Selecting For Extremists: Evaluating the Impact of Judicial Selection Methods on Ideology

Brett Parker (Stanford University)

Abstract

U.S. states display striking heterogeneity in their choices of judicial selection method. Researchers have produced dozens of papers exploring how the most commonly-used selection methods—partisan election, nonpartisan election, merit selection, and unconstrained gubernatorial appointment—affect the ideology and behavior of the judges they produce. Nevertheless, these articles have studied only a limited set of outcome variables. Most published work concerns either how selection method affects (1) judicial responsiveness to public opinion or (2) ideological direction. To my knowledge, however, no empirical work explores how choice of selection method impacts ideological extremity. This paper fills that gap. Using generalized propensity score matching, fixed effects counterfactual estimators, and synthetic controls to conduct causal inference, I examine whether some methods of selection produce more moderate (or extreme) judges than others. In order to do so, I extend and supplement the Bonica-Woodruff dataset on state supreme court judges, which features a commonly used measure of judicial ideology extracted from political donations. I find consistent evidence that judges picked by unconstrained gubernatorial appointment are more extreme on average than those selected by other methods. However, I do not find substantial evidence of differences between other pairs of selection methods.

Word Count: Approximately 9,228 (Excluding Appendices)

Prepared for the APSA Annual Meeting, Sept. 2022

Keywords: Judicial Selection, State Politics, Ideological Extremity

1. Introduction[1]

When it comes to selecting governors and state legislators, the fifty states exhibit extraordinary homogeneity. Nearly every state employs partisan elections for these offices and there is little serious consideration of adopting a different system.[2] This uniformity stands in stark contrast to the lack of consensus about how to pick judges. Twenty-two states use commission-assisted appointment ("merit selection" or "merit commissions"), thirteen states have nonpartisan elections, eight states use partisan elections, five deploy unconstrained gubernatorial appointment, and two have legislative appointment (Bonica & Sen, 2021).[3] Even this taxonomy oversimplifies the situation, as selection mechanisms vary further within these categories.

Perhaps unsurprisingly, interest groups, politicians, and academics have expended considerable energy debating which selection method is best (e.g., Geyh, 2019; Czarnezki, 2005). Much of this discussion revolves around the normative question of what kind of judges our society should have. Just as important, though, is determining whether the various methods actually produce different types of jurists.

The phrase "different types of jurists" is capacious—it could include variations in experience, gender, responsiveness to public opinion, racial background, and ideology. For example, some studies conclude that judges chosen via merit selection have more liberal tendencies than others (Fitzpatrick, 2017). Other research indicates that judges picked in nonpartisan contests are more likely to take popular positions on high salience issues than judges selected through other means (Canes-Wrone et al., 2014; Caldarone et al., 2009). Finally, older work suggests that merit selection produces greater racial and gender diversity

[1]I gratefully acknowledge advice and feedback from Adam Bonica, Justin Grimmer, Tobias Nowacki, Xiqing Xu, and participants at the Stanford Political Science Department's Internal Workshop.

[2]Nebraska is the only state with a nonpartisan legislature, a status it obtained more than eighty years ago (Masket & Shor, 2015).

[3]While these sources generally label Ohio and Michigan as nonpartisan states, both jurisdictions have partisan primaries for state supreme court, and so their elections are more accurately considered partisan (Kritzer, 2018).

on the bench (Henry, 1985; but see Goelzhauser, 2019).

This paper addresses an unexamined aspect of this larger question: whether some methods of judicial selection produce more *ideologically extreme* judges than others. Existing work attempts to determine the relationship between selection method and the direction of ideology, but largely ignores the magnitude of this ideological lean (Bonica & Sen, 2021; Fitzpatrick, 2017). This lack of research represents an important gap in our knowledge. From a practical perspective, policymakers would certainly be interested in knowing whether a particular selection method produces moderates or extremists. Minority parties in state legislatures might be willing to live with the majority's choice of selection method if they know the method will limit the number of committed ideological opponents on the bench. Meanwhile, majority parties may push to maximize the ideological lean of courts while they hold power.

From the perspective of political theory, the ideological extremity of judges also matters. The power a particular political system allocates to the judiciary may depend on how judges are likely to use it. In the United States—where both federal and state judges often have the last word in determining what the law means —residents might prefer moderate judges who are less likely to direct their substantial power towards unpopular ends. By contrast, in a country that generally practices parliamentary supremacy (like the United Kingdom) an ideologically uncompromising judiciary might be essential to pushing back against laws that disadvantage minorities. To the extent that a nation or state relies on a specific type of judicial attitude for its political system to function, it must know what means of selection produce that sort of judge.

In order to determine the relationship between ideological extremity and selection methods, this paper examines the ideology of judges sitting on state courts of last resort. In doing so, I make use of the Bonica-Woodruff dataset on state supreme court judges, which exploits a reliable measure of ideology derived from campaign donations (Bonica & Woodruff,

2015). The original dataset runs from 1990-2012 and includes ideology scores for more than 95% of high court judges. I was able to expand the dataset backwards to 1980 and forward to 2016, increasing the coverage from 23 years to 37 years. Doing so allows me to examine several selection method transitions that would otherwise go unexplored. Expanding the dataset decreases completeness slightly (to about 92%—donor records become more scarce for 1980s judges), but the tradeoff is well worth it.[4] These ideology scores provide a reasonably clear picture of how selection methods affect ideological extremity.[5]

Armed with this data, I use three econometric techniques in an effort to draw causal inferences. The form of the independent variable (categorical) presents some difficulty, because well-known inference techniques often assume a binary or continuous IV. Accordingly, my primary analyses rely on one of the few methods designed for categorical IVs: generalized propensity score matching (GPSM) (Zhou et al., 2021; Lopez & Gutman, 2017). GPSM allows me to make pair-wise comparisons for any two selection methods. I supplement these results with tests using various Fixed Effect Counterfactual Estimators (FEct) developed by Liu et al. (2022) and the synthetic control method (SCM) popularized by Abadie et al. (2015; 2010). When applying FEct, I iteratively recode my independent variable to be binary, examining each method as the "treatment" in turn. For the SCM tests, I focus on the states that changed selection method, comparing the initial mode of selection to the subsequent one.

These tests consistently indicate that judges selected by unconstrained gubernatorial appointment are more extreme than those picked by other methods. The GPSM, FEct, and SCM analyses all indicate that when a governor selects judges without assistance, the judges

[4]In Appendix F, I consider what impact the missing data might have on my results. After calculating lower bounds for my effect estimates, I conclude that the missing data is not a serious concern.

[5]I should note upfront that this paper is concerned only with initial selection methods for open seats, not with appointment to subsequent terms or mid-term vacancies. This focus on initial selection methods is undoubtedly an oversimplification. However, given the wide variety of initial selection methods and the further diversity of retention menthods and mid-term vacancy procedures, it would be next to impossible to account for more than the initial method of selection within a plausible causal inference framework. In addition, I do not examine legislative selection given its limited prevalence.

tend to be about 0.2 units more extreme than those selected by alternative means (on a scale of 0-2). There is some limited evidence that merit commissions may produce more extreme judges than those selected by electoral methods—however, the point estimates are relatively small and do not consistently reach any conventional level of statistical significance.

The remainder of this paper is arranged as follows. First, I examine what previous scholars have written about how different selection methods affect judicial ideology. I consider both theoretical accounts that seek to explain why various selection methods produce certain types of judges, as well as empirical work that endeavors to quantify those relationships. After doing so, I discuss my data and methodology. This section describes my process informally but refers interested readers to more technical treatments of these subjects as necessary. Next, I present my conclusions from the empirical tests, beginning with the GPSM results before turning to the FEct analyses. Finally, I consider the ramifications of my findings and conclude.

2. Background

a. Theoretical Perspectives

Theoretical discussions of judicial selection methods generally do not explicitly consider whether some mechanisms produce more ideological judges than others. Nevertheless, some existing scholarship indirectly suggests that judicial elections may produce particularly extreme judges.

For instance, increasingly fractious campaigns for judicial office may produce extreme jurists by persuading moderates not to run. Spending on state supreme court campaigns from 2000-2009 more than doubled the previous decade's total, and races have become vitriolic (Kowal, 2016). Faced with a bruising campaign and personal attacks, non-ideologues may not find it worthwhile to run. Political scientists have found evidence that moderates self-select out of the candidate pool in other contests (e.g., Hall, 2019; Thomsen, 2014). It seems plausible that moderate potential judges could do the same.

Moreover, commentators have frequently raised concerns that candidates in both types of judicial elections—partisan and nonpartisan—feel pressure to favor campaign contributors while on the bench (Lamphier, 2011; Failinger, 2005). Though genuine quid pro quo corruption appears uncommon, judges may nevertheless feel some psychological pressure when one of their contributors appears before them in court (a common occurrence). To be sure, responsiveness to contributors does not necessarily imply extremity. However, campaign donors are substantially more ideological than the general public, so if judges do subconsciously prefer those donors, those judges may adopt more extreme stances (La Raja & Schaffner, 2015).

One countervailing consideration is the possibility that the public exerts a moderating influence on judicial candidates. Some scholars vigorously contend that only elites—not ordinary voters—have polarized (e.g., Fiorina, 2016). If the public is indeed centrist, elections may force candidates to the center (as compared to appointive methods). However, for this mechanism to operate, the median voter theorem would need to apply to judicial elections. While there is some evidence that judicial candidates converge on high-salience issues like capital punishment (Parker, n.d.), this moderation is likely the exception, rather than the rule. In all likelihood, public awareness of judicial elections—and the issues involved—is simply too low for judges to be particularly responsive to the broader electorate (Burnett & Tiede, 2015; Lim & Snyder, 2015).

b. Empirical Investigations

Empirically oriented scholars have long sought to determine how selection methods can affect ideology. Much of this research focuses on how specific selection methods influence judicial rulings in highly salient issue-areas. In particular, researchers have been interested in whether elections produce greater responsiveness to public opinion than various appointive methods. For example, several investigators have explored the relationship between selection mechanism and judicial rulings in capital cases. One set of scholars—Brandice Canes-Wrone,

Tom Clark, and Jason Kelly—have argued that nonpartisan judicial elections produce pro-death penalty appellate judges (Canes-Wrone et al., 2014). Likewise, Professors Paul Brace and Brent Boyea find that selecting judges via election rather than appointment impacts judicial behavior in capital cases (Brace & Boyea, 2008). Researchers have also posited a connection between selection method and ideology when it comes to criminal sentencing (Gordon & Huber, 2007), abortion (Caldarone et al., 2009), and marijuana cases (Nelson, 2014). However, elections do not always appear to produce greater responsiveness—Canes-Wrone and coauthors find little evidence of a relationship between public opinion and judicial ideology in the area of environmental law (Canes-Wrone et al., 2018).

Less common are efforts to move beyond issue-level ideology and assess the relationship between selection method and overall ideology. The reason for this paucity of studies is intuitive. While it is feasible for a small team of researchers to obtain and categorize all state supreme court rulings in a particular issue-area, it is nearly impossible to do the same for the entire universe of cases across all fifty states. As a consequence, research examining overall ideology requires developing and applying indirect measures.

One creative attempt to discern the relationship between selection methods and ideology belongs to Brian Fitzpatrick (2017). Fitzpatrick hypothesized that judges selected by merit commission would be more liberal than the publics they served; he did not expect to find significant differences between the ideologies of the public and those of judges selected via unconstrained gubernatorial appointment or partisan election. To assess these claims, he used the political party each judge most frequently donated to as a proxy for ideology. While this measure proved a blunt instrument—Fitzpatrick was only able to categorize judges as "Democrat," "Republican," or "Unknown"—it did provide a means for him to compare the ideology of the judiciary to the ideology of the public (which he measured using federal and state election outcomes). The results of Fitzpatrick's empirical analyses supported his hypotheses.

The same year as Fitzpatrick's study, Adam Bonica and Maya Sen also published an article examining the relationship between selection method and judicial ideology (Bonica & Sen, 2017b). Like Fitzpatrick, they relied on campaign donations to evaluate the ideology of judges. However, their measure provides more precise estimates of ideology. It runs from -2 (most liberal) to 2 (most conservative) with individual ideologies measured to several decimal points. I discuss Bonica and Sen's method for arriving at these estimates in Section 3(a), since I use the same measure of ideology here. For now, it is enough to observe that Bonica and Sen's fine-grained measure of ideology allows for more nuanced conclusions than Fitzpatrick's data. Ultimately, the authors determine that merit commissions and nonpartisan elections produce less ideologically-based selection; judicial ideology largely replicates the ideology of the underlying attorney pool. Meanwhile, unconstrained gubernatorial appointment and partisan elections result in ideology-conscious selection. Accordingly, judges are more likely to share the ideology of the state's average elected politician.

Given this diversity of perspectives, it is hard to generate clean hypotheses about how each selection method will affect the ideological extremity of judges. Nevertheless, a few tentative predictions seem reasonable. First, it appears likely that unconstrained gubernatorial appointment will produce particularly extreme judges. Extensive empirical research has documented that political elites generally and governors specifically have become ideologically polarized (e.g., McCarty et al, 2006; Warner, n.d.). The vast majority of governors are squarely liberal or conservative, meaning that relatively few prefer moderates to more ideological jurist (Warner, n.d.). All states with unconstrained gubernatorial appointment require some additional body (often a small committee) to confirm the governor's choice, but governors can usually find a well-qualified state supreme court candidate who fits their general ideological preferences without difficulty. Accordingly, unconstrained gubernatorial selections will likely be the most ideologically extreme.

Second, partisan elections will also probably generate more extreme judges. Judges who

win their seats in these elections generally have to prevail in partisan primaries in order to be nominated—accordingly, it is fair to assume that candidates for these offices will typically share the ideological preferences of their co-partisans (Hirano & Snyder, 2014). Contests between explicitly partisan judges should consequently pit candidates against each other who reflect the median ideologies of the primary electorates, with the winner holding non-moderate views (Andreottola, 2021). Though there is some evidence that judicial candidates may adopt the views of the median general election voter—rather than the median primary voter—on particularly high salience issues like capital punishment (Parker, n.d.), it still seems probable that they hold party-conforming views on most other subjects (particularly if moderates choose not to run).

Meanwhile, if Bonica and Sen are correct—that judges in states with nonpartisan elections and merit selection will generally resemble the ideology of the median attorney—we would expect those methods to produce more moderate jurists. As Fitzpatrick accurately observed, attorneys in most states tend to be liberal. However, as Bonica and Sen demonstrate in their book, the median lawyer in most states is closer to the ideological center than the median politician (Bonica & Sen, 2021). As such, while merit selection and nonpartisan elections may produce more liberal judges, we might also expect those judges to be less extreme than those selected through other means.

3. Methods and Data

a. Measuring Judicial Ideology

Essential to evaluating the impact of selection methods on judicial extremity is a measure of ideology. Without a metric that locates judges from different states in the same ideological space, it would be effectively impossible to compare all four selection methods in the same analysis.

Fortunately, political scientist Adam Bonica (2014) has developed a robust method of measuring judicial ideology without looking at cases: campaign finance scores (hereinafter

"CFscores"). CFscores are a measure of ideology derived from an individual's campaign contributions. While the actual process of estimating a person's CFscore is complex, the basic idea is surprisingly simple. It relies on the notion that you can determine an individual's ideological leanings by examining the politicians they donate to. If I contribute only to Elizabeth Warren, Bernie Sanders, and Alexandria Ocasio-Cortez, I am probably quite liberal—by contrast, if I give exclusively to Ted Cruz, Mitch McConnell, and Kevin McCarthy, I am likely a conservative (Bonica & Sen, 2017b).

In this paper, I exploit a preexisting dataset of CFscores for high court judges compiled by Bonica and collaborator Michael Woodruff (Bonica & Woodruff, 2015). Their data contains ideology scores for over 95% of state supreme courts judges between 1990-2012. Bonica and Woodruff estimate about 40% of these scores from the judges' personal contributions. They obtain scores for an additional 31% of judges via the ideologies of those who contributed to those judges' campaigns. For another 24%, they use ideology scores for the appointing governor or legislature to estimate the judge's preferences. Nevertheless, all scores ultimately derive from campaign finance contributions and Bonica and Woodruff use the same -2 to 2 scale for each measure of ideology, making the scores commensurable. The authors also demonstrate strong correlations between each set of scores, and other researchers have made use of their data (e.g., Wilhelm et al., 2020; Kane, 2018; Brown, 2018).[6]

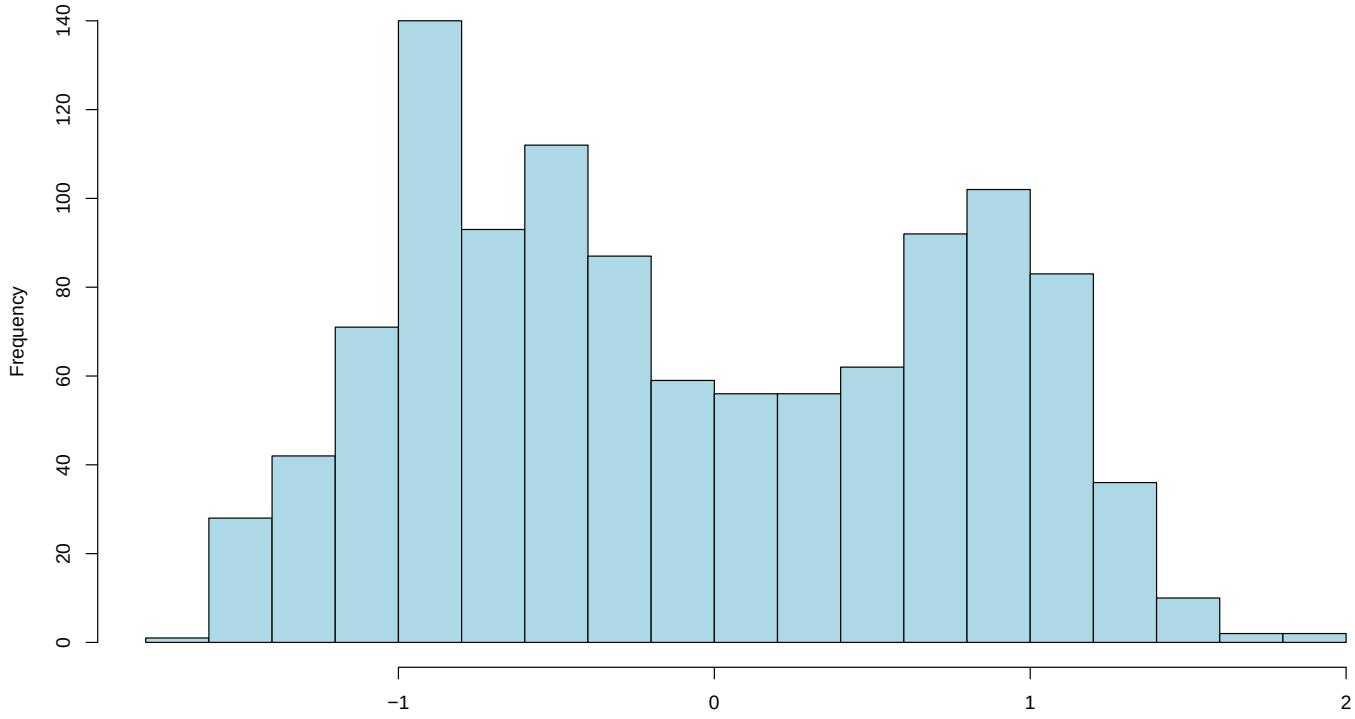
In an effort to supplement this dataset, I extend the time period covered back to 1980 and foward to 2016. In order to do so, I independently examined Bonica's DIME database in search of CFscores for judges not included in the original data.[7] I was able to locate CFscores for most of those judges, making my extended dataset 92 percent complete. I consider the potential impact of the missing 8 percent of observations in Appendix F, and conclude that it likely has little impact on my results.

[6]The Bonica-Woodruff scores also correlate strongly with other measures of judicial ideology (Bonica & Sen, 2017a).

[7]I describe this search process in Appendix A.

I plot the distribution of CFscores in Figures 1. The distributions is bipolar, with a slightly larger concentration of data on the liberal end of the spectrum. These results are encouraging, as they largely comport with previous research on the ideological distribution of state high court judges (Bonica & Sen, 2017b.)

Figure 1: Distribution of Ideology Scores



b. Causal Inference

My primary method of causal inference in this paper is generalized propensity score matching (GPSM).[8] GPSM is a technique that extends propensity score matching to data with a categorical independent variable (Lopez & Gutman, 2017). Like ordinary propensity score matching, GPSM relies on two assumptions: common support (that is, every unit has a non-zero probability of being assigned to each treatment) and the absence of unobserved confounders (Lopez & Gutman, 2017; Angrist & Pischke, 2009).

[8]For ease of reference, I refer this technique as if it were a traditional matching operation. In reality, I apply GPSM through a weighting analogue to pair matching, implemented via *PSweight* in R (see generally Zhou et al., 2021; Li & Li, 2019; Li & Green, 2013).

The first assumption is highly plausible here—both my propensity score estimates (discussed in the results section) and common sense indicate that every state could theoretically adopt any of the judicial selection mechanisms discussed in this paper. The second assumption is much more contentious. In estimating propensity score vectors, I use year, partisan composition of the State House and State Senate, the party of the incumbent governor, whether the state has unified government, and the region the state belongs to. Given that (1) changes to judicial selection method come in waves (Bonica & Sen, 2021), (2) that these changes come at the behest of political actors in the state government, and (3) the considerable correlation between selection method and region, the absence of unobserved confounders is a potentially plausible assumption. However, it is not an unassailable one, which is why I use Fixed Effect Counterfactual Estimators (FEct) to supplement these results.

FEcts are a class of estimators developed to address problems with staggered adoption panel designs (Liu et al., 2022). A flurry of research suggests that two-way fixed effects (TWFE)—a workhorse in modern causal inferece (Imai & Kim, 2020)—do not produce reliable results when units adopt the treatment at difference points in time (de Chaisemartin and d'Haultfoeuille, 2020; Imai & Kim, 2019; Blackwell & Glynn, 2018). These findings have inspired a series of innovations in panel data methods, which include FEcts. Some of the new estimators in this class include the two-way fixed effect counterfactual, the interative fixed effects counterfactual, matrix completion methods, and generalized synthetic controls (Liu et al., 2022; Xu, 2017). Unfortunately, most of the work in this area is still focused on binary treatments, meaning that FEcts currently have limited use for categorical independent variables. For that reason, I iteratively recode my independent variable as binary when conducting my FEct analyses.

In addition to my GPSM and FEct tests, I examine seven comparative case studies (Mississippi, Arkansas, North Carolina, Connecticut, Utah, New Mexico, and Tennessee) using the synthetic control method. The basic idea underlying the synthetic control is

straightforward in the abstract (if computationally challenging). The researcher aims to select a weighted average of control units that closely resemble the treated unit on outcome variable predictors in the pretreatment period (Abadie, 2021). This "synthetic control" provides a reasonable assessment of what would have occurred in the treated unit had it not implemented the treatment. Unlike regression techniques, the SCM can safely incorporate the value of lagged dependent variables as predictors—by doing so, it implicitly incorporates information about unobserved confounders that affect the value of the dependent variable (Abadie et al., 2015; Kaul et al. 2015; Ferman & Pinto 2016; Bifulco et al. 2017; Botosaru & Ferman 2019; Renberg 2020). However, the SCM is generally only useful for comparative case studies in which one unit opts into treatment and numerous others remain untreated. Accordingly, it is a less flexible tool than GPSM. In creating my synthetic controls, I generally use the same covariates as I do in propensity score estimation, but I also follow Parker (2021) by using three lagged years of the dependent variable (mean absolute CFscores).

c. My Data

As mentioned above, my measure of ideology comes from Bonica-Woodruff (2015) and Bonica's DIME database. Since each judge received only a single ideology score (even if she served multiple years), I have selected court-year as my unit of analysis. For most of the tests I perform, the outcome variable is the average absolute ideology score of the judges sitting on a given court in a given year—the theoretical range of that variable is 0-2. On occasion, I check my results using median absolute ideology in a court-year as the outcome variable. However, since this measure is strongly correlated with my primary dependent variable, I report these (often redundant) robustness tests sparingly in the body of the paper.

To obtain data on state selection methods by year, I relied primarily on data compiled by the American Judicature Society and the Brennan Center. I supplemented these resources as necessary with data from other academics working in the field (e.g., Bonica & Sen, 2021). The tests I run below also include several covariates, including the partisan composition

of state legislatures and partisan control of the governor's mansions by state and year. I acquired data on state legislatures from the Institute for Public Policy and Social Science Research[9] and on state governors from Jacob Kaplan (2021).

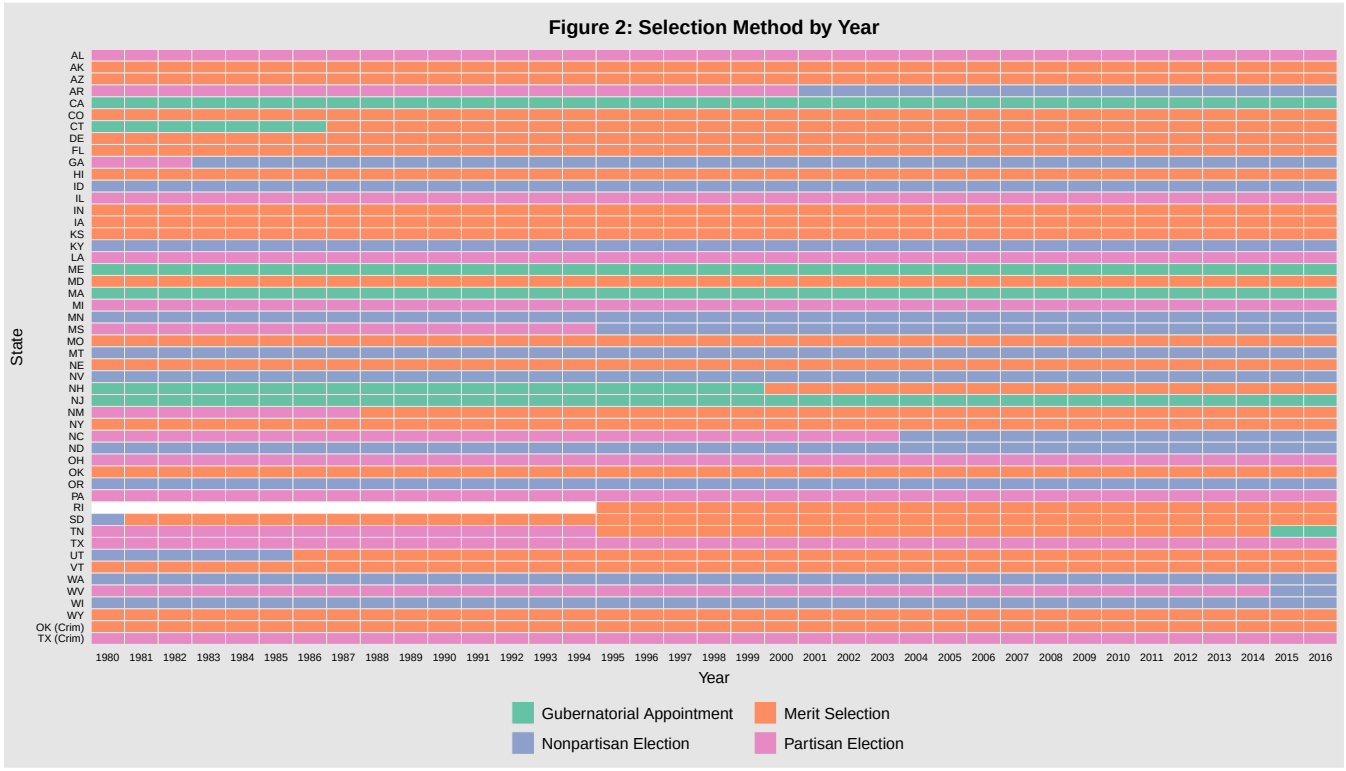
Table 1 provides summary statistics for my data, while Figure 2 visually depicts selection method by state and year. The ideology results are notably consistent with Bonica & Sen's (2017b) predictions. Gubernatorial appointment and partisan elections introduce partisanship into the selection process, and so we would expect the judges ultimately selected to reflect the underlying state ideology. The states with gubernatorial appointment—California, Maine, New Jersey, Massachusetts, and New Hampshire (through 1999)—are either Democratic leaning or solidly blue. Unsurprisingly, they produce the most liberal judges, on average. Meanwhile, the states that used partisan elections during my sample period—Alabama, Louisiana, Texas, West Virginia, Illinois, Michigan, Ohio, Pennsylvania, Arkansas (through 2000), North Carolina (through 2003), and Tennessee (through 1994)—tend to be more conservative. Predictably, they end up with the most conservative judges. Meanwhile, states with merit selection or nonpartisan elections end up in the middle, with the average judge just slightly left of center. Bonica & Sen (2017b) expected those methods to produce similar types of judges, because both leave a smaller role for ideology in selection, allowing the distribution of judges to look like the distribution of the attorney pool generally.

[9]Specifically data compiled by Cark Klarner in 2013.

Table 1: Summary Statistics

Variable	*Partisan Elections*	*Nonpartisan Elections*	*Merit Selection*	*Gubernatorial Appointment*
Number of Observations				
Judge-Years	**2954**	**2817**	**4775**	**1196**
Governor				
Proportion of Years with Democratic Governor	**0.42**	**0.54**	**0.49**	**0.39**
Proportion of Years with Independent Governor	**0**	**0.01**	**0.01**	**0.05**
Proportion of Years with Republican Governor	**0.58**	**0.45**	**0.49**	**0.56**
Legislature				
Mean Democratic Senate Proportion	**0.59**	**0.51**	**0.5**	**0.6**
Mean Democratic House Proportion	**0.6**	**0.51**	**0.51**	**0.59**
Proportion of Years with Unified Government	**0.44**	**0.56**	**0.46**	**0.46**
Ideology				
Mean Absolute Value of CFscore	**0.66**	**0.72**	**0.74**	**0.71**
Mean CFscore	**0.15**	**-0.15**	**-0.15**	**-0.25**

In terms of extremity, the summary statistics do not suggest substantial differences between the methods. States with judges selected in partisan contests had, on average, the least extreme judges, while states with merit selection had the most extreme judges. However, the difference between these two methods is only abut 0.08 units on a 0-2 scale. Of course, these naive differences cannot distinguish between the impact of selection methods and the influence of other factors—that task requires more extended analysis.



4. Results

a. Generalized Propensity Score Matching (GPSM)

I begin with the results from GPSM. I implemented this strategy using the *PSweight* package in R (Zhou et al., 2021), after first cross-validating to select the variables in my propensity score model. Figure 3 plots the propensity scores for each possible selection method given the actual method for that state. These graphs provide strong evidence for the common support assumption—there is considerable propensity score overlap for each possible selection method.

Figure 3: Propensity Score Overlaps

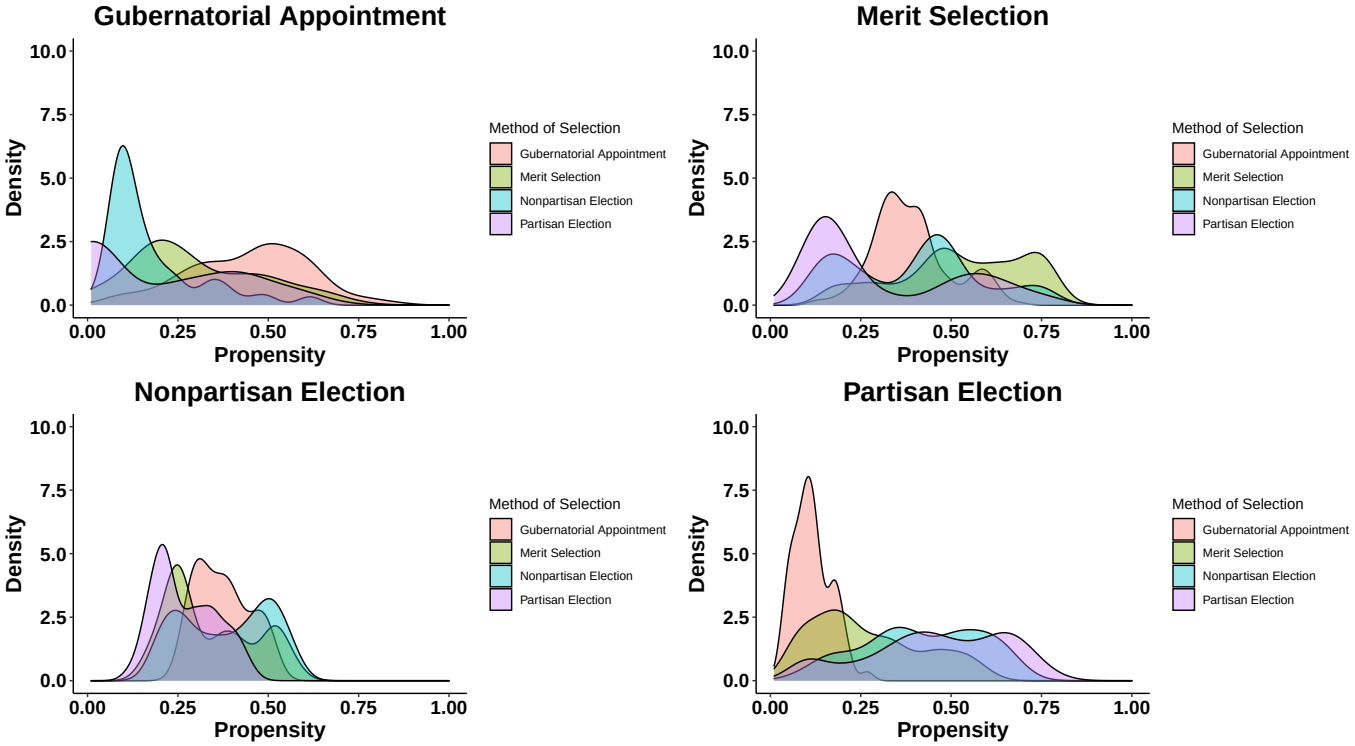


Table 2 provides my primary results for this matching analysis. Column (1) lists the methods compared. The estimand reported in column (2) is the ATM: average treatment effect in the matched sample (Greifer & Stuart, 2021; Zhou et al., 2021). The standard errors in column (3) come from the empirical sandwich variance (Zhou et al., 2021), while column (4) provides p-values rounded to three decimal places.

Table 2: Generalized Propensity Score Matching Results (Mean Judge Extremity)

Pair of Methods	*ATM Estimate*	*SE*	*p-value*
Merit Selection - Gubnatorial Appointment	**-0.25*** **	**0.07**	**0**
Nonpartisan Elections - Gubernatorial Appointment	**-0.21*** **	**0.07**	**0.002**
Partisan Elections - Gubernatorial Appointment	**-0.21*** **	**0.06**	**0.001**
Nonpartisan Elections - Merit Selection	**0.04**	**0.04**	**0.313**
Partisan Elections - Merit Selection	**0.04**	**0.03**	**0.131**
Partisan Election - Nonpartisan Elections	**0**	**0.04**	**0.923**

Variables contributing to propensity scores: Year, Party of Governor, State Senate Composition, State House Composition, Unified Gov.

ATM obtained using PSweight package in R.

p < 0.9 = *, p < 0.95 = **, p < 0.99 = ***

The results are stark—gubernatorial appointment appears to produce considerably more extreme judges than other methods of selection. On a 0-2 scale, moving to gubernatorial appointment from some other method is associated with increases in extremity ranging from 0.21 to 0.25—all three estimates are significant at the $p < 0.01$ level. For context, the standard deviation of mean absolute extremity is approximately 0.24. That is, moving to gubernatorial appointment from a different selection method appears to increase extremity by approximately a full standard deviation.

Meanwhile, the table does not reveal any substantial differences in judge extremity among any pair of methods that does not include gubernatorial appointment. There appear to be no substantial differences between merit commissions and either form of election, nor do partisan elections produce more extreme judges than non-partisan elections.

Little changes when I substituted median court extremity in as the dependent variable (Table 3). The point estimates for pairwise comparisions involving gubernatorial appointment were still large and negative, indicating once again that the method appears to produce particularly extreme judges. The associated p-values do decrease, which makes sense given

that the standard errors attached to these estimates are more than double the ones reported in Table 2. Indeed, only one of the p-values falls below 0.05 in two-tailed tests—however, the others do approach or exceed significance at the $p < 0.1$ level.

Table 3: Generalized Propensity Score Matching Results (Median Judge Extremity)

Pair of Methods	*ATM Estimate*	*SE*	*p-value*
Merit Selection - Gubnatorial Appointment	**-0.35****	**0.16**	**0.033**
Nonpartisan Elections - Gubernatorial Appointment	**-0.26**	**0.16**	**0.112**
Partisan Elections - Gubernatorial Appointment	**-0.3***	**0.16**	**0.058**
Nonpartisan Elections - Merit Selection	**0.08***	**0.04**	**0.056**
Partisan Elections - Merit Selection	**0.04**	**0.03**	**0.174**
Partisan Election - Nonpartisan Elections	**-0.04**	**0.05**	**0.34**

Variables contributing to propensity scores: Year, Party of Governor, State Senate Composition, State House Composition, Unified Gov.

ATM obtained using PSweight package in R.

p < 0.9 = *, p < 0.95 = **, p < 0.99 = ***

Unlike in Table 2, one other pair of method does produce a significant result. Row (4) indicates that nonpartisan elections produce slightly more extreme judges than merit commissions do. The point estimate of 0.08 achieves significance at the $p < 0.1$ level in a two-tailed test, and just barely misses significance at the $p < 0.05$ level. However, given that the companion value in Table 2 is small and non-significant, it would be unwise to draw conclusions about the nonpartisan election-merit commission diad from this estimate alone. Indeed, the synthetic control analyses I conduct later in this article counsel against rejecting the null hypothesis of no difference when it comes to these two methods.

b. Fixed Effect Counterfactual Estimators (FEct)

The GPSM results strongly suggest that unconstrained gubernatorial appointment is distinctive among selection methods when it comes to producing ideologically extreme judges. My FEct analyses yield the same conclusion.

The *fect* package in R (created by Yiqing Xu and coauthors) allows applied researchers to implement three recently developed estimators: the Two-way Fixed Effects Counterfactual, the Interactive Fixed Effects Counterfactual (iFe), and a matrix completion estimator (MC). These estimators are analytically complex, but are in principle comparable to the traditional Two-Way Fixed Effects estimator (Liu et al., 2022). They have two main advantages over the TWFE estimator. First, they do not assume a constant treatment effect over time; and second, they allow for straightforward evaluation of the main identification assumption (no pretrend).

In principle, researchers should chose the estimator that best ensures the identification assumption is met. Fortunately, the *fect* package will select the optimal estimator for the user via cross-validation—in this case, the iFe and MC estimators are best suited to my analyses. I explore four types of transitions between selection methods using these estimators:

(1) Other Method to Nonpartisan Election

(2) Other Method to Merit Commission

(3) Unconstrained Gubernatorial Appointment to Other Method

(4) Partisan Election to Other Method

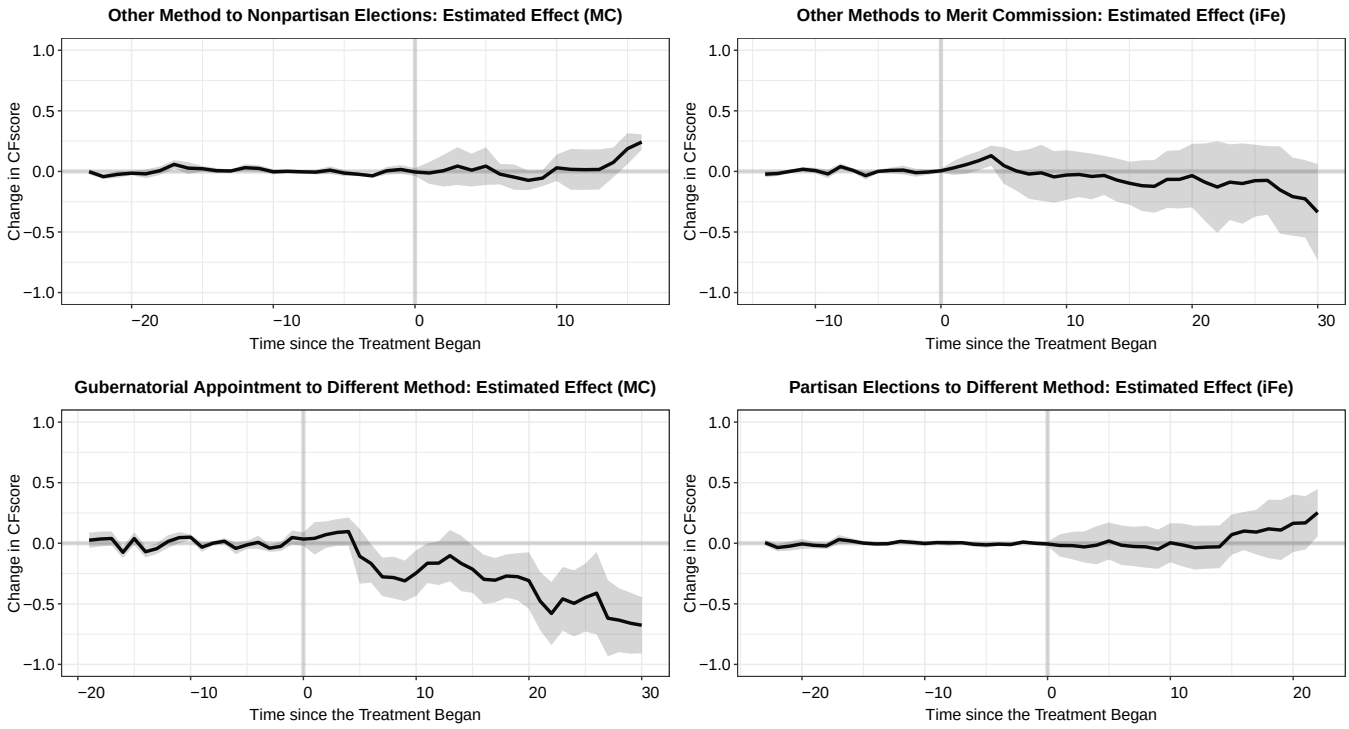
As one might notice, the direction of change is not constant across these analyses. For nonpartisan elections and merit commissions, I examine the transition *to* these methods from some other selection institution. When it comes to unconstrained gubernatorial appointment and partisan elections, I explore shifts *away* from those methods. The structure of my data dictates this approach—between 1980 and 2016, nonpartisan elections and merit commissions became increasingly popular, while unconstrained gubernatorial appointment and partisan elections declined.[10]

I present the results of these analyses in Figure 4. The x-axis plots the time since the treatment began, while the y-axis plots the dynamic treatment effect. Unsurpisingly

[10]There are some indications that partisan elections and unconstrained gubernatorial appointment may experience a resurgence in the coming years (see Bonica & Sen, 2021 for details).

given my GPSM results, only the shift away from gubernatorial appointment produces significant results. Abandoning gubernatorial appointment is associated with a substantial decline in mean extremity for a given court-year—the ATT over the post-treatment period is approximately -0.24 ($SE = 0.09$, $p < 0.01$).[11]

Figure 4: Changes in Extremity Using Fixed Effects Counterfactual Estimators



While the top right and bottom right panels both hint that merit commissions and partisan elections (respectively) might be associated with a decrease in extremity, the ATTs for those analyses do not approach conventional standards of statistical significance. The ATT associated with adopting a merit commission is approximately -0.05 ($SE = 0.09$, $p = 0.60$), and the effect associated with eliminating partisan elections is about 0.02 ($SE = 0.06$, $p = 0.72$).

While useful, these FEct estimates should be treated with a note of caution. The dataset for these estimates includes 11 states that switched election methods and 36 that

[11] All models include unified government as a covariate, as well as court and year as the equivalent of fixed effects. Every estimate reported in this section weights treated observations equally.

did not. This limited data makes it difficult to verify that the identification assumptions for the various FEct methods hold. I run robustness checks for these estimates in Appendix D—while all models perform reasonably well when it comes to the placebo tests described in Liu et al. (2022), they fare worse in the non-parametric pretrend tests those authors suggest. Liu et al. note that those non-parametric tests are "clearly a conservative" standard; they require that "all pretreatment periods generate significant results" when testing for equivalence to zero (p. 10). Visual inspection of the graphs in Figure 4 do not reveal any obvious monotonic pretreatment trends, which gives me confidence that the estimates are worth reporting. Nevertheless, they should not be received uncritically.

c. Synthetic Controls

My synthetic control analyses examine seven sets of selection method transitions:

(1) The transition from partisan elections to nonpartisan elections in Arkansas;
(2) The transition from partisan elections to nonpartisan elections in North Carolina;
(3) The transition from partisan elections to nonpartisan elections in Mississippi;
(4) The transition from partisan elections to a merit commission in New Mexico;
(5) The transition from partisan elections to a merit commission in Tennessee;
(6) The transition from nonpartisan elections to a merit commission in Utah; and
(7) The transition from unconstrained gubernatorial appointment to merit commission in Connecticut.

While other states switched selection methods during the period under study, applying the SCM to units with relatively few pre-treatment or post-treatment years can create inference issues (Abadie, 2021). Accordingly, I only examine transitions with at least six years of pre-treatment and post-treatment data.[12] Appendix C breaks down the composition

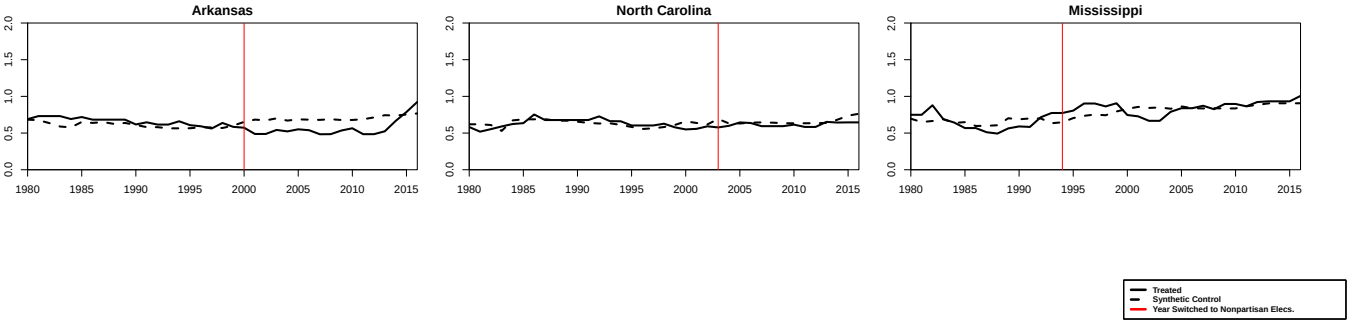
[12]While New Hampshire—which transitioned from unconstrained gubernatorial appointment to a merit commission in 2000—meets these criteria, it is not amendable to synthetic control analysis because its pre-treatment state supreme court was more extreme than that of any potential donor state. However, in Appendix E, I apply a recently developed technique (synthetic difference-in-differences) to this case (Arkhangelsky et al., 2021).

of each synthetic control by donor state for interested readers.

Figure 5 contains the synthetic control plots for each state that switched from partisan elections to nonpartisan elections. In each panel, the solid black line represents the mean absolute CFscore for the supreme court of the state under study. The dotted black line depicts the synthetic control, and the solid red line marks the year the state decided to change selection methods.

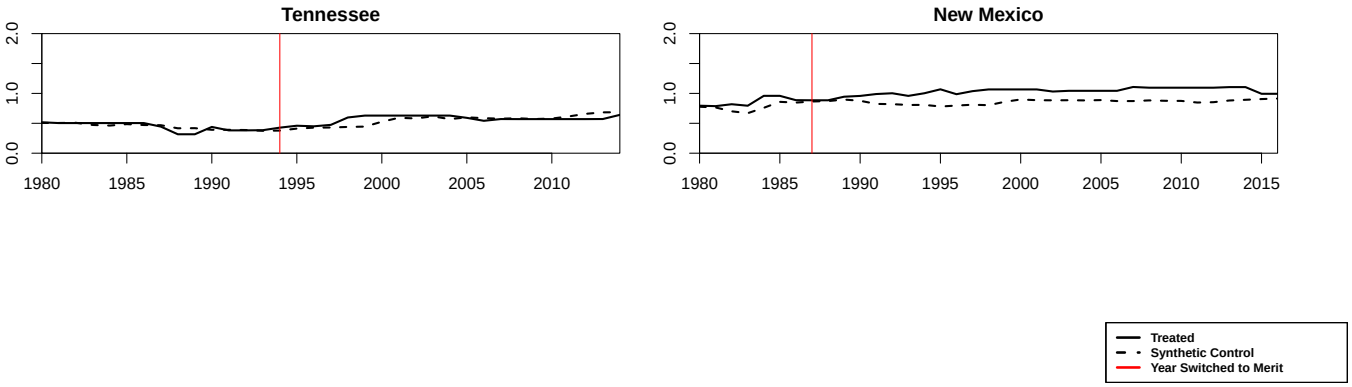
In the center and right panels of Figure 5 (North Carolina and Mississippi), the synthetic control and treatment state appear quite similar in the post-treatment period. Visually, there appears to be no substantial change in mean extremity associated with the switch. When it comes to Arkansas, though, a treatment effect appears plausible. For the vast majority of the post-treatment period, the average absolute extremity of the sythentic control is noticeably greater than that of Arkansas.

Figure 5: Synthetic Control Plots for States Switched from Partisan to Nonpartisan Elections



The results for Tennessee and New Mexico are similarly inconsistent. While both transitioned from partisan elections to merit selection, only the New Mexico plot shows any possible change in extremity. The New Mexico panel suggests that merit commissions are associated with increased extremity; the Tennessee panel indicates no meaningful shifts.

Figure 6: Synthetic Control Plots for States Switched from Partisan Elections to Merit Commission



Figures 7 and 8 examine two other types of transitions: the move from nonpartisan elections (Figure 7) and unconstrained gubernatorial appointment (Figure 8) to merit commissions. The former shift (which occurred in Utah) had no obvious impact, while the latter change (in Connecticut) appears to have led to a substantial decrease in extremity. It is not clear, though, whether this decrease is distinguishable from statistical noise.

Figure 7: Synthetic Control Plots for States Switched from Nonpartisan Election to Merit Commissions

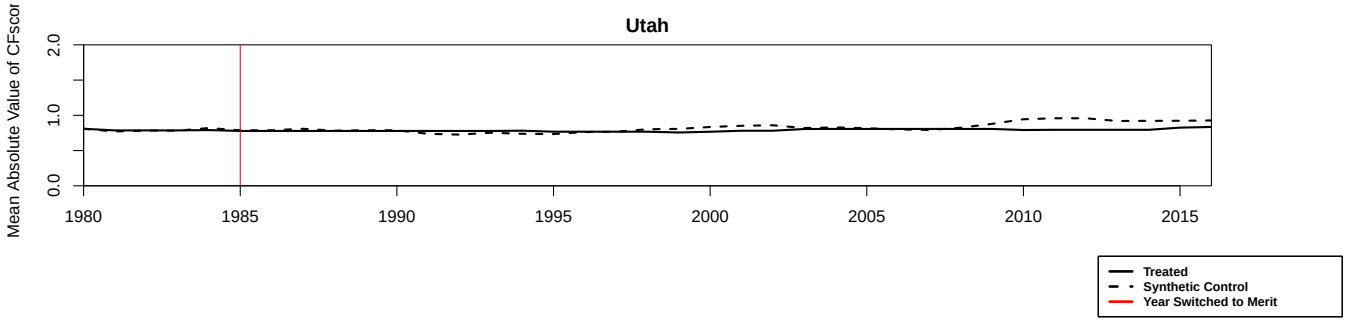
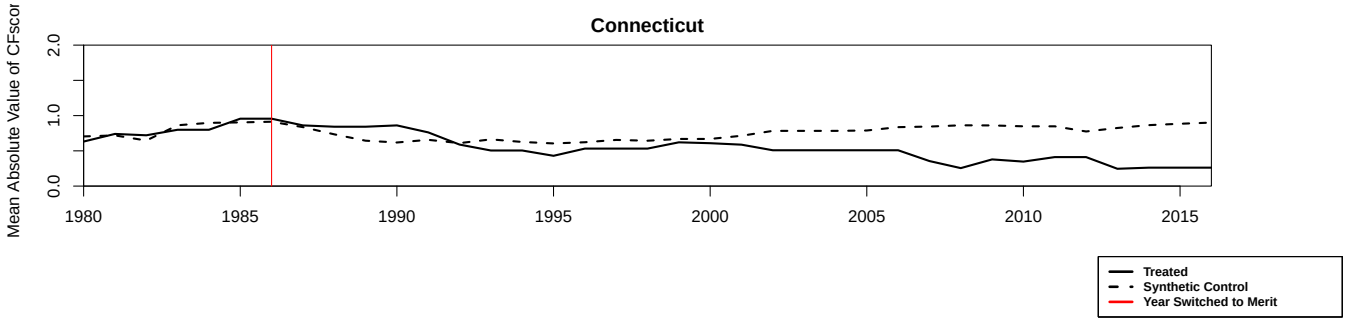


Figure 8: Synthetic Control Plots for States Switched from Gubernatorial Appointment to Merit Commissions



Fortunately, statisticians have developed a more precise measure of statistical significance for synthetic controls than mere visual inspection. This solution involves "iterative placebo tests" (e.g., Abadie et al., 2011). The basic idea underlying this technique is to re-run the synthetic controls iteratively using each control state as a "placebo" treated state. This process provides an estimate of the distribution of synthetic control plots that might occur by random chance.

To obtain estimated treatment effects for both the original synthetic control and each placebo iteration, I calculate average difference between the synthetic control and the treated unit for both the pre-treatment period and the post-treatment period. By subtracting the former from the latter, I can obtain an effect estimate (e.g., Parker, 2021). If the true treatment effect is statistically significant, it should be more negative (or positive) than 95% of the "placebo" treatment effects. The entirety of this process is sometimes referred to as a permutation test. I report the results of these analyses in Table 4.

Table 4: Iterative Placebo Tests

State	Estimated Treatment Effect	Placebos With Larger Treatments Effects	Placebos With Smaller Treatments Effects	Proportion of Placebos With Larger Treatment Effects
Partisan to Nonpartisan Elections				
Arkansas	**-0.18**	**5**	**3**	**0.62**
North Carolina	**-0.04**	**4**	**4**	**0.50**
Mississippi	**0.02**	**3**	**5**	**0.38**
Partisan Elections to Merit				
Tennessee	**0.02**	**4**	**4**	**0.50**
New Mexico	**0.10**	**2**	**6**	**0.25**
Nonpartisan Elections to Merit				
Utah	**-0.04**	**6**	**3**	**0.67**
Gubernatorial Appointment to Merit				
Connecticut	**-0.23**	**3**	**1**	**0.75**

** = p < 0.1, ** = p < 0.05, *** = p < 0.01*

The Connecticut case permutation results largely reinforce the conclusion drawn early in this paper. The point estimate for the switch from gubernatorial appointment to merit commission is in the expected direction, and appropriately of the same magnitiude as the results from the GPSM and FEct analyses. The p-value of the point estimate is only 0.25—however, the lack of significance is likely because of the limited power of the test. For the

point estimate to reach significance at the $p < 0.05$ level, the estimated effect size would be to be -0.36—appropriately 150% of the standard deviation in mean absolute ideology. Given this context, the results of the permutation test tend to reinforce my previous conclusion: gubernatorial appointment appears to produce particularly extreme judges.

Likewise, the other pairwise comparisons provide no reason to doubt my earlier results. The point estimates for Arkansas and New Mexico are relatively large (-0.18 and 0.10, respectively). However, neither reaches conventional levels of statistical significance. Moreover, there is no evidence from either the GPSM or FEct analyses suggesting that nonpartisan elections decrease extremity or than merit commissions increase it. Indeed, the FEct estimates suggest the opposite.[13]

5. Discussion

The analyses I perform in this paper suggest two primary conclusions: first, it appears that unconstrained gubernatorial appointment produces more extreme judges than those selected by other methods; and second, other pairs of methods do not seem to create substantial differences in extremity.

The distinctiveness of unconstrained gubernatorial appointment seems clear in hindsight. The alternatives—nonpartisan elections, merit selection, and partisan elections—all require the participation of non-extremists in selection. The median voter in almost every state will not be an extremist, and merit commissions are typically (if not invariably) set up to involve a number of voices. By contrast, an ideologically-committed governor can act alone in nominating an extremist judge—she need not receive permission from others before doing so. True, the governor still must get that nominee confirmed by another body, but she has an agenda-setting advantage—if the confirming entity rejects her choice, she can simply nominate another ideologue.[14]

[13] Another common test for significance with synthetic controls is the RMSPE ratio test—I report the results of these analyses in Appendix B.

[14] It goes without saying that not every governor is an extremist, and even some extremist governors will

These findings have the potential to inform public policy. Both politicians and members of the public are likely to have preferences when it comes to the ideological commitment of their jurists. Most people—at least prima facie—would probably say they prefer more moderate judges. However, ideological moderation is not always a virtue. Many revered Supreme Court decisions in the twentieth century—such as *Brown v. Board of Education* and *Miranda v. Arizona*—were determined expansions of constitutional rights that more moderate judge may well have opposed. The appropriate level of ideological extremity for judges is a question of political theory, and one that I do not purport to answer in this paper.

However, it is difficult to overlook the applicability of my findings to the federal selection of judges. The federal process is essentially a form of unconstrained gubernatorial appointment. The president faces no ex ante restrictions on who she might select to fill federal judicial vacancies.[15] While it is by no means certain that my results generalize to the federal context, they do at least suggest that the federal government's selection method produces particularly extreme judges. Given eroding public confidence in the Supreme Court (Barnes & Kim, 2021), it might be worthwhile to consider reforming the way federal judges ascend to the bench. Altering the process to produce more moderate judges may restore some faith in the judiciary as a neutral arbitrator. Elections for federal judge would likely be untenable (the salience of these elections would probably produce results similar to congressional polarization), but it still might be worth exploring other alternatives to the status quo. For example, my results indicate that merit selection produces more moderate judges than unconstrained appointment. Merit commission composition differs widely across states, and some types of arrangements might be well-suited for federal use. This subject (and its implications for federal selection) is a promising area for future research.

I should also acknowledge the limitations of this study. By its terms, it attempts causal

prefer more moderate judges. However, even a few governors bent on nominating ideological purists could be enough to substantially raise the average extremity of judges selected via unconstrained gubernatorial appointment.

[15]Federal judges need only be U.S. citizens—the Constitution imposes no other restrictions on who may ascend to the bench.

inference with multiple non-ordered treatments, an econometrically challenging proposition. While I make every effort to reinforce my conclusions with multiple inference strategies, these results should be viewed with caution. In addition, this research is limited in scope. It tests only the impact of judicial selection methods used to fill initial full-term vacancies. Given the causal inference challenges this work already presents, it would have been extremely difficult to determine how these methods interact with (1) reappointment methods and (2) methods for filling interim vacancies. That reality does not imply other aspects of selection are unworthy of study; it was simply not feasible to address them in this paper. However, the field would greatly benefit from more research into these mechanisms.

These caveats notwithstanding, this paper represents a robust first effort at answering an important question. It move beyond assessing the directional impact of selection methods and seeks to understand how those methods affect the ideological commitment of the judges they produce. This information will be useful to any state considering changing selection method, as well as to those engaged in the long-running debate over which selection method should prevail.

References

Abadie, Alberto (2021) "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," 59 *J. of Economic Literature* 391.

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

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

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2015) "Comparative Politics and the Synthetic Control Method," 59 *American J. of Political Science* 495.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager (2021) "Synthetic Difference-in-Differences," 111 *American Economic Review* 4088.

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

Andreottola, Giovanni (2021) "Signaling Valence in Primary Elections", 126 *Games & Economic Behavior* 1.

Barnes, Robert and Seung Min Kim (Sept. 26, 2021), "Supreme Court Observers See Trouble Ahead as Public Approval of Justices Erodes," *Washington Post*.

Bifulco, Robert, Ross Rubenstein, and Hosung Sohn (2017) "Using Synthetic Controls to Evaluate the Effect of Unique Interventions: The Case of Say Yes to Education," 41 *Evaluation Review* 593.

Blackwell, Matthew and Adam N. Glynn (2018) "How to Make Causal Inferences with

Time-Series Cross-Sectional Data under Selection on Observables," 112 *American Political Science Review* 1067.

Bonica, Adam (2014) "Mapping the Ideological Marketplace," 58 *American J. Political Science* 367, 367.

Bonica, Adam and Maya Sen (2021) *The Judicial Tug of War.* Cambridge, MA: Cambridge University Press.

Bonica, Adam and Maya Sen (2017) "A Common-Space Scaling of the American Judiciary and Legal Profession," 25 *Political Analysis* 114.

Bonica, Adam and Maya Sen (2017) "The Politics of Selecting the Bench from the Bar: The Legal Profession and Partisan Incentives to Introduce Ideology into Judicial Selection," 60 *J. Law & Economics* 559.

Bonica, Adam and Michael J. Woodruff (2015) "A Common-Space Measure of State Supreme Court Ideology," 31 *J. Law, Economics & Organizations* 472.

Botosaru, Irene, and Bruno Ferman (2019) "On the Role of Covariates in the Synthetic Control Method," 22 *Econometrics J.* 117.

Brace, Paul and Brent D. Boyea (2008) "State Public Opinion, the Death Penalty, and the Practice of Electing Judges", 52 *American J. Political Science* 360.

Brown, Adam R. (2018) "The Role of Constitutional Features in Judicial Review," 18 *State Politics & Policy Quarterly* 351.

Burnett, Craig M. and Lydia Tiede (2014) "Party Labels and Vote Choice in Judicial Elections," 43 *American Politics Research* 232.

Caldarone, Richard P., Brandice Canes-Wrone, and Tom S. Clark (2009) "Partisan Labels and Democratic Accountability: An Analysis of State Supreme Court Abortion Decisions," 71 *J. of Politics* 560.

Canes-Wrone, Brandice, Tom S. Clark, and Amy Semet (2018) "Judicial Elections, Public Opinion, and Decisions on Lower-Salience Issues," 15 *J. Empirical Legal Studies* 672.

Canes-Wrone, Brandice, Tom S. Clark and Jason P. Kelly (2014) "Judicial Selection and Death Penalty Decisions," 108 *American Political Science Review* 23.

Czarnezki, Jason J. (2005) "A Call for Change: Improving Judicial Selection Methods," 89 *Marquette Law Review 169.*

de Chaisemartin, Clement and Xavier d'Haultfoeuille (2020) "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," 110 *American Economic Review* 2964.

Failinger, Maria A. (2005) "Can a Good Judge Be a Good Politician?: Judicial Elections From a Virtue Ethics Approach," 70 *Missouri Law Review* 433.

Ferman, Bruno, and Christine Pinto (2016) "Revisiting the Synthetic Control Estimator," *MPRA Paper No. 86495.*

Fiorina, Morris (2016) *Has the American Public Polarized?* Stanford, CA: Hoover Institution.

Fitzpatrick, Brian T. (2017) "The Ideological Consequences of Selection: A Nationwide Study of the Methods of Selecting Judges," 70 *Vanderbilt Law Review* 1729.

Geyh, Charles G. (2019) *Who Is to Judge?* New York, NY: Oxford University Press.

Goelzhauser, Greg (2019) *Judicial Merit Selection.* Philadelphia, PA: Temple University Press.

Gordon, Sanford C. and Gregory A. Huber (2007) "The Effect of Electoral Competition on Incumbent Behavior," 2 *Quarterly J. Political Science* 107.

Greifer, Noah and Elizabeth A. Stuart (2021) "Choosing the Estimand When Matching or Weighting in Observational Studies." *arXiv:2106.10577.*

Hall, Andrew B. (2019) *Who Wants to Run?: How the Devaluing of Political Office Drives Polarization.* Chicago, IL: University of Chicago Press.

Henry, M.L. (1985) *The Success of Women and Minorities in Achieving Judicial Office.* New York, NY: American Express Foundation.

Hirano, Shigeo and James M. Snyder (2014) "Primary Elections and the Quality of Elected Officials," 9 *Quarterly J. Political Science* 473.

Imai, Kosuke and In Song Kim (2019) "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" 63 *American J. of Political Science* 467.

Imai, Kosuke and In Song Kim (2020) "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data," 29 *Political Analysis* 405.

Kane, Jenna B. (2018) "Informational Need, Institutional Capacity, and Court Receptivity: Interest Groups and Amicus Curiae in State High Courts," 71 *Political Research Quarterly* 881.

Kaplan, Jacob (2021) "United States Governors 1775-2020," *OpenICPSR.*

Kaul, Ashok, Stefan Klobner, Gregor Pfeifer, and Manuel Schieler (2015) "Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates," *MPRA Paper No. 83790.*

Kowal, John F. (2016) *Judicial Selection for the 21st Century* Brennan Center for Justice.

Kritzer, Herbert M. (2018) "Judicial Elections in the 2010s," 67 *DePaul Law Review* 387.

Lamphier, Elizabeth K. (2011) "Justice Run Amok: Big Money, Partisanship, and State Judiciaries," 2011 *Michigan State Law Review* 1327.

La Raja, Raymond J. and Brian Schaffner (2015) *Campaign Finance and Political Polarization: When Purists Prevail.* Ann Arbor, MI: University of Michigan Press.

Li, Liang and Tom Greene (2013) "A Weighting Analogue to Pair Matching in Propensity Score Analysis," 9 *International Journal of Biostatistics* 1.

Li, Fan and Fan Li (2019) "Propensity score weighting for causal inference with multiple treatments," 13 *Annals of Applied Statistics* 2389.

Lim, Claire S.H. and James M. Snyder (2015) "Is More Information Always Better? Party Cues and Candidate Quality in U.S. Judicial Elections," 128 *J. Public Economics* 107.

Liu, Licheng, Ye Wang, and Yiqing Xu (2022), "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data," ___ *American J. Political Science* ___.

Lopez, Michael J. and Roee Gutman (2017) "Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas," 32 *Statistical Science* 432.

Masket, Seth and Boris Shor (2015) "Polarization without Parties: Term Limits and Legislative Partisanship in Nebraska's Unicameral Legislature," 15 *State Politics & Policy Quarterly* 67.

McCarty, Nolan, Keith Poole and Howard Rosenthal (2006) *Polarized America.* Cambridge, MA: MIT Press.

Nelson, Michael J. (2014) "Responsive Justice," 2 *J. Law & Courts* 117 (2014).

Parker, Brett (2021) "Death Penalty Statutes and Murder Rates: Evidence From Synthetic Controls," 18 *J. Empirical Legal Studies* 488.

Parker Brett (n.d.) "Do Partisan Judicial Elections Cause Convergence in Capital Cases? Empirical Evidence from Close Contests," *CELS Working Paper.*

Renberg, Kristen M. (2020) "The Impact of Retention Systems on Judicial Behavior: A Synthetic Controls Analysis of State Supreme Courts," 41 *Justice System J.* 292.

Thomsen, Danielle M. (2014) "Ideological Moderates Won't Run: How Party Fit Matters

for Partisan Polarization in Congress," 76 *J. of Politics* 786.

Warner, Seth (n.d.) "Measuring Executive Ideology and its Influence," *APSA Working Paper*.

Wilhelm, Teena, Richard L. Vining Jr., Ethan D. Boldt, and Bryan M. Black (2017) "Judicial Reform in the American States: The Chief Justice as Political Advocate," 20 *State Politics & Policy Quarterly* 135.

Xu, Yiqing (2017) "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," 25 *Political Analysis* 57.

Zhou, Tianhui, Guangyu Tong, Fan Li, Laine E. Thomas, and Fan Li (2021) "PSweight: An R Package for Propensity Score Weighting Analysis," _ *The R J.* _.

Appendices

A. Data Collection

As noted in the body of this article, I began this project using the Bonica-Woodruff data on state judicial ideology. The Bonica-Woodruff data contains measures of ideology for approximately 95% of state high court judges sitting between 1990 and 2012. While this dataset provided an excellent starting point for my paper, it covered only a handful of changes in selection methods. Judicial selection institutions were relatively volatile in the 1980s, when five states switched from one method to another. In order to exploit these changes in selection method for inference purposes, I expanded the Bonica-Woodruff dataset back to 1980 (as well as forward to 2016).

Expanding the dataset meant collecting ideology scores for judges that did not sit between 1990 and 2012. In order to do so, I needed to search through Adam Bonica's DIME database, which contains estimated ideology scores for approximately 14.7 million donors between 1979 and 2014. For each missing judge, I conduct a manual search (assisted by R) for that judges name, as well as any plausible nicknames, misspellings, or truncations. Many contribution records include (1) a state of residence, (2) an occupation, (3) a local address, and (4) an employer. I used various combinations of this data in order to verify that a particular contribution record belonged to a judge with that name. In some cases, the judge in question would list their occupation as "judge" or "justice" of the relevant court, making identification easy. For others, I had to scour online records, such as obituaries, to determine whether employment or location information could identify a contribution as belonging to a specific judge.

Unsurprisingly, not every judge had a CFscore (particularly judges who served in the 1980s). For these individuals, I checked to see if they were appointed by a governor with a known CFscore (following Bonica and Woodruff, 2015). Between these two methods, I was able to achieve approximately 92% completeness for the sample period.

B. RMSPE Ratios

Given the limited power of the permutation tests, it makes sense to check my results using an alternative inference strategy in the synthetic control toolbox: root mean squared prediction error (RMSPE) ratios. One recent paper (Parker, 2021) that deploys RMSPE ratios describes their utility as follows:

> "One alternative way of conducting causal inference with synthetic controls is to calculate root mean squared prediction errors (RMSPE)—defined as the square root of the mean squared difference between the [outcome variable] for the synthetic control and actual data—for pre and posttreatment periods for each of the placebo states and for the actual [treated] state. If [the treatment] were to cause [a change] in the [outcome variable], we would expect the ratio of postperiod RMSPE to preperiod RMSPE to be higher in the [treated] states than in the placebo states."

Put otherwise, the ratio of postperiod RMSPE/preperiod RMSPE provides an alternative estimate of the magnitude of the treatment effect. Some scholars actually prefer this estimand to the one obtained from the permutation tests, as the former limits the influence of outlying placebo treatment effect estimates if the synthetic control assigned to that placebo state fits the state poorly.

In figures B.1 through B.4, below, I report the results of these RMPSE ratio tests. None of these tests substantially alter the conclusions I drew in the main text. As there, the analyses for Arkansas and Connecticut stand out, but they do not quite produce statistically significant results.

Figure B.1: Ratios of Post-Treatment RMSPE to Pre-Treatment RMSPE (Partisan to Nonpartisan Elections

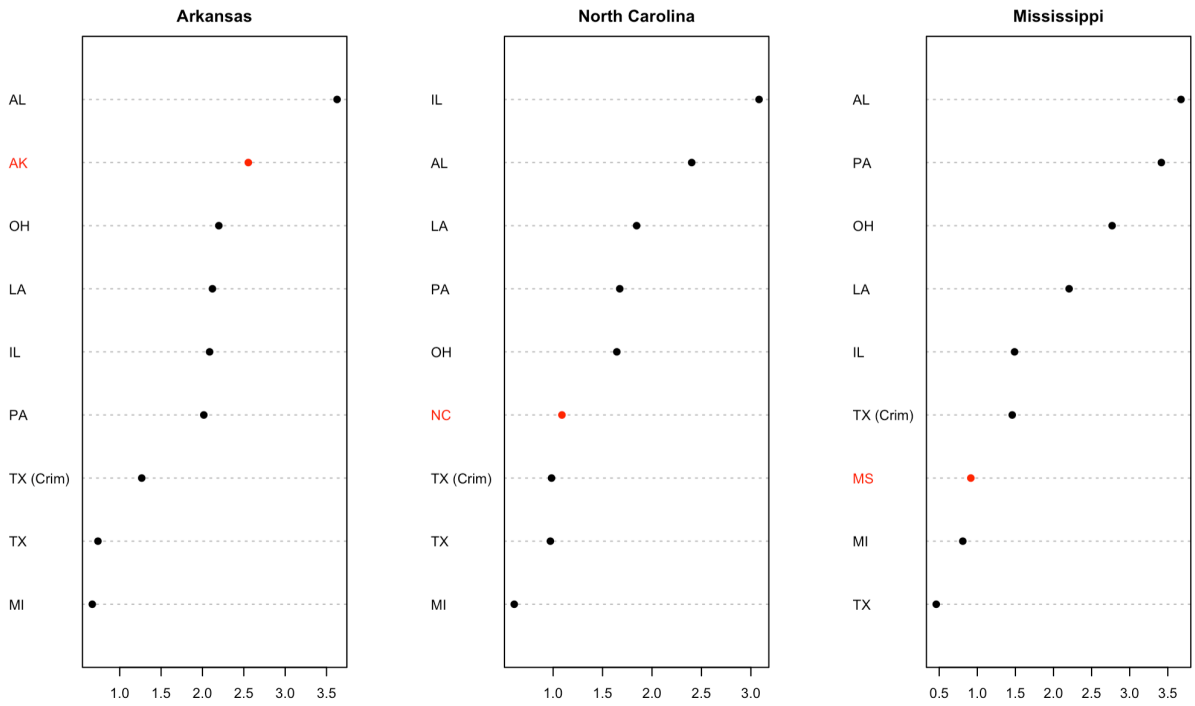


Figure B.2: Post-Period RMSPE / Pre-Period RMSPE (Partisan to Merit)

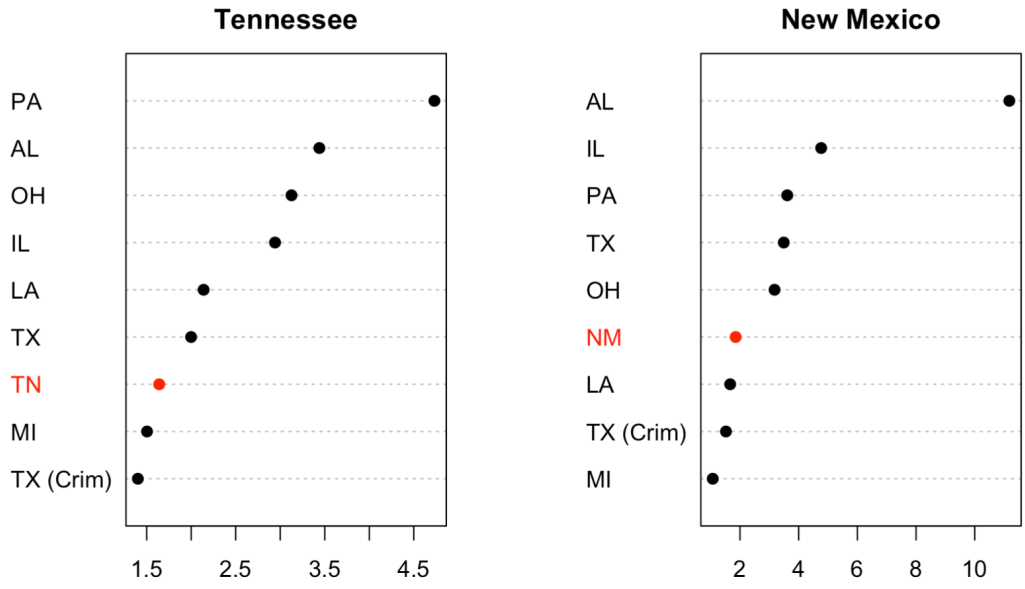


Figure B.3: Post-Period RMSPE / Pre-Period RMSPE (Nonpartisan Election to Merit)

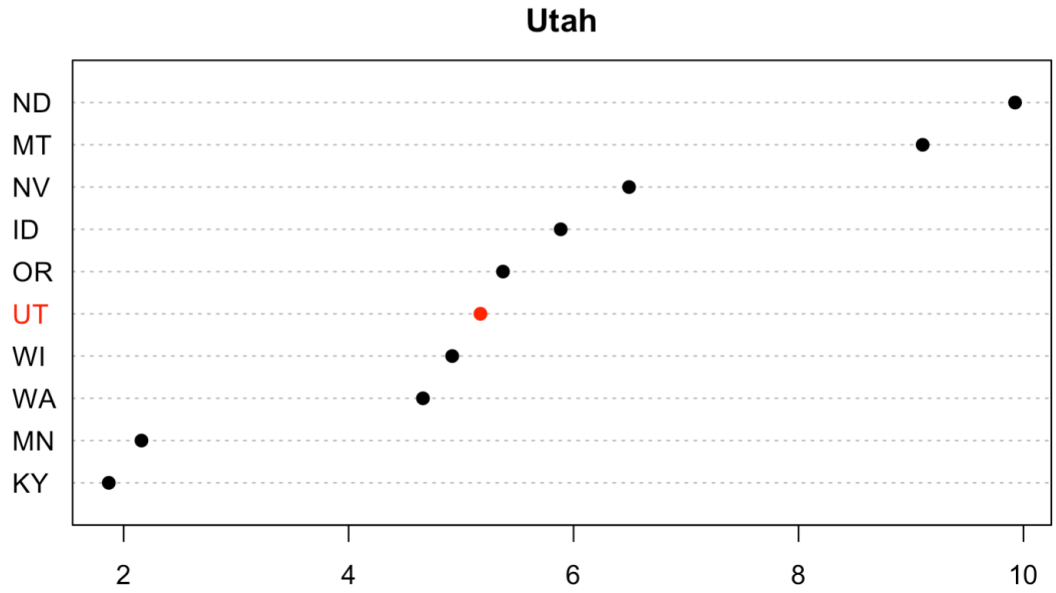
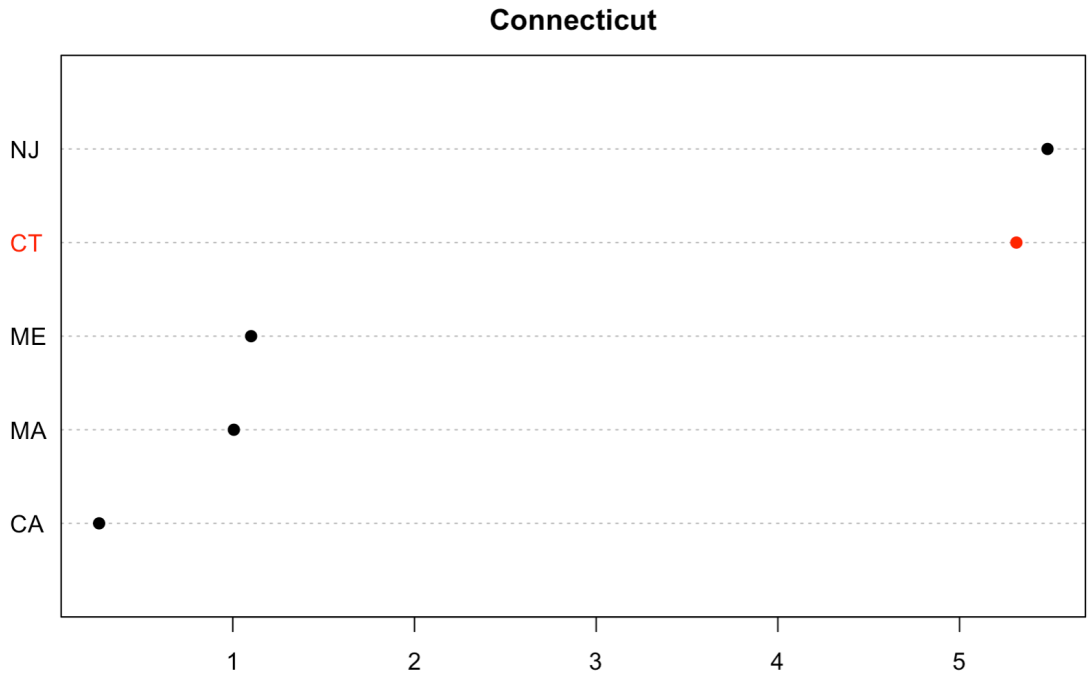


Figure B.4: Post-Period RMSPE / Pre-Period RMSPE (Gubernatorial Appointment to Merit)



C. Composition of Synthetic Controls

In this Appendix, I report the composition of the synthetic controls discussed in the body of the paper.

Table C.1: Donor States for Partisan Election Synthetic Controls

Donor State	*Arkansas*	*North Carolina*	*Mississippi*	*Tennessee*	*New Mexico*
AL	**0.28**	**0**	**0.25**	**0.57**	**0**
IL	**0**	**0**	**0**	**0.07**	**0**
LA	**0**	**0**	**0**	**0**	**0**
MI	**0.39**	**0.09**	**0**	**0**	**0.63**
OH	**0**	**0**	**0**	**0.07**	**0**
PA	**0.22**	**0.61**	**0**	**0.29**	**0**
TX	**0.11**	**0**	**0.75**	**0**	**0.2**
TX (Crim)	**0**	**0.3**	**0**	**0**	**0.16**

Table C.2: Synthetic Utah Donors

Donor State	*Utah*
ID	**0.08**
KY	**0.11**
MN	**0**
MT	**0.61**
NV	**0**
ND	**0**
OR	**0**
WA	**0.18**
WI	**0.02**

Table C.3: Synthetic Connecticut Donors

Donor State	*Connecticut*
CA	**0.46**
MA	**0**
ME	**0.54**
NJ	**0**

D. FEct Robustness Checks

This Appendix discusses the FEct robustness checks mentioned in the main text. I begin by providing the results of the placebo tests recommended by Liu et al. (2022). The idea behind the placebo tests is to pretend that the treatment of interest began several periods before actual onset. According to Liu et al., the magnitude of the estimated treatment effect for these placebo estimates should be close to zero (p. 9).

Figure D.1 contains plots of these placebo tests with the placebo periods highlighted in blue. For all four models, I fail to reject the null hypotheses of treatment effects equivalent to zero ($p = 0.72$, $p = 0.75$, $p = 0.80$, $p = 0.28$ for the non-partisan election analysis, merit commission analysis, gubernatorial appointment analysis, and partisan election analysis, respectively).[16] This result is encouraging as it suggests that the identification assumptions underlying the FEct etimates are at least plausible. The results are more tenuous when the test is inverted (i.e., when a non-zero ATT becomes the null hypothesis)—the relevant p-values for those equivalence tests are $p = 0.30$, $p = 0.24$, $p = 0.21$, $p = 0.62$. Ideally, these values would fall below 0.05, but given the paucity of data and the conservative nature of the test, they do not appear to invalidate the results reported in the body of the paper.

[16]For all robustness checks, I set the expected effect size at half a standard deviation of the residualized untreated outcome (see Liu et al., (2022), p.10 for details).

Figure D.1: Placebo Tests for FEct Estimates

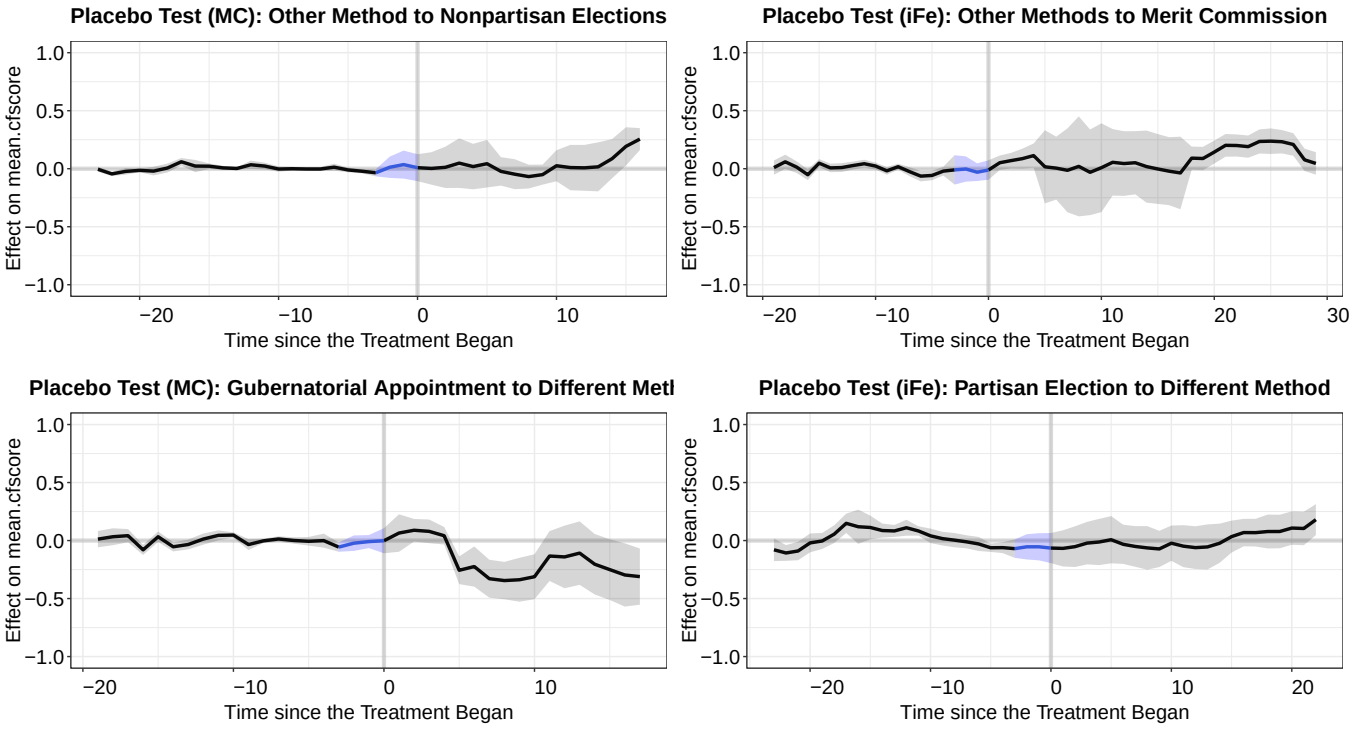
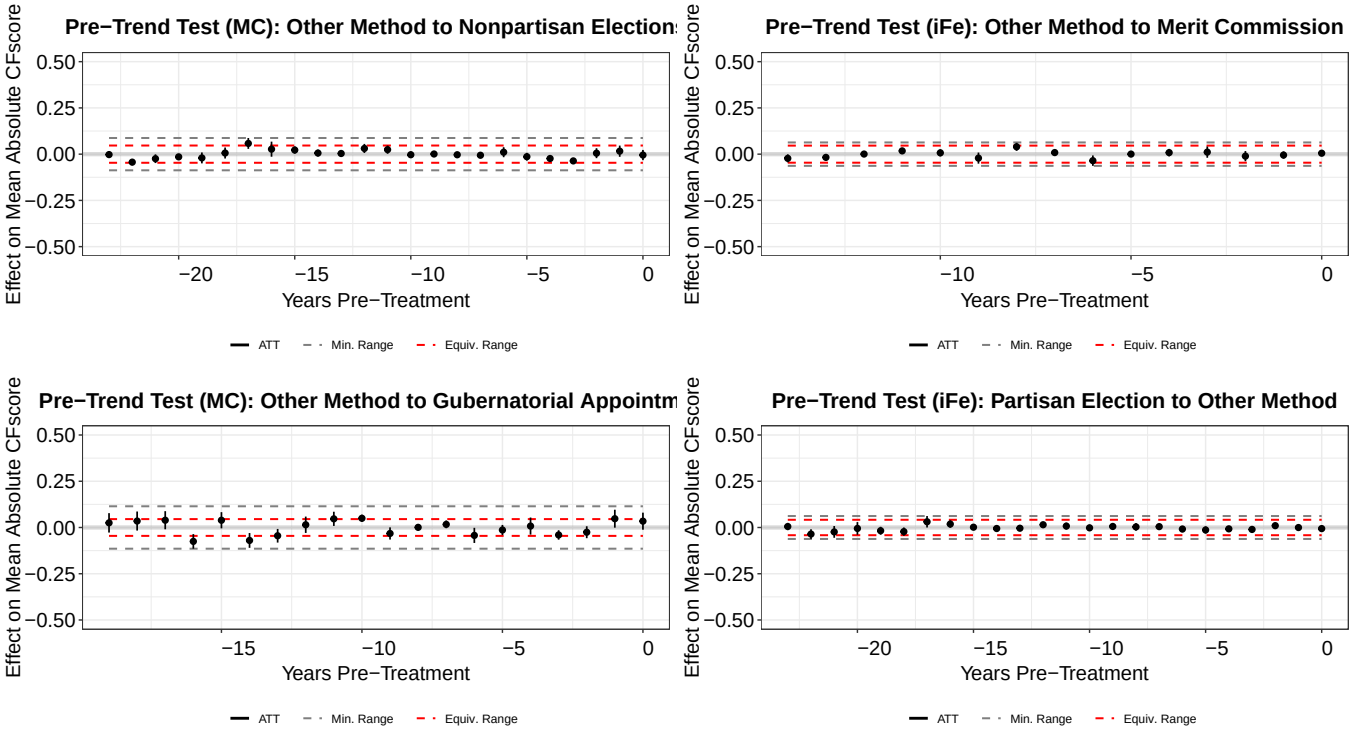


Figure D.2 presents the results of the second robustness check recommended by Liu et al., pretrend tests. This highly conservative test starts with a null hypothesis that the average of the residuals for any pre-treatment period is non-zero, and rejects that null only if "tests for all pretreatment periods generate significant results" (p.10). Ideally, every dot-and-whisker in Figure D.2 should fit within the red dotted lines (the equivalance range)—otherwise, I would fail to reject the null of a non-zero pre-trend for at least one period.

Figure D.2: Pre−trend Tests for FEct Estimates



Unsurprisingly given my limited data, I cannot reject the null of no pre-trend for any of the FEct analyses I perform in this paper. Given the strictness of this test, I do not treat this outcome as fatal to my results. However, it is fair to say that I cannot prove that the identification assumptions underlying the FEct analyses definitively hold.

E. Synthetic Difference-in-Differences

Dmitry Arkhangelsky and coauthors introduced the synthetic difference-in-differences estimator in the *American Economic Review* in 2021. Unlike the synthetic control, the synthetic difference-in-difference estimator can accomodate situations in which the value of the treated unit exceeds (or falls below) the value of each donor unit during the pre-treatment period. Such was the case for New Hampshire in my data. New Hampshire switched from unconstrained gubernatorial appointment to a merit commission in 2000; in the pre-treatment period, the average extremity of its state supreme court exceeded that of all other states with unconstrained gubernatorial appointment.

I apply the synthetic difference-in-differences estimator to New Hampshire in Figure E.1. For good measure, I also perform the same analysis for Connecticut, which made the same transition as New Hampshire post-1986 (Figure E.2). The estimated effects are in line with those reported in Tables 2-4 and Figure 4 (*-0.17* and *-0.38*), though the confidence intervals are massive (New Hampshire: (-0.67, 0.33); Connecticut: (-0.96, 0.20); both calculated using the placebo method from Arkhangelsky et al. (2021)).

Figure E.1: Synthetic Difference−in−Differences: Synthetic New Hampshire v. Actual New Hampshire

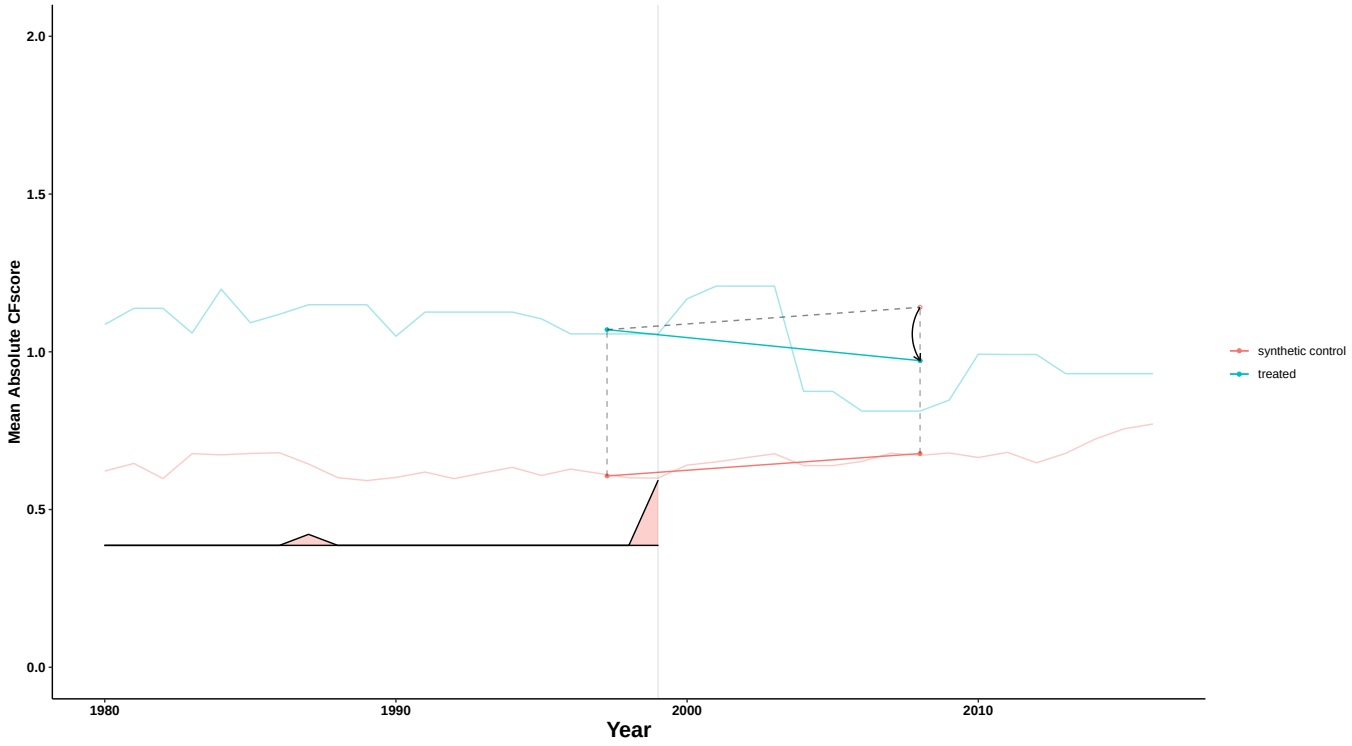


Figure E.2: Synthetic Difference−in−Differences: Synthetic Connecticut v. Actual Connecticut

F. Missing Data

As the main text observes, I do not have ideology measures for approximately 8% of state high court judges from the 1980-2016 period. Should missingness be correlated with selection method, these absent judges could threaten my primary conclusion: that unconstrained gubernatorial appointment produces more extreme judges than other methods of selection.

To guard against this possibility, I recalculate the results from Table 2, under the assumption that all missing data is maximally biased against my primary conclusion. That is, I assumed that every missing judge selected by merit commission, partisan election, or nonpartisan election had an ideology score of -2 or 2 (that is, those judges were as extreme as possible). Likewise, I assume that every missing judge from a state with unconstrained gubernatorial appointment had an ideology score of 0 (that is, those judges were as moderate as possible). This procedure should provide an extremely conservative estimate of the effect of unconstrained gubernatorial appointment.

Table F.1: Missing Data Analysis (GPSM Results for Mean Judge Extremity)

Pair of Methods	***ATM Estimate***	***SE***	***p-value***
Merit Selection - Gubnatorial Appointment	**-0.25*** **	**0.07**	**0**
Nonpartisan Elections - Gubernatorial Appointment	**-0.16****	**0.07**	**0.024**
Partisan Elections - Gubernatorial Appointment	**-0.13****	**0.06**	**0.035**

Variables contributing to propensity scores: Year, Party of Governor, State Senate Composition, State House Composition, Unified Gov.

ATM obtained using PSweight package in R.

p < 0.9 = *, p < 0.95 = **, p < 0.99 = ***

As Table F.1 indicates, my primary results barely change. As expected, the effect sizes decrease in magnitude, but they remain large, significant, and in the same direction. Accordingly, missing data does not appear to be a serious threat to the substantive conclusions drawn in the body of this paper.