Legislative Studies Quarterly

Comparative Legislative Research Center
The University of Iowa

LUKE KEELE
Penn State University
NEIL MALHOTRA
COLIN H. MCCUBBINS
Stanford University

Do Term Limits Restrain State Fiscal Policy? Approaches for Causal Inference in Assessing the Effects of Legislative Institutions

Scholars of state politics are often interested in the causal effects of legislative institutions on policy outcomes. For example, during the 1990s a number of states adopted term limits for state legislators. Advocates of term limits argued that this institutional reform would alter state policy in a number of ways, including limiting state expenditures. We highlight a number of research design issues that complicate attempts to estimate the effect of institutions on state outcomes by addressing the question of term limits and spending. In particular, we focus on (1) treatment effect heterogeneity and (2) the suitability of nonterm-limit states as good counterfactuals for term-limit states. We compare two different identification strategies to deal with these issues: differences-in-differences (DID) estimation and conditioning on prior outcomes with an emphasis on synthetic case control. Using more rigorous methods of causal inference, we find little evidence that term limits affect state spending. Our analysis and results are informative for researchers seeking to assess the causal effects of state-level institutions.

Scholars of state politics often study how legislative institutions affect policy outcomes (e.g., Besley and Case 2003). In other words, do the “rules of the game” influence the behavior of political actors and their choices? In recent years, particular interest has been paid to the effects of legislative term limits on fiscal policy. The 1990s witnessed increased popular pressure to deprofessionalize legislatures via term limits, with a corollary claim that term limits might restrain the growth of government with the return of the “citizen legislator.” Proponents claimed term limits, like tax reform initiatives, would help put an end to wasteful government spending, reduce the amount of pork being passed through the legislature, and curtail the size and scope of the government. Specifically, term limits

DOI: 10.1111/lsq.12016

were supposed to remove the incentive for legislators to procure funds for interest groups, since legislators would no longer be beholden to lobbyists within the representative's short time in office. While term limits were never enacted at the national level in the United States, advocates were successful at implementing term limits in state legislatures, with 20 states imposing limits between 1990 and 2000.[1] We have now observed over a decade of fiscal policy outcomes since the imposition of term limits, allowing us to assess if term limits have restrained spending.

The variation in the adoption of term limits at the state level presents an obvious opportunity to comparatively study the effects of term limits on fiscal policy. However, analyzing the population of states presents challenges as well. Although the states share cultural and historical similarities, the political, economic, social, and demographic heterogeneity at the state level is enormous. Thus, one challenge we always face in comparative analyses of state legislative institutions is one of selection. That is, some states selected into this institutional reform, and the variables that predict selection may also predict budgetary decisions. For instance, if fiscally conservative states were more likely to adopt term limits, it may appear that term limits cause states to spend less. However, the results are simply due to lower-spending states selecting into term limits to begin with. Given that term limits were first championed by the conservative grassroots, it is possible that these activists were able to successfully implement term limits—primarily via ballot initiatives—with a sympathetic conservative electorate. On the other hand, it is also possible that fast-growing states (for example, those in the West and Sun Belt with initiative processes) were both more likely to implement term limits and were spending more due to economic growth. In this case, we might spuriously observe that term limits have increased the size of state budgets. Because term limits are not randomly assigned, one can construct various stories of this sort to explain away any observed relationship between term limits and spending. We detail these threats to inference formally below.

Although we are not the first to consider the question of how state legislative term limits affect fiscal policy, we approach the question with a focus on the research design needed to overcome the challenges of assessing nonrandomized institutional reforms such as term limits. Principally, we argue that it is critical to make appropriate counterfactual comparisons and to account for heterogeneous effects. We conceive of term limits as unique interventions (or treatments) that should be studied in a way that mimics the paradigm of randomized assignment but should also be partially treated as case studies. That is, we depart from extant practice and do not pool all states together in a single statistical model.

Instead, we analyze each state's adoption of term limits as a unique case study. At the same time, we attempt to mimic as-if random assignment by creating a "synthetic" control unit for each term-limited state. We detail the methodology below but briefly, the idea is to construct a control state by weighting a set of nonterm-limit states based on observed pretreatment characteristics such that the pretreatment trajectories in spending are similar between the treated state and its synthetic control (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010). This approach has two main advantages that help us rule out alternative explanations. First, the synthetic control analysis allows us to assess whether the treated and control states were on similar fiscal paths before the adoption of term limits. Second, this approach takes treatment heterogeneity seriously. A standard panel-data analysis rarely examines whether effects differ across units. Instead, we allow the effect for each state to vary. This allows us to combine the benefits of quantitative hypothesis testing with the texture afforded by a qualitative analysis. We compare this approach against the difference-in-difference (DID) method, the technique used in previous studies of state legislative term limits (and the more standard approach in the state politics and policy literature).

Applying this methodology, we find little evidence that term limits affected state budgets. For nearly all of the 14 current term-limit states, posttreatment spending levels are statistically indistinguishable between the term-limit states and their synthetic controls. That is, attempting to model the hypothetical experiment shows that *comparable* states that did and did not implement term limits exhibited similar fiscal trajectories after the treatment state adopted term limits. Moreover, in states that repealed term limits, we see no evidence that spending changed after the repeal, further ruling out the presence of a causal effect of term limits on spending. Our findings stand in contrast to existing studies that conclude that term limits significantly affected budgets.

Hence, the aim of the article is twofold. In addition to exploring a substantively important question in the state politics and policy literature, we raise an important methodological issue about research design. Pooling together all states to examine the effect of policy interventions is a good first start, but it may not be sufficient for two reasons. First, the full set of untreated states do not always represent a good counterfactual for what outcomes would have looked like if treated states did not experience the policy intervention. Analyzing all states together in a standard regression framework often masks these issues. Second, the effect of institutional structures may vary across states, and treatment heterogeneity is not easily modeled using standard approaches. Indeed, applying these techniques to studying the fiscal effects of term limits leads us to contradict many of the supposedly strong results in extant literature. Therefore, we hope this research will push scholars to employ more rigorous tests when analyzing the effects of state-level institutions.

This article is organized as follows. We first provide a theoretical overview for why we might expect term limits to affect spending. We then review the existing empirical literature on the effect of legislative term limits on fiscal policy. We then describe two different approaches for answering this empirical question with an emphasis on clearly delineating the assumptions behind each approach. Finally, we describe the data, present the results, and discuss their implications for the study of political institutions and policy outcomes.

Background and Theoretical Overview

We begin with an overview of term-limit laws and how they vary from state to state. We later argue that this variation in term-limit laws must be accounted for in the research design. We divide term limits into two broad categories: consecutive and lifetime. Under consecutive term limits, a legislator is limited to serving a particular number of years in a single chamber. Upon reaching the limit in one chamber, a legislator may run for election in the other chamber or leave the legislature. After a period of time (usually two years), the clock resets on the limit, and the legislator may run for his/her original seat and serve up to the limit again. With lifetime limits, on the other hand, once a legislator has reached the limit, he or she may never again run for election to that office, which is obviously more restrictive. States also adopted limits of varying lengths. Table 1 contains basic details about the 14 states that currently have legislative term limits.[2] Utah and Louisiana were the only two states where term limits did not become law through the initiative process.[3] Table 1 also contains a list of the six states that adopted term limits but later repealed them. The standard approach of coding term limits as a dummy variable may be misguided since the policies vary. More importantly, term limits may interact with existing state characteristics in unique ways. Thus, it may be impossible to recover a single, meaningful estimate of the term-limit "treatment effect."

Why should we expect term limits to alter state fiscal policy? The original proponents of term limits constructed loose arguments about career politicians, entrenched in government, who were more likely to support an expansion of the public sector and thus increase state spending (Ehrenhalt 1991; Fund 1990; Payne 1991; Will 1992). This logic is based on the assumption that lobbyist capture increases with each additional term. According to Bandow, "what special interests fear most is a

TABLE 1
Variation in State Term-Limit Statutes

State	Year Enacted	Year Repealed	House Limit	Senate Limit	Lifetime Ban	Who Repealed
Maine	1993	–	8	8	No	–
California	1990	–	6	8	Yes	–
Colorado	1990	–	8	8	No	–
Arkansas	1992	–	6	8	Yes	–
Michigan	1992	–	6	8	Yes	–
Florida	1992	–	8	8	No	–
Ohio	1992	–	8	8	No	–
South Dakota	1992	–	8	8	No	–
Montana	1992	–	8	8	No	–
Arizona	1992	–	8	8	No	–
Missouri	1992	–	8	8	Yes	–
Oklahoma	1990	–	12	12	Yes	–
Louisiana	1995	–	12	12	No	–
Nevada	1996	–	12	12	Yes	–
Idaho	1994	2002	8	8	No	Legislature
Massachusetts	1994	1997	8	8	No	State Supreme Court
Oregon	1992	2002	6	8	No	State Supreme Court
Utah	1994	2003	12	12	No	Legislature
Washington	1992	1998	6	8	No	State Supreme Court
Wyoming	1992	2004	12	12	No	State Supreme Court

continuing influx of freshmen, who neither know nor care to learn the rigged rules of the game" (1995). The political science literature on legislative institutions can assist us in fleshing out this logic. Careerism may require increased spending due to the need to build constituency support needed for reelection (Fiorina 1989; Mayhew 1974). Accordingly, severing the "electoral connection" may reduce wasteful spending that arises due to "common pool" problems inherent to geographical districting (Weingast, Shepsle, and Johnson 1981). Additionally, term limits weaken the bargaining power of state legislatures relative to the governor (Kousser 2005). Since governors are state oriented and not district oriented, they have less incentive to incorporate geographically targeted distributive benefits into the budget.

Further, as Bandow (1995) and Kousser (2005) explicate, the reduction of spending may partially be due to how term limits deprofessionalize legislatures. Experienced, professional representatives may better understand the inner workings of the legislature and as a result more efficiently navigate the pitfalls that might otherwise prevent them from passing legislation. The presence of term limits prevents representatives from learning how to pass the types of legislation that would increase spending. Professionalism may also moderate any effects of term limits. Citizen legislatures have always experienced high levels of turnover (Squire 1988); therefore, we might expect term limits to have the smallest effects in these states. Conversely, term limits greatly disrupted professionalized legislatures by substantially increasing turnover and limiting advancement opportunities (Kousser 2005). If the goal of term limits was to deprofessionalize legislatures, then citizen bodies have in effect already been "treated" before the imposition of term limits.

There is also empirical evidence that supports these theoretical arguments. Surveys of state legislators show that those in term-limited states spend less time worrying about bringing home pork (Carey, Niemi, and Powell 1998, 2000). Similarly, Moore and Steelman (1994) find that those spending a longer time in the legislature support greater spending (although see Aka et al. (1996) for an alternative account). Bandow (1995), Cain and Kousser (2004), and Kousser (2005) have all concluded that term limits have altered legislator incentives such that spending bills include fewer particularistic funds targeted at interest groups.

However, there are also reasons why we might expect term limits to increase spending. By weakening and deprofessionalizing the legislature, term limits may make legislators more reliant on interest groups and lobbyists that push for greater particularistic spending (Kousser 2005). Additionally, by severing the electoral connection, term limits may produce less competent legislators both through selection and treatment effects. With respect to selection, term limits hinder accountability by reducing the incentives for good performance (Alt, de Mesquita, and Rose 2011). With respect to treatment, term limits also prevent members from gaining experience and learning on the job. Lower competence may produce more inefficient budgets. The relationship between term limits and spending may also vary by the type of spending. Herron and Shotts (2006) show that when pork is very socially inefficient, term limits may counterintuitively increase spending. If the representative wants to reduce future pork in other districts, term limits decrease the incentive to reduce future spending moderation. Herron and Shotts (2006) also suggest that there may be heterogeneous effects according to the type of spending. For instance, term limits may strongly impact line items where lawmakers may have more discretion (e.g., transportation spending).

Of course, these mechanisms are not mutually exclusive. In fact, they could counteract one another to produce no net increase in spending. Also, there may be substitution effects. If pork is a cheaper credit-claiming device and term limits reduce the incentive to claim credit, then spending may just be transferred from geographically based expenditures to more general forms. Thus, the direction of the treatment effect is uncertain.

Previous Empirical Evidence

We focus on two recent articles that have suggested that legislative term limits increase spending and harm state fiscal performance. Both articles: (1) pool states in regression analyses and (2) apply some technique of causal inference. Analyzing state fiscal data from 1977 to 2001, Erler (2007) finds that states with term limits have higher spending than states without them. This analysis employs difference-in-difference to identify the causal effect of term limits. Erler (2007) estimates that the imposition of term limits increased state spending by \$36.80–\$59.80 per capita (depending on the specification) and that these estimates are very reliable with *t*-statistics ranging from 2 to 3. For some line items, the effect is up to \$677 per capita. However, this approach cannot account for spuriousness due to time-varying confounders. For example, if states with high economic growth—and an increasing public sector to accommodate that growth—were the ones that adopted term limits, then a difference-in-difference approach will fail to uncover the causal effect, particularly if economic growth is not measured precisely.

Lewis (2012) finds that term limits decrease state bond ratings, a measure of state fiscal performance. In addition to estimating standard regression models pooling states and years, Lewis attempts to resolve the causal inference problem via instrumental variables (IV) regression. Three variables are used as instruments for the presence of term limits in a state: (1) whether there was a term-limits ballot measure; (2) party competition prior to the implementation of term limits; and (3) citizen ideology prior to the implementation of term limits. An important requirement for an IV estimator is that the "exclusion restriction" holds. If the exclusion restriction does not hold, then the estimates from the IV estimator can be severely biased. As Sovey and Green note in their primer on instrumental variables for political scientists: "For Z_i (the instrument) to be valid, however, it must transmit its influence on the outcome solely through the mediating variable X_i" (2011, 189–90). In other words, the instrument has to be plausibly "random" in that it is not correlated with the error term. Valid instruments include random assignment in experiments where there is noncompliance (e.g., Gerber and Green 2000). In the absence of random assignment, the author must explain why we should expect the exclusion restriction to hold. As Sovey and Green explain: “It should be noted that the plausibility of the exclusion restriction hinges on argumentation; it cannot be established empirically” (2011, 190).

Lewis does not justify why these instruments satisfy the exclusion restriction. It appears that all three clearly violate the exclusion restriction. The initiative process (as shown by California’s example) can have a direct negative effect on fiscal performance unrelated to term limits because the public often votes for inflexible budget constraints that worsen deficits. Party competition may also have a direct effect on fiscal performance if the lack of a strong opposition party disincentivizes good management or if there is a relationship between partisan control and fiscal performance (Bartels 2008). Finally, citizen ideology affects state expenditures through many means besides term limits. A more conservative citizenry likely elects conservative legislators, who will push for lower spending regardless of whether term limits are in place. Hence, because the exclusion restriction (even for a single instrument) is violated, the IV estimates are biased. This is why IV estimators, which may be intuitively appealing, are extremely difficult to implement outside of randomized experiments. Unless the assumptions are clearly met, IV estimators can do more harm than good (Angrist and Pischke 2009).

Other articles have examined the related question of whether gubernatorial term limits alter state spending. Besley and Case (1994) find that per capita spending was higher under term-limited governors but later found that this effect has declined over time (Besley and Case 2003). Alt, de Mesquita, and Rose (2011) also found that gubernatorial term limits increased spending and decreased fiscal performance. In an analysis that examines state legislative experiments as case studies, Kousser, McCubbins, and Moule (2008) show that state tax and expenditure limits are largely ineffective and that state officials often circumvent them. As explained in the next section, our research design has important advantages over those in the extant literature.

Research Design and Methodological Overview

Here, we are interested in estimating a parameter that reflects the amount of state-spending changes due to the presence of term limits. Our main concern is finding a research design that provides us with a plausible identification strategy. Informally, a parameter is said to be “identified” if the confidence interval for that parameter shrinks to a single point as the sample size increases to infinity. All research designs, at least implicitly, make an assumption or a set of assumptions about identification. Only when these identification assumptions hold can we give estimated quantities causal interpretations. Outside of experiments, we must invariably rely on strong untestable assumptions. Below, we introduce the notation we use to formalize our identification assumptions. We then outline two different research designs that we might use for identification.

One way to understand the issues that underlie identification is to use the potential outcomes framework (Holland 1986). The potential outcomes framework aids in defining the correct counterfactual for any statistical analysis. Here, we think of term limits as a treatment that is administered at the state level, and we wish to observe whether the treatment changes state fiscal policy. More formally, let there be $J+1$ states that could receive a treatment, and let Y_{it} be the potential outcome (i.e., level of expenditure) for state i at time t. Each state has two potential outcomes. The first potential outcome is Y_{it}^T if state i is exposed to the intervention at time T_0. The second potential outcome is Y_{it}^C if state i is not exposed to the intervention at time T_0. Here, we assume that only the first state of the $J+1$ states receives the treatment. D_i is an indicator variable that is 1 if state i received the treatment but is 0 otherwise. We define the observed outcome as a function of the potential outcomes and observed treatment status: $Y_{it} = D_i Y_{it}^T + (1 - D_i) Y_{it}^C$.

Let $\mathbf{X}_{it}$ be a matrix of observed and unobserved pretreatment characteristics for all $J+1$ states. If $\mathbf{X}_{it}$ is independent of Y_{it}^T and Y_{it}^C, then at T_0: $Y_{it}^T = Y_{it}^C$. Assuming this is true,

$$\alpha_{it} = Y_{it}^T - Y_{it}^C, \tag{1}$$

and α_{it} is the unit causal effect for state i at time T_0+1 if the unit is exposed to the treatment. If we assume linearity and additivity, we can also rewrite Equation (1) as

$$Y_{it}^T = Y_{it}^C + \alpha_{it} D_i. \tag{2}$$

Based on this equation, we wish to estimate: $\alpha_{1,T_0+1}, \ldots, \alpha_{n,T_0+t}$.

As is always the case when estimating causal effects, we face a missing data problem. For the treated state, we do observe the outcome before and after the intervention at T_0, but we do not observe Y_{it}^C at T_0+1 for this state since it is a counterfactual quantity: it is how a treated state would behave if it did not receive the treatment. If we could conduct an experiment, we would randomize the application of D_i such that some states receive the treatment and some did not. Randomization of the treatment ensures that the outcomes are independent of the treatment, or in formal notation: $\{Y_{it}^T, Y_{it}^C \perp\!\!\!\perp D_i\}$. Under a randomized treatment, the inference is straightforward since the observed and unobserved baseline variables contained in $\mathbf{X}_{it}$ for both treatment and control groups will balance in expectation. That is, states in the control group should be no different from states in the treatment group other than differences due to random error. As a consequence, the treatment would be independent of these baseline variables. For example, the average expenditure among states that did not receive the treatment would serve as an estimate of Y_{it}^C, and using least squares, we could simply regress the outcome on D_i for a consistent estimate of α_{it}. Observational data, however, produces a variety of complications. The problem is that while Y_{it}^T is observed, Y_{it}^C is not. But to estimate α_{it}, we must also estimate Y_{it}^C. It is accurate estimation of Y_{it}^C that will identify our estimate of the treatment effect, α_{it}. One could argue that identification is a particular challenge with state-level data. For example, to estimate the term-limit treatment effect for California, we must find a state or set of states that will serve as the counterfactual (i.e., California without term limits). This is not a quantity that is easily imagined. Next, we outline two strategies for estimating Y_{it}^C. These two identification strategies rely on important assumptions about what is a valid counterfactual. We first review difference-in-difference.

Difference-in-Difference

Perhaps the most common identification strategy used when dealing with large aggregate units like states is that of differences-in-differences. The logic behind the difference-in-difference estimator is based on using fixed effects to make the treatment and control groups as similar as possible. Consider the following fixed effects model for Y_{it}

$$Y_{it} = \phi Z_{it} + \delta_t + \alpha_i + v_{it}, \tag{3}$$

where

$$D_{it} = \begin{cases} 1 & \text{if state receives treatment in period } t, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The terms δ_t and α_i denote time, and unit-specific effects, respectively, and v_{it} is an individual-transitory shock that is mean zero in each time period. The unit-specific effects can be eliminated through first differencing:

$$\Delta Y_{it} = \phi Z_{it} + (\delta_t - \delta_{t-1}) + v_{i1} - v_{i0}. \tag{5}$$

If we restrict the model to only two time periods and so long as treatment only occurs in the second time period such that $Z_{i1} = 0$ for all units in period 1 and $Z_{i2} = 1$ for the treated and 0 for the nontreated in period 2, we can drop the t subscript from the last equation and estimate

$$\Delta Y_{it} = \phi Z_{it} + \delta + v_i, \tag{6}$$

using ordinary least squares (OLS). Here the treatment effect is

$$\hat{\phi} = \Delta Y^{Z=1} - \Delta Y^{Z=0}. \tag{7}$$

This is called the difference-in-difference model since one estimates the over-time change in control and treatment groups and then takes the difference of the two over-time differences. One can show that the DID estimate of the treatment effect can be estimated with the following OLS regression:

$$Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 D_i + \beta_3 T_{it} \times D_i + \varepsilon_{it}. \tag{8}$$

In the above equation, $T_{it} = 1$ in the posttreatment period and $D_i = 1$ if the unit is in the treatment group. Therefore $D_{it} \times T_i$ equals 1 for treated units in the posttreatment period, β_3 is the DID estimate of the treatment effect, and ε_{it} is a disturbance term (Cameron and Trivedi 2005). The difference-in-difference model is a significant improvement over a simple regression since it allows for the presence of unobserved, *time-invariant* confounders, but it also comes with an important assumption.

To identify the treatment effect in the DID model, we must assume the following

$$P(D_i(1) = 1 \mid v_{it}) = P(D_i(1) = 1). \tag{9}$$

This assumption implies that for $t = 0, 1$ the unobservables $v_{i1} - v_{i0}$ are mean independent of $D_i(1)$ (Abadie 2005). What does this assumption imply about the empirical process under observation? Assuming Equation (9) holds, the analyst must assume that the differences between treatment and control would have stayed constant in the absence of treatment. That is, absent the treatment, the average outcomes of the treated and control groups would have followed *parallel* paths over time. In short, the dynamics for one group cannot differ from the dynamics for the other group. We can make a conditional identification assumption if we observe pretreatment **X** variables that are related to the outcome dynamics. Typically such adjustments are made by entering the time-varying **X** variables into Equation (8). Equation (9) is needed to identify the DID estimate as a valid causal-effect estimate and cannot be relaxed. This assumption is untestable with any configuration of the data.

One method for evaluating the parallel-paths assumption is to plot the treated and control outcomes in the pretreatment period. If the paths of the treated and control outcomes appear to be roughly parallel before the treatment takes effect, that provides indirect evidence that the assumption is plausible. If, however, the trends in the pretreatment period diverge significantly, that obviously is indirect evidence against the assumption.

The DID approach is widely used when trying to estimate causal effects for state-level treatments and outcomes. The DID method does, however, have a number of limitations. Under the DID approach, we assume that controlling for state and year means creates appropriate counterfactuals for treated states, but there is no obvious diagnostic for understanding whether the counterfactual comparisons in this approach are really appropriate.

Second, the DID model has limited means of accounting for treatment heterogeneity. That is, we might expect that the effects of term limits are not the same for all treated units. There are good reasons to think this may be the case. First, there are key differences in term-limit types that we might expect to induce treatment heterogeneity. As we noted in Table 1, some states chose to make term-limit bans for life. That is, once a legislator hits the term limit, he or she cannot run for state legislative office again. In other term-limit states, however, legislators might switch back and forth from the upper and lower chamber and avoid having to leave state legislative politics for good once the limit has been reached. In the second case, we might expect term limits to be far less successful at changing state fiscal policy. Under this form of term limits, legislators may be able to maintain spending as they simply use their new position in either the upper or lower chamber to maintain spending as before. Table 1 also shows differences in the number of terms legislators are allowed to serve in term-limit states. We might expect a negative treatment effect to be larger in states which allow fewer terms. Second, once term limits are passed they become part of a variety of preexisting institutional arrangements—such as budget limits—that may alter how even similarly structured term limits operate within each state. Further, upon adoption of term limits, states may have preexisting budgetary or economic climates that may interact with the institutional reform. Finally, as mentioned above, term limits may have a greater influence in professionalized legislatures where turnover is already low preintervention.

While the basic DID model can allow for some heterogeneity by interacting the treatment-effect indicator with each of the fixed effects, this often fails in practice due to an incidental parameters problem. The DID approach cannot typically estimate effects for each state individually, especially while controlling for time-varying covariates. However, doing so provides nuance that can be lost in estimating a regression pooling all states together.

Finally, recent work in economics demonstrates that variance estimation for DID models presents a number of challenges (Donald and Lang 2007). There is little difficulty if we can assume each of the units in the DID model is generated with independent and identically distributed (IID) data. With state-level data, however, we must assume that the data are not IID since there should be strong within-state correlations. When such clustering is present, the asymptotics for DID variance estimation rely on the number of groups and not the number of observations used for each group mean (one of the motivations behind panel-corrected standard errors; Beck and Katz 1995). If there are within-group correlations, then the reported *t*-statistics will be too high, which of course implies that one will tend to find an effect of the treatment even if none exists (Donald and Lang 2007). Donald and Lang (2007) also note that corrections such as the cluster command in Stata or other sandwich variance estimators also depend on a large number of groups, and when the number of groups is small, the distribution of the test statistic for the treatment effect is generally unknown. In the context of states, we have perhaps 50 unique data points but often less when trying to account for heterogeneity.

Another solution is to use a large number of units across longer time periods. However, Bertrand, Duflo, and Mullainathan (2004) demonstrate that standard corrections for serial correlation are ineffective for the DID model. They find that a block bootstrap is required to fully correct for serial correlation. Of course using a larger number of units in the DID model increases the risk of causal heterogeneity. As such, variance estimation in DID models present serious challenges.

Identification Based on Past Outcomes

Next, we outline an alternative to the DID approach. This identification strategy can be implemented in two ways: either with standard regression models or with a nonparametric approach called synthetic case control. This approach, like DID, relies on a specification assumption, but here the specification assumption takes a very different form. With the most common specification-based identification strategy, we assume that we observe all relevant covariates that account for treated and control differences other than the treatment (Barnow, Cain, and Goldberger 1980). We can write this assumption formally as

$$Y_{it}^{T}, Y_{it}^{C} \perp\!\!\!\perp D_i \mid \mathbf{X}.$$

Under this assumption, potential outcomes are independent of treatment once we condition on observed covariates. This is a specification assumption since it depends on correctly specifying a statistical model so that the potential outcomes are independent of treatment.

We can alter the specification assumption in the following way to make it more plausible:

$$Y_{it}^{T}, Y_{it}^{C} \perp\!\!\!\perp D_i \mid \mathbf{X}, Y_{it-1}.$$

Here, we condition on not only relevant covariates but also on at least one lag of the outcome. Why does it make sense to also condition on past outcomes? Past outcomes are a function of both unobservables as well as observable covariates. Thus, conditioning on past outcomes allows us to indirectly condition on unobservables. The simplest way to implement this identification strategy within a statistical framework is to estimate a regression model with a treatment indicator, lag (or lags) of the outcome, and other relevant covariates on the right-hand side of the model. The difficulty with a regression approach is that it requires strong functional form assumptions, and it is difficult to judge whether the estimated counterfactual fits the treated units prior to treatment. Next, we discuss a nonparametric method for estimating treatment effects that uses past outcomes for identification. This method is well suited to contexts where the units of analysis are states.

Synthetic Case Control

Synthetic case control is a statistical method designed to estimate treatment effects when the study units are large aggregates likes states (Abadie, Diamond, and Hainmueller 2010, 2011; Abadie and Gardeazabal 2003). Here, we estimate Y_{it}^{C}, the counterfactual, with a weighted combination of states that do *not* have term limits. This weighted average of states is meant to serve as a synthetic state without term limits that we can compare to a state with term limits. The logic is that while no single control state may be a good counterfactual for a treated state, we can leverage multiple control states and weight them appropriately to construct a better counterfactual that more closely approximates the treated state. In short, the assumption is that a treated state can be much more accurately approximated by a weighted combination of untreated states than by any single untreated state alone or the unweighted universe of untreated states.

More specifically, let J be the number of states that do not have term limits, and $\mathbf{W} = \{w_1, \ldots, w_j\}'$ be a vector of nonnegative weights which sum to 1. Each scalar element of $\mathbf{W}$ is the weight for state j in the synthetic control. Different weights in $\mathbf{W}$ produce a different synthetic control, so the weights must be chosen so that the synthetic state most closely resembles the treated state before the adoption of term limits. Let $\mathbf{X_1}$ be a $K \times 1$ vector of predictors for state fiscal policy before the term-limit intervention for the state that adopts term limits. Next, $\mathbf{X_0}$ is a $K \times J$ matrix of the same predictor variables for the J states that do not have term limits. To construct the synthetic control, we minimize $(\mathbf{X_1}\text{–}\mathbf{X_0W})'\mathbf{V}(\mathbf{X_1}\text{–}\mathbf{X_0W})$, where $\mathbf{V}$ is a diagonal matrix with nonnegative elements that reflect the relative importance of the different predictors of expenditures. This minimization is subject to the following constraints: $w_j \geq 0\ (1 = 1, 2, \ldots, J)$ and $w_1 + \ldots + w_j = 1$. These constraints prevent extrapolation beyond support in the data. The result is a vector of optimal weights $\mathbf{W}^*$ which defines the combination of nonterm-limit states which best resemble a particular state with term limits before that state adopted term limits.

Next, we define $\mathbf{Y_1}$ as a $T \times 1$ vector of a measure of state fiscal policy for a state with term limits, where T represents the number of the time periods (i.e., years) under observation, and $\mathbf{Y_0}$ is a $T \times J$ matrix of the same outcomes for the J nonterm-limited states over the same time period. We approximate the over-time path of state fiscal policy for the synthetic control state for comparison to the treated state in question. This counterfactual outcome, here, expenditures for the synthetic state is $\mathbf{Y_1}^* = \mathbf{Y_0W}^*$. Once the counterfactual is formed, we simply compare $\mathbf{Y_1}$ and $\mathbf{Y_1}^*$. Much like in a classic event study, differences between the two within an event window following the enactment of term limits is evidence of a treatment effect, while no differences between the two is evidence against a treatment effect.

The synthetic case control is a nonparametric estimation method. Like other nonparametric estimation methods, such as nonparametric regression, the output after estimation is graphical. To assess whether the treatment had an effect, we compare, in a plot, the over-time level of the outcome before and after the period that a state adopted terms along with the outcome for the counterfactual synthetic state. The outcomes for the pretreatment time period should be approximately the same in the plot (if not, then the method has been unable to find an appropriate counterfactual). If adoption of term limits changes the level of the outcome, the treated unit should diverge from the synthetic control unit after term limits have been adopted. We can also summarize the treatment effect in a more parametric fashion by estimating the average mean or median difference across treated and control outcomes in the posttreatment time period. That is, we take the difference between the treated and synthetic outcomes in each posttreatment period and then take the average for this set of differences.

As outlined above, conditioning on preintervention outcomes allows us to indirectly condition on unobservables, since we know that preintervention outcomes are at least in part a function of unobservables. One worry is that the effect of unobservables may evolve over time. The solution is to condition on a greater number of preintervention outcomes to capture any temporal evolution in those effects. Abadie, Diamond, and Hainmueller (2010) use a linear factor model to argue that as the number of preintervention periods becomes large, the bias in the synthetic control estimator goes to 0. In sum, identification hinges on the insight that only states that are both alike in terms of observed covariates *and* in terms of preintervention outcomes should produce similar temporal paths on the outcome over extended periods of time. Thus, unlike with DID, we can account for time-varying effects in the unobservables.

This identification strategy also has a testable assumption. If identification holds, then $Y_{it}^T = Y_{it}^C$ should hold at T_0 and other earlier time periods. That is, we can verify that the counterfactual state is nearly identical to the treated state *before* treatment. Only if $Y_{it}^T = Y_{it}^C$ holds at T_0 can we hope that some confounder is not the cause of posttreatment differences. If we observe that the treated state and the synthetic control unit have similar behavior over extended periods of time prior to the intervention, we can be more confident that a difference in the outcomes following the enactment of term limits serves as evidence that the intervention altered the treated unit's fiscal policy.

Inference. Standard large-sample inferential techniques are poorly suited to analyses of aggregate data like states since the data are not a random sample drawn from any known population and have strong within-unit correlations. Moreover, the sample sizes are small since we are confined to a subset of the 50 states. For these reasons, inference for the synthetic control method proceeds via a series of placebo tests akin to permutation or randomization tests (Abadie, Diamond, and Hainmueller 2010; Abadie and Gardeazabal 2003). For the placebo test, one of the states in the control group from which the synthetic unit is constructed is used as the treated unit. We then apply the synthetic control method to this unit, repeating this process for every unit in the control group. This gives us a set of estimates that compare the control units to one another, creating a set of placebo estimates. We then compare all the placebo estimates to the estimate for the true treated unit, which allows us to observe whether the outcome for the treated unit is large or small relative to the estimates for all of the states in the control group. If the treated-unit effect is large relative to the estimates from the control units alone, we might conclude that the treatment altered the outcome. The result is an approximately exact inference regardless of the number of control units or the number of time periods.[4] To avoid difficulties with poor-fitting placebo estimates, we use the ratio of the preintervention mean-squared prediction error (MSPE) to the postintervention MSPE as a test statistic for each unit (Abadie, Diamond, and Hainmueller 2010). That is, one way to evaluate the gap between a term-limit state and the estimates obtained from the placebo runs is to look at the distribution of the MSPE ratios in the post/preterm-limit time periods. If the MSPE ratio for the treated state is large compared to the placebo units, the probability of observing such an outcome should be quite small. This allows us to calculate an exact *p*-value for the estimate of each treated state. Given the rather small sample size, we use the more liberal 0.10 threshold rather than the standard 0.05 level before deciding whether we are able to reject the null hypothesis, as the small samples may not have sufficient power to make inferences at the more stringent significance level. However, given that the relationship between term limits and expenditures is theoretically ambiguous, we apply two-tailed tests.

Lags and DID

One obvious alternative to using either lags or DID in isolation is to mix the two strategies. It would seem relatively simple to use fixed effects but also include one or more lags into a regression model. Unfortunately, OLS estimates are inconsistent in this setting (Nickell 1981). Moreover, while various solutions exist, each requires fairly strong assumptions for identification that are often hard to justify. Angrist and Pischke (2009), however, note an interesting bracketing property of these two approaches. They prove that if the past outcomes model is correct, but the analyst uses DID, the estimate of the treatment effect will be too large since the unestimated lag parameter will be additive with the treatment effect through the error term. Conversely, if the DID model generated the data, and the analyst estimates a model that conditions on past outcomes, this will generate a correlation between the treatment and the lagged outcome which will bias the treatment effect downward.[5] One can, therefore, view the estimates from these two models as bounding the causal effect of interest. We can use these bounds along with the synthetic case-control results to assess the reliability of the various estimation approaches.

Analysis Plan and Data

As we outlined above, we have two different approaches to identification for estimating the effect of term limits. In our analysis, we attempt to leverage agreement across methods. To that end, we begin the analysis with DID estimation. We complement the DID estimates with regression-based estimates that include two lags of the dependent variable on the right-hand side of the model. Assuming the functional form of the models is appropriate, the DID and lagged outcome estimates provide bounds on the term-limit treatment-effect estimate. Next, we conduct an analysis with synthetic case control. For this analysis, we present selected fully nonparametric results via figures. The full set of nonparametric results is presented in the online appendix. Parametric summaries based on the analysis output is presented in tables.

The data we use in these analyses were gathered from various sources including the *Statistical Abstract of the United States*, *State Government Tax Collections*, *Fiscal Survey of States*, and *The Book of the States* for years 1977 through 2007. The outcome measure we focus on is total state-government expenditures, which we measure in real per-capita terms.[6] We also examine several other outcome measures. These measures include spending by subcategories including education, health and hospitals, transportation, and welfare, as well as expenditures as a percentage of state income.

For both methods (DID and synthetic case control), we include a series of covariates based on prior empirical research that explain variance in levels of state spending. These measures include: state population, yearly population growth in percentages, population density, gross state product per capita, total federal grants per capita, the state unemployment rate, the number of federal civilian employees per capita, the number of federal military employees per capita, total state and local government employees per capita, number of seats in the upper and lower chambers, state level Republican share of the two-party vote in presidential elections, partisan control of the state government, the Squire index of legislative professionalism (as reported in Squire 2007), whether the state has the initiative process, whether the state has a tax and expenditure limit, and whether the state has a debt limit. A more detailed description of the data can be found in the appendix. For the regression-based estimates, these variables are used on the right-hand side as controls. For the synthetic case-control models, these measures are used to create the synthetic control unit for each treated state.

While these explanatory variables are important, we expect the past outcomes to play an important role in identification for the synthetic case-control method. As such, we used lagged outcomes in the estimates presented below. For the regression-based estimates, we include two lags of the outcome on the right-hand side of the model. For the synthetic case-control analysis, we used all lags with the exception of the last two before the enactment period.

Alaska and Hawaii are both omitted from the dataset, as is conventional in the literature due to their unique fiscal arrangements (Erler 2007; Gilligan and Matsusaka 1995; Primo 2006). As mentioned above, we also dropped Nebraska because it is missing data on key covariates as it has a unicameral, nonpartisan legislature.

An important point to note is that we define the initiation of term limits based on when they are adopted, not when the first set of legislators is term limited out of office. We follow previous research arguing that legislators are forward-looking and change their behavior in an anticipatory fashion (Carey, Niemi, and Powell 2000; Erler 2007). From a research design perspective, it is much better to err on the side of coding term limits as being in place earlier rather than later. This is because if there is indeed a treatment effect that occurs earlier and the treatment is specified to take place on a later date, then the researcher has then conditioned on posttreatment values, which can induce bias (Rosenbaum 1984). Nonetheless, we also replicated all analyses using the year the first legislators departed due to term limits to mark the beginning of the treatment period. However, we could not include all term-limit states in this analysis because there are sometimes not enough posttreatment years following the implementation date.

TABLE 2
The Effects of Term Limits on State Expenditures
Via Regression Models

DID Estimate No Controls	Lag Estimate No Controls	DID Estimate With Controls	Lag Estimate With Controls
–22.69	0.97	–5.56	2.14
(11.94)	(1.90)	(10.17)	(3.18)

Note: Models with lags include two lags of the dependent variable on the right-hand side of the regression model. Standard errors clustered by state in parentheses. Models with covariates include all measures listed in the Analysis Plan and Data section. Outcome is total expenditures in real per capita dollars.

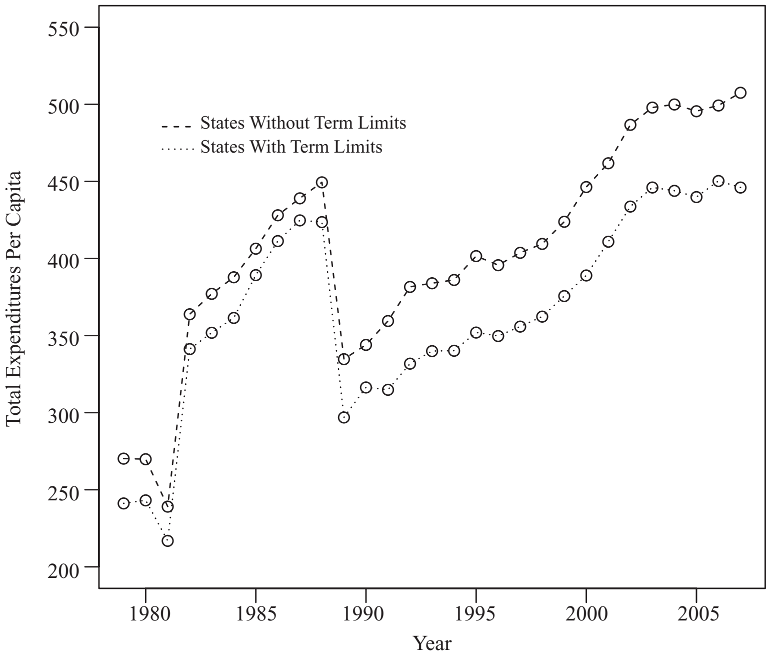
Results

We begin by presenting the regression-based estimates. Table 2 contains estimates based on regression models with a DID specification as well as models with two lags of the outcome on the right-hand side. We present the term-limit effect estimate based on a model without the control variables and a model with all the control variables included. We estimate the model both ways in order to understand how much the estimate depends on the basic DID or lag specification as opposed to control variables.

We observe that only the DID treatment-effect estimate without covariates approaches statistical significance ($p = 0.064$), implying that term limits reduced spending by \$23 per capita. However, as we discussed above the DID estimate might be too large (in magnitude), while the estimate based on lags might be too small. Thus, we interpret the two estimates as a bound on the true causal effect. We see that the estimate based on lags implies that term limits increased spending by nearly \$1 per capita. This implies that the bounds on the term-limit effect are –\$23 per capita and \$1 per capita. Hence, we cannot conclude that term limits altered spending since the bounds bracket 0. As shown in the third and fourth columns of Table 2, the estimates from the models with the full specification are slightly different, but still bracket 0. These estimates, however, do not account for possible treatment heterogeneity. Within the DID specification, we attempted to account for heterogeneity by interacting the treatment effect with the fixed effects. This failed, however, due to high collinearity in the models.

Before turning to the synthetic case-control results, we present some diagnostic evidence for the DID estimates by plotting past trends in

FIGURE 1
Pretreatment Trends for Real Per Capita Expenditures



spending across states with and without term limits across the time period of interest. If the trends in spending are roughly parallel before term limits go into effect, that will strengthen the plausibility of the DID estimates. We simply averaged total expenditures within the treated and control categories by year. As shown in Figure 1, there is no sign of any divergence in spending levels across states with term limits and states without term limits before the enactment of term limits up until 1989. This lack of divergence does lend some credence to the DID estimates. However, the general pattern undermines the broad causal hypothesis that term limits decreased spending. We see an obvious gap in spending levels across the treated and control groups in 1989, but this would appear to be too soon for a plausible term-limits effect. In 1990, three states enacted term limits, but most states did not enact term limits until 1992 or later. This would imply that state legislators in these three states decreased spending in the year before enactment enough to cause the large decrease observed in Figure 1. This also suggests that using synthetic case control to account for treatment heterogeneity is worthwhile.

We present results from the synthetic case-control results in two ways. In Table 3, we present a parametric summary. These tables contain an effect estimate which is the mean difference between the treated unit and its synthetic control in the posttreatment period. We also include the *p*-value from the placebo test and the mean-squared prediction error in the pretreatment period. The MSPE from the pretreatment period provides us with a basic form of model fit, since the smaller this number is, the better the synthetic control approximates the treated unit. The tables contain this information for all six outcomes. Synthetic control plots for all analyses as well as the complete set of weights are included in the online appendix.

We begin with a discussion of the effect on total expenditures. First, we note that some states are better approximated by the synthetic control unit than others. As an example, we compare the MSPE for Arkansas and Maine, where the values are 20.89 and 5.20, respectively. So here, Maine has a pretreatment fit to the synthetic control that is approximately four times better compared to Arkansas. As such, it is much easier to approximate Maine as a weighted combination of other states than it is to approximate Arkansas. Thus, we would be much more confident in any effect we observe for Maine than for Arkansas. While we generally observe negative estimates of the effects of term limits on total spending, that is not always the case as Arkansas, Missouri, and Ohio on average had higher spending than their synthetic controls. The size of the negative effects varies substantially as well. While a number of estimates suggest that total spending fell by approximately $17 per capita, several other states have estimates that are over $40 per capita.

We find that only Arizona and Nevada have estimates that are statistically significant. As explained earlier, we conducted placebo tests in order to make inferences for each treated state and obtain exact *p*-values. To reprise, we apply the synthetic case-control method to the control states one by one, where each control state becomes the treated unit in the analysis. These placebo estimates should be 0 since the control states never implemented term limits. This provides us with a distribution where we can then compare the estimated effect to these placebo estimates. If the treated estimate in this distribution is substantially different than the other cases (i.e., it displays one of the largest effects in the distribution), then we can conclude that the treatment has had an effect. Conversely, if the treated estimate is not unusual compared to the placebo estimates, then we conclude that the effect is not distinguishable from 0. These exact *p*-values have a specific interpretation. Take Louisiana, for example, where the exact *p*-value is 0.30 in Table 3. If one were to relabel the term-limited state in the dataset at random among Louisiana and the

TABLE 3
The Effects of Term Limits on State Expenditures by Category

Effect from Date of Enactment	Total Expenditures			Education Spending			Exp. to Income		
	Effect[a]	Exact p-value	Pre-treatment MSPE	Effect[a]	Exact p-value	Pre-treatment MSPE	Effect[a]	Exact p-value	Pretreatment MSPE
Arizona	–49.48	0.09	7.44	–22.92	0.20	6.58	0.10	0.27	0.03
Arkansas	17.24	1.00	20.89	20.24	0.13	5.14	0.05	0.13	0.01
California	–17.65	0.70	9.74	–3.64	0.92	5.76	1.83	0.50	0.70
Colorado	–42.08	0.25	7.79	–10.99	0.52	5.46	–0.03	0.68	0.01
Florida	–17.74	0.56	12.26	–16.58	0.13	4.04	0.19	0.15	0.05
Louisiana	–19.09	0.30	12.25	–3.54	0.96	7.72	–0.08	0.16	0.03
Maine	–17.57	0.12	5.20	–17.63	0.13	4.62	–0.01	0.67	0.00
Michigan	–7.81	0.58	8.81	33.37	0.04	3.77	0.01	0.92	0.05
Missouri	3.33	1.00	22.77	3.86	0.91	3.77	0.03	0.96	0.04
Montana	–8.45	1.00	16.81	3.20	0.85	7.63	0.01	0.95	0.01
Nevada	–80.20	0.07	17.40	–5.63	0.89	10.34	0.03	0.43	0.01
Ohio	17.89	0.65	8.42	4.23	0.31	2.37	0.13	0.31	0.04
Oklahoma	–26.69	0.48	8.05	9.96	0.70	4.06	–0.06	0.42	0.01
South Dakota	–42.77	0.48	13.66	–19.32	0.16	6.04	–0.02	0.71	0.01
Effect from Date of Implementation									
Arizona	–21.95	0.63	24.45	–10.91	0.59	10.65	0.08	0.35	0.03
Arkansas	9.05	1.00	24.01	15.24	0.23	7.25	0.04	0.17	0.01
California	–15.27	0.46	9.37	2.06	0.85	11.09	1.83	0.50	0.70
Colorado	–21.23	0.70	17.98	–8.83	0.38	5.97	–0.01	0.96	0.02
Florida	–8.19	0.27	12.70	–16.02	0.27	8.96	0.09	0.62	0.08
Maine	–12.13	0.35	6.07	–23.04	0.09	5.19	–0.01	1.00	0.00
Michigan	–6.47	0.32	11.23	23.99	0.09	4.80	0.06	0.69	0.05
Montana	–18.52	0.58	15.62	–3.22	0.96	10.83	0.00	1.00	0.01
Ohio	17.17	0.15	8.91	4.40	0.38	2.93	0.06	0.77	0.06
South Dakota	–26.43	0.26	17.04	–11.82	0.23	6.64	–0.02	0.63	0.01
Repeal States									
Idaho	–24.37	0.42	13.11	2.62	0.88	6.60	0.03	0.40	0.00
Massachusetts	–4.66	0.44	11.58	–5.43	0.36	3.88	–0.13	0.04	0.02
Oregon	19.42	0.72	9.56	–4.78	0.81	6.63	0.05	0.50	0.02
Utah	1.00	0.89	12.59	1.36	0.27	3.92	0.04	0.20	0.01
Washington	–19.61	0.14	7.47	2.43	0.56	5.49	0.02	0.32	0.02
Wyoming	8.60	1.00	115.95	17.61	1.00	76.24	–0.02	0.96	0.02

(*continued on next page*)

TABLE 3
(continued)

Effect from Date of Enactment	Health Expenditures			Transportation			Welfare		
	Effect[a]	Exact *p*-value	Pre-treatment MSPE	Effect[a]	Exact *p*-value	Pre-treatment MSPE	Effect[a]	Exact *p*-value	Pretreatment MSPE
Arizona	0.96	0.96	3.72	–5.03	0.79	5.95	–0.19	1.00	7.35
Arkansas	0.11	0.15	1.38	1.09	0.67	2.37	–6.89	0.36	1.35
California	0.29	0.37	1.84	–3.69	0.96	4.56	–10.14	0.96	8.32
Colorado	0.04	0.14	0.60	–3.35	0.50	1.00	–10.12	0.80	2.17
Florida	0.32	0.67	0.85	2.15	0.82	1.85	–3.68	1.00	4.22
Louisiana	0.32	0.35	2.23	–5.60	0.46	4.83	–20.51	0.15	4.72
Maine	0.25	0.28	2.90	–0.33	0.75	1.26	0.42	0.89	2.01
Michigan	0.14	0.20	1.34	–0.32	0.93	1.67	–23.38	0.52	5.22
Missouri	0.18	0.38	0.87	0.13	0.95	2.27	–1.51	0.67	1.13
Montana	0.07	0.11	1.19	–2.07	0.93	6.40	–22.64	0.19	3.48
Nevada	0.68	0.68	6.26	–9.45	0.30	4.20	–23.99	0.33	7.54
Ohio	0.64	0.79	1.36	0.62	1.00	2.94	–10.32	0.78	2.70
Oklahoma	0.29	0.33	1.79	0.39	1.00	1.22	–19.98	0.89	9.56
South Dakota	0.68	0.78	1.98	–1.16	0.89	4.45	–14.03	0.16	1.51
Effect from Date of Implementation									
Arizona	0.93	0.92	4.80	–1.44	0.86	5.64	0.19	1.00	6.57
Arkansas	0.39	0.45	1.92	0.07	0.42	3.25	–1.58	0.94	2.22
California	0.25	0.30	2.29	–2.46	1.00	5.05	–7.50	1.00	11.24
Colorado	0.39	0.39	2.29	–1.64	0.39	1.91	–7.49	0.48	4.35
Florida	0.21	0.44	1.05	1.21	0.36	2.19	–1.47	1.00	4.48
Maine	0.43	0.48	3.53	0.86	0.50	1.34	–1.38	0.89	2.40
Michigan	0.11	0.13	1.97	0.27	0.78	1.61	–9.07	0.96	8.56
Montana	0.11	0.17	1.67	–0.77	1.00	6.79	–14.26	0.22	3.41
Ohio	0.50	0.68	1.89	0.19	1.00	2.54	0.12	0.96	4.62
South Dakota	0.93	0.96	2.45	1.38	0.85	4.94	–9.50	0.50	4.63
Repeal States									
Idaho	0.29	0.32	3.49	–3.23	0.32	2.90	–6.54	0.96	3.90
Massachusetts	0.07	0.22	1.13	10.91	0.12	3.28	–17.47	0.44	7.53
Oregon	0.14	0.22	1.16	–5.42	0.28	3.31	2.25	0.78	3.29
Utah	0.36	0.43	1.86	0.84	0.50	4.81	–10.82	0.43	3.07
Washington	0.25	0.26	1.98	–3.55	0.61	2.88	–7.37	0.25	2.17
Wyoming	0.96	0.96	11.40	24.22	1.00	42.80	–4.53	1.00	8.52

Note: [a]Cell quantity represents the mean difference between the treated and synthetic control spending in the post-treatment time periods. If term limits decrease spending we would expect this quantity to be negative.

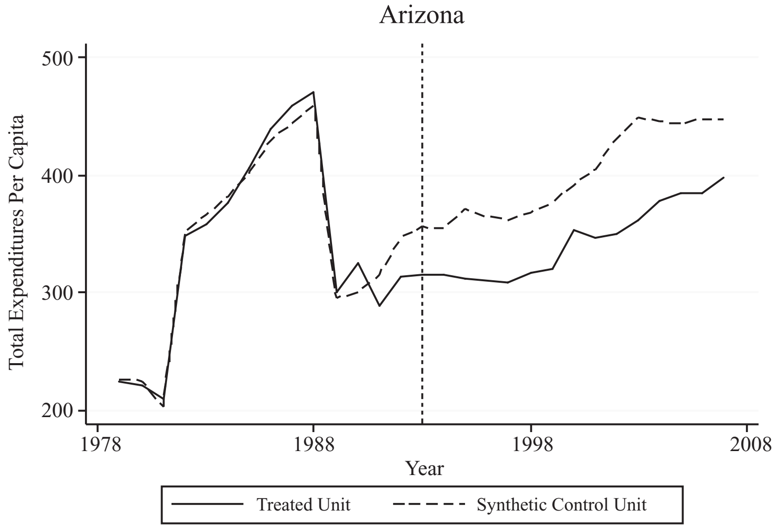
control states, the probability of obtaining results where the estimated effect is as large as it is for Louisiana is 0.30. In other words, the treated estimate is the eighth largest estimate out of the 27 estimates (Louisiana plus 26 placebo estimates).[7] The *p*-value is therefore 8/27, or 0.30. As shown in the second column of the table, the *p*-value is below the 0.10 threshold only for Arizona and Nevada.

Using the date term limits went into effect instead of the date enacted does little to change our inferences. As shown in the middle section of Table 3, for the total spending category, we generally find that the *p*-values increase when we use the implementation date as the date of treatment onset. We obtain similar results for the other dependent variables. Even using the more liberal standard of 0.10 for the test of the null hypothesis, we rarely observe statistically significant effects for the other spending outcomes. With adjustments for multiple testing, we would clearly be unable to find any significant treatment effects. Six out of 180 tests exhibit statistically significant *p*-values. By chance alone, we would expect to find 18 significant *p*-values. Further, the direction of the effect is not clear; across the 180 tests, we observe four negative and significant effects of term limits on spending and two positive and significant effects. We should also note that the three states that enacted term limits in 1990 did not exhibit statistically significant results. This suggests that spending dropped in term-limit states before most states adopted term limits.

We examine Arizona and Nevada in greater detail since these are the only two states that exhibited statistically significant effects for the total expenditures outcome. In Figures 2 and 3, we plot a solid black line for the treated units (Arizona and Nevada, respectively), and a dashed line for the estimated synthetic control. We also mark the year of enactment with a dashed vertical line. In both states, spending for the state with term limits is clearly lower than for the synthetic control. In Figure 4, we summarize the results for the Arizona placebo test. Here we plot the average gap between Arizona and its synthetic control with a thick black line. We also plot this difference for the 27 placebo estimates from the control states (thin gray lines). The figure makes clear that few of the placebo estimates exhibit more dramatic changes compared to their synthetic controls after the imposition of term limits in 1990. As a contrast, we also plot the results for Montana, a typical state where we did not observe a term-limit effect (see Figure 5). Here, we observe that the synthetic control unit does a good job of approximating spending in Montana in the pretreatment time period. However, the synthetic unit never diverges from the path of Montana.

By some objective measures, we observe an effect of term limits in both Arizona and Nevada. The difficulty in both cases (but especially

FIGURE 2
Arizona Synthetic Case Control Analysis Output

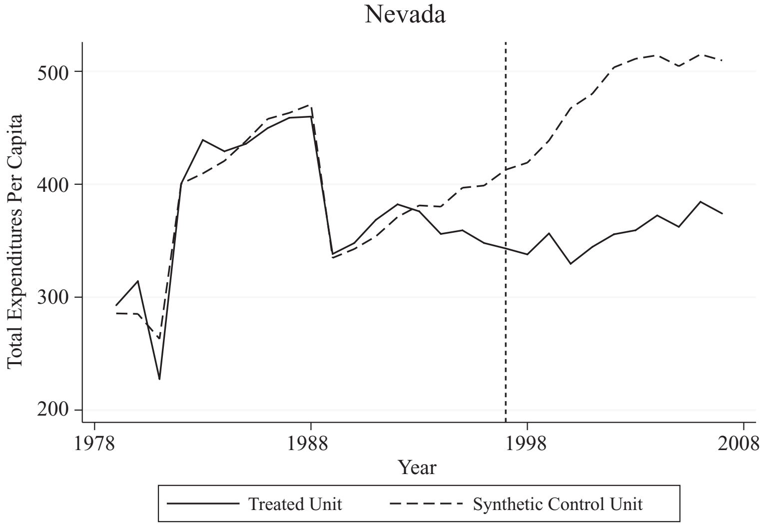


Note: Dashed vertical line represents date of enactment.

with Nevada) is that the change in spending occurs before term limits were enacted, consistent with the trends illustrated in Figure 1. This suggests that while both states did have lower spending compared to a weighted combination of states without term limits, the timing of the change in spending does not fit the causal hypothesis. That is, it is impossible for term limits to lower spending before term limits were enacted. What is more likely is that the enactment of term limits reflects a broader conservative policy agenda that included lower spending and later resulted in term limits. It is important to note that only by visualizing the data in this way can we see what is going on. Estimating a regression model with a plethora of controls obscures the clear pattern in the data.

Finally, we discuss the states that enacted term limits but later repealed them. If term limits affect spending, we might speculate that in a repeal state spending dropped (increased) and then rebounded (dropped) once the term limits were removed. With the exception of Oregon, we do not generally observe this pattern. Instead, there are little differences between the term-limit states and their synthetic controls before or after enactment or repeal. A typical result from Massachusetts is displayed in Figure 6. Here, we observe little difference between the term-limit state and the synthetic control before or after enactment or repeal.

FIGURE 3
Nevada Synthetic Case Control Analysis Output



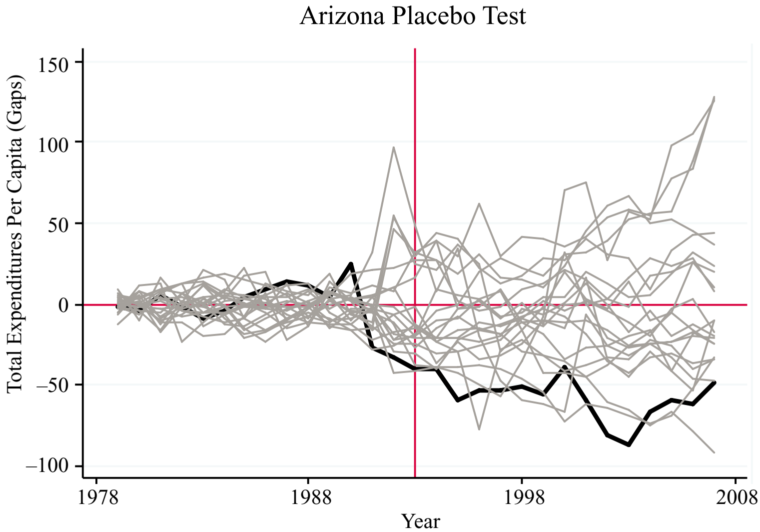
Note: Dashed vertical line represents date of enactment.

The results also provide little support for the hypothesized heterogeneity discussed above. The effects of term limits are not consistently stronger in states where the limits are more stringent (e.g., fewer terms, lifetime limits). Second, the treatment effects are not stronger for line items which can be considered more particularistic and where lawmakers have potentially more discretion, such as transportation. Third, the effects are not strongest in the most professional legislatures. Indeed, for the five most professional legislatures that enacted term limits at some point—California, Michigan, Ohio, Florida, Massachusetts—the treatment effects are substantively small.

Discussion

We find no consistent, systematic evidence that term limits significantly decreased spending (as suggested by commentators and advocates) or increased spending (as suggested by prior academic research). As explained above, there are theoretical reasons to believe that term limits should both increase and decrease spending, and these effects may net to 0. Alternatively, any effect of term limits may be dwarfed by more

FIGURE 4
Arizona Synthetic Case Control Analysis Output—Placebo Test

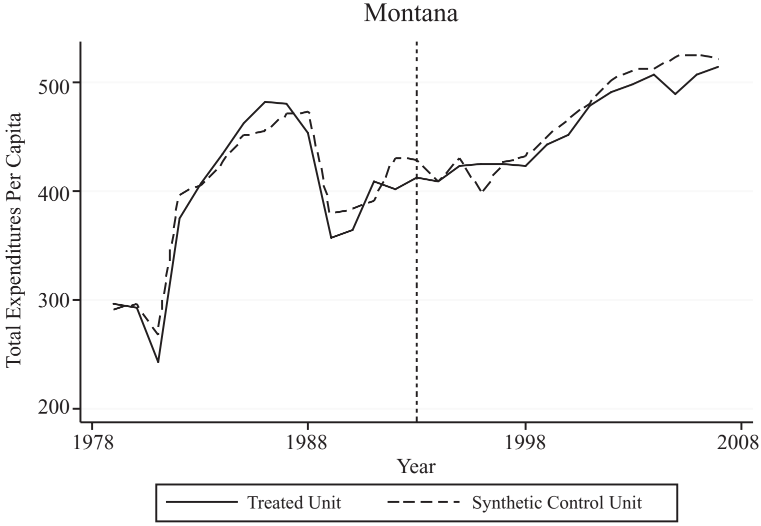


Note: Solid vertical line represents date of enactment.

important factors. One key driver of state spending is macroeconomic conditions. State spending grew strongly during years of strong growth in the 1990s and dropped when the economy faltered. The recent fiscal crisis further underscores that term limits did little to change spending patterns relative to general economic trends. One might argue that if term limits had provided a curb on state spending, states with term limits should have been better prepared to deal with the downturn. Term-limit states such as California, Nevada, and Arizona have fared no better and perhaps even worse than states such as Illinois and New York which do not have term limits.

Future substantive research can address questions about term limits using the methodological approach presented in this article. Other dependent variables assessing fiscal policy can be addressed, perhaps with a more thorough theoretical account of what types of spending should be affected by term limits. External institutions such as the courts and the federal government may constrain budgetary decisions at the state level, and it is important to consider these constraints in future work. Further, we found that any spending changes that we did observe appeared to have occurred before the enactment of term limits. We conjectured that this

FIGURE 5
Montana Synthetic Case Control Analysis Output



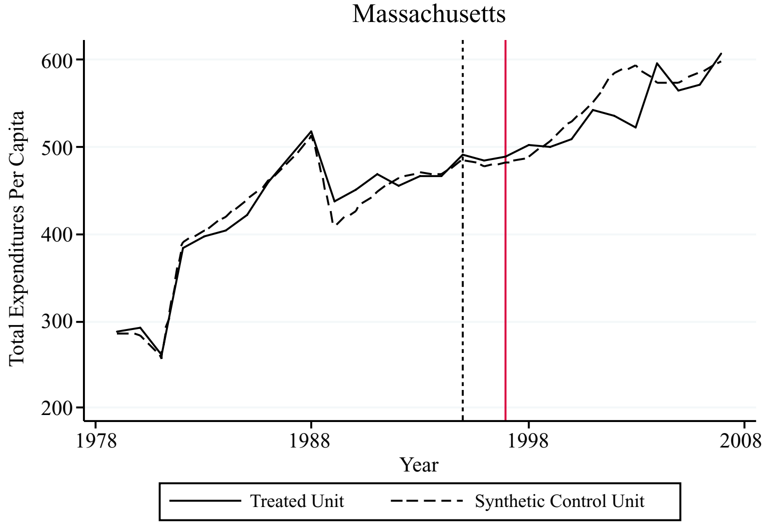
Note: Dashed vertical line represents date of enactment.

was due to a general conservative movement in the states but subsequent research may be able to isolate the distal cause more precisely.

Our results are consistent across two different identification strategies: differences-in-differences and conditioning on past outcomes. Moreover, for the past-outcomes approach we used both a parametric model and the nonparametric synthetic case-control method. The key methodological insight to be gleaned from our study is that although it is tempting to perform “large-n” analyses by pooling cases, treatment heterogeneity must be taken seriously by researchers of state politics. For political units like states, it is difficult to believe that the response to a large institutional change like term limits will be uniform across states. First, the exact form of term limits varied from state to state. Moreover, even if the exact same form of term limits was adopted by each state, the highly differentiated political cultures and institutions would with high probability alter the effects of the policy change. While the study of state politics focuses on common variation across states, one might argue that heterogeneity is more likely to be the rule rather than the exception.

Our study contributes both methodologically and substantively to the study of how state-level political institutions influence public policy. The synthetic case-control approach complements the standard

FIGURE 6
Massachusetts Synthetic Analysis Output



Note: Dashed vertical line represents date of enactment and solid vertical line indicates date of repeal.

difference-in-difference model. With DID models, it is difficult to assess the quality of the counterfactual comparison implied by the model. In our approach, we are able to evaluate the fit between the treated state and the synthetic control in the pretreatment period. Synthetic case control allows the analyst to know when we can be confident that we can construct good counterfactuals. While the "bread and butter" of state politics research is cross-state comparisons, at times those comparisons may not be valid. Because institutional reforms are not randomly assigned to states, we must be more careful in constructing appropriate counterfactuals, leaning on the logic of experimentation as much as possible.

Luke Keele <ljk20@psu.edu> *is Associate Professor of Political Science, Penn State University, 211 Pond Lab, University Park, PA 16802. Neil Malhotra* <neilm@stanford.edu> *is Associate Professor at the Graduate School of Business, Stanford University, 655 Knight Way, Stanford, CA 94305. Colin H. McCubbins* <cmccubbi@stanford.edu> *is a Ph.D. candidate in political science, Stanford University, 616 Serra Street, Stanford, CA 94305.*

APPENDIX

Data Summary and Sources

Dependent Variables. We predict six dependent variables. Five dependent variables are state expenditures figures divided by state population (measured in dollars): total state expenditures, state education expenditures, state health expenditures, state transportation expenditures, and state welfare expenditures. The sixth dependent variable is total state expenditures divided by state income. All spending figures are adjusted for inflation using the Consumer Price Index (CPI) and expressed in 2004 dollars. In cases where expenditure figures were missing, missing values were linearly interpolated. Expenditures data are taken from various editions of the *Statistical Abstract of the United States* (U.S. Census Bureau). State income data are from the Bureau of Economic Analysis (U.S. Department of Commerce).

Treatment Variable. Information on the presence of state legislative term limits and the years of enactment and implementation were obtained from the National Council of State Legislatures and corroborated by Erler (2007).

Variables Used to Construct Synthetic Controls. *Population* represents the total state population based on the revised July estimates. *Population Growth* is the change in current population from the previous year divided by the total population of the previous year. *Population Density* is presented in persons per square mile and was calculated by using the aforementioned *Population* variable divided by the state's total land area. These variables are collected from the *Statistical Abstract of the United States* (U.S. Census Bureau). *Democratic Control* is a binary variable indicating that Democrats control the governorship and both chambers of the legislature. *Divided Government* is a binary variable with a "1" representing a divided state government (i.e., one political party does not have control over both legislative chambers and the governorship) and "0" otherwise. *House Seats* and *Senate Seats* are the number of seats in the lower and upper chamber, respectively. Data for these four variables were obtained from various editions of *The Book of the States*. *Per Capita Gross State Product* is inflation adjusted and measured in millions of dollars. Data are taken from the Bureau of Economic Analysis. *Per Capita Federal Grants* represents the total amount of inflation-adjusted federal aid given to a state divided by total state population, expressed in millions of dollars. *Per Capita Federal Civilian Employment* denotes the total number of civilian federal jobs in a state divided by state population, while *Per Capita Federal Military Employment* is the total number of military-related federal jobs in a state divided by state population. *Per Capita State and Local Government Employment* indicates the total number of state and local government jobs divided by state population. Data are from the Bureau of Economic Analysis. *Legislative Professionalism* is measured using the Squire index, which averages legislator salary, days in session, and staff resources. Squire index scores are taken from Squire (2007). We also coded whether the state has in place a *Debt limit* (using data from Wagner and Elder 2004) and whether the state has a *Tax and Expenditure Limit* (using data from Kousser, McCubbins, and Moule 2008). *State Unemployment Rate* is the percent of the state labor force that is unemployed and is not seasonally adjusted. The 2003 and 2004 figures account for revised population estimates. Data were obtained from the Bureau of Labor Statistics. *Initiative* indicates whether the state has an initiative process. Data were obtained from the Initiative and Referendum Institute (Matsusaka 2004).

NOTES

Authors are in alphabetical order. We thank Jens Hainmueller and Thad Kousser for valuable feedback. Keshav Dimri, Terrence Reilly, and Amanda Barth provided excellent research assistance. All errors are our own. A previous version of the article was presented at the 2010 Annual Meeting of the Western Political Science Association in San Francisco, CA.

1. The Supreme Court declared Congressional term limits to be unconstitutional in *U.S. Term Limits, Inc. v. Thornton* (1995) and a key plank of the Contract with America to limit House members to six terms did not achieve the two-thirds majority necessary to pass a constitutional amendment.

2. Following Erler (2007), we exclude Nebraska, which has term limits, from our analysis because it has a unicameral, nonpartisan legislature, preventing us from including variables for partisan control and lower-chamber legislature size as pretreatment covariates.

3. State limits were adopted by the state legislature in Utah to preempt an initiative that had not yet qualified for the ballot. However, the state legislature later repealed these limits.

4. The inference for the placebo test is exact in the sense that we are comparing the treated unit to the population of control units. The inference is not exact in the sense that tests such as Fisher's exact test are since randomization has not occurred to ensure exchangeability of the units.

5. See Angrist and Pischke (2009, 246–47) for a formal proof.

6. For a small number of states and years, we linearly interpolated the data to fill in missing values.

7. One placebo run is dropped due to poor pretreatment fit.

REFERENCES

Abadie, Alberto. 2005. "Semiparametric Difference-in-Difference Estimators." *Review of Economic Studies* 75 (1): 1–19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2011. "Synth: An R Package for Synthetic Control Methods in Comparative Case Studies." *Journal of Statistical Software* 42 (13): 1–17.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93 (1): 112–32.

Aka, A., W.R. Reed, D. Schansberg, and Z. Zhu. 1996. "Is There a Culture of Spending in Congress?" *Economics and Politics* 8 (3): 191–209.

Alt, James, Ethan Bueno de Mesquita, and Shanna Rose. 2011. "Disentangling Accountability and Competence in Elections: Evidence from U.S. Term Limits." *Journal of Politics* 73 (1): 171–86.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Bandow, Doug. 1995. "Real Term Limits: Now More Than Ever." http://www.cato.org/pubs/pas/pa-221.html.

Barnow, B.S., G.G. Cain, and A.S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies*, Vol. 5, ed. E. Stromsdorfer and G. Farkas. Vol. 5 San Francisco, CA: Sage.

Bartels, Larry M. 2008. *Unequal Democracy*. Princeton, NJ: Princeton University Press.

Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to Do (And Not to Do) With Time-Series Cross-Section Data." *American Political Science Review* 89 (3): 634–47.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249–75.

Besley, Timothy, and Anne Case. 1994. "Does Electoral Accountability Affect Economic Policy Choices? Evidence from Gubernatorial Term Limit." *Quarterly Journal of Economics* 110 (3): 769–98.

Besley, Timothy, and Anne Case. 2003. "Political Institutions and Policy Choices: Evidence from the United States." *Journal of Economic Literature* 41 (1): 7–73.

Cain, Bruce E., and Thad Kousser. 2004. *Adapting to Term Limits: Recent Experiences and New Directions*. San Francisco, CA: Public Policy Institute of California.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Carey, John M., Richard G. Niemi, and Lynda W. Powell. 1998. "The Effects of Term Limits on State Legislatures." *Legislative Studies Quarterly* 23 (2): 271–300.

Carey, John M., Richard G. Niemi, and Lynda W. Powell. 2000. *Term Limits in State Legislatures*. Ann Arbor: University of Michigan Press.

Donald, Stephen G., and Kevin Lang. 2007. "Inference With Differences-In-Differences and Other Panel Data." *The Review of Economics and Statistics* 89 (2): 221–33.

Ehrenhalt, A. 1991. *The United States of Ambition: Politicians, Power, and the Pursuit of Office*. New York: Random House.

Erler, H. Abbie. 2007. "Legislative Term Limits and State Spending." *Public Choice* 133 (3): 479–94.

Fiorina, Morris P. 1989. *Congress: Keystone to the Washington Establishment*. 2nd edition ed. New Haven, CT: Yale University Press.

Fund, John H. 1990. "Term Limitation: An Idea Whose Time Has Come." *Cato Policy Analysis 141*. Cato Institute.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Personal Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (3): 653–63.

Gilligan, Thomas W., and John G. Matsusaka. 1995. "Deviations from Constituent Interests: The Role of Legislative Structure and Political Parties in the States." *Economic Inquiry* 33 (3): 383–401.

Herron, Michael C., and Kenneth W. Shotts. 2006. "Term Limits and Pork." *Legislative Studies Quarterly* 32 (3): 383–403.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Kousser, Thad. 2005. *Term Limits and the Dismantling of State Legislative Professionalism*. Cambridge: Cambridge University Press.

Kousser, Thad, Mathew D. McCubbins, and Ellen Moule. 2008. "For Whom The Tel Tolls: Can State Tax and Expenditure Limits Effectively Reduce Spending?" *State Politics and Policy Quarterly* 8 (4): 331–61.

Lewis, Daniel. 2012. "Legislative Term Limits and Fiscal Policy Performance." *Legislative Studies Quarterly* 37 (3): 305–28.

Matsusaka, John G. 2004. *For The Many Or The Few: The Initiative, Public Policy, and American Democracy*. Chicago: Chicago University Press.

Mayhew, David R. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale University Press.

Moore, Stephen, and Aaron Steelman. 1994. "Term Limits: An Antidote to Federal Red Ink?" Cato Institute *Briefing Paper 21*. http://www.cato.org/pubs/briefs/bp-021.html.

Nickell, Stephen. 1981. "Biases In Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–26.

Payne, James L. 1991. *The Culture of Spending*. San Francisco, CA: ICS.

Primo, David M. 2006. "Stop Us Before We Spend Again: Institutional Constraints on Goverment Spending." *Economics & Politics* 18 (3): 269–97.

Rosenbaum, Paul R. 1984. "The Consequences of Adjusting For a Concomitant Variable That Has Been Affected By The Treatment." *Journal of The Royal Statistical Society Series A* 79 (1): 41–48.

Sovey, J. Allison, and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55 (1): 188–200.

Squire, Peverill. 1988. "Career Opportunities and Membership Stability in Legislatures." *Legislative Studies Quarterly* 13 (1): 65–82.

Squire, Peverill. 2007. "Measuring State Legislative Professionalism: The Squire Index Revisited." *State Politics and Policy Quarterly* 7 (2): 211–27.

Wagner, Gary A., and Erick M. Elder. 2004. "The Role of Budget Stabilization Funds in Smoothing Government Expenditures over the Business Cycle." *Public Finance Review* 33 (4): 439–65.

Weingast, Barry R., Kenneth A. Shepsle, and Christopher Johnson. 1981. "The Political Economy of Benefits And Costs: A Neoclassical Approach to Distributive Politics." *Journal of Political Economy* 89 (4): 642–64.

Will, George F. 1992. *Restoration: Congress, Term Limits, and the Recovery of Deliberative Democracy*. New York: Free Press.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website: