The Three Prongs of a Jurisprudential Regimes Test: A Response to Kritzer and Richards

Jeffrey R. Lax Columbia University, New York
Kelly T. Rader Columbia University, New York

In "Legal Constraints on Supreme Court Decision Making: Do Jurisprudential Regimes Exist?", we explored whether one can say that jurisprudential regime change occurred in Supreme Court decision making—whether key legal precedents led to changes in how justices voted. We found that the standard test, a Chow test of coefficient change, used in Kritzer and Richards's research design, is strikingly overconfident in finding that a change has occurred in voting across cases before and after the precedent. Rather than making a Type-1 error of finding regime change when none exists 5% of the time, the test does so sometimes close to 100% of the time, even though the data has been randomly shuffled so that no systematic difference can exist between the before and after cases.

We appreciate Kritzer and Richards's open-mindedness about our inquiry into their findings and are grateful for their thoughtful responses (now and while we worked on our original paper). We also appreciate the opportunity to clarify our findings, respond to their arguments (old and new), and to present supplemental results that answer, we hope, the questions they raised.

Kritzer and Richards state that our paper focuses on only one of the three prongs of their argument, the significance test of regime change, whereas they now clarify that the second and third elements are actually of great (greater?) importance. They have also supplemented their findings with a more sophisticated test of regime change. We discuss each of these three components, including their new statistical evidence, below.

Statistical Tests of Regime Change

In their response to our findings, Kritzer and Richards present stronger evidence of regime change than in their original papers. The Wald tests they now use, for which they present tests of joint significance for the full sets of case factor variables, can take clustering into account in part. As they point out, however, the problem remains that even Wald tests cannot take into account clustering by justice, by case, and by term simultaneously. As with the Chow tests used originally, they require strong parametric assumptions of error distributions (specifically, that observations across the clustered groups are independent). One must assume what sort of clustering can exist, assume that other forms of clustering do not exist, and accept how the clustering algorithm operationalizes these assumptions. Since the randomization tests we performed did not require us to make such arbitrary assumptions about the existence of or exact nature of clustering, we stand by our original conclusion that the evidence for jurisprudential regime change does not reach statistical significance. (And, as we noted in our original footnote 8, the Wald test tends to be overconfident even when the error structure is correctly modeled.)

Including clustering at one level is, we agree, likely better than setting it aside completely, as in the standard Chow test. On the other hand, their new results show that of the 12 individual variables tested for change over the three papers we analyzed, eight are significant if no clustering is taken into account, seven are if votes are clustered by justice alone, and six are if votes are clustered by case alone. Limiting the analysis to those justices that served before and after the break to avoid contamination due to personnel change, significant effects are found for seven with no clustering, six clustering by justice alone, and three clustering by case alone. That results differ given which type of clustering is modeled confirms our original intuition that one should worry about being limited to including only one form of clustering (let alone setting aside clustering completely). Note that we find that none of these differences are significant in the randomization tests for justices who cast sufficient votes before and after the predicted regime break.

The Journal of Politics, Vol. 72, No. 2, April 2010, Pp. 289–291 doi:10.1017/S0022381609990739
© Southern Political Science Association, 2010
ISSN 0022-3816

Kritzer and Richards also express concern that among the 1,500 random shuffles of data, too high a proportion of our random splits may have been substantially aligned with the actual breakpoint they examined—thus leading to positive Chow test results because the cases we randomly label "after" would be highly correlated with the true treatment "after." This is not a problem, however, when using a randomization test. Suppose that a given random shuffle put many true "after" years in one set of randomly selected years. The proper split, with *all* the treated years kept together (that is, with the biggest and cleanest difference between the two samples of votes) should show an even higher Chow test result than this one, in which there is mixing of non-"treated" and "treated" data. The proper split, if "after" has an effect, would still be in the top percentile of such effects among random shuffles. In fact, if the effect were large enough and noise small enough, it would beat the test statistics of all the random shuffles (each of which would have some nontreated years mixed in with the treated years). Moreover, we have replicated our results using a larger set of shuffles.

Still, to put this concern to rest, we repeated our analysis constraining the shuffles so that the percentage of years labeled "before" that are actually true "after" years is the same as the percentage of years labeled "after" that are true "after" years. Now, even if true "after" years are different than true "before" years, the subsets of data *labeled* "before" and "after" are as similar as possible, and as different from the true split as possible.[1] Thus, the standard Chow tests on these shuffled data should not, given that the null hypothesis is true by construction, show a positive Chow test result more than 5% of the time for the 95% significance level. We compared the Chow test statistic of the true split of the data to the distribution of Chow test statistics produced by these constrained shuffles. The *p*-values generated by the constrained randomization tests are almost exactly the same as in the final column of our paper. Not one significance test result is different. Even for these constrained shuffles, type-1 errors occur strikingly high percentages of the time, at almost exactly the same rate as for the unconstrained shuffles. All of these results are available upon request.

Sensitivity Tests

The second element of Kritzer and Richards's research design is to conduct a sensitivity analysis of their finding that regime change has occurred at a specific time break. They do so by showing that the Chow statistic at that break (associated with a precedent they specify) is large relative to many other (though not all) sequential splits of the data at alternative breaks. We agree that this is more convincing than *only* presenting the Chow test result at the specified break, but note that the fact that the specified break's Chow statistic is higher than that of many other sequential breaks does not mean that it is itself statistically significant. The randomization test results show that the true breaks are not statistically significant. Each "true" break produces a Chow statistic that is actually quite low compared to even those produced by meaningless splits of the data, where there is no systematic difference between the two sets of cases. We also note, as we explained in footnote 3, that one reason other break points might not yield high Chow test statistics is that, by virtue of smaller sample towards one or the other side of the break, they might have noisier subsamples.

Kritzer and Richards also argue that it is not surprising that when one splits the years into odd and even samples one gets a significant Chow test result. While it might not be surprising to see splits of the data correlated to the "true" split at the precedent show a similarly significant effect, it is surprising to us that meaningless splits, without any correlation to the hypothesized breakpoint, would do so. To be sure, the odd-even split is only one such shuffle, but it is one without any meaningful difference across vote samples. This pattern of false positives turns out to be the norm for Chow tests on Supreme Court vote data. Yes, as Kritzer and Richards state, "various splits could be statistically significant, even without a basis in theory," but only 5% of meaningless splits should be, at the 95% confidence level. At the very least, our opinion is that if one wishes to argue that there is a larger test statistic at the hypothesized breakpoint than at many other points, then one should omit unfounded claims of statistical significance.

Substantive Patterns

The third element is substantive pattern matching. We certainly agree that the directional predictions for the key variables matter, that the case factors associated with the key precedent are the ones that change, and that they change in the predicted direction. On the other hand, Kritzer and Richards, both in their original papers, and in their supplemental analysis now, present tests of joint significance

[1] Note: this is not the correct way to conduct a randomization test, for which the shuffles should be truly randomly drawn from the set of all possible shuffles, as in our original analysis.

of the full set of variables and subsets thereof, for which no directional test is possible, and not just tests of individual specific variables. Certainly they meant these joint tests to add to their argument, yet such tests turn out to be highly overconfident. To be sure, so do the tests of individual variables, albeit less so, in our analysis.

We also agree that one should not worship statistical significance. Kritzer and Richards do find estimated effects that are in the correct direction in four applications of the jurisprudential regimes design. As they note in their response, however, in their two unpublished applications, the estimated effects are in the wrong directions. In the four published applications, one can say that there is a statistical correlation for some variables that is consistent with substantive expectations, but that we cannot distinguish this correlation from statistical noise, given our findings. We should certainly not, we agree, make the mistake of *accepting* the null simply because findings do not reach statistical significance. We are not saying we can conclude there is *no* such effect with statistical confidence. Perhaps these data are simply too noisy for us to detect such effects, as we suggested in our original conclusion.

The bottom line is that we cannot say with statistical confidence that there are regime changes at the points predicted by jurisprudential regime theory.[2] We agree that statistical significance is not sufficient to draw conclusions. However, we might disagree about the extent to which it should be considered a necessary part of the jurisprudential regimes research design, along with direction predictions for key variables and sensitivity checks. We have now supplemented our original randomization test evidence with an alternative constrained-randomization analysis suggested by Kritzer and Richards. This confirmed our original conclusion: of the substantively motivated regime tests that allow for membership change, none were statistically significant. While sensitivity analysis and pattern checking are, we agree, necessary components of the jurisprudential regimes research design, statistical significance tests are surely at the heart of jurisprudential regimes analysis as well.

Manuscript submitted 11 November 2009
Manuscript accepted for publication 9 December 2009

References

Leoni, Eduardo L. 2009. "Analyzing Multiple Surveys: Results from Monte Carlo Experiments." Working Paper.

Rader, Kelly T. 2009. "Randomization Tests and Inference with Grouped Data." Working Paper.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93: 133–38.

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.

[2] Smaller points in response to Kritzer and Richards' comments: (a) We agree that Kritzer and Richards do not argue that precedents bind justices, but instead argue that the justices act as though precedents bind them, which is why we put bind in quotation marks. (b) We do include variables in our tables they now clearly state as not connected to the substance of regime change, but followed them in so doing. For example, the threshold variable they now argue is irrelevant was clearly labeled a jurisprudential variable in their original test, playing a role in the joint test of three variables. (c) Kritzer and Richards write that, in their original paper, they were "primarily interested in comparing the effects of content-based and content-neutral regulations compared to the baseline category in the period after *Grayned*, which is not considered in the randomization test results the authors [Lax and Rader] present.)" The variables on content-based and content-neutral capture the effect relative to the baseline in our analysis, as in theirs, with these being represented by dummy variables relative to the omitted category whose effects are captured in the intercept. We regret any confusion as to this point, but we have followed standard practice (as did they when setting up the data analysis we replicated). (d) Because the clustering method attains its asymptotic properties when the number of clusters is large relative to the number of observations within clusters, it is unclear that when clustering by justice that there are enough groups for clustering to perform properly (Leoni 2009; Rader 2009; Wooldridge 2003).