Legal Constraints on Supreme Court Decision Making: Do Jurisprudential Regimes Exist?

Jeffrey R. Lax Columbia University
Kelly T. Rader Columbia University

The founding debate of judicial politics—is Supreme Court decision making driven by law or politics? —remains at center stage. One influential line of attack involves the identification of jurisprudential regimes, stable patterns of case decisions based on the influence of case factors. The key test is whether the regime changes after a major precedent-setting decision, that is, whether the case factors are subsequently treated differently by the Supreme Court justices themselves so that they vote as though constrained by precedent. We analyze whether binding jurisprudential regime change actually exists. The standard test assumes votes are independent observations, even though they are clustered by case and by term. We argue that a (nonparametric) ''randomization test'' is more appropriate. We find little evidence that precedents affect voting.

The founding debate of judicial politics—is Supreme Court decision making driven by law or politics? —remains at center stage today. In the first half of the twentieth century, the realist movement stressed the role of personal choice in judging. Now, a common view in political science, that of Spaeth and Segal (1999) and Segal and Spaeth (2002), is even more extreme: law has little or no influence over the case votes of Supreme Court Justices.

Richards and Kritzer (2002) pioneered a bold challenge to this view, arguing that law does affect these votes, that justices do make decisions as though bound by law. Their claim is that this influence can be detected by studying ''jurisprudential regimes,'' stable patterns of case decisions in a given area before and after key precedents are established. That Supreme Court decisions affect the decisions of *lower* court judges (vertical stare decisis) is not controversial. Their argument, however, goes much further. They argue that precedent effectively constrains the justices themselves (horizontal stare decisis), that key precedents induce new jurisprudential regimes in Supreme Court decision making. In short, they find that law can trump politics even in the Supreme Court.

Applying this clever research design, Kritzer and Richards present evidence that key precedents have caused jurisprudential regime change in four areas of the law: *Grayned v. Rockford* (1972) in freedom of expression doctrine (Richards and Kritzer 2002); *Lemon v. Kurtzman* (1971) in religious establishment doctrine (Kritzer and Richards 2003); *Illinois v. Gates* and *Segura v. U.S.* (1984) in search-and-seizure doctrine (Kritzer and Richards 2005); and *Chevron v. Natural Resources Defense Council* (1984) in administrative law (Richards, Smith, and Kritzer 2006)—hereafter, KR1, KR2, KR3, and KR4.

This growing and frequently cited body of work is dramatically shifting conventional wisdom about the Court. The central message—that the justices are bound by their own precedents or act as though they are—is aimed directly at the heart of the portrait of the Court as purely ideological, and this line of work is a large part of the empirical evidence against the hegemony of the ideology-dominant view of the Supreme Court. If correct, such findings would mark considerable progress in the elusive quest, spanning at least four decades, to demonstrate empirically that ''law matters'' in Supreme Court decision making. A conclusion this bold deserves close attention.

In this paper, we ask whether jurisprudential regime change actually exists. Part of our inquiry is conceptual: should we think of Supreme Court decision making as long periods of routine law application punctuated by structural breaks when a big precedent is announced? Part of our inquiry is methodological: does the standard test used to identify regime change actually work?

The specific statistical test used for regime change is whether the coefficients in a regression are

The Journal of Politics, Vol. 72, No. 2, April 2010, Pp. 273–284 doi:10.1017/S0022381609990855

ISSN 0022-3816

significantly different across two subsets of data according to a Chow test. This test assumes that each observation (here, a vote of a given justice in a given case in a given Court term) is independent from every other observation. That is, there is no correlation of the votes of the justices within a case or within a specific term. If this is not true, then one can have a falsely inflated sense of how much independent information is in the data, possibly understating standard errors, overstating confidence in determining the influence of case factors on votes, and overstating confidence that the influence of case factors has changed.

We explore whether these problems actually arise when testing jurisprudential regime change by applying a ''randomization test.'' Like the standard test, this test determines the degree of confidence with which we can state that the observed result is a meaningful finding by comparing the observed result to a distribution of results that we would observe by chance alone. Usually, we require that the effect be in the top 5% of this distribution, so that we can say at the 95% confidence level that there is a treatment effect and that we would not observe a test result this large if the null hypothesis of no effect were true.

The difference is in how we generate the reference distribution of results under the null hypothesis. The ''textbook'' distribution is a theoretical one, generated mathematically and requiring a set of rather explicit assumptions about the shape of the errors in the data and thus the correlations of votes within cases or terms. In contrast, a randomization test does not need to make binding assumptions about correlations between votes. Rather, the data themselves tell us what to expect if the null hypothesis is true, and so any findings will not be contingent on such assumptions.

To perform a randomization test, we first shuffle the data to break any systematic relationship between the treatment (being ''after'' the precedent) and the outcome (justice vote) to see what we would observe when there is no treatment effect. We drop the correct labels ''before'' and ''after'' and randomly reassign them to years. We then note what Chow test statistic we get from this random shuffle. We repeat this many times to get the distribution of test statistics we would observe when the null hypothesis (no treatment effect) is true by construction—that is, when there is no meaningful difference in the influence of case factors between the ''before'' years and the ''after'' years. We use this empirically generated reference distribution in place of the textbook distribution. As in any statistical test, if the test statistic lies in the top (say) 5% tail of the reference distribution, that is, when the test statistic from the real data is sufficiently unlikely to occur by chance, we say there is a statistically meaningful effect.

In KR1–3, the standard jurisprudential regimes test turns out to be highly overconfident, sometimes shockingly so, in concluding that regime change has occurred. The chance of concluding that there is an effect even though there is none (a false positive, or ''Type I error'') is generally far higher than the target 5% associated with 95% confidence. In short, evidence of jurisprudential regime change is much weaker than it first appears. Next, we briefly assess the current debates on the impact of law and then discuss regimes tests in more detail.

Law in the Court

One of the newest lines of attack in the law/politics debate is the jurisprudential regimes approach. This approach builds on fact-pattern analysis, the study of judicial decision making in a given area of law, and the specific case factors that drive such decision making. Fact-pattern analysis has a long history, dating to Kort (1957), and much work demonstrates how the presence (or absence) of case factors can successfully predict judicial decisions, a relationship that holds across both issue areas and courts (for an overview of this literature, see Kastellec and Lax 2008).

Scholars are usually interested in whether a particular case factor has a significant effect on the probability that a court or justice will rule in one direction or the other, where direction is typically measured dichotomously as liberal or conservative. Coefficients on case factors can be thought of as weights, measuring how much each factor ''pushes'' or ''pulls'' a particular case towards a conservative or liberal classification (Kastellec 2010).[1]

The jurisprudential regimes research design takes this one step further. A jurisprudential regime is said to ''structure Supreme Court decision making by establishing which case factors are relevant for decision making and/or by setting the level of scrutiny or balancing the justices are to employ in assessing case factors (i.e., weighing the influence of various case factors)'' (Richards and Kritzer 2002, 305). When an important precedent is announced, this is said to restructure the relevancy and weight of these case factors (factors such as the location of a search, whether a law regulating speech was content neutral,

[1]Kastellec and Lax (2008) show that case selection can drastically affect such estimates.

whether a law affecting the establishment of religion had a secular purpose, etc.). Kritzer and Richards assess regime change by comparing the influence of case factors before and after a key precedent is established. The weights the justices apply to certain case factors should change in response to the precedent, so that they act as though they are bound by it. If the influence of case factors changes significantly after the precedent—in particular, the specific case factors at the heart of the precedential holding—then this is evidence that the precedent (and thus "law") has affected the behavior of the justices.

KR1–4 do find such changes, using a Chow test to show that a statistically significant change in coefficients has occurred. Many scholars have employed the regimes approach in other applications (e.g., Bartels and O'Geen 2008; Benesh and Martinek 2005; Buchman 2005; Luse et al. 2007; Martinek 2008; Scott 2006;).

Segal and Spaeth (2003) raise three concerns about the jurisprudential regimes approach. First, they criticize the Chow test on the full set of variables, which attributes meaning to changes in variables that have little to do with the alleged regime change. However, Kritzer and Richards do test specific predictions for some key case factors. Second, Segal and Spaeth argue that these regimes could be "attitudinally" created, due to membership change on the court, rather than to the effects of a regime-setting precedent. Kritzer and Richards deal with this in part by focusing some tests solely on those justices who continue on the Court before and after the precedential decision.[2]

Third, Segal and Spaeth note that if the actual treatment of case variables is consistently changing in a given direction as the Court moves policy towards the right (or left) then we will find a "break" wherever we split the data. Indeed, Kritzer and Richards do conduct a sensitivity analysis of their results for a series of alternative annual break points in both KR1 and KR3 and find statistically significant results even when the "wrong" breakpoint is tested.[3] This is *not* itself, however, evidence against the standard test. In an ideal randomized experiment, if one accidentally mixed a few members of the treatment group into the control group (or vice versa) when assessing the effects of treatment, it would not be surprising if one still found treatment effects—the label "treatment" would still be highly correlated with actual treatment.

There is, however, one heretofore unexplored issue with the jurisprudential regimes test, as we discuss in the next section.

Testing Regime Change

The Chow test, used to determine significance of regime changes between the two groups of votes, assumes that the errors have the same variance (homoskedasticity) in the two groups and that the errors are independently distributed (Chow 1960). That is, in this context, it assumes that there is no relationship among votes within the same case (no clustering of errors by case) or among votes cast within the same Court term (no clustering by term, no autocorrelation) or among votes cast by the same justice (no clustering by justice). If this assumption is not met, then we are acting as though we have a much larger number of independent observations than we indeed have.[4]

This independence assumption seems hard to accept on substantive grounds. Whatever influences the vote of a given justice in a given case beyond the measured variables will likely influence the votes of other justices. They do not vote in a vacuum, but do interact with each other. Idiosyncratic features of the case at hand could easily push all justices in the same direction. At the docket level, the justices often coordinate cases in a similar issue area in the same term, moving the law in a particular direction to clarify or develop doctrine, so that we also might expect votes to be correlated within a given term.

Technically, all jurisprudential regimes findings in KR1–4 are predicated on the assumptions of uncorrelated votes by case or term or justice.[5] Even so, there is no reason to assume a priori that the effects of not strictly satisfying the necessary

[2]If the votes of continuing justices are themselves affected by new justices, due to interactions between justices or similar influences, then the Segal-Spaeth critique still has some bite.

[3]The test result associated with the "true" breakpoint does tend to be one of the highest among all such results. On the other hand, one reason this might be so is that the true breakpoint lies roughly in the *middle* of the time span. In the sensitivity analyses, particularly when the breakpoint is near the ends of the range, there is very little data in one sample (either "before" or "after") so that results are likely noisier in that sample, yielding a worse log likelihood (degree of fit)—and so a lower chi-square statistic. In any case, it is irrelevant if the observed test statistic is high relative to those from alternative time breakpoints, if, as we show, it is not significant relative to the correct reference distribution.

[4]It is also theoretically possible that negative intra-cluster correlation could cause us to understate the information in the data, but Arceneaux and Nickerson (2007) note this rarely arises in practice.

[5]KR4 (which we have not analyzed) shows robust standard errors for the logit coefficients, but this cannot affect the Chow test because, as discussed below, it based only on log likelihood and does not make use of the coefficients' standard errors.

assumptions will taint findings to any meaningful extent. Whether they do so, and how much, are empirical questions which can be addressed in one of two ways. First, one could move beyond the Chow test and make explicit assumptions about the distribution of errors across cases and terms (e.g., clustering, robust errors, GEE, etc.), but conclusions would then rest on these being the correct parameterization of the errors. It is by no means clear which theoretical remedy would be the correct one. The second approach, an empirically driven remedy rendering such parametric assumptions unnecessary, is a randomization test.[6]

Randomization Tests. Randomization tests were developed by Fisher (1935) to test treatment effects in an experimental setting. They are now used widely in biology (e.g., Manly (1997)) and to a growing extent in the economics and business literature (e.g., Kennedy (1995)), but remain rare in political science. For a separate overview, see Moore et al. (2003).

Like standard tests, randomization tests ask what distribution of test statistics we would observe if the null hypothesis of no effect (here, no change in case factor weights) were true. We can then see how likely it is that the observed test statistic would have occurred by chance, whether the size of the regime change is larger than that expected by random chance were there really no regime change. Similar standards for statistical significance apply. We might inquire whether the observed value lies in the top 5% of values that would occur by chance, so that we can say at the 95% confidence level that it did not occur by chance but rather represents a real change.

The difference is that, rather than rely on the textbook distribution of such values, theoretically derived, but assumption constrained, the randomization test derives the appropriate distribution of test statistics for the very data in question. It generates the reference distribution empirically by breaking the systematic relationship between treatment and outcome, randomly shuffling the data so that the null hypothesis is true by construction.

To perform the randomization test, we first observe the test statistic of interest from the actual data. This could be any test statistic such as a coefficient and its *t*-statistic, but here it is the chi-square value from a Chow test. We then randomly shuffle the data in a way that is consistent with the null hypothesis of no effect, so that we *know* there is no systematic relationship between the treatment and the outcome. We repeat this many times. Each time we note the test statistic of interest, so that we generate an entire reference distribution of test statistics consistent with the null hypothesis being true. We then calculate the *p*-value associated with the observed test statistic by locating it within this reference distribution, assessing the probability that a value this high would have been observed by chance alone.

This approach to figuring out the correct reference distribution for a test statistic is particularly useful when, as here, we do not know the exact structure of the correlation of votes within cases or terms, or where the theoretical standard errors could not easily be derived. Rather than base the test on extreme or arbitrary assumptions, we use the structure of the data themselves to tell us how confident we can be in test results. It does not require us to assume a distribution of test statistics, nor does it require the usual assumptions of normality and homoscedasticity, inter alia. This is a nonparametric way of deriving the correct standard errors and test-statistic distributions.[7]

Randomization tests still do require one assumption, exchangeability: if the null hypothesis is true, if treatment has no effect, then observed outcomes across observations should be similar (conditional on confounding covariates) no matter what the level of the treatment of interest. This is a weaker condition than the standard assumption of independent and identically distributed errors (that is, i.i.d. implies exchangeability but not vice versa). Note that if the randomization test distribution matches the textbook distribution, then, on these data, the two tests are equivalent, which would suggest that any violations of the usual assumptions did not taint the standard Chow test in this instance. If they do not match, then this usually indicates the standard assumptions do not hold (Moore et al. 2003, 57).

Before explaining how a randomization test for regime change can be set up, we must first discuss the Chow test in more detail.

Chow Tests. The Chow test for regimes (see Richards and Kritzer 2002, 319), compares two values: $-2 \times$ log likelihood of the logit regression including all cases and a dummy variable for "after" cases (call it the *nested model*) and $-2 \times$ log likelihood of a model which adds an interaction term between the "after" dummy and each of the other variables (call it

[6]To be clear, while randomization tests will deal with the standard test's problems with the assumptions discussed above, they do not resolve all other critiques such as history threats, selection bias that changes over time, or continuous changes that only appear to be structural breaks.

[7]Randomization tests use actual data, not ranks, so have higher power than other nonparametric tests (Edgington 1987; Manly 1997).

the *full model*). The chi-square value is the difference between these values.[8]

The larger the chi-square value, the greater the evidence that regime change occurred. But is this value statistically distinguishable from random variation? If all Chow assumptions hold, we could assess significance by comparing the chi-square we observe to the textbook chi-square distribution with the appropriate degrees of freedom and desired confidence level. The degrees of freedom will be the difference in degrees of freedom between the nested and full models, that is, the number of variables being tested. To run a test of change in a subset of variables, we compare a model with all interactions included to that of a model dropping the interaction terms to be tested.

A simple example will illustrate how this works. Suppose that votes are based on three case factors, denoted by x_1, x_2, and x_3. We have the votes of each justice in each case, and we create a dummy variable *After* which takes a value of 1 when the vote is after the key precedent and 0 otherwise. Suppose the regime change involves the last two of these variables in particular, so that we have a special interest in whether the influence of those two case factors changed. We begin with the full model:

$$\begin{aligned}\Pr(y_i = 1) = \ & \text{logit}^{-1}(\alpha + (\beta_1 \times x_1) + (\beta_2 \times x_2) \\ & + (\beta_3 \times x_3) + (\beta_4 \times \mathit{After}) \\ & + (\beta_5 \times \mathit{After} \times x_1) + (\beta_6 \times \mathit{After} \times x_2) \\ & + (\beta_7 \times \mathit{After} \times x_3))\end{aligned} \quad (1)$$

To test the full set of variables, we compare the $-2 \times$ log likelihood of the full model to the $-2 \times$ log likelihood of a nested model which drops the interaction terms connected to β_5, β_6, and β_7.

To test the second and third case factors together, we compare the full model to a nested model including β_5 but omitting β_6 and β_7.

Finally, to test the third factor alone, we compare the full model to a nested model omitting only the third interaction term, β_7.

[8]The Chow test *only* makes use of the log likelihood function, which does not incorporate the standard errors of the coefficients, but only the estimated coefficients and observed values. There is therefore no direct way of incorporating clustered or robust standard errors into it. The Wald test can do so in part. Unlike the Chow test, it does not assume error variance is the same across samples or that errors are normally distributed. The Wald test does, however, require explicit parametric assumptions about error distributions, is not invariant to how precisely the hypothesis is formulated (even when specifications are equivalent), and can have type I error rates higher than the desired critical values (Green 2003, 110, 133).

We then compare each observed chi-square value to the reference distribution for the desired confidence level and degrees of freedom (3df, 2df, and 1df, respectively).

The reference distribution used in regime change studies to date is the textbook chi-square distribution with its array of theoretical assumptions. Using this, we successfully replicated each of the tables in KR1–3, with the exception of the continuing-justices sample in KR2.[9]

An odd experiment. As we noted earlier, it could be that the standard test works fine for regime change, that the data at hand come close enough to the theoretical ideal so that any errors are small, even trivial. Before we move to our randomization tests, we conduct a small "experiment." Suppose we break the relationship between treatment (being after the breakpoint) and outcome (justice vote) by discarding the labels "before" and "after"—and instead divide up the data into even and odd terms of the Court. If we then performed the standard test for a difference in case factor influences, would we find a significant difference between the influences of case factors in even terms versus odd terms?

It turns out that we do. We repeat the key KR1 tests, but with the treatment "odd" instead of "after." The standard Chow test for the treatment "odd" on the full set of variables in KR1 yields a chi-square value of 74.1 (22 df), significant at $p < .001$. When we limit the comparison to the three variables in KR1 associated substantively with the new jurisprudential regime, we get a chi-square of 37.8 (3 df), again significant at $p < .001$. If we test each of the three variables individually, we get chi-square values of 21.1, 15.2, and 1.29 (1 df), and only the last is not significant at $p < .001$. That is, the standard jurisprudential regimes test would conclude that Supreme Court justices use a different legal regime in odd years than they do in even years. It would also conclude that the effect operated through the key jurisprudential variables governing freedom of expression cases.

Note this is *not* like the sensitivity analysis of Kritzer and Richards in which they tried other cutoffs for the label "after." Since those other divisions into control group and treatment group maintained a strong correlation with the true treatment "after," it should be not surprising nor worrying that positive test results still occurred. Here, however, the "after" years are split between the two groups "odd" and

[9]Kritzer verified that there were printing errors. We substituted the corrected findings.

"even," as are the "before" years. Here, there is *no* relationship between assignment to the treatment group "odd" and assignment to the treatment group "after." Thus, an actual effect of "after" cannot explain why we find odd/even jurisprudential regimes with a high degree of statistical confidence using the standard jurisprudential regimes test.

It is possible that there really is such an election year effect in freedom of expression cases, but this does not seem likely. Or, this could be a fluke. Or, it might be that we cannot use the textbook test straight off the shelf. The randomization test will tell us.

Randomization Test for Jurisprudential Regime Change. There are various ways to shuffle the data so as to break the relationship between the treatment of interest and the observed outcome (see Kennedy 1995), but Kennedy and Cade (1996) show that shuffling the treatment variable (here, the label "after" the regime change) is sufficient in the multivariate context so long as inferences are based on the distribution of test statistics and not the distribution of coefficients themselves (see also O'Gorman 2005).[10]

In randomized experiments with small sample sizes, it is possible to shuffle the data to reflect all possible permutations, in which case the randomization test is called an exact test (or permutation test). When the number of permutations is too large for an exact test, random sampling is used instead to generate an approximate distribution, with 1,000 shuffles shown to be sufficient for 95% confidence (Manly 1997).[11] Further, the shuffling method chosen should be consistent with the research design of the actual study. Here, that means that we randomly shuffle which terms (years) are given the treatment label "after," thus preserving any correlations within cases and within terms (some terms could still get labeled correctly, but only by chance). This is sometimes called "block randomization" (Donohue and Wolfers 2006, Manly 1997).

Using this method, we can investigate the following: What is the distribution of chi-square statistics that results when the null hypothesis is true, as it is here in the randomization test by construction? What value of the chi-square test is truly associated with conventional levels of confidence (e.g., 95%) for these data? How often would the traditional test make a Type I error, finding a significant change even though there is no change? How does each observed chi-square test result compare with the distribution of values from the randomization test? That is, what is the true *p*-value of the observed jurisprudential regime change? Finally, given this *p*-value, is the observed regime change larger than what we would expect by random chance alone? Is the observed regime change truly significant?

Results and Discussion

For a longer description of the data used, see KR1-3 themselves.[12]

KR1. Did *Grayned* (and its companion case) establish a new regime for Freedom of Expression cases, centering on content neutrality? We begin by running the standard Chow test for each of the relevant variables sets. There are 22 variables to predict voting in this area (including judicial ideology), of which three are considered key jurisprudential variables associated with the regime change. Does the regulation at hand meet the Threshold (does the regulation involve state action or abridge expression, or fall short of this threshold)? Is it Content Based or Content Neutral? The tests are for All Variables, Jurisprudential Variables, and each of the three individual jurisprudential variables.

Sections A and B of Table 1 show the results for KR1, with the tests for all justices shown at the top and for the continuing justices below that. The second column shows the significance determined in the original paper. Some of the feasible tests for KR1 (and KR2) were not included in those papers (indicated by "na" in the table), but are performed here.[13] The third column shows the observed chi-square test statistic (with the "after" treatment assigned as in KR), along with the degrees of freedom (these capture how many interaction terms are being

[10]This is the approach of recent social science applications (Donohue and Wolfers 2006, Erikson, Pinto, and Rader 2010, Helland and Tabarrok 2004).

[11]We take 1,500 shuffles, being sure only to use those with the correct number of degrees of freedom for comparing the observed test statistic. In some shuffles, a variable drops out because it does not vary in either the "before" portion or the "after" portion of the data. More than 1,000 shuffles always remain.

[12]We dropped the interaction terms for one variable in KR1 and for one in KR2, where the variable never takes the value 1 in the true "before" sample of each, so that no change could be registered in the true Chow test. These variables dropped out of Kritzer and Richards's regressions in KR1 and KR2.

[13]For some variables, Kritzer and Richards did not test whether there was a significant difference in the weights across time periods; rather, they only showed that the coefficient was significant in one time period and insignificant in the other. This alone need not mean that there is a significant difference between coefficients, or as Gelman and Stern put it, "the difference between 'significant' and 'not significant' is not itself statistically significant" (2006, 328).

TABLE 1 Jurisprudential Regime Results. See text for explanation. (* .05, ** .01, and * .001)**

Variable Set	Signif. in Original	Chi-Sq (Deg. of Freedom)	Signif. in Standard Test	Prob. Type I Error	Randomization Test *p*-value	Signif. in Rand. Test
A. Freedom of Expression: *Grayned v. Rockford* (KR1) All Justices (N = 4,986)						
All	***	124.7 (22)	***	.99	.24	
Jurisprudential	***	46.7 (3)	***	.70	.03	*
Threshold	na	18.9 (1)	***	.50	.10	
Content Based	na	22.4 (1)	***	.39	.03	*
Content Neutral	na	25.3 (1)	***	.36	.02	*
B. Freedom of Expression: *Grayned v. Rockford* (KRl)—Continuing Justices (N = 3,056)						
All	***	113.2 (22)	***	.99	.18	
Jurisprudential	***	21.8 (3)	***	.56	.11	
Threshold	na	13.8 (1)	***	.39	.09	
Content Based	na	8.0 (1)	**	.34	.15	
Content Neutral	na	7.9 (1)	**	.24	.09	
C. Establishment of Religion: *Lemon v, Kurtzman* (KR2)—All Justices (N = 743)						
All	***	35.3 (6)	***	.64	.04	*
Jurisprudential	na	14.3 (3)	**	.42	.19	
Purpose	na	.1 (1)		.34	.86	
Neutrality	na	.2 (1)		.31	.81	
Monitoring	na	14.0 (1)	***	.18	.001	***
D. Establishment of Religion: *Lemon v. Kurtzman* (KR2)—Continuing Justices (N = 398)						
All	***	12.0 (6)		.40	.43	
Jurisprudential	na	5.0 (3)		.34	.52	
Purpose	na	.2 (1)		.30	.82	
Neutrality	na	.5 (1)		.25	.68	
Monitoring	na	4.8 (1)	*	.14	.10	
E. Search and Seizure: *Illinois v. Gates, Segura v. U.S.* (KR3)—All Justices (N = l,763)						
All	***	73.0 (12)	***	.97	.12	
House	*	4.4 (1)	*	.52	.49	
Business	**	9.9 (1)	**	.54	.30	
Person	***	10.1 (1)	**	.58	.33	
Car	***	19.7 (1)	***	.56	.13	
Full Search		1.2 (1)		.48	.69	
Warrant		.4 (1)		.34	.75	
Probable Cause	***	18.8 (1)	***	.29	.01	*
Incident Arrest	***	12.7 (1)	***	.41	.10	
After Arrest	**	8.3 (1)	**	.34	.17	
After Unlawful		3.2 (1)		.20	.23	
Exceptions		.5 (1)		.20	.64	
Attitude	**	9.5 (1)	**	.43	.18	
F. Search and Seizure: *Illinois v. Gates, Segura v. U.S.* (KR3)—Continuing Justices (N = l,170)						
All	***	46.7 (12)	***	.91	.26	
House		3.3 (1)		.43	.46	
Business	**	7.0 (1)	**	.44	.30	
Person	*	5.3 (1)	*	.43	.36	
Car	**	8.0 (1)	**	.44	.26	
Full Search		1.1 (1)		.38	.65	
Warrant		0.0 (1)		.15	.99	
Probable Cause	*	5.4 (1)	*	.17	.12	
Incident Arrest	***	17.6 (1)	***	.43	.04	*

TABLE 1 (Continued)

Variable Set	Signif. in Original	Chi-Sq (Deg. of Freedom)	Signif. in Standard Test	Prob. Type I Error	Randomization Test *p*-value	Signif. in Rand. Test
After Arrest	**	9.8 (1)	**	.39	.14	
After Unlawful		2.6 (1)		.15	.24	
Exceptions	*	4.0 (1)	*	.21	.20	
Attitude		1.8 (1)		.27	.46	

tested for joint significance). The fourth column shows the full set of significance findings using the standard Chow test for each relevant variable set. Note that all the KR1 tests are significant at .01 or better according to the standard test.

Next, we turn to the randomization test results. The fifth column shows how often we get a finding of ''significant at .05'' even though the null hypothesis of no systematic treatment effect is true. If the textbook test operated on this data ''as advertised,'' we would see a column of .05 values (remember that 95% confidence means that we still make Type I errors 5% of the time). We do not. That the error rates are too high indicates that the conditions for the standard test do not hold here.

In the ''all variables'' test, we get a false positive result 99.8% of the time. This explains why the odd-even year division showed a significant result—almost *any* random division of years would. We make Type I errors for the set of jurisprudential variables 70% of the time, and at high rates for the individual variables (36% or more). The error rates for the continuing justices are better, but also far from the target 5% rate. At best, we still make Type I errors 24% of the time, for the Content Neutral test.

Of course, just because we are too likely to make a Type I error does not mean we have actually done so. The observed chi-square result could still be significant even when compared to the distribution of chi-square values generated by the randomization test. The sixth column in the Table shows the randomization test *p*-values associated with the observed chi-square statistics in the third column, and the seventh column shows the significance levels. These results indicate that the standard test is indeed overconfident in this application. The observed chi-square value for the All Variables test is 124.7, which is significant at only 76% confidence level, far short of conventionally accepted levels and far short of the 99.9% confidence level if the standard distribution applied. The shift in the joint influence of the three Jurisprudential variables after *Grayned* for the set of all justices is still significant, at a *p*-value of .03.

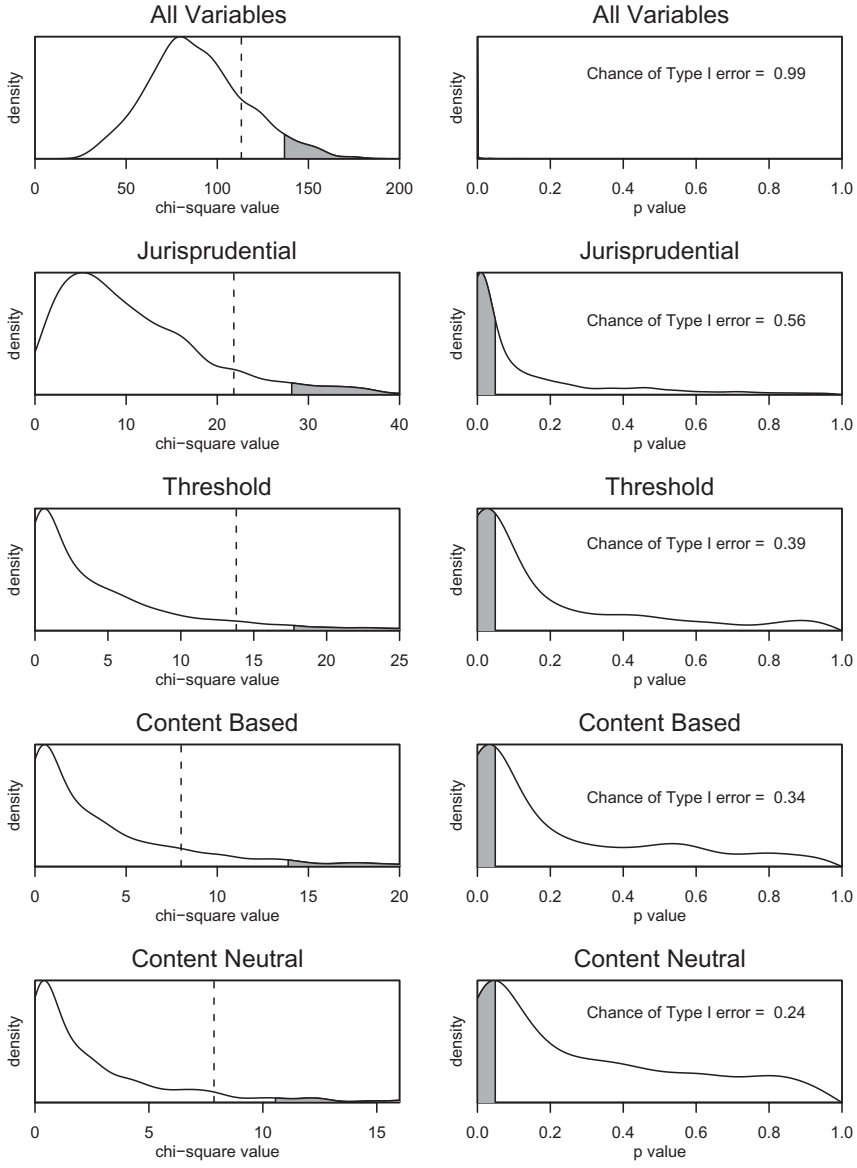
However, if, as Kritzer and Richards suggest, we limit our comparison to those justices serving both before and after *Grayned* so that we can rule out the confounding influence of membership change, we no longer can say at the 95% confidence level that there was such a change, as shown in the last two columns. We do come close to being able to say so at the 90% level, and two of the three individual variables do reach the 90% level.

The randomization test results for continuing *Grayned* justices are also shown graphically in Figure 1. The right-hand panels show the distribution of *p*-values we observe when we apply the Chow test to each random shuffle. We shade in the region for values lower than .05. Since we graph the density of values, the total area under the curve is 1, and the area of the shaded region reveals how often we would make a Type I error. (These values correspond to those in the fifth column of Table 1.) Again, at the 95% confidence level, only 5% of *p*-values should be less than or equal to .05.

The left-hand panels in Figure 1 show the distribution of chi-square values under the null hypothesis for each of the five tests for continuing justices. Note that they are indeed shaped like chi-squared distributions for the given degrees of freedom. The shaded region shows the 5% tail of each distribution, with the observed chi-square statistic (from the true data) indicated by a dashed line. That the values fall outside the tails means that we do not find a change at the 95% confidence level.

A final point to note: the textbook chi-square value for the All Variables test is 33.9 (22 degrees of freedom), which should be the boundary of the 5% tail if the standard distribution applied—but it is nowhere near the 5% tail of the corrected distribution. For the full set of justices, the actual 5% cut-off is 136.8. The textbook chi-square values for the Jurisprudential and individual variable tests would be 7.8 (3 degrees of freedom) and 3.8 (1 degree of freedom), respectively. Again, these are far from the true 5% cut-off of 28.1 for the set of Jurisprudential variables and the cut-offs for the three individual variables, which are 17.7, 13.9, and

FIGURE 1 ***Distributions of Chi-Square and p Values, Grayned v. Rockford, Continuing Justices.*** **Left-hand panels show, for each variable or variable set, the distribution of chi-square test statistics generated using randomization tests, with the 5% tail shaded and the observed test statistic shown by the dashed line (significant at 95% if within the shaded region). Right-hand panels show the distribution of *p*-values we observe when we apply the Chow test to each random shuffle. The area of the shaded region shows how often these values are less than or equal to 5%, revealing the Type I error probability.**



10.5, respectively. This shows why the confidence levels generated under by the standard tests were so overstated.

KR2. Besides judicial ideology, there are six variables in KR2, of which three are considered key jurisprudential variables associated with the regime change, forming the well-known Lemon Test for Establishment Clause cases: Purpose (the purpose of the statute or practice, secular or other), Neutrality (towards religion), and Monitoring (whether an excessive entanglement between government and religion is involved). Results are shown in sections C and D of Table 1. Using the standard test and the full set of justices, we would think the All Variables, Jurisprudential Variables, and Monitoring tests were significant at the 99% confidence level or better. These results are partially upheld when using the randomization test distribution of chi-square values: the true confidence for All Variables is

lower and the Jurisprudential Variables are not collectively significant. However, when we limit the comparison to the seven justices casting a nontrivial number of votes before and after *Lemon*, none of the tests are significant at 95% level, though the change in Monitoring is significant at 90% level. We cannot be confident that the results are not driven by membership change alone.

KR3. There are 12 variables included for search-and-seizure cases: the location of the search (House, Car, etc.), the degree of the search (Full or partial), whether a Warrant was issued, the lower court finding of Probable Cause, details of the arrest, whether Exceptions to the warrant requirement applied, and judicial ideology. Results are shown in sections E and F of Table 1. The textbook test would show vast regime change, across almost every variable. Only four individual variables fail to reach significance in the full set of justices; five do for the continuing-justice tests. However, Type I errors abound, so that in the end, of 26 possible tests, only two are significant: we can say at the 95% confidence level that the influence of Probable Cause on all justices changed and that the influence of Incident to Arrest on continuing justices changed (the latter is not a variable that is substantively related to the regime change in question). No other test reaches even the 90% confidence level. The changes we would find in case factor influences "before" and "after" key precedents are generally in line with what we would expect to see by chance alone.

What about odd-even regimes? The randomization test would (correctly, in our view) not conclude that the justices follow biannual cycles. None of the five chi-square values generated—for All Variables, Jurisprudential Variables, or the three individual jurisprudential variables—reaches significance at 95% using the randomization test.

A comment on randomization tests. The standard tests for regime change turned out to be far less accurate than anyone could have anticipated. That we found problems with the standard tests here suggests the diagnostic usefulness of randomization tests. In this particular context, they helped to uncover problems with what by all appearances was a very reasonable approach for testing regime change. They helped to uncover problems that remained hidden even given Kritzer and Richards' close attention to detail and their all-too-unusual willingness to subject their findings to scrutiny for robustness and sensitivity.

We would suggest that randomization tests might be used more widely in judicial politics and in political science in general. Once too "costly" to run in terms of computer time, they are now quite feasible. They represent a viable alternative to parametric corrections for error correlations and the like, given that such corrections require explicit assumptions about error distributions that randomization tests render unnecessary. They also give test distributions tailored for the particular data being studied. Of course, randomization tests are more difficult to implement than standard tests or tests with parametric error corrections, and so may not be worth the effort in cases in which a researcher has strong theoretically driven beliefs about error distributions.

Conclusion

Do jurisprudential regimes exist? Our answer is no... and yes.

We find only weak evidence that major Supreme Court precedents affect the way the justices themselves vote in subsequent cases. The observed regime changes are generally no larger than what we should expect from random chance. We cannot rule out at standard levels of confidence that observed regime changes are more than noise.[14]

In total, there were 46 tests of regime change. Of these, using the standard test, 31 would appear positive for regime change. The randomization test showed that only 7 tests are actually positive at the 95% confidence level. Setting aside results that might be induced by membership change alone, evidence of regime change was even more limited. There were 23 tests that included only justices who continue to serve after the precedent. Of these, 14 were positive under the standard test. Under the randomization test, only 1 test was positive at the 95% confidence level, and it was not a test of a case factor substantively tied to the regime change in question.[15]

The standard test for jurisprudential regime change, requiring textbook assumptions not met in Supreme Court vote data, has led us astray in the law versus politics debate. And so the search for the elusive effects of law on Supreme Court voting must continue.

Yet this does not mean that the jurisprudential regimes concept is wrong. Kritzer and Richards are clearly correct that the justices want to see cases

[14]To be sure, we cannot say with reasonable certainty that there are no such regime changes—we fail to reject the null hypothesis, which is not the same as accepting it.

[15]Given that we are doing multiple tests, were they independent, we would expect that about two of the 46 tests to be significant at the 95% level by chance alone (one of the 23 continuing-justice tests).

decided in a structured way. Arguably, even the most ideological justice would decide cases in a structured fashion, sorting out how cases will be decided depending on what the facts of the cases are—even if how those facts are weighted is based on her preferences and not traditional notions of law or the precedents handed down by the Court itself (Lax 2007). In this sense, jurisprudential regimes certainly exist. We simply cannot say with much confidence that precedents "bind" future votes, that precedents cause regime change other than that caused by the shifting membership of the Court. Given that the the regime changes tested in KR1–3 are tied to rather important precedents, if we find no such effects here, we might be tempted to conclude that regime change does not really occur at all.

But there is an additional concern. Statistical fact-pattern analyses, widely used in judicial politics, can only be pushed so far. Saying that there is a generally stable doctrinal structure at work in a given body of law (whether constructed in whole or in part by ideological preference alone) does not mean that we can figure out that doctrinal structure in terms of relatively simple sets of case factors. It does not mean that a short list of case factors can fully flesh out the nuances of a legal doctrine.

What can such analyses tell us about what doctrines govern an area of the law? If the goal is to predict votes in the cases the justices actually take based on such factors then this approach often works well (e.g., Segal 1984). But this does not mean that we can ascertain the doctrine the justices want to see applied more generally in cases they do not take by inspecting statistical patterns in the cases they do take. Prediction is not always explanation. Saying that votes are relatively predictable is not the same as saying we can figure out the full underlying doctrine by using statistical fact pattern analysis.

The problem, in our view, lies in the conception of Supreme Court decision making at the heart of *some* fact-pattern analyses. The implicit picture is that of simple case sorting, or routine law application. A case comes before the Court, is assessed according to some rule or balancing test, and is sorted into a "yes" or "no" bin accordingly, like separating spoons and forks from a pile of silverware. But most Supreme Court cases are not spoons or forks. Being a Supreme Court justice is like holding a "spork" and trying to decide whether it is more like a spoon or a fork.[16] That is, most Supreme Court cases are unusual or novel in some way. They are about law creation, development, or modification (whether ideologically driven or not).[17]

If the Court took all cases, then, yes, the observed patterns of decision making could tell us everything we needed to know about how case factors influence decisions. But if they are taking cases that do not fit well, or if each case is taken to shift the jurisprudential regime in some way, we should not expect to uncover their underlying doctrinal preferences by focusing only on their votes in the cases they choose to hear. If each case shifts the law, then it might be problematic to lump together decades of cases on one side or the other of a dividing line—each case may have been taken because it did *not* fit the established doctrine of its day. The hypothesis that key precedents change the behavior of sitting justices could even be true, but we should not expect to find such effects by testing a narrow set of case factors. We might need to focus more on doctrine as described in opinions instead of focusing only on dichotomous votes and case factors. Indeed, we might better observe the Supreme Court's desired legal doctrine by examining *lower* court cases, to the extent that such cases are representative and straightforward applications of law.

In short, it may not make sense to use the standard jurisprudential regimes test (or even the upgraded test we use here) to find regime change if most Supreme Court cases are meant to be, and likely are, regime changing. The problem is not that jurisprudential regime change does not occur, but that there is too much of it for us to find it in this way.

Acknowledgments

We thank Mark Richards, Bert Kritzer, Jeff Segal, and Joe Ignagni for graciously sharing their data with us, and in particular Mark and Bert for their generous assistance in replicating their work. We also thank Charles Cameron, Bob Erikson, and John Kastellec for helpful comments.

Manuscript submitted 30 November 2008
Manuscript accepted for publication 12 March 2009

[16]Sporks are the pronged spoons common to lunchtime disposable packets. We thank the movie *Wall-E* and its title character for this metaphor.

[17]They might get *some* forks and spoons. Sometimes the justices take cases to correct simple errors made by the lower courts. Also, they take some cases to rectify noncompliance by lower courts.

References

Arceneaux, Kevin, and David W. Nickerson. 2009. "Modeling Certainty with Clustered Data: A Comparison of Methods." *Political Analysis* 17 (2): 177–190.

Bartels, Brandon L., and Andrew O'Geen. 2008. "Is Legal Change Revolutionary or Evolutionary? The Foundations and Consequences of Jurisprudential Regimes in Supreme Court Decision Making." Presented at the 2008 annual meeting of the American Political Science Association.

Benesh, Sara C., and Wendy Martinek. 2005. "Jurisprudential Regimes and the State Supreme Courts: A New Way to Measure Supreme Court Compliance." Presented at the annual meeting of the Southern Political Science Association.

Buchman, Jeremy. 2005. "Jurisprudential Regimes and Obscenity Doctrine." Presented at the annual meeting of the American Political Science Association.

Chow, Gregory C. 1960. "Tests of Equality Between Two Sets of Coefficients in Two Linear Regressions." *Econometrica* 28 (3): 591–605.

Donohue, John J., and Justin Wolfers. 2006. "Uses and Abuse of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review* 58: 791–835.

Edgington, E. S. 1987. *Randomization Tests*. New York: Marcel Dekker.

Erikson, Robert S., Pablo M. Pinto, and Kelly T. Rader. 2010. "Randomization Tests and Multilevel Data in State Politics." *State Politics and Policy Quarterly* 10 (2). Forthcoming.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh, Scotland: Oliver and Boyd.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between Significant and Not Significant is not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31.

Green, William H. 2003. *Econometric Analysis*. 5th ed. New York: Prentice Hall.

Helland, Eric, and Alexander Tabarrok. 2004. "Using Placebo Laws to Test More Guns, Less Crime." *Advances in Economic Analysis and Policy* 4: 1–7.

Kastellec, Jonathan P. 2010. "The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees." *Journal of Empirical Legal Studies*. forthcoming.

Kastellec, Jonathan P., and Jeffrey R. Lax. 2008. "Case Selection and the Study of Judicial Politics." *Journal of Empirical Legal Studies* 53 (3): 407–46.

Kennedy, Peter E. 1995. "Randomization Tests in Econometrics." *Journal of Business and Economic Statistics* 13: 85–94.

Kennedy, Peter E., and Brian S. Cade. 1996. "Randomization Tests for Multiple Regression." *Communications in Statistics: Simulation and Computation* 34: 923–36.

Kort, Fred. 1957. "Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the 'Right to Counsel' Cases." *American Political Science Review* 51 (1): 1–12.

Kritzer, Herbert M., and Mark J. Richards. 2003. "Jurisprudential Regimes and Supreme Court Decisionmaking: The Lemon Regime and Establishment Clause Cases." *Law and Society Review* 37 (4): 827–40.

Kritzer, Herbert M., and Mark J. Richards. 2005. "The Influence of Law in the Supreme Court's Search-and-Seizure Jurisprudence." *American Politics Research* 33 (1): 33–55.

Lax, Jeffrey R. 2007. "Constructing Legal Rules on Appellate Courts." *American Political Science Review*. 101 (3): 591–604.

Luse, Jennifer, Geoffrey McGovern, Wendy L. Martinek, and Sara C. Benesh. 2009. "'Such Inferior Courts.': Compliance by Circuits with Jurisprudential Regimes." *American Politics Research* 37 (1): 75–106.

Manly, Brian F. J. 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd ed. London: Chapman Hall.

Martinek, Wendy. 2008. "Compliance and Jurisprudential Regimes: The Case of Search and Seizure Decision Making." Presented at the annual meeting of the American Political Science Association.

Moore, David S., George P. McCabe, William M. Duckworth, and Stanley L. Sclove. 2003. *The Practice of Business Statistics Companion Chapter 18: Bootstrap Methods and Permutation Tests*. New York: W. H. Freeman.

Richards, Mark J., and Herbert M. Kritzer. 2002. "Jurisprudential Regimes in Supreme Court Decision Making." *American Political Science Review* 96 (2): 305–20.

Richards, Mark J., Joseph L. Smith, and Herbert M. Kritzer. 2006. "Does Chevon Matter?" *Law & Policy* 28 (4): 444–69.

Scott, Kevin M. 2006. "Reconsidering the Impact of Jurisprudential Regimes." *Social Science Quarterly* 87 (2): 380–94.

Segal, Jeffrey A. 1984. "Predicting Supreme Court Cases Probabilistically: The Search and Seizure Cases, 1962-1981." *American Political Science Review* 78 (4): 891–900.

Segal, Jeffrey A., and Harold J. Spaeth. 2002. *The Supreme Court and the Attitudinal Model Revisited*. New York: Cambridge University Press.

Segal, Jeffrey A., and Harold J. Spaeth. 2003. "Reply to the Critics of the Supreme Court Attitudinal Model Revisited." *Law and Courts* 13 (3): 31–8.

Spaeth, Harold J., and Jeffrey A. Segal. 1999. *Majority Rule or Minority Will: Adherence to Precedent on the U.S. Supreme Court*. Cambridge: Cambridge University Press.

Jeffrey R. Lax is Assistant Professor, Department of Political Science, Columbia University, NY, NY, 10027.

Kelly T. Rader is Doctoral Candidate, Department of Political Science, Columbia University, NY, NY, 10027.