learning contract under the provisions of this section shall receive a state income tax credit for a portion of the released time given to the employee to study for the tests. The application for the tax credit shall be supported with attendance documentation provided by the department for **adult education** and literacy."

[31] E.g. 1993 Mississippi SB 2720: "For the purposes of this article, except as otherwise provided, the term 'gross income' means and includes the income of a taxpayer derived from salaries, wages, fees or compensation for service, of whatever kind and in whatever form paid, including income from

The bottom half of Table 4 reports the revenue-relevant phrases by topic for sales taxes. Topic 8 (panel a) has to do mainly with automobiles. These phrases often crop up in sales tax statutes to define what types of vehicles and fuels are exempt from sales taxation.[32] These phrases affect revenues through their influence on the exemptions.

Topic 12 (panel b) is related to retail trade. The phrases in this topic appear frequently in sales tax legislation, for example to describe which retailers must collect sales tax.[33] Note again the inclusion of "such distributor": just as we saw with income tax, adding the word "such" tends to increase revenue.

We see the same trend in Topic 14 (panel c). Both "such dwelling" and "such transaction" are predicted to increase sales tax revenues. Seeing all of these phrases together is suggestive that clarifying language tends to increase tax collections. This suggests a role for good legal writing in the efficient implementation of tax policies. Meanwhile, "valuable consideration" is often used to define what constitutes a taxable sales transaction.[34]

Finally, Topic 19 (panel d) has phrases related to health care. Compare this set of phrases to that selected for income tax; it is the same topic, but a different set of phrases are chosen as relevant. This shows that the rankings are picking out different phrases for different revenue sources, which makes intuitive sense. These phrases are going to mostly be related to sales tax exemptions for health care services. However, they can also be used for classifications related to non-profit status.[35]

The rest of the word clouds for income tax and sales tax, and the word clouds for corporate tax, are in the appendix. While some of the topics are not interpretable, the ones listed in this section are suggestive. Overall, they suggest that the 2SLS estimates are measuring a strong impact on revenue of tax expenditures: exemptions, deductions, and credits. This is consistent with the view that the tax code has an important impact

---

governmental agencies and subdivisions thereof; or from professions, **vocations**, trades, businesses, commerce or sales, or renting or dealing in property, or reacquired property."

[32]E.g. 2007 California SB 774: "There are exempted from the taxes imposed by this part the gross receipts from the sale of, and the storage, use, or other consumption in this state of, by a qualified person any of the following. . . . any motor fuel or mixture of motor fuels that is . . . Advertised, offered for sale, suitable for use, or used as a motor fuel in an **internal combustion engine**."

[33]E.g. 23 VAC 210-630: "The preceding paragraph establishes when a **fuel dealer** must collect tax at the time of sale, and it does not establish any rule of exemption for consumers."

[34]E.g. Oklahoma Code 68-1352: "'Sale' means the transfer of either title or possession of tangible personal property for a **valuable consideration** regardless of the manner, method, instrumentality, or device by which the transfer is accomplished in this state."

[35]E.g. Nebraska Reg. 1-090: "A nonprofit organization operating any of the following facilities that are licensed under the Health Care Facility Licensure Act is only exempt on purchases for use at the facility. . . . A health clinic, when one or more hospitals, or the parent corporations of the hospitals, own or control the health clinic for the purpose of reducing the **cost of health** services..."

on tax revenues by changing the legal definition of the tax base.

7 Effect of political control on tax policy

This section describes the empirical strategy for measuring the effect of political control on tax policy. Subsection 7.1 describes the research design. Subsection 7.2 reports the results on tax rates and revenues. Subsection 7.3 provides descriptive statistics on the tax law phrases that are related to political party control.

7.1 Empirical strategy

There are many ways one could try to measure the effect of political control on state tax policy. One could look at the number of years of political party control, for example. To keep things simple, this papers estimates the sign of the average change from one party to the other.

The empirical approach for identifying the effect of political control on tax rates and tax code language is a panel data design similar to a regression discontinuity (Lee and Lemieux, 2010). This approach has gained traction in political economy through the use of electoral votes as the forcing variable, with a cutoff at 50 percent of the popular votes (e.g. Lee et al., 2004). Leigh (2008) and Beland (2015) document causal effects on state policy of barely electing a Democratic (rather than Republican) governor. Warren (2009) and De Magalhães and Ferrero (2015) take the analogous approach to state legislatures, using the number of legislative seats belonging to the political parties as the forcing variable. Warren (2009) shows that there is a positive local treatment effect of a Democratic legislature on the total tax burden.

Caughey et al. (2015) show that an RD using seat shares in the legislature as the forcing variable is associated with covariate imbalance. Therefore I do not use a standard RD with small bandwidths around the threshold. The regressions include all the observations. To control for the type of variation that RD's are designed to control for, I include polynomials above and below the cutoff. These regressions are designed to isolate the variation from going from minority Democrat to majority Democrat.

Let D_{st} be an indicator variable or set of indicator variables for stronger Democratic control in state s at period t. This could include an indicator equaling one for a Democrat-controlled lower chamber, for example. Let d_{st} be the vote-share variable(s) (in percentage points) determining D_{st}, with associated polynomial(s) $f(d_{st})$ for use in

RD-type regressions. The empirical analysis uses the party in charge of the legislative chambers, the governorship, and an index for the number of these governing bodies that are controlled.

The estimating equation is a panel data regression with polynomials in the forcing variables. For outcome variable y_{st}, estimate

$$y_{st} = \alpha_{st} + D'_{st}\rho + f(d_{st}) + \epsilon_{st}$$

where α_{st} may include state and year fixed effects. For $f(d_{st})$, specifications include linear or quadratic polynomials. Again I cluster standard errors by state.

7.2   Effect of political control on tax revenues and tax rates

This subsection provides estimates for the effect of political control on tax policy outcomes besides the tax code. I provide estimates of the effect of political control on marginal tax rates and tax revenue. I also estimate the share of the revenue effect due to the rate structure. This analysis adds to the previous literature (Chernick, 2005; Reed, 2006; Leigh, 2008) by providing separate estimates for income tax, corporate tax, and sales tax.

The first question is whether political parties change marginal tax rates when they come into office. Let m^r_{st} be the top marginal tax rate for source r in state s at time t. The effect of political party control on the marginal rate is obtained from

$$m^r_{st} = \alpha_{st} + \rho^r_m D_{st} + f(d_{st}) + \epsilon^r_{st} \tag{8}$$

where α_{st} includes state and year fixed effects by revenue source.

The second question is whether party control is associated with changes in tax revenues as a share of income. This involves estimating ρ_g from Subsection 3.2 for each revenue source r. Let government revenue be given by g^r_{st}. The empirical model is

$$g^r_{st} = \alpha_{st} + \rho^r_g D_{st} + f(d_{st}) + \epsilon^r_{st} \tag{9}$$

where α_{st} include state and year fixed effects by revenue source. In these regressions, $D_{st} \in [0,3]$ is defined as an index of Democrat control, equaling the number of governing bodies (the legislative houses and the governorship) that Democrats control. A tied legislature adds one half to this index. The $f(d_{st})$ term includes linear polynomials above and below the cutoffs, for both legislatures and the governorship.

Table 5: Effect of Political Control on State Tax Policy

	(1) Marginal Tax Rate	(2) Tax Revenue (including rates)	(3) Tax Revenue (net of rates)
Effect of Democrat Power			
Income Tax	0.0384	0.0460	0.0134
	(0.0782)	(0.0811)	(0.0765)
Sales Tax	-0.0766	-0.176	-0.157
	(0.0644)	(0.114)	(0.110)
N	3091	3091	3091
State-Source FE's	Yes	Yes	Yes
State-Source-Rate FE's			Yes

Estimates for effect of Democrat Control index on the marginal tax rate and tax revenue, separately by tax source. Observation is a state-source-year. Regressions include linear polynomials in the forcing variables for both houses and governor, separately for values above and below the cutoffs. Outcome variables are standardized so coefficients can be interpreted as changes in the standard deviation of the outcome variable. Standard errors in parentheses, clustered by state. + p<0.1, * p<0.05, ** p<0.01.

The third question is what share, if any, of ρ_g^r is due to changes in the rate structure. This is not given by the estimate for ρ_m^r from (8), which just gives the effect of political control on the marginal rate. The rate structure is complex, with multiple rates and brackets, so one cannot estimate the share of the revenue change due to the rate structure, ρ_τ^r, strictly from the marginal rate. Instead, this quantity is obtained as follows. First, estimate $\hat{\rho}_g^r$ from (9) as previously described. Second, estimate $\ddot{\rho}_g^r$ from (9) using state-rate-source fixed effects, as described in Subsection 4.1. This provides an estimate for the effect of political control on revenues purged of any effects from the rate structure. Then the share of the revenue due to tax rates is obtained from $\rho_\tau^r = \hat{\rho}_g^r - \ddot{\rho}_g^r$.

Table 5 reports estimates for the effect of Democrat control on marginal tax rates and tax revenues. First, Column 1 reports the effect of party control on the marginal tax rate. If tax rates are the most important component of fiscal policy, then changing political parties should be associated with a change in the marginal tax rate. As can be seen in Column 1, there is no statistical effect of party control on the tax rate.

Next, Columns 2 through 3 show the effect of political control on tax revenues, with and without state-source-rate fixed effects. The estimates are noisy, and not statistically significant. The coefficients are positive for income tax and negative for sales tax.

As expected, the coefficients are smaller when fixed effects for the rate structure are included. These coefficients are used in the computation of U in Subsection 8.3 below.

7.3 Tax code language associated with political control

This section discusses the method and provides summary statistics associated with the the effect of political control on tax code language. As with Subsection 6.6, the individual phrase coefficients are not treated as precisely estimated. Instead, the goal is to construct a rough ranking of the political party differences for use of phrases in tax legislation. Then these scores can be used to analyze the political economy of state fiscal policy, as done in Section 8 below.

The estimating equation is a phrase-wise panel data regression. The set of outcomes is the vector of tax code language features $\mathbf{x}_{st}^r$. There are separate regressions for each source r, with the goal of testing whether different political parties have different priorities for the incidence of tax liability. Formally, estimate

$$x_{st}^{ir} = \delta_{ir} D_{st} + f(d_{st}) + \epsilon_{st}^{ir}, \forall i, r$$

for each phrase, to get the average effect of Democrat control D_{st} on the use of phrase i for tax code provisions related to source r.

Table 6 reports samples of phrases associated with Democrat and Republican control for the same selection of topics used in Table 4. A positive t-statistic is associated with Democrats; a negative t-statistic is associated with Republicans. Unlike Gentzkow and Shapiro (2010), these phrases are not clearly partisan. This reflects that the text of legislation is not as politicized as floor debate speech.

The top half of Table 6 reports the phrases that Democrats and Republicans prefer to use on income tax legislation, with the bottom half doing so for sales tax legislation. These particular phrases don't play a large role in the analysis but show which types of policies the parties spend time on legislating. For example, it seems that Republicans spend more time on health care, while Democrats spend more time on education.

One notable example is the issue of "home health care" (Topic 19). Health care services are an important but somewhat controversial target for tax expenditures, as a deduction for income tax and an exemption for sales tax. Recent press articles have detailed how tax-cutting Republicans tend to favor these exemptions and deductions.[36] A second notable example is the inclusion of "groceries" in sales tax legislation (Topic

---

[36]E.g. Pennsylvania State Capital Newsfeed, Rep. Will Tallman, Aug. 27 (2015).

Table 6: Phrases with a Significant Relation to Political Party Control

Personal Income Tax

Topic 3		Topic 19	
Democrat Phrases	*Republican Phrases*	*Democrat Phrases*	*Republican Phrases*
written design	other pension	physical health	**home health care**
rate of wage	employee organization	first aid	response person
period of employ	plan or system	medical	such commitment
service as member	age of sixty	service and supply	private practice
become member	compensation	school of medicine	federal social security act
normal retirement date	patrolmen	convincing evidence	epileptic
attainment of age	retirement contribution		foster care

Sales Tax

Topic 8		Topic 12	
Democrat Phrases	*Republican Phrases*	*Democrat Phrases*	*Republican Phrases*
drive	wagon	use	swim
commission of motor	licensed motor vehicle	put	stockyard
drive vehicle	thirty feet	other means	ink
respective jurisdiction	livery	material	hook
vehicle	vehicle or trailer	firework	wild
trip	passenger motor vehicle	apply to sale	fur
operating motor vehicle	clearance	**groceries**	prohibit the use

12). Democrats have long favored exempting groceries from sales tax, although Republicans are generally opposed.[37] This is a clear example of a redistribution-focused tax expenditure.

For a more detailed discussion of tax code phrases, see Appendix A.3. That section discusses phrases identified by the regressions as having both a political impact and a revenue impact. The appendix discusses examples of where those phrases may be found in the statutes, and also provides examples of court cases construing the language in revenue-relevant caselaw.

8   The effective tax code and the politics of redistribution

This section analyzes the role of the effective tax code in how political parties implement preferred redistributive policies. I provide two methods. In Subsection 8.1, I construct predicted changes in revenue by state-year-source using the tax code features, and test how that predicted measure responds to changes in political control. In Subsection 8.2, I focus on the granularity of the language features, relating the average revenue impact of a phrase to the average political impact on a phrase, separately by revenue source. Subsection 8.3 provides a discussion.

8.1   Testing for the effect of political control on textually predicted tax revenue

This section reports estimates for the effect of political control on the predicted revenue changes from tax legislation. I construct a metric for the predicted change in tax revenue based on the effective tax code. I then estimate the effect of changes in political party control of state government on this metric.

The metric is constructed as follows. For each state, year, and revenue source, define

$$\tilde{g}_{st}^r = \sum_{i=1}^{p} x_{rt}^{ir} \frac{\hat{\beta}_{ir}}{\hat{\sigma}_{ir}}$$

where $\hat{\sigma}_{ir}$ give the standard error for the 2SLS estimate $\hat{\beta}_{ir}$. Only phrases with a strong first-stage F-statistic in the 2SLS framework are included. This can be understood

---

[37]E.g. "Alabama House Democrats make creating jobs a priority," Nov. 2. (2011), quoting a House Democrat this way: ""It's not like people have a choice about eating. The grocery tax is unfair, immoral and it has to go."

as the predicted tax revenue change in a state-year, weighted by the precisions of the estimated effects of each phrase.

Then I regress

$$\tilde{g}_{st}^r = \alpha_{st} + \phi_r D_{st} + f(d_{st}) + \epsilon_{st}$$

to obtain the effect of Democrat control, $\hat{\phi}_r$, on the predicted tax revenue change from the effective tax code. I cluster standard errors by state.

Because the statute text gives the flow of legislation, the outcome variable is first-differenced. This will eliminate bias from time-invariant state characteristics. The term α_{st} includes source-year fixed-effects and and state-source trends (Bertrand et al., 2004). The source-year fixed effects control for bias associated with time-varying national trends in the outcome variable. The state-source trends are designed to account for preexisting trends in the outcome variable that may be correlated with treatment.

The term $f(d_{st})$ includes linear or quadratic polynomials in the forcing variables (vote share for governor, seat shares for the legislatures), separately interacted with each revenue source, and separately for observations above and below the cutoff. This allows for the model to flexibly control for vote and seat shares. All observations are included, rather than only observations near the cutoff as would be done in a standard RD. Including these time-varying forcing variables is designed to control for other political institutions and factors that may affect the tax code text.

For D_{st}, I include three specifications. First, I include an index for Democrat control of state government that counts the number of bodies controlled by Democrats, from zero to three. Second, I break out the governor separately from the legislature, where the Legislative Power index is the number of legislatures controlled by Democrats. Third, I include separate regressors for each legislature. In the case of a tied legislature, that adds one-half to the index (or is a one-half instead of zero-one in the indicators).

The regression results are reported in Table 7. These regressions analyze the combined effects of changes in Democratic control on the tax revenue text. The regressions look at the within-state effect of changes in political control to the three government bodies. The regressions included corporate taxes, but those are not reported here because there were no significant effects.

Columns 1 and 4 look at the aggregate effect of Democratic power in state government. There is a significant positive effect on text-predicted income tax revenue, and a significant negative effect on text-predicted sales tax revenue. When Democrats take control of an additional wing of state government, there is a 0.14 standard deviation predicted increase in income tax revenues due to tax code changes, and a 0.07 standard

Table 7: Effect of Party Control on Text-Predicted Tax Revenue

	(1)	(2)	(3)	(4)	(5)	(6)
Income Tax						
Democrat Power	0.0992**			0.144**		
	(0.0337)			(0.0478)		
Legislative Power		0.0822+			0.120	
		(0.0480)			(0.0771)	
Democrat Governor		0.140+	0.147*		0.180*	0.187*
		(0.0724)	(0.0720)		(0.0838)	(0.0814)
Dem. Upper House			0.0985			0.113
			(0.0610)			(0.0949)
Dem. Lower House			0.0514			0.0950
			(0.104)			(0.132)
Sales Tax						
Democrat Power	-0.0324			-0.0677*		
	(0.0254)			(0.0311)		
Legislative Power		-0.0388			-0.0865*	
		(0.0284)			(0.0382)	
Democrat Governor		-0.0158	-0.0179		-0.0434	-0.0527
		(0.0442)	(0.0444)		(0.0538)	(0.0530)
Dem. Upper House			-0.0362			-0.121+
			(0.0458)			(0.0604)
Dem. Lower House			-0.0509			-0.0990+
			(0.0460)			(0.0536)
State-Source FD's	X	X	X	X	X	X
Source-Year FE's	X	X	X	X	X	X
State-Source Trends	X	X	X	X	X	X
Forcing Var Polys				X	X	X

Estimates from regressing an index for Democrat control on the predicted revenue change based on the text, as described in Subsection 8.2, separately by tax source. $N = 3,588$ observations, state-source-year. Columns 4 through 6 include linear polynomials in the forcing variables for both houses and governor, separately for values above and below the cutoffs. Outcome variables are standardized so coefficients can be interpreted as changes in the standard deviation of the outcome variable. Standard errors in parentheses, clustered by state. + $p<0.1$, * $p<0.05$, ** $p<0.01$.

deviation decrease in the predicted sales tax revenues due to tax code changes.

Columns 2 and 5 look at the separate effects of the governor and the legislatures, where Legislative Power is the number of legislatures controlled by Democrats. This shows that both the legislature and the governor have positive estimated effects on income-tax-increasing tax code language. Only the effect of the governor is individually significant, however. The sales-tax effects are wholly driven by the legislature.

Columns 3 and 6 include all three bodies as separate regressors. In the case of income tax, all three bodies contribute materially to the effect in terms of magnitudes. Again, only the governor effect is individually statistically significant. In the case of sales tax, both the upper house and lower house of the legislature have a statistically significantly negative estimated effect. The governor again has no effect.

Table 8 provides two additional specifications to probe the robustness of the results. First, in Columns 1 through 3 the regressions include lagged covariates for state gross domestic product and state expenditures on financial administration. These are two major economic and political factors that may be correlated with tax code changes and tax revenue collections. These do not change the results. Columns 4 through 6 add the lagged dependent variable to test for further confounding trends. This also does not change the results.

Furthermore, the results are robust to the inclusion of non-interacted linear or quadratic polynomials in the forcing variables (rather than interacted). Adding an interacted quadratic polynomial strengthens the sales tax effect but weakens the income tax effect. Adding state-source fixed effects in addition to the state-source first-differences, which can be seen as double-differencing the outcome, does not affect the income tax effect but weakens the sales tax effect (without changing the sign). Using more lags in the covariates variables, and/or using the current-period values, also does not change the results. Finally, adding the federal-circuit average of state GDP also does not change the results.

The Democrat Power index can be interpreted as the predicted change in standard deviations of revenue from Democrat control of an additional wing of state government. This means that moving from full Republican control to full Democrat control is associated with a 0.42 standard deviation increase in income tax revenues due to the tax code. This roughly translates to a 34.7% increase in income taxes as a share of personal income, and an additional \$1.96 billion in income tax revenues (in 2007 dollars) in the average state.

For sales tax, moving from full Republican control to full Democrat control is as-

Table 8: Party Control and Text-Predicted Tax Revenue (Additional Specifications)

	(1)	(2)	(3)	(4)	(5)	(6)
Income Tax						
Democrat Power	0.138**			0.145**		
	(0.0458)			(0.0418)		
Legislative Power		0.107			0.120+	
		(0.0735)			(0.0680)	
Democrat Governor		0.186*	0.190*		0.182*	0.189*
		(0.0775)	(0.0763)		(0.0807)	(0.0794)
Dem. Upper House			0.130			0.164+
			(0.0879)			(0.0818)
Dem. Lower House			0.0738			0.0708
			(0.134)			(0.128)
Sales Tax						
Democrat Power	-0.0829*			-0.0780*		
	(0.0326)			(0.0310)		
Legislative Power		-0.106**			-0.100*	
		(0.0396)			(0.0419)	
Democrat Governor		-0.0503	-0.0596		-0.0477	-0.0567
		(0.0579)	(0.0575)		(0.0499)	(0.0497)
Dem. Upper House			-0.143*			-0.155**
			(0.0606)			(0.0559)
Dem. Lower House			-0.105+			-0.0777
			(0.0590)			(0.0617)
State-Source FD's	X	X	X	X	X	X
Source-Year FE's	X	X	X	X	X	X
State-Source Trends	X	X	X	X	X	X
Forcing Var Polys	X	X	X	X	X	X
Lagged Covariates	X	X	X	X	X	X
Lagged Dep. Var.				X	X	X

Estimates from regressing an index for Democrat control on the predicted revenue change based on the text, as described in Subsection 8.2, separately by tax source. $N = 3,588$ observations, state-source-year. Columns 4 through 6 include linear polynomials in the forcing variables for both houses and governor, separately for values above and below the cutoffs. Outcome variables are standardized so coefficients can be interpreted as changes in the standard deviation of the outcome variable. Standard errors in parentheses, clustered by state. + $p<0.1$, * $p<0.05$, ** $p<0.01$.

sociated with a predicted 0.21 standard deviation decrease in sales tax revenues due to the tax code. That corresponds to a 25.9% decrease in sales taxes as a share of sales receipts income. In 2007 dollars, the average state loses $1.73 billion in sales tax revenue based on tax code changes.

To show these results a different way, Figure 12 plots the change in tax-predicted revenue before and after a political takeover of the state legislatures. The top figure shows the lower chamber effect, while the bottom figure shows the upper chamber effect. The purple line gives the trend in text-predicted revenue for income tax, while the orange line does so for sales tax. Note that the houses often change party control the same year or in nearby years, which explains the similarity of the trend. Republican takeovers are also included in the graph – with the sign of the outcome variable reversed so that the treatment is treated symmetrically. Excluding Republican takeovers results in a similar trend.

These graphs show that after a change in political control, the text features of income tax legislation change in a way that would predict increasing revenues. Conversely, the text features of sales tax legislation change in way that would predict decreasing revenues. These graphs support the idea that the political parties put different types of language into the tax code when they are in power, in such a way that Democrats increase revenues from income tax but decrease revenues from sales tax.

8.2   Assessing the granularity of the redistributive consequences of tax code language

This section complements the previous section by looking at language features directly. The approach tests for differences in how political parties use phrases based on their predicted revenue consequences. The goal is to assess the level of textual subtlety that is driving the effects. Democrats may be selecting broadly different policies and topics than Republicans, or they may be making specific textual substitutions within the same topics. The approach in this section is to provide evidence on this question.

For each phrase i in the vocabulary P, I have a set of statistics from the previous sections. First, I have a t-statistic for the 2SLS effect of phrase i on revenue from source r, $\tilde{\beta}_i^r$. Second, I have a t-statistic for the effect of Democrat control on frequency of phrase i for tax legislation on source r, $\tilde{\delta}_i^r$. To test whether the language used by political parties is systematically related to the revenue consequences of that language,

Figure 6: Dynamic Effect of Democrat Control on Text-Predicted Revenue

(a) Dynamic Effect of Democrat Takeover of Lower House

(b) Dynamic Effect of Democrat Takeover of Upper House
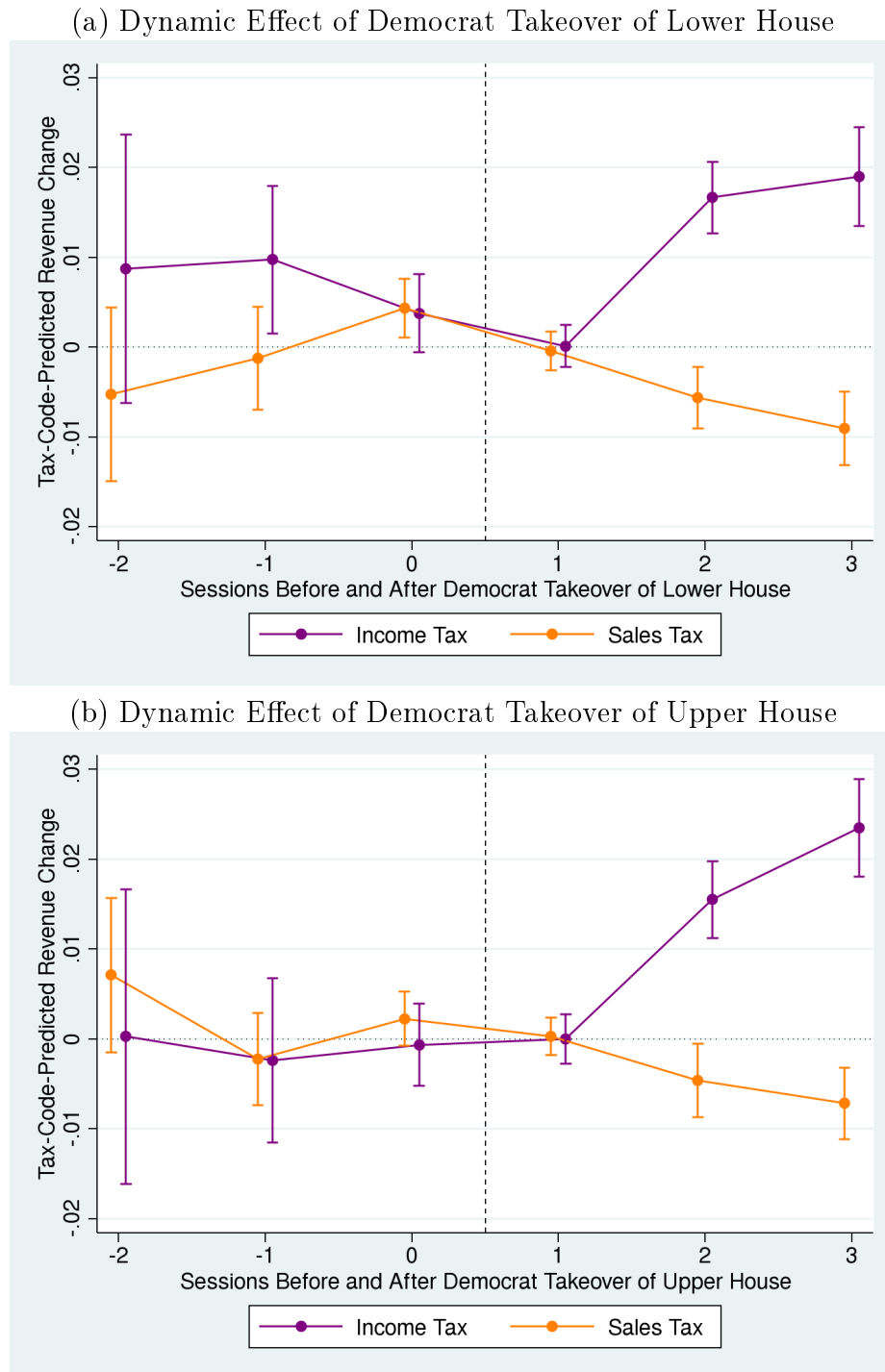
Event study graphs for change in text-predicted revenue before and after Democratic takeover of the lower legislature (panel a) and upper legislature (panel b), respectively. The vertical axis is the metric for state-predicted revenue $\tilde{g}$, as described in the text. The horizontal axis is years before and after a change in political control. Republican takeovers are also included, with the sign of the outcome variable reversed.

I regress

$$\tilde{\beta}_i^r = \alpha + \psi_r \tilde{\delta}_i^r + \epsilon_i^r, \forall r \in R \tag{10}$$

to estimate $\hat{\psi}_r$. A positive $\hat{\psi}_r$ means that relative to Republicans, Democrats tend to use revenue-increasing phrases on revenue source r. A negative $\hat{\psi}_r$ means that Democrats tend to use revenue-decreasing phrases on revenue source r. The regression is weighted by the average frequencies of the phrase observations.

Topics are constructed using the k-means clustering method described in Subsection 6.6. The topics are used, firstly, to cluster standard errors by 50 topics.[38] Next, a varying number of topic fixed effects are added to the regression. Then the regression obtains the within-topic relationship of Democrat control and the revenue effect of language. The goal is to assess the subtlety of the tax code differences that lead to the observed effects. If the effect is killed off after adding a few topics, that suggests the party-control effect is driven by choices across broad topics or policies. If the effect remains after adding fixed effects for a large number of topics, that means the effect is driven by highly specific choices between closely related words.

There are 8,923 phrases in the vocabulary. This means, for example, that with 100 topic fixed effects, each topic will have 89 words on average. As one adds more topics, we are looking at small groups of words on average – around 9 words each for 1000 topics, for example. With 4000 topics, many words will have their own topic, and the topics that do remain will be groups of closely related words and phrases. If there is still a significant language effect with this many topics, we can say that the fiscal policy differences between Republicans and Democrats are embodied in highly specific language choices in the tax code.

Table 9 reports the regression coefficients from (10). The column specifications gradually add more fixed effects for topics. As before, we generally see that Democrats prefer revenue-increasing language on income taxes, but revenue decreasing language on sales taxes. For sales tax, the effect persists for up to 2000 topics. For income tax, the effect persists for up to 4000 topics. Above those thresholds, the number of topics is large enough that the effects go to zero.

These results support the view that the policy effects of tax code language are encoded in relatively specific choices of legal wording. For sales tax, the effects are still significant with 2000 topics – that is, the effects come from the within-topic choices between 4 to 5 phrases on average. Income tax legislation is even more granular – the effects are still significant for 4000 topics. This means that the effects of income

---

[38]The basic results are statistically significant with at least 10 topic clusters.

Table 9: Granularity of the Revenue-Politics Relation of Tax Code Language

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Income Tax Effect	0.0528+	0.0621*	0.0646*	0.119**	0.0739**	0.114**	0.139*	0.0804
of Dem Power	(0.0274)	(0.0243)	(0.0241)	(0.0317)	(0.0221)	(0.0376)	(0.0604)	(0.0724)
Sales Tax Effect	-0.0802**	-0.0647*	-0.0714**	-0.0688*	-0.109*	-0.0254	-0.0614	-0.110
of Dem Power	(0.0251)	(0.0258)	(0.0225)	(0.0336)	(0.0470)	(0.0630)	(0.0650)	(0.106)
# of Topic Fixed Effects	-	10	100	1000	2000	3000	4000	5000
Mean Words per Topic	8923	892	89.2	8.92	4.46	2.97	2.23	1.78

Estimates from regressing the revenue effect of a phrase (beta) on the party effect on a phrase (delta), separately by tax source. Outcome variables and explanatory variables are standardized. $N = 8,923$ phrases with strong first-stage F-statistics. Standard errors in parentheses, clustered by 50 phrase topics. Regressions weighted by average frequency of the phrase. $+$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

tax legislation come from the within-topic choices between 2 to 3 phrases on average. The redistributive fiscal policies implemented by the political parties in the U.S. states consist of highly specific choices in the tax code.

Identifying these subtle differences would likely be difficult for researchers taking a more standard approach of subjectively coding discrete policy changes. The natural language processing tools are needed. Moreover, with such small clusters of phrases having an important association with the politics of redistribution, it may be useful for researchers and policymakers to analyze these phrases more systematically. This demonstrates the usefulness of natural language processing tools in the analysis of the tax code.

8.3   Discussion

Personal income taxes are progressive taxes. Sales taxes are regressive taxes. If Democrats prefer more redistribution, then one would expect them to increase income taxes but decrease sales taxes. As shown in Subsection 3.2, they do not change the major tax rates. However, as we see here, they do change the tax code in line with this intuition. These results are consistent with the idea that the tax code plays an important role in the political economy of fiscal policy in the U.S. states. This may reflect that because tax rates are salient, political bargaining is difficult and tends to stalemate. Instead, political parties have to implement redistributive policies in the specifics of legislation, which allow for tradeoffs across different issues.

These results are related to the evidence in Finkelstein (2009), who found that toll rates were difficult to increase when they were salient and known to drivers, but could be increased when the toll rates became less salient. In the case of state governments, politicians who are interested in changing redistributive policy will have trouble doing so by changing the rates. Because the major rates are so salient for voters, it is politically costly to change them. On the other hand, changing the text of the tax code is less politically costly, since these textual features are not salient to voters.

9 Conclusion

This paper has examined the role of the tax code in the political economy of fiscal policy in the U.S. states. The approach combines prediction, the focus of most machine-learning methodologies, with identification, the traditional concern of applied econometrics. I used a data-driven method to extract the effective tax code – those text features of legislation that have a causal impact on tax collections. The paper then showed which phrases are related to changes in political control. Democrat control of state government is associated with a preference for tax code language that is predicted to increase the progressivity of the state tax system. The tax code, rather than the rate rate, is the more important fiscal policy tool in the U.S. states. Work on state tax policy cannot limit attention to changes in tax rates.

This paper's analysis has focused on the positive questions of how the tax code affects revenues and how political parties differ in the language they insert into the tax code. In future work one could use these methods to analyze the equity and welfare consequences of tax code features. An example of this analysis is provided in the appendix, which uses the method to find replacement phrases that are predicted to increase tax revenues.

A natural extension of this project is in linking the text features of the tax code to other text data. For example, it would be important to understand the role of courts in legal tax avoidance. Second, it would be interesting to measure connections between legislative text and newspaper text, to see how media attention influences the salience of tax code reform.

This approach has the potential to open up a new area for research in political economy and public finance. Economists tend to view economic systems through national accounts and other numerical data sets. Yet complex economies will not run well without a complex corpus of statutes regulating it, and a well-managed system of courts

enforcing those laws as written. A data-driven approach to legal text will help uncover the impact of written laws on the real economy.

As natural language processing technology improves, there will be a growing set of tools for lawyers and legislators to use for designing legislation that more effectively implements desired policy goals. This method is not limited in use to tax legislation. It could be applied to any set of legal documents with a defined quantitative policy goal. For example, exogenous variations in criminal laws could be analyzed for their effects on crime rates. Exogenous variations in contract laws could be analyzed for their effects on transaction efficiency. And "laws" in this context include not just legislation but court cases and administrative regulations.

References

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2014). Democracy does cause growth. Technical report, National Bureau of Economic Research.

Allingham, M. and Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of public economics*, 1(3-4):323–338.

Alm, J. and Borders, K. (2014). Estimating the tax gap at the state level: The case of georgia's personal income tax. *Public Budgeting and Finance*, 34(4):61–79.

Andreoni, J., Erard, B., and Feinstein, J. (1998). Tax compliance. *Journal of economic literature*, 36(2):818–860.

Ash, E., Morelli, M., and Van Weelden, R. (2015). Elections and divisiveness: Theory and evidence. *NBER*.

Atkinson, A. B. and Stiglitz, J. E. (1976). The design of tax structure: direct versus indirect taxation. *Journal of public Economics*, 6(1):55–75.

Bai, J. and Ng, S. (2008). *Large dimensional factor analysis*. Now Publishers Inc.

Bai, J. and Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(06):1577–1606.

Balla, S. J. (2001). Interstate professional associations and the diffusion of policy innovations. *American Politics Research*, 29(3):221–245.

Bartik, T. J. (1991). Who benefits from state and local economic development policies? *Books from Upjohn Press*.

Beland, L.-P. (2015). Political parties and labor market outcomes: Evidence from us states. *American Economic Journal: Applied Economics*.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Berry, F. S. and Berry, W. D. (1990). State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review*, 84:395–415.

Berry, F. S. and Berry, W. D. (1992). Tax innovation in the states: Capitalizing on political opportunity. *American Journal of Political Science*, 36(3):pp. 715–742.

Berry, F. S. and Berry, W. D. (1994). The politics of tax increases in the states. *American Journal of Political Science*, 38(3):pp. 855–859.

Berry, W. D. and Baybeck, B. (2005). Using geographic information systems to study interstate competition. *American Political Science Review*, 99(04):505–519.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust

differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1):249–275.

Bertrand, M., Pan, J., and Kamenica, E. (2013). Gender identity and relative income within households. Technical report, National Bureau of Economic Research.

Besley, T. and Case, A. (1995). Does electoral accountability affect economic policy choices? evidence from gubernatorial term limits. *The Quarterly Journal of Economics*, pages 769–798.

Besley, T. and Case, A. (2003). Political institutions and policy choices: evidence from the united states. *Journal of Economic Literature*, pages 7–73.

Bird, R. C. and Smythe, D. J. (2008). The structure of american legal institutions and the diffusion of wrongful-discharge laws, 1978-1999. *Law and Society Review*, 42(4):833–864.

Bommarito, M. J., Katz, D. M., and Isaacs-See, J. (2011). An empirical survey of the population of united states tax court written decisions. *Virginia Tax Review*, 30(2).

Burman, L. and Christopher Geissler, E. J. T. (2008). How big are total individual income tax expenditures, and who benefits from them? *The American Economic Review*, 98(2):79–83.

Cabral, M. and Hoxby, C. (2012). The hated property tax: salience, tax rates, and tax revolts. Technical report, National Bureau of Economic Research.

Cai, H. and Liu, Q. (2009). Competition and corporate tax avoidance: Evidence from chinese industrial firms*. *The Economic Journal*, 119(537):764–795.

Caner, M. (2009). Lasso-type gmm estimator. *Econometric Theory*, 25(01):270–290.

Caner, M. and Zhang, H. H. (2014). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics*, 32(1):30–47.

Carp, R. A. (1972). The scope and function of intra-circuit judicial communication: A case study of the eighth circuit. *Law and Society Review*, pages 405–426.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Case, A. C., Rosen, H. S., and Hines, J. R. (1993). Budget spillovers and fiscal policy interdependence. *Journal of Public Economics*, 52(3):285 – 307.

Caughey, D., Warshaw, C., and Xu, Y. (2015). The policy effects of the partisan composition of state government.

Chang, A. C. (2014). Tax policy endogeneity: evidence from r&d tax credits.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

Chernick, H. (2005). On the determinants of subnational tax progressivity in the u.s. *National Tax Journal*, 58(1):pp. 93–112.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., et al. (2017). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Chetty, R. (2009). Is the taxable income elasticity sufficient to calculate deadweight loss? the implications of evasion and avoidance. *American Economic Journal: Economic Policy*, 1(2):31–52.

Chetty, R. and Hendren, N. (2013). The economic impacts of tax expenditures: Evidence from spatial variation across the u.s.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *The American Economic Review*, 99(4):1145.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

De Magalhães, L. and Ferrero, L. (2015). Separation of powers and the tax level in the us states. *Southern Economic Journal*.

Denny, M. J., O'Connor, B., and Wallach, H. (2015). A little bit of nlp goes a long way: Finding meaning in legislative texts with phrase extraction. Microsoft Research NYC, http://dirichlet.net/.

Desai, M. (2005). The degradation of reported corporate profits. *The Journal of Economic Perspectives*, 19(4):171–192.

Dubé, J.-P., Fox, J. T., and Su, C.-L. (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267.

Dynarski, S. and Scott-Clayton, J. (2006). The cost of complexity in federal student aid: Lessons from optimal tax theory and behavioral economics. Technical report, National Bureau of Economic Research.

Fajgelbaum, P., Morales, E., Serrato, J. C. S., and Zidar, O. (2015). State taxes and

spatial misallocation. Technical report, Working Paper.

Feldstein, M. (1999). Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics*, 81(4):674–680.

Finkelstein, A. (2009). E-ztax: Tax salience and tax rates. *Quarterly Journal of Economics*, 124(3).

Gamage, D. and Shanske, D. (2011). Three essays on tax salience: Market salience and political salience. *Tax L. Rev.*, 65:19.

GAO, G. A. O. (2003). Internal revenue service, challenges remain in combating abusive tax shelters, reprinted in gao rep. *No. 04-104T*.

Gautier, E. and Tsybakov, A. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. Technical report, National Bureau of Economic Research.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Gentzkow, M., Shapiro, J. M., and Sinkinson, M. (2014). Competition and ideological diversity: Historical evidence from us newspapers. *The American Economic Review*, 104(10):3073–3114.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2015). Measuring polarization in high-dimensional data: Method and application to congressional speech.

Givati, Y. (2009). Resolving legal uncertainty: The unfulfilled promise of advance tax rulings. *Virginia Tax Review*, 29:137.

Goldin, J. (2015). Optimal tax salience. *Journal of Public Economics, forthcoming*.

Goldin, J. and Homonoff, T. (2013). Smoke gets in your eyes: Cigarette tax salience and regressivity. *American Economic Journal: Economic Policy*, 5(1):302–336.

Gordon, R. and Li, W. (2009). Tax structures in developing countries: Many puzzles and a possible explanation. *Journal of public Economics*, 93(7):855–866.

Gordon, R. H. and Kopczuk, W. (2014). The choice of the personal income tax base. *Journal of Public Economics*, 118:97–110.

Graetz, M. (2007). Tax reform unraveling. *The Journal of Economic Perspectives*, 21(1):69–90.

Graetz, M. J. (1995). Paint-by-numbers tax lawmaking. *Columbia Law Review*, 95(3):609–682.

Griswold, E. N. (1944). The need for a court of tax appeals. *Harvard Law Review*, 57(8):pp. 1153–1192.

Guo, J., Che, W., Wang, H., and Liu, T. (2014). Revisiting embedding features for simple semi-supervised learning. In *Proceedings of EMNLP*, pages 110–120.

Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 26(4).

Hansen, S., McMahon, M., and Prat, A. (2014). Transparency and deliberation within the fomc: a computational linguistics approach.

Heen, M. L. (1996). Plain meaning, the tax code, and doctrinal incoherence. *Hastings LJ*, 48:771.

Hertel-Fernandez, A. and Kashin, K. (2015). Capturing business power across the states with text reuse. In *annual conference of the Midwest Political Science Association, Chicago, April*, pages 16–19.

Hettich, W. and Winer, S. (2005). *Democratic choice and taxation: A theoretical and empirical analysis*. Cambridge Univ Pr.

Hinkle, R. K. (2015). Into the words: Using statutory text to explore the impact of federal courts on state policy diffusion. *American Journal of Political Science*.

Holcombe, R. (1998). Tax policy from a public choice perspective. *National Tax Journal*, 51(2):359–371.

Holtzblatt, J. and McCubbin, J. (2003). Whose child is it anyway? simplifying the definition of a child. *National Tax Journal*, 56(3):701–718.

Hoover, C. A. (1982). Deference to federal circuit court interpretations of unsettled state law: Factors, etc., inc. v. pro arts, inc. *Duke Law Journal*, pages 704–732.

Howard, C. (1999). *The hidden welfare state: Tax expenditures and social policy in the United States*. Princeton University Press.

Jansa, J. M., Hansen, E. R., and Gray, V. H. (2015). Copy and paste lawmaking: The diffusion of policy language across american state legislatures.

Jelveh, Z., Kogut, B., and Naidu, S. (2015). Political language in economics.

Jensen, J., Naidu, S., Kaplan, E., Wilse-Samson, L., GERGEN, D., ZUCKERMAN, M., and SPIRLING, A. (2012). Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech [with comments and discussion]. *Brookings Papers on Economic Activity*, pages 1–81.

Katz, D. M. and Bommarito II, M. J. (2014). Measuring the complexity of the law: the united states code. *Artificial Intelligence and Law*, 22(4):337–374.

Kleven, H., Knudsen, M., Kreiner, C., Pedersen, S., and Saez, E. (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica*, 79(3):651–692.

Kleven, H. and Kopczuk, W. (2011). Transfer program complexity and the take-up of social benefits. *American Economic Journal: Economic Policy*, 3(1):54–90.

Kopczuk, W. (2001). Redistribution when avoidance behavior is heterogeneous. *Journal of Public Economics*, 81(1):51–71.

Kopczuk, W. (2005). Tax bases, tax rates and the elasticity of reported income. *Journal of Public Economics*, 89(11):2093–2119.

Krishna, A. and Slemrod, J. (2003). Behavioral public finance: tax design as price presentation. *International Tax and Public Finance*, 10(2):189–203.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic L*, 48(2):281–355.

Lee, D. S., Moretti, E., and Butler, M. J. (2004). Do voters affect or elect policies? evidence from the us house. *Quarterly Journal of Economics*, 119(3):807–859. Times Cited: 91 Lee, DS Moretti, E Butler, MJ 91.

Leigh, A. (2008). Estimating the impact of gubernatorial partisanship on policy settings and economic outcomes: A regression discontinuity approach. *European Journal of Political Economy*, 24(1):256–268.

Levy, O. and Goldberg, Y. (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Levy, O., Goldberg, Y., and Ramat-Gan, I. (2014). Linguistic regularities in sparse and explicit word representations. *CoNLL-2014*, page 171.

Likhovski, A. (2004). The duke and the lady: Helvering v. gregory and the history of tax avoidance adjudication. *Cardozo Law Review*, 25.

Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288.

Livingston, M. (1995). Practical reason, purposivism, and the interpretation of tax statutes. *Tax. L. Rev.*, 51:677.

Logue, K. (2007). Optimal tax compliance and penalties when the law is uncertain. *Va. Tax Rev.*, 27:241.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in*

neural information processing systems, pages 3111–3119.

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The review of economic studies*, pages 175–208.

Mooney, C. Z. and Lee, M.-H. (1995). Legislative morality in the american states: The case of pre-roe abortion regulation reform. *American Journal of Political Science*, pages 599–627.

Nelson, M. A. (2000). Electoral cycles and the politics of state tax policy. *Public Finance Review*, 28(6):540–560.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 165(1):70–86.

Ordower, H. (2010). The culture of tax avoidance. *Saint Louis University Law Journal, Vol. 55, 2010, Saint Louis U. Legal Studies Research Paper No. 2010-06.*

Paul, D. (1997). Sources of tax complexity: How much simplicity can fundamental tax reform achieve, the. *NCL Rev*, 76:151.

Persson, T. and Tabellini, G. E. (2002). *Political economics: explaining economic policy.* MIT press.

Plesko, G. (2007). Estimates of the magnitude of financial and tax reporting conflicts. Technical report, National Bureau of Economic Research.

Pomeranz, D. (2011). No taxation without information.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Poterba, J. and Sinai, T. (2008). Tax expenditures for owner-occupied housing: Deductions for property taxes and mortgage interest and the exclusion of imputed rental income. *The American Economic Review*, 98(2):84–89.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Reed, W. R. (2006). Democrats, republicans, and taxes: Evidence that political parties matter. *Journal of Public Economics*, 90(4):725–750.

Rork, J. C. (2003). Coveting thy neighbors' taxation. *National Tax Journal*, pages 775–787.

Schizer, D. (2005). Enlisting the tax bar. *Tax L. Rev.*, 59:331.

Serrato, J. C. S. and Zidar, O. (2014). Who benefits from state corporate tax cuts? a local labor markets approach with heterogeneous firms. Technical report, National Bureau of Economic Research.

Shaviro, D. (1990). Beyond public choice and public interest: A study of the legislative

process as illustrated by tax legislation in the 1980s. *University of Pennsylvania Law Review*, pages 1–123.

Shaviro, D. (1992). An economic and political look at federalism in taxation. *Michigan Law Review*, pages 895–991.

Shaviro, D. (2004). Rethinking tax expenditures and fiscal language. *Tax Law Review*, 57:187.

Slemrod, J. (2004). The economics of corporate tax selfishness. Technical report, National Bureau of Economic Research.

Slemrod, J. (2005). The etiology of tax complexity: Evidence from us state income tax systems. *Public Finance Review*, 33(3):279.

Slemrod, J. and Bakija, J. (2008). Taxing ourselves: a citizen's guide to the debate over taxes. *MIT Press Books*, 1.

Slemrod, J., Blumenthal, M., and Christian, C. (2001). Taxpayer response to an increased probability of audit: evidence from a controlled experiment in minnesota. *Journal of Public Economics*, 79(3):455–483.

Slemrod, J. and Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. *Handbook of public economics*, 3:1423–1470.

Solan, L. and Dean, S. (2007). Tax shelters and the code: Navigating between text and intent. *26 Va. Tax Rev. 879*, 2006:879.

Surrey, S. S. (1957). The congress and the tax lobbyist: How special tax provisions get enacted. *Harvard Law Review*, pages 1145–1182.

Tollison, R. D. (1988). Public choice and legislation. *Virginia Law Review*, pages 339–371.

Vasconcellos, R. P. (2007). Vague concepts and uncertainty in tax law: The case of comparative tax judicial review.

Warren, P. L. (2009). State parties and taxes: A comment on reed in the context of close legislatures. *Available at SSRN 1144057*.

Weisbach, D. (1999). Formalism in the tax law. *The University of Chicago Law Review*, pages 860–886.

Weisbach, D. (2002). An economic analysis of anti-tax-avoidance doctrines. *American Law and Economics Review*, 4(1):88–115.

Yu, M., Zhao, T., Dong, D., Tian, H., and Yu, D. (2013). Compound embedding features for semi-supervised learning. In *HLT-NAACL*, pages 563–568.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,

67(2):301–320.

Zucman, G. et al. (2015). The hidden wealth of nations. *University of Chicago Press Economics Books*.

A   Appendix

A.1   Vector Representation of Words and Documents

The algorithm for representing the linguistic meaning of words and phrases as data is called Word2Vec, a machine-learning model developed by Google researchers (Mikolov et al., 2013). The model is inspired by Harriss's distributional hypothesis that words in similar contexts have similar meanings. Recent work in natural language processing has made progress in representing words as dense vectors, culminating in the skip-gram with negative sampling training method, better-known as Word2Vec.

Levy and Goldberg provide an accessible introduction to Word2Vec. The model assumes a corpus of words $x_1, x_2, ..., x_n$, each drawn from vocabulary V_x. Each word is observed in an associated context, which is an ordered set of the words appearing in an l-sized window around the word: $\{x_{i-l}, ..., x_{i-1}, x_{i+1}, ..., x_{i+l}\}$. The standard window used in NLP tasks is $l = 5$, which is used in my analysis. The vocabulary of contexts (a very long list of all possible combinations of preceding and succeeding words in the corpus) is given by V_c.

Each word x has an associated vector $\mathbf{x} \in \mathbb{R}^d$, where d is the dimensionality of the word vector space. A standard choice in the NLP literature is $d = 300$, which also gives good results in this dataset. Next, each context has an associated vector $\mathbf{c} \in \mathbb{R}^d$, which plays a role in training the model but is not used further in the analysis.

In Word2Vec, an adjacency matrix of collocations, where each entry in the matrix A_{ij} is the number of times word i appears within l words of word j. This high-dimension $|W| \times |W|$ matrix is then factored into a pair of matrices of dimension $|W| \times |C|$ and $|C| \times |W|$, where the vector space C can be understood as the latent "contexts" of the word. Taking the first matrix, we get a mapping *vec* between words and points in a $|C|$-dimensional vector space. Words that are "similar" are located near each other in context space, in that they tend to be surrounded by similar words. By looking at the sequences of words that occur before and after a particular word (the "contexts" of the word), Word2Vec "learns" which other words in the vocabulary could fit into the same context.[39]

Word2Vec has several desirable features for this paper's purposes. First, it can be trained in eight hours on the corpus of statutes. Once trained, it can quickly compute similarity statistics between phrases and documents. Importantly, the vector dimen-

---

[39]See "A Word is Worth a Thousand Vectors" by Chris Moody (2015), available at `multithreaded.` `stitchfix.com/blog/2015/03/11/word-is-worth-a-thousand-vectors/`.

sions encode information about the underlying relations between words. This is why analogies work:

$$vec['\text{corpor\_incom\_tax}'] - vec['\text{corpor}'] + vec['\text{individu}'] \approx vec['\text{individu\_incom\_tax}']$$

This example shows that the word dimensions are encoding semantic information about types of taxes.

The Word2Vec model is implemented in Python's gensim package. I train the model on the processed statutes for 1963 through 2010 in random sequence. For parameters, I select $C = 300$ dimensions. This is the default and works well on Wikipedia. I choose a context window of $l = 5$, which means that Word2Vec learns relations within five words of each other. This is also the default.

In the trained model, each phrase p is represented as a vector $\vec{p} = vec[p] = \frac{1}{|words(p)|} \sum_{w_i \in words(p)} vec[w_i]$, with a value between -1 and 1 for each of $d \in \{1, 2, ..., 300\}$ dimensions. While it may appear that we lose a lot of information by taking the mean, recall that our phrases are already filtered so as to be noun phrases and word phrases. "Similarity" between a phrase p and q is computed using cosine similarity between the vectors for those phrases:

$$\text{sim}(\vec{p}, \vec{q}) = \frac{\vec{p} \cdot \vec{q}}{||\vec{p}|| \cdot ||\vec{q}||}.$$

This metric is between -1 and 1, with higher numbers meaning the words are more similar (Levy et al., 2014). For example, $\text{sim}('\text{democrat}', '\text{republican}') = 0.86$.

In future work one could analyze the vectors directly rather than the phrases. It may turn out that some dimensions encode important information for certain parts of tax law, such as defining the base, enforcement, or the style of legal writing. A document (statute) can be represented as a vector – as the mean or sum of the constituent phrase vectors. Then one could measure the effect of treatments on vector dimensions rather than phrase frequencies. In some unreported exploratory work I have found that some dimensions are correlated with higher tax revenue across bases, for example, or with changes in political party. This may provide better measures at the document level than using phrase frequencies.

A.2   Factor IV Approach

This appendix section describes the factor IV approach to estimating the first stage. As discussed in the text, I obtained similar second-stage results using this approach

(the Section 8 results looking at the effects of political control on language), but the out-of-sample PLS prediction was worse. That said, a potential problem with Lasso is if the sparsity assumption fails. This can occur if the true Γ is actually a dense matrix. Lasso will wrongly exclude many elements of Γ. If the included instruments are correlated with the excluded ones, those elements of Γ will also be inconsistent.

An alternative dimension reduction method that addresses this problem is PCA (Bai and Ng, 2010). PCA projects high-dimensional data down to a lower-dimensional space while retaining as much information as possible. Formally, PCA finds the $n \times p$ projection matrix $\tilde{Z}$ that solves

$$\min_{\tilde{Z}} ||\tilde{Z}Z - Z||^2,$$

The columns of $\tilde{Z}$ are principal components, which are orthogonal to each other and are ordered by their explanatory power for Z. Taking the first few components of $\tilde{Z}$ is a convenient way to reduce the dimensionality of Z while preserving as much information as possible. Since each row in $\tilde{Z}$ is a linear combination of a row in Z, the reduced matrix inherits any exogeneity properties of the original matrix. Moreover, the components included in $\tilde{Z}$ are orthogonal to any excluded components, which solves the exclusion issue of correlated instruments we faced with Lasso.[40]

In the baseline implementation, I included enough components to explain 90 percent of the variance in Z. In the baseline specification this required 205 components. To select among these components, I again used Lasso to estimate (6) but with the PCA components $\tilde{Z}$ as the instruments rather than the original Z matrix.

Figure 7 shows a heat map (bivariate histogram) in which each observation is a phrase-component pair (i, j). The horizontal axis is the correlation between the instrument phrase z^i and the component $\tilde{z}^j$. The vertical axis is the t-statistic for the first-stage effect of component $\tilde{z}^j$ on endogenous phrase x^i; that is, the element $\hat{\gamma}_{ij}$ in the matrix of first-stage coefficients $\hat{\Gamma}$. This shows that the components that are correlated with particular instrument phrases also tend to have stronger effects on those same endogenous phrases. This supports the idea that language diffusion is occurring through preference for phrases in the same judicial circuit.

---

[40]In the empirical analysis, PCA is implemented with Python's scikit-learn package, using the truncated singular value decomposition algorithm.

Figure 7: Instrument Phrases Have a Stronger Effect on Own Endogenous Phrase
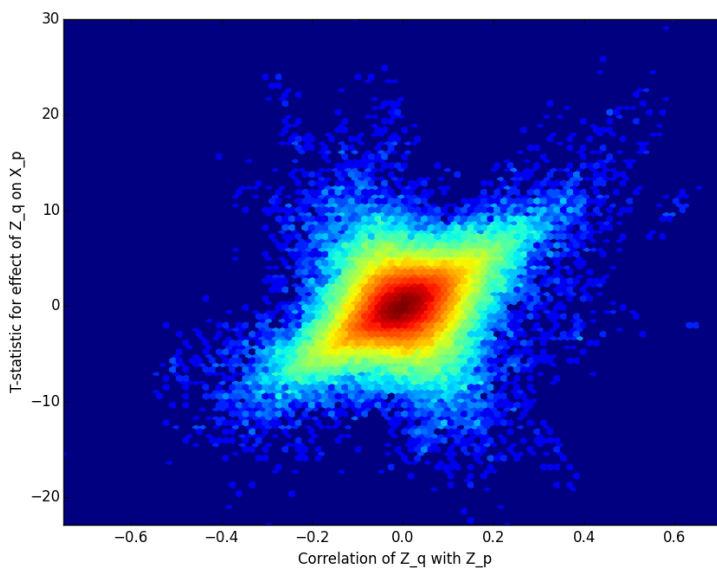
A.3 Example phrases with related court cases

A foremost issue in this paper is how tax code features are used to implement redistributive fiscal policy. Text features that have the effect of broadening the income tax base would serve to increase the progressivity of the tax code. If used preferentially by Democrats, that would be consistent with Democrats using these phrases to implement a progressive redistributive policy. Analysis of these phrases quickly procured two examples: "old age" and "fire-fighter."

The term "old age" is used in income tax provisions related to age-related exemptions. For example, 2005 Pa. ALS 40 states that "The term 'compensation' shall not mean or include: payments commonly recognized as **old age** or retirement benefits paid to persons retired from service after reaching a specific age or after a stated period of employment." In this case the term is evidence of a deduction, but the 2SLS estimate for "old age" indicates that it is associated with increased income tax revenues on average. Moreover, this phrase is associated with Democrat political control. The Commonwealth Court of Pennsylvania construed this clause in a 1987 opinion (*Bickford v. Commonwealth*, 111 Pa. Commonw. 246), finding that a pension plan from a private employer was not covered by this clause and therefore was taxable. In this case the clause did not decrease revenues generated. However, in *Pugliese v. Township of Upper*

St. Clair, 660 A.2s 155 (1995), the same court held that a similar corporate incentive plan (with a longer deferral) was exempt from taxation.

The term "fire-fighter" is also predicted to increase income tax revenue and is associated with Democrats. An example of a statute where it may appear is 2006 Al. ALS 352, providing that "the following exemptions from income taxation shall be allowed to every individual resident taxpayer: The first $8,000 of any retirement compensation, retirement allowances, pensions and annuities, or optional allowances, received by any eligible **fire-fighter**." An Alabama case construing this type of clause is *Ex parte Melof*, 735 So.2d 1172 (1999), wherein the Supreme Court of Alabama held that firefighters could be given special tax treatment in spite of a state constitutional amendment forbidding special tax treatment for public sector workers.

These cases are good examples of the indeterminacy and unpredictability of how statutory language will be construed by courts. Before these cases were decided, a researcher interested in coding the policies in these provisions would have had difficulty deciding where they would apply. The phrases demonstrate that the machine learning method can effectively identify revenue-relevant tax code language using a data-driven approach.

A.4 Decomposition of the party effect on tax revenues

This appendix uses the notation in the model (Subsection 3.2) to compute the share of revenue changes due to party control that can be assigned to the various components. In particular, although $\frac{\partial g}{\partial u_j}$ and $\frac{\partial u_j}{\partial D}$ cannot be estimated, the summation U can be computed as

$$U = \rho_g - \rho_\tau - \sum_{i=1}^{p} \beta_i \delta_i,$$

assuming the effects are separable. These estimates can be used to compute the relative importance of the tax rate, the tax code, and other policies in the implementation of fiscal policy in the U.S. states.

The estimates for ρ_g and ρ_τ were computed in Subsection 7.2. The language effect term $\sum_{i=1}^{p} \beta_i \delta_i$ was computed in Subsection 8.1. That provides enough information to compute U.

The estimates from the empirical section, along with their relation to U, are reported in Table 10. The table highlights that the tax code has a larger effect than the tax rate. The best estimates for U are -0.131 and -0.089, respectively. These unobserved policies are comparable in importance to the effect of the tax code. They suggest that

Table 10: Decomposition of Party Control Effects on Government Revenue

Effect Component	Notation	Income Tax	Sales Tax
Revenue	ρ_g	0.046	-0.176
Tax Rate	$\rho_\tau = \rho_g - \tilde{\rho}_g$	0.033	-0.019
Tax Code	$\sum_{i=1}^{p} \beta_i \delta_i$	0.144**	-0.068*
Unobserved Policies	$U = \rho_g - \rho_\tau - \sum_{i=1}^{p} \beta_i \delta_i$	-0.131	-0.089

Decomposition of political control effects on government revenue, separately for Income Tax and Sales Tax. Units are in standard deviations. Values computed in previous sections.

the unobserved policies implemented by Democrats – besides tax rates and the tax code – are associated with reduced tax revenues for both income tax and sales tax.

A.5 Substituting Phrases to Increase Tax Revenues

In this appendix, I use the machinery to try to re-write statutes to increase tax revenues. I iterate through phrases in statutes and find closely related words or phrases that are predicted to increase tax capacity. These substitutions are credible because the predicted changes in revenue are derived from the instrumental variables estimates described previously.

The method for phrase substitution works as follows. Consider a given document, which is a list of phrases, indexed by p. For each p, search the nearby words in the Word2Vec space. In these small clusters, the phrases q are closely related and sometimes synonymous. I take the first-stage F-statistics, coefficients β_q, and standard errors for each phrase in the cluster. Then I attempt to make replacements if there is an improvement in predicted tax capacity.

To find possible replacements, I look for any q such that $\beta_q > \beta_p$, then make pairwise comparisons based on the phrase statistics. First, I filter out any q with weak F-statistics. Next, let σ_p^2 and σ_q^2 equal the standard errors for β_p and β_q. I use a Wald test statistic to compute whether they are significantly different:

$$W(p, q) = \left(\frac{\beta_q - \beta_p}{\sqrt{\sigma_p^2 + \sigma_q^2}} \right)^2.$$

This test statistic follows an F-distribution. If I cannot reject the null that $\beta_q = \beta_p$,

exclude q from the list of possible replacements. If there are any phrases remaining that satisfy these criteria, I choose q that results in the largest predicted improvement in tax capacity, that is, the highest β_q.

To illustrate this approach, I analyze the 4000 phrases in the vocabulary with the highest cosine similarity to "tax." For each phrase in this subset, I assess the twenty most similar phrases. Of these twenty, I exclude any phrases with cosine similarity less than 0.5. I further skip any potential replacement where the first-stage F-statistic for p or q is below 5. Finally I exclude any proposed replacements where $W(p,q) < 10$. If all q are excluded, no replacement is made. If any q are left, I select the one with the highest β_q. This turns out to be a relatively conservative specification, resulting in about 2% of phrases replaced.

Table 2 reports the set of proposed replacements for these phrases. I have roughly organized them into groups of related phrases. As can be seen immediately, many of these replacements don't make a lot of intuitive sense. Machine-learning methods are necessarily imperfect and will pick up a lot of nonsensical relations. That said, there are some replacements recommended here that deserve a closer look. As can be seen in the predicted revenue change column, making these substitutions could result in significant increases in revenue.

In particular, the recommendation to replace "failure" with "such failure," and "person or firm" with "such person or firm," both make intuitive sense from a statutory interpretation perspective. By adding "such," these replacements increase the clarity of a tax statute and likely increase revenue by reducing avoidance.

The other replacements are not as intuitive but still deserve discussion. First, naturally enough, there are a couple of accounting-related suggestions. "Submit report" and "certified public accountant" are predicted to increase revenue, which is consistent with an effect of better record-keeping (Gordon and Li, 2009).

Next there is a collection of phrases related to businesses and corporations. Anchoring tax liabilities to the "principal place" of business rather than where the "business is located" appears to increase revenue. For the insurance-related phrases, changing phrases about insurance premiums to those about life credits seems to make a difference.

There are also suggestive phrases related to collections and enforcement. "Collect revenue" is preferred to "collect rate," while "amount of revenue" is preferred to "tax revenue." These might be better specifications of the collections process. The replacements for appeals and penalties likely reflect that the structure of these statutes have

Phrase	Replacement	Rev. ($M)
Accounting, Business, Insurance, and Debt		
annual report	submit report	0.87
audit book	certified public accountant	2.79
business	business state	2.88
business is located	principal place	4.24
corporate	corporate law	1.72
corporate limit	said city	0.87
failure	such failure	1.41
file article	certificate of incorporation	0.89
incorporate	corporate law	1.72
operating business	business	1.12
person or firm	such person or firm	2.67
person or partnership	such person or firm	2.79
purpose author	corporate purpose	3.28
bank corporate	federal deposit insurance	2.13
inheritance tax	executor	0.51
insurance premium	credit life	0.93
premium finance	credit life	1.29
state unemployment	unemployment trust	2.28
Collections and Enforcement		
anticipated revenue	adopt budget	2.17
collect rate	collect revenue	3.20
subject tax	tax	0.92
tax jurisdiction	state tax	0.76
tax revenue	amount revenue	3.25
total contribution	contributor	0.94
appeal	notice appeal	1.72
become delinquent	enforce collect	1.69
enforce	other remedy	1.43
fix penalty	penalty violation	3.76
same penalty	fail neglect	4.21
annual tax	sufficient pay interest	4.12
bond interest	payment of interest	2.36
other obligations	other obligations issued	1.49
person liability	liability	1.27
rate of interest	maximum rate	1.10
rate per annum	exceed forty year	2.75
sink	payment of interest	2.06
sink fund	payment of interest	1.45
such installments	equal installments	3.16
such rate	exceed forty year	1.70
sufficient pay	sufficient pay interest	5.87
Miscellaneous		
addition to power	have power	3.00
be in effect	be effect	2.05
code title	annotated	0.50
not inconsistent	not inconsistent herewith	4.44
nothing herein	provided however nothing	2.85

Phrase	Replacement	Rev. ($M)
Dates		
annum	cent annum	0.45
annum from date	cent month	4.33
calendar	import	0.44
day after receipt	receipt of request	3.81
expiration year	expiration date	3.09
file	file within day	2.51
final determination	file within day	2.25
first day month	date retire	2.38
first month	last day	1.47
succeed calendar	last day	3.99
such estimate	next fiscal year	2.37
sunday legal holiday	following day	3.55
thirty-first day	june year	2.82
thirty-first day december	last day	1.54
twelfth	twenty-seventh	1.17
Amounts		
cent	one-half cent	1.22
exceed sum	exceed	0.71
per	one-half	1.21
proportion	same ratio	2.56
proportion amount	same proportion	2.42
seven-tenth	cent	1.06
three-fourth		
Local Issues		
additional levy	author levy	3.07
certified state	local district	2.26
charge	fee service	1.11
county auditor	county clerk	0.37
county general fund	fund of county	5.06
county township	city village	1.3
fee be paid	fee	3.86
gas water	heat power	4.3
hospital district	director district	1.44
proposition	majority voter	1.65
question	such proposition	1.39
record deed	file for record	2.14
record in office	county record	1.85
referendum	such referendum	1.67
royalty	school land	1.14
said license	payment fee	1.64
territory	include territory	3.94
town	such town	1.15
Sales Tax		
said vehicle	motor	2.43
tax stamp	retail dealer	2.34
purchase price	trade-in	1.65

List of phrases, proposed replacements, and predicted increase in revenue due to the replacement (in millions of dollars). Top 4000 phrase that are most similar to "tax."

Table 11: Examples of Replaced Phrases

an important impact on collections.

In the debt category, the phrase "payment of interest" seems to matter a lot. This is likely related to paying interest on delinquent taxes. Changing "rate of interest" to "maximum rate" increases revenues, perhaps reflecting a higher rate of interest paid on delinquent taxes.

Another striking observation is the large number of suggested replacements for dates. This emphasizes that the timing of tax obligations is an important tool for legal avoidance (Slemrod and Bakija, 2008). The technical phrasing of statutes is important for facilitating this type of avoidance. Similarly, the technical phrasing related to amounts (e.g., "proportion" versus "same ratio") can have large impacts on tax liability.

The three other groups of phrases – Miscellaneous, Local Issues, and Sales Tax – are perhaps less intuitive. As mentioned, this is a machine learning method that will produce some results that are not useful. This emphasizes that any suggested replacements will require verification by lawyers and policymakers.