Competing Approaches to Predicting Supreme Court Decision Making

Andrew D. Martin, Kevin M. Quinn, Theodore W. Ruger, and Pauline T. Kim

Political scientists and legal academics have long scrutinized the U.S. Supreme Court's work to understand what motivates the justices. Despite significant differences in methodology, both disciplines seek to explain the Court's decisions by focusing on examining past cases. This retrospective orientation is surprising. In other areas of government, for example, presidential elections and congressional decision making,[1] political scientists engage in systematic efforts to predict outcomes, yet few have done this for court decisions. Legal academics, too, possess expertise that should enable them to forecast legal events with some accuracy. After all, the everyday practice of law requires lawyers to predict court decisions in order to advise clients or determine litigation strategies.

The best test of an explanatory theory is its ability to predict future events. To the extent that scholars in both disciplines seek to explain court behavior, they ought to test their theories not only against cases already decided, but against future outcomes as well. Employing two different methods, we attempted to predict the outcome of every case pending before the Supreme Court during its October 2002 term and compared those predictions to the actual decisions. One method used a statistical forecasting model based on information derived from past Supreme Court decisions. The other captured the expert judgments of legal academics and professionals. Using these two distinct methods allows us to test their predictive power not only against actual Court outcomes, but also against each other.

In comparing a statistical model with actual legal experts, we do not join the stylized debate between "attitudinalism" and "legalism." Rather than arraying the justices in simple ideological space, our statistical model relies on information from past cases to discern patterns in the justices' votes based on observable case characteristics, and to construct classification trees to predict outcomes based on the characteristics of the pending cases. Conversely, our legal experts considered many factors beyond "the law" in making their predictions. Although they read legal materials, such as court opinions and the parties' briefs, legal experts also took account of factors such as the justices' policy preferences and ideologies.

The critical difference between the two methods of prediction lies not in the law/politics dichotomy, but in the nature of the inputs used to generate predictions. The statistical model looked at only a handful of case characteristics, each of them gross features easily observable without specialized training. The legal experts, by contrast, could use particularized knowledge, such as the specific facts of the case or statements by individual justices in similar cases. The statistical model also differed from the experts in explicitly taking into account every case decided by this natural court prior to the 2002 term. No individual could have such comprehensive knowledge of the Court's output for the last eight terms, and so the experts necessarily relied on fewer (albeit more detailed) observations of past Court behavior.

Not surprisingly, these different decision-making processes often resulted in divergent predictions in particular cases. More unexpectedly, for the 2002 term, the statistical model more accurately predicted case outcomes, while the experts did slightly better overall at predicting the votes of individual justices. The model's success in predicting outcomes was in large part due to its relatively greater success in predicting the votes of the five most conservative justices on the Court, including the pivotal Justice O'Connor. These results raise interesting questions about when a global,

Andrew D. Martin (admartin@wustl.edu) is associate professor of political science at Washington University, St. Louis; Kevin Quinn (kquinn@latte.harvard.edu) is assistant professor at Harvard University, Department of Government; Theodore W. Ruger (truger@law.upenn.edu) is assistant professor at the University of Pennsylvania Law School; and Pauline T. Kim (kim@wulaw.wustl.edu) is professor of law at Washington University School of Law. *The authors thank Michael Cherba, Nancy Cummings, David Dailey, Alison Garvey, Nick Hershman, and Robin Rimmer for their assistance. Their project is supported in part by National Science Foundation grants SES-0135855 and SES 0136679. The foundation bears no responsibility for the results or conclusions.*

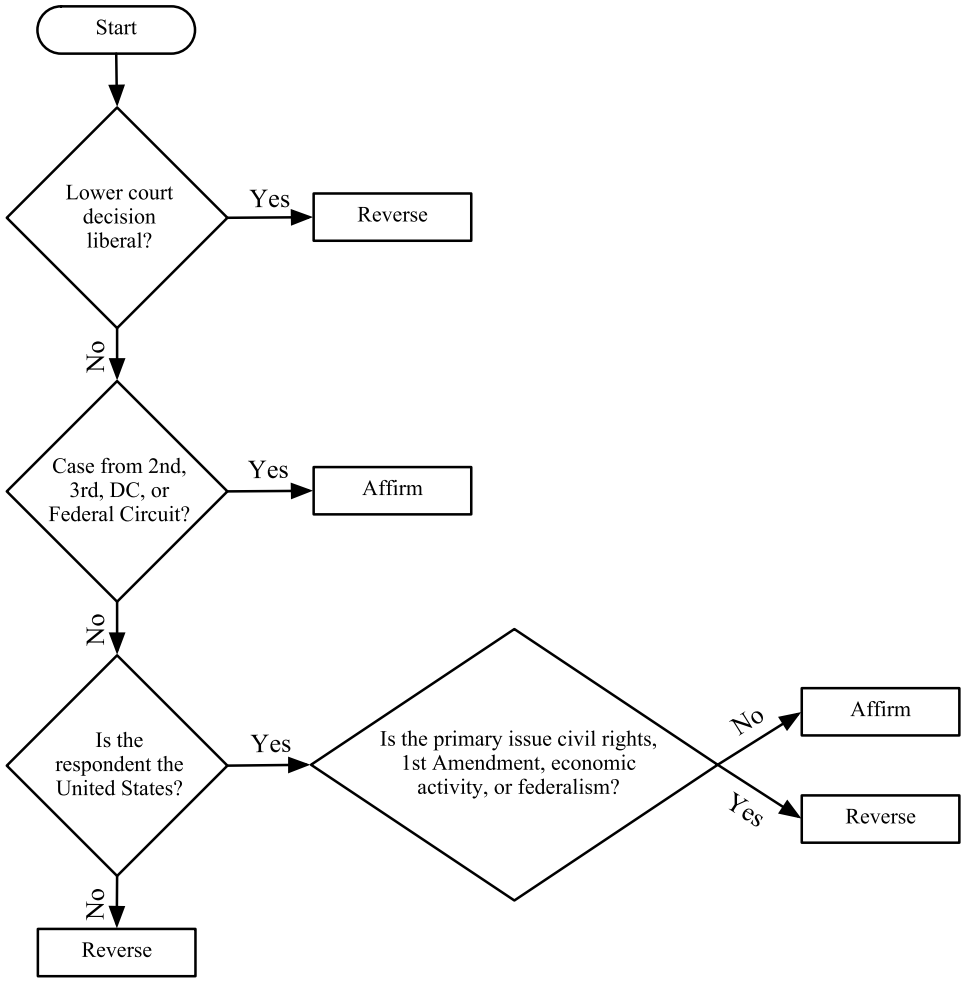
quantitative approach has a comparative advantage over particularized, expert knowledge in predicting case outcomes, and vice versa.[2]

The Statistical Model

Our approach begins with data from all 628 cases previously decided by this natural court.[3] We use these data to estimate classification trees, which were selected with a view to optimally forecasting case outcomes. The set of estimated classification trees are thus the "statistical model,"[4] which is used to predict the outcomes of each case.[5]

For each case, we collected six observable characteristics to be used as explanatory variables, each of which had to be available before oral argument so they could be used for forecasting: (1) the circuit of origin for the case;[6] (2) the issue area of the case, coded from the petitioner's brief using Spaeth's protocol; (3) the type of petitioner (e.g., the United States, an injured person, an employer); (4) the type of respondent; (5) the ideological direction of the lower court ruling, also coded from the petitioner's brief using Spaeth's protocol; and (6) whether or not the petitioner argued the constitutionality of a law or practice.

Figure 1
Estimated classification tree for Justice O'Connor for forecasted nonunanimous cases



Unlike most statistical work in political science, these variables were not chosen for explicitly theoretical reasons. Rather, they were chosen based on their availability and plausible relationship to Supreme Court decision making. Some variables, such as the ideological direction of the lower court decision, likely capture attitudinal factors,[7] while others, such as the type of petitioner, are less strongly tied to extant theory. This forecasting exercise was not one of theory testing, but rather one of determining whether systematic patterns can be uncovered using these variables.

For each case that we predicted for this study, we used the estimated classification trees to generate a forecast prior to oral argument. In figure 1 we present the estimated classification tree for Justice O'Connor.[8] To illustrate how a classification tree works, consider *Grutter v. Bollinger* (2003)—the law school affirmative action case. In this case the U.S. Court of Appeals for the Sixth Circuit held that the law school's consideration of race to achieve a diverse student body served a compelling state interest. We code this case as a liberal decision in the civil rights issue area. Proceeding down the tree, the lower-court decision is liberal, and we thus (incorrectly) forecast reversal. Had the lower-court decision been conservative, we would have progressed further down the decision tree. Note that the same prediction holds in the undergraduate affirmative action case (*Gratz v. Bollinger* [2003]) because the characteristics of the case are the same; in this case, the statistical model correctly forecast Justice O'Connor's vote. These estimated classification trees should not be interpreted causally; in figure 1, for example, it would be incorrect to infer that the fact that the Second Circuit issued a conservative decision would *cause* Justice O'Connor to affirm. Rather, for some reason (likely due to the agenda process), Justice O'Connor

tends to affirm cases decided in a conservative direction by the Second Circuit.

The Legal Experts

The study's other method of prediction sought to capture the judgments of legal experts. Experts are distinguished from nonexperts by extensive training and experience in the relevant domain; they also have the ability to perceive meaningful patterns that cannot easily be coded into a statistical model and to structure their knowledge around principle-based schema. Often their judgments are based on qualitative analyses. The process underlying the judgments of our legal experts thus differed considerably from that underlying the model.

Because no metric exists to measure expertise precisely, we recruited participants much the way anyone might look for expert assistance: we researched their writings, checked their training and experience, and relied on referrals from knowledgeable colleagues. The 83 individuals who participated easily qualify as "experts," having written and taught about, practiced before, and/or clerked at the Supreme Court, and developed special knowledge in one or more substantive fields of law.[9] Collectively, they form an accomplished group of 71 academics and 12 appellate attorneys. Of this group, 38 clerked for a Supreme Court justice, 33 hold chaired professorships, and 5 are current or former law school deans.

We asked experts to predict a case or cases within their areas of expertise. Experts assigned to the same case did not communicate with one another about their predictions. We requested their forecasts prior to oral argument, assuring them that their individual predictions and the cases to which they were assigned would not be revealed. Their predictions took the form of an "affirm or reverse" choice for the Court as a whole and for each justice. The bluntness of this binary choice precluded discussion of the kind of doctrinal nuance and partial holdings that are important to lawyers. However, limiting experts to a binary choice was necessary for direct comparison with the output from the statistical model.

Experts were free to consider any sources of information or factors they thought relevant to making their prediction. We provided a copy of the lower court opinion and citations to the parties' Supreme Court briefs, but did not limit experts to these materials. After they had returned their prediction for a particular case, we surveyed them regarding what factors were important to their decision. Nearly 90 percent of the experts completed at least one survey.

Results

We posted all of the statistical model and expert forecasts on the project Web site (http://wusct.wustl.edu) prior to oral argument. After decision, case outcomes and individual justice votes were coded "affirm" or "reverse." Cases that were vacated and remanded or reversed even in part were coded as "reverse." We use 68 cases to analyze the case outcome forecasts[10] and 67 to analyze individual vote predictions.[11]

Table 1
Model and expert forecasts of case outcome for decided cases

	Case outcome forecast		
	Correct	Incorrect	Total
Model	51 (75.0%)	17 (25.0%)	68 (100.0%)
Experts	101 (59.1%)	70 (40.9%)	171 (100.0%)

Note: Table is based on 68 cases. The unit of analysis is the case-prediction. Row percentages are in parentheses. The estimated (conditional maximum likelihood) odds ratio is 2.073 (p = 0.025, Fisher's exact test).

As seen in table 1, the model correctly forecast 75.0 percent of ultimate case outcomes, while the experts' predictions had an accuracy rate of only 59.1 percent. In forecasting individual justice votes, the experts did marginally better than the model. The experts' vote predictions were 67.9 percent accurate, compared to 66.7 percent correct for the model. See table 2.

In tables 1 and 2 we treat each expert independently, summarizing the results by aggregating all expert predictions. As an alternative method of comparison,[12] we take the predictions of a majority of the experts on a particular case to generate an "expert consensus" forecast. Using the "expert consensus" forecast results in somewhat similar results, depending upon how we treat inconclusive forecasts.[13] See table 3.

For all of the tables we report the estimated (conditional maximum likelihood) odds ratio, and a p-value from Fisher's Exact Test.[14] The null hypothesis is that the proportion of correct predictions by the model and the experts is the same. In table 1, which treats all expert predictions independently, the difference in forecasting accuracy of the model

Table 2
Model and expert forecasts of individual justice's votes for decided cases

	Justice vote forecast		
	Correct	Incorrect	Total
Model	400 (66.7%)	200 (33.3%)	600 (100.0%)
Experts	1015 (67.9%)	479 (32.1%)	1494 (100.0%)

Note: Table is based on 67 cases. The unit of analysis is the justice-case-prediction. Row percentages are in parentheses. Some justices did not vote on some cases, and are thus not included. The estimated (conditional maximum likelihood) odds ratio is 0.943 (p = 0.571, Fisher's exact test).

Table 3
Model and expert consensus forecasts of case outcomes for decided cases

	Case outcome forecast			
	Correct	Incorrect	Inconclusive	Total
Model	51 (75.0%)	17 (25.0%)	0 (0.0%)	68 (100.0%)
Expert consensus forecasts	40 (58.8%)	21 (30.9%)	7 (10.3%)	68 (100.0%)

Note: Table is based on 68 cases. Row percentages are in parentheses. The expert consensus forecast is based on the predictions of the majority of experts on a particular case. If only 2 experts predicted a given case and their predictions disagreed, the expert consensus forecast is listed as "inconclusive." Treating inconclusive forecasts as incorrect, the estimated (conditional maximum likelihood) odds ratio is 2.088 (p = 0.067, Fisher's Exact Test). Alternatively, we could assume that if a third prediction had been obtained, the distribution of correct predictions would mirror the overall distribution of correct expert prediction, resulting in 64.7% of the expert consensus predictions being correct. Using this assumption, the estimated odds ratio is 1.630 (p = 0.262, Fisher's exact test).

and the experts is statistically significant. The model clearly beats the experts on these cases, although a different result might well be obtained in a different term with another group of experts.

Of greater significance than the winner of this particular contest is the model's considerable success in forecasting case outcomes. Why did the experts not do as well in predicting case outcomes? The explanation may well lie in their relative inability to predict the votes of Justice O'Connor. Figure 2 graphs the proportion of correctly predicted votes by the model and the experts, for each justice. As seen in the figure, the experts did worst at predicting O'Connor's votes among all the justices, and considerably worse than the model. Because O'Connor is often thought of as the Court's pivotal justice, the experts' relative lack of success in predicting her votes meant a poorer showing than the model in forecasting case outcomes.

Figure 2 arrays the justices along the vertical axis in order of increasing conservatism as estimated for the 2001 term by Martin and Quinn.[15] Focusing on the proportion of each justice's votes that the experts were able to predict correctly reveals an interesting pattern. The proportion of correct predictions forms a sideways V-shape, indicating that the experts were most accurate at predicting the votes of the most ideologically extreme justices and least successful at forecasting the votes of the centrist justices. This pattern is consistent with attitudinalist explanations of justices' votes, and legal experts' implicit acceptance, at least in part, of an attitudinalist explanation. In fact, a solid majority of legal experts reported that the justices' policy preferences and their conservative or liberal ideologies were important factors in making their predictions.[16]

We have also sorted our results by issue area using Spaeth's coding protocol.[17] Figures 3 and 4 display the proportion of correctly predicted case outcomes and justice votes for issue areas with five or more cases in our sample. These figures suggest that the relative success of the two methods varies significantly depending upon the issue area. Given the small number of cases in each category, these comparisons are obviously quite sensitive to the category definitions and the coding decisions in individual cases. Nevertheless, the relative success of the experts in certain issue areas suggests that the particularized knowledge to which they have access may provide a comparative advantage in predicting outcomes of certain types of cases.

The experts did significantly better than the model in predicting the judicial power cases (see figs. 3 and 4). These cases generally involved narrow technical issues of procedure in which the rule of decision was unlikely to have much impact outside the legal system itself. For example, *Breuer v. Jim's Concrete,* in which all three experts correctly predicted a 9–0 affirmance and the model predicted a 5–4 reversal, raised the question of whether statutory language conferring concurrent jurisdiction in state and federal courts

Figure 2
Model and expert forecasts of votes for decided cases (n = 67), by justice.

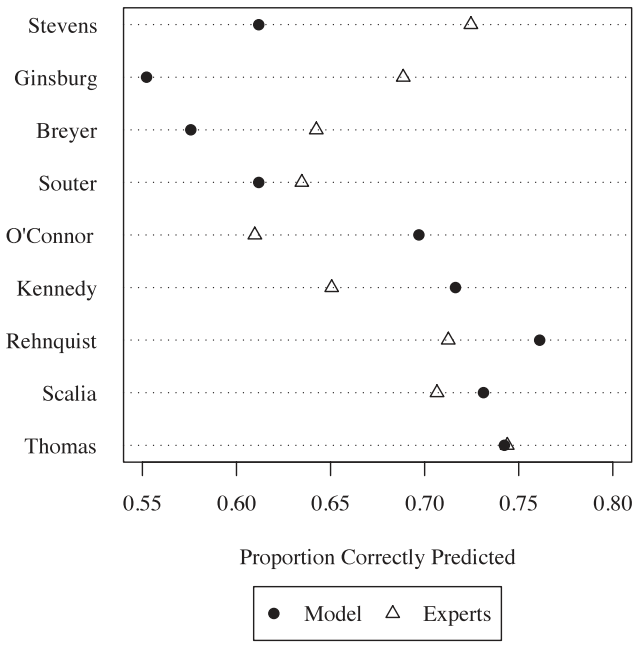
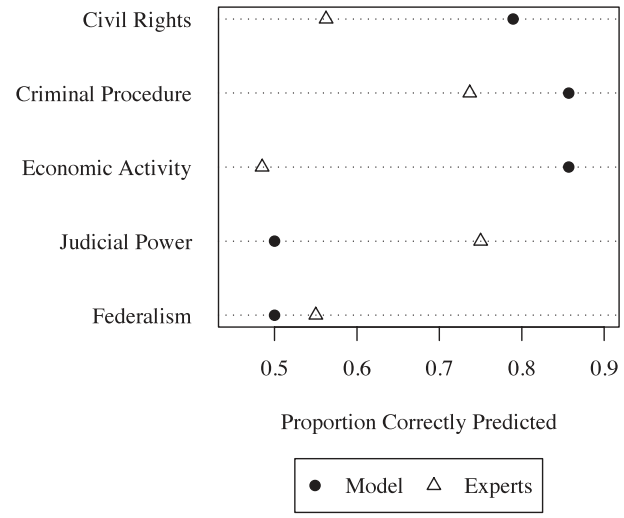
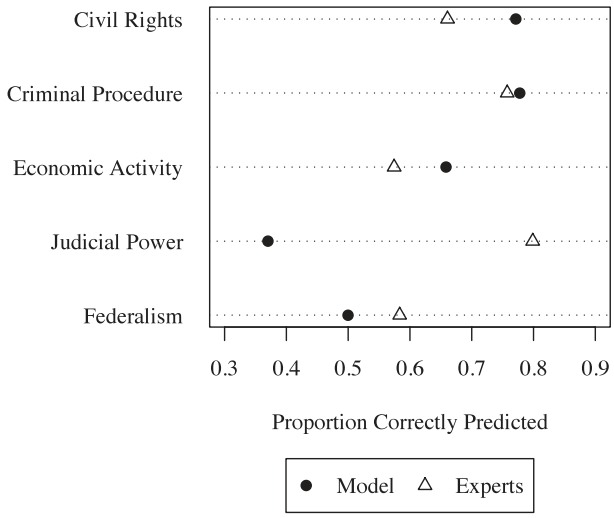


Figure 3
Model and expert forecasts of case outcomes for decided cases selected by issue area



Note: Issue areas are coded by Spaeth (2003) and are mutually exclusive. The issue categories are: civil rights (n = 14), criminal procedure (n = 14), economic activity (n = 16), judicial power (n = 8), and federalism (n = 5).

Figure 4
Model and expert forecasts of justice votes for decided cases selected by issue area



Note: Issue areas are coded by Spaeth (2003), and are mutually exclusive. The issue categories are: civil rights (n = 14), criminal procedure (n = 14), economic activity (n = 16), judicial power (n = 8), and federalism (n = 5).

barred removal to federal court of an action initiated by the plaintiff in state court. It is likely that these cases, more than most, turned on highly particularized features of the case—perhaps conventional "legal" factors such as statutory text and *stare decisis*—that the experts were able to recognize and incorporate into their decision-making process. The model, limited to the gross features of the case, likely missed the very specific factors on which these outcomes turned. The experts' relatively better performance in predicting justices' votes in the federalism cases also suggests that these votes often turned on factors not easily captured by the blunt case characteristics relied upon by the model.

The experts' survey responses clarify the difference in inputs between the two methods of prediction. Three of the six variables used in the model—circuit of origin, identity of the petitioner, and identity of the respondent—were deemed unimportant by large majorities of experts in the predicted cases (64 percent, 69.1 percent, and 75.5 percent, respectively). And unlike the model, the legal experts put great weight on legal authority—primarily in the form of prior Supreme Court opinions—in making their predictions. Although the statistical model's "issue area" variable incorporates some legal factors, it could only influence predictions in a fairly blunt way, in contrast to the legal experts' focus on particular cases, or even particular statements in past cases.

The two methods of prediction are perhaps most similar in taking account of the justices' preferences. The statistical model relied on a variable intended to capture the ideological direction of the lower-court opinion. In fact, the lower-court direction variable entered into the classification trees for all justices except Stevens (who was forecast to vote liberal on nearly every case), and for the small subset of cases forecast to be unanimous in the conservative direction. Similarly, substantial majorities of the experts responded that the justices' policy preferences and their conservative or liberal ideologies were important factors in their forecasts.

Conclusions and Implications

We designed this study primarily to compare two methods of assessing and predicting Supreme Court decision making. We found—somewhat to our surprise—that for this particular term the outcome predictions of the statistical model were more accurate than those of the experts, despite the fact that the experts were slightly better at forecasting the votes of individual justices. Critically, the model did significantly better than the experts at predicting the votes of Justices O'Connor, Kennedy, and Rehnquist, and this fact, coupled with the importance of those three justices in the ideological makeup of the current Supreme Court, explains much of the statistical model's success. Of course, the number of predictions is small enough that these differences may not hold across future terms, and additional iterations of this project are necessary to form firm conclusions.

With this caveat in mind, it is possible to discuss a few implications of our results. The two prediction techniques we used differ more dramatically in methodology than in underlying theory, and one methodological distinction is particularly stark. In calculating probabilities, the model incorporated each of the 628 cases that the sitting natural Court has decided in the past, and assigned each result equal empirical weight. The legal experts, by contrast, did what lawyers are trained to do—focus on a smaller subset of cases and give predominant analytical weight to a few leading cases. Law professors who study the Court retrospectively likewise focus on a handful, or at most a few dozen, important cases, rather than assess every case in a given term or given issue area. Even if an individual could take into account every case in a given term, much less a decade, the limits of human cognition would prevent an expert from assigning equal analytical weight to every case, as the model does. But there is a predictive benefit from a broad observation of past Supreme Court behavior, and there may be a similar benefit even in retrospective analyses from looking at a greater swath of Supreme Court decisions than typically considered by legal specialists.

A second difference in methodology has greater bearing on the traditional law-politics debate about the factors that underlie Supreme Court decision making. Although the model incorporated a large number of past results in its analysis, it took no account of the explanations the Court itself gave for those decisions. Nor did it take into account specific precedent or relevant statutory or constitutional text. The model is essentially nonlegal in that the factors used to predict decisions—the circuit of origin, the type of the petitioner and respondent, and so forth—are indifferent to law. If patterns relative to these observable factors capture legal differences (grounded in text, history, or precedent) at all, they do so only incidentally and with extreme generality.

Lawyers, politicians, and policy makers more generally care as much or more about the form that the law takes as about which party wins in a particular case. This concern raises the question whether a statistical model could ever forecast cases in terms of the legal principles they announce instead of a simple "affirm" or "reverse" outcome. In theory, the answer is yes, with a significant caveat. Statistical forecasting is possible whenever there is reliable training data. If a reliable and replicable scheme existed to code for legal arguments across a variety of case types, it could be applied to a large number of past cases and used to develop a model capable of predicting justices' choices among legal principles. The difficulty, of course, lies in developing such a coding system.

Conversely, although our experts were not limited to considering only "legal" factors, such as text and precedent, they did rely consistently on such factors. In addition, they were able to capture, in a manner that the model could not, factual idiosyncrasies in particular cases, and to consider how particular law might interact with particular facts to affect outcomes. Still, the overall predictive success of the model was higher—which suggests that for many cases an expert's ability to identify and analyze legal factors is not much help in predicting the Court's behavior.

Of interest in this regard is a subset of cases—those within the broad subject area of "judicial power"—where the experts did do well relative to the model. If one assumes that Supreme Court decision making is multifaceted—that it is an aggregation of the justice's preferences, of strategic interaction with other branches and within the Court, and of the constraints of text or precedent[18]—it is natural to also assume that in certain cases some factors matter more than in others. One motivation for this study was to determine whether there are some kinds of cases where observable "legal" factors are more important for the prediction of outcomes. Our results in the judicial power cases suggest that this is an area where "legal" factors are important.

There may be another implication of this project beyond academia—in the world of litigants before the Supreme Court and (if our model is extended) before other courts. Expected outcomes play a huge role in litigants' decisions to press an appeal and in settlement negotiations, and to the extent this statistical model (or an improved future version) can reliably forecast judicial outcomes, it may benefit practicing attorneys and their clients.

Finally, we hoped through this explicitly interdisciplinary experiment to create a project of interest to the two groups of scholars who study the Supreme Court most closely, and thereby to enhance the gradually increasing dialogue between our two disciplines. This symposium is one realization of that aim, and we hope that the discussion continues with insight from others in both political science and law.

Notes

A complete reference list for the entire symposium appears on pp. 791–93, below.

1 Lewis-Beck and Rice 1992; Poole and Rosenthal 1991.

2 We offer a necessarily abbreviated description of the two methods of prediction and our major findings. More detailed analyses, along with a complete description of the statistical model, may be found in our longer article in the May 2004 issue of the *Columbia Law Review,* Ruger et al. 2004.

3 The statistical forecasting model uses classification tree analysis and past voting behavior of the current Supreme Court justices to compute probable outcomes of decisions in the October 2002 term. Breiman et al. 1984.

4 Our "model" for this study is actually eleven distinct classification trees. The first two predict whether, based on the explanatory variables, a case is likely to be either a unanimous "liberal" decision or a unanimous

"conservative" decision. If a unanimous decision is predicted in one direction then the forecast for that case is complete. However, if neither initial model predicts a unanimous decision (or if both do, in opposite directions), then the forecast is based on nine justice-specific tree models that forecast the direction of the vote of each justice. We allow the predicted votes of some justices to enter into the decision trees of some other justices to allow for interdependence. The specific trees used for forecasting were those with the best out-of-sample forecasting properties.

5 As discussed on the project Web site (http://wusct.wustl.edu), there was a bug in the code used to produce the forecasts given the classification trees. The original posting on the Web site thus did not correctly reflect the predictions of the statistical model. The forecasts currently on the Web site are the correct forecasts. A technical description of the bug and replication information is available on the project Web site.

6 Lower court decisions from a state court or a three-judge federal district court were coded as arising from the federal circuit encompassing the state.

7 Segal and Spaeth 2002.

8 The estimated classification trees for the other justices are available on the project Web site and in an appendix to Ruger et al. 2004.

9 We list the names of the participating experts in an appendix to Ruger et al. 2004. We note with gratitude that the experts' participation was an entirely volunteer effort, and their substantial intellectual generosity quite literally made this project possible.

10 Of the 76 cases in which the Court heard oral argument, we excluded 8 from our analysis. We excluded 3 cases because they were dismissed without opinion (*Abu-Ali Abdur'Rahman v. Ricky Bell, Ford Motor Co. v. McCauley,* and *Nike, Inc. v. Kasky*) and 2 because they were affirmed by an evenly divided Court, with no information about individual votes (*Borden Ranch Partnership v. U.S. Army Corps of Engineers and Dow Chemical v. Stephenson*). We excluded 3 additional cases due to intractable coding ambiguities (*Virginia v. Black, Green Tree Financial Corp. v. Bazzle, and National Park Hospitality Assn. v. Dept. of Interior*). See Ruger et al. 2004.

11 We excluded *Chavez v. Martinez* from our vote analysis only, due to coding difficulties caused by ambiguous concurrences by some of the justices.

12 Treating individual expert predictions independently might be misleading, because the same number of experts did not predict each case, meaning that the weight of the machine's forecasts differs depending upon the number of experts who also predicted that case.

13 Cases with only two experts with opposite predictions result in an inconclusive "expert consensus" forecast. We could treat these inconclusives as incorrect (resulting in a 58.8 percent success rate for the experts), exclude them altogether (resulting in a 65.6 percent success rate), or assume that if a third prediction had been obtained in all these cases, the distribution of correct predictions would mirror the overall distribution of correct expert predictions (resulting in a 64.7 percent success rate).

14 Fisher 1935.

15 Martin and Quinn 2002.

16 Experts were sent a survey following receipt of their prediction in a particular case. Experts predicting more than one case received a survey for each case. In all, approximately 90 percent of our experts returned at least one survey, and we received responses for 65 percent of the expert predictions made during the term. The survey presented a list of factors that might be considered relevant to predicting Court decision-making and asked experts how important each factor was to his or her prediction in a given case. Respondents were asked to rate the factors on a 5 point Likert scale (1 = not at all important; 5 = very important). Over two-thirds (67.5 percent) of the respondents indicated that the "policy preferences of the justices on the specific issue presented" was an important factor (4 or 5), and over half (54.3 percent) responded that "the conservative or liberal ideologies of the individual justices" was an important factor (4 or 5).

17 These issue codes are based on the VALUE variable in Spaeth 2003.

18 See, for example, Epstein and Knight 1998.