Journal of Empirical Legal Studies
Volume 5, Issue 3, 407–446, September 2008

Case Selection and the Study of Judicial Politics

Jonathan P. Kastellec and Jeffrey R. Lax[*]

One complication in studying the Supreme Court and the judicial hierarchy is that the Court's docket is now nearly completely discretionary. Although the Justices' strategies in picking cases affect the observations we make and the inferences we draw, this is rarely taken into account in studies of judicial politics. In this article, we study how case selection can affect our inferences within judicial politics, including those about decision making in the Supreme Court itself (such as whether law constrains the Justices) and throughout the judicial hierarchy (such as whether lower courts comply with Supreme Court doctrine). We use simulation analysis to show that the inferential problems raised by the Court's case selection range from moderate to severe. At stake are substantive conclusions within some of the most important and controversial debates in judicial politics.

I. Introduction

One complication in studying the Supreme Court and the judicial hierarchy is that the Court's docket is now nearly completely discretionary. The Justices themselves choose which cases they will hear, and it is universally recognized that they eschew less weighty cases, choosing instead to take hard, important,

[*]Address correspondence to Jonathan P. Kastellec, Department of Political Science, Columbia University, 420 W. 118th St., New York, NY 10027; email: JPK2004@columbia.edu. Kastellec is Ph.D. candidate; Lax is Assistant Professor, Department of Political Science, Columbia University.

Earlier versions of this article were presented at the Annual Meeting of the American Political Science Association, August 31–September 3, 2006, and the Annual Meeting of the Conference on Empirical Legal Studies, November 9–10, 2007. We would like to thank Michael Bailey, Cheryl Boudreau, Theodore Eisenberg, Andrew Gelman, Shigeo Hirano, David Klein, Dan Klerman, Eduardo Leoni, Justin Phillips, Kirk Randazzo, and Rebecca Weitz-Shapiro for helpful comments and suggestions, and Jeffrey Segal, Donald Songer, Charles Cameron, David Klein, and Robert Hume for generously sharing their data with us. Replication data sets and statistical code are available at ⟨http://www.columbia.edu/~jpk2004/selection_replication.html⟩.

or controversial cases (Easterbrook 1982; Perry 1991; Kritzer & Richards 2002).[1] They can select cases in order to develop particular legal doctrines or to rectify noncompliance in the lower courts, to name just two further possibilities.

The precise selection strategy employed by the Justices will affect the set of Supreme Court cases we observe in a given area of the law in a given time period. Legal scholars and others have long been concerned about making generalizations about law, legal practice, and judicial behavior from the Supreme Court given that the cases the Justices hear are not representative (Edwards 1985; Cross 1997; Friedman 2006). Regardless of the particulars of case selection, it hardly seems likely that the Court chooses cases randomly. Accordingly, the Court's selection process raises the potential for selection bias in the inferences we draw from its cases.

Since Heckman's (1979) seminal article, social scientists have been aware that selection bias may materially affect model estimation and substantive inferences (see, e.g., Grier et al. 1994; Poe & Meernik 1995; Timpone 1998; Hart 2001; Ashworth et al. 2008).[2] However, this possibility has not been incorporated into standard research designs in judicial politics. Indeed, the problem of selection bias is usually noted only in passing, with conclusions drawn as though case selection were random.[3]

Consider a simple example of why selection matters for studying judicial behavior. Over the 20th century, the reversal rate on the U.S. Courts of Appeals in published decisions was roughly 30 percent (Songer et al. 2000:105). In contrast, the Supreme Court reversed around 60 percent of cases it heard in the same period (Segal & Spaeth 2002:263). One possibility is that Supreme Court Justices simply have a higher propensity to reverse lower courts than courts of appeals judges, but the more likely explanation is that the Justices use their discretionary docket to select cases to reverse, while appellate court judges frequently hear routine cases. Although this

[1] Indeed, the Judiciary Act of 1925 was realized directly as a result of the Justices' professed desire to not have to decide hundreds of "run-of-the-mill" cases each term (Hartnett 2000).

[2] Even if sample selection does not substantially affect estimation, selection models can help establish the robustness of traditional estimation techniques (Allee & Huth 2006).

[3] By definition, the one area of research where selection *has* been incorporated is that of certiorari (cert)—the Court's decision whether to hear a particular case or not. Caldeira et al. (1999) and Cameron et al. (2000), for instance, account for the cases the Court chooses *not* to hear in their analyses. This is almost never done in more general studies of judicial decision making.

discrepancy in well-known, it helps illustrate that selection bias is relevant for evaluating even simple patterns of judicial behavior.

Other examples turn up when we think about litigants or the courts of appeals directly. Studying only cases that go to trial and ignoring settled cases may produce a biased sample of the underlying population of cases (Priest & Klein 1984; Eisenberg 1990). Similarly, taking into account litigants' decision to appeal can affect our understanding of decision making on the U.S. Courts of Appeals (Clermont & Eisenberg 2001). Finally, studying only published cases to the exclusion of unpublished decisions may lead to either an incomplete or misleading assessment of a particular court's output (Siegelman & Donohue 1990; Merritt & Brudney 2001; Law 2005).

To obtain a full measure of the possible selection bias at the Supreme Court level, consider the typical path a case must take to be decided by the Supreme Court. The parties must be engaged in a dispute that is not settled before a trial court's decision. One party must appeal following that decision (by no means a costless action), again without a settlement before the appellate court decision. One party must then bear the costs of a cert petition, which must compete against thousands of other petitions for the Supreme Court to actually grant cert. Each step in this process creates different selection biases, which may be compounded as a case moves through the judicial process.

It is diffcult to think of a major debate within judicial politics that is not potentially touched by this problem. Scholars rely on observed Supreme Court cases to study the preferences of Supreme Court Justices (Grofman & Brazill 2002; Martin & Quinn 2002; Segal & Spaeth 2002), external constraints on their decision making (Spiller & Gely 1992; Segal 1997; Bergara et al. 2003), the treatment of case facts (Segal 1984; McGuire 1990; Hagle 1991; Ignagni 1994), the role of law (George & Epstein 1992; Richards & Kritzer 2002), and ideological change on the Court (Epstein et al. 1998, 2007b; Martin & Quinn 2007). At stake, therefore, are many of the most important and controversial substantive debates in judicial politics.

In this article, we present the first investigation (to the best of our knowledge) of how selection bias resulting from the Supreme Court's case selection might affect the inferences we draw using common research designs in judicial politics. We ask with what confidence we can trust inferences that do not take selection bias into account. We explore studies of the Supreme Court's decision making itself, as well as research that extrapolates from the cases the Court hears to those it does not in order to make inferences about the behavior of other legal actors. The former include such issues as the Supreme Court's choice of legal doctrine and the role of ideology. The latter includes the study of compliance by lower courts with Supreme Court decisions. The link between them is that they use fact-pattern analysis, a research method central to much scholarship in judicial politics. We demonstrate that research designs that use Supreme Court decisions to extrapolate general preferences as to case facts and issues are prone to biased results, undermining our confidence in their conclusions.

To examine the possible effects of selection bias, we run simulations using Fourth Amendment case data. Using these simulations, we can specify the true preferences of the Supreme Court over case decisions and the treatment of various case facts, and then compare these to the inferences one would draw from observed case decisions, given an array of case-selection strategies. We also can assess the validity of inferences regarding lower court compliance based on observed Supreme Court cases. Specification of true preferences and the use of simulations allow us to assess the output of standard research designs, which cannot assess estimates against true parameter values because they are unknown.

It may seem obvious that selection effects are going to exist, yet they have not been taken into account. Moreover, our results demonstrate that the potential biases introduced by case selection are not minor, as one might suspect given the paucity of consideration to the issue in the judicial politics literature. Instead, they range from moderate to severe.

Key substantive conclusions are directly called into question. Among the inferences affected are which case facts matter, and how much; changes in Supreme Court doctrine; the liberalism of the Court and individual Justices; and the degree of compliance in the lower courts. We show that inferences as to all of these can be quite wrong. Our results show that selection bias *cannot* simply be set aside as it has been in most work. We show that commonly used research designs are flawed, undercutting the conclusions drawn within such work These results sound the alarm that without serious attention to these issues, serious mistakes of inference can be made.

The article proceeds as follows. We begin by laying out a framework for exploring case selection, focusing on two specific research areas to ground our general results. (We also raise an empirical puzzle about lower court compliance, which we call the "noncompliance puzzle.") Section 3 presents our data and methods, which rely on simulations of case selection given lower court case data. In Section 4, we discuss our results and their implications. Section 5 concludes with thoughts on how to move forward given our findings.

II. Cases, Facts, and Outcomes

To gain leverage on the effects of case selection, we employ fact-pattern analysis: the study of judicial decision making in a given area of law in the context of specific case facts.[4] Building on the foundation established by Kort (1957), Segal (1984) and other scholars have demonstrated how the presence or absence of certain case facts can successfully predict judicial decisions, a relationship that holds both across issue areas and across courts.[5] How these facts come together to affect decisions is central to the study of judicial decision making.

Fact-pattern analysis is primarily used in two ways. First, scholars are interested in whether a particular case fact has a significant effect on the probability that a court will rule in one direction or the other, where direction is typically measured dichotomously as liberal or conservative.[6] Given the dichotomous nature of the dependent variable, logistic regression or probit is usually used to estimate the coefficients on each case fact. These coefficients can be thought of as weights on the various facts, measuring how much the fact in question "pushes" or "pulls" a particular case toward receiving a conservative or liberal classification.

Second, scholars can use these fact weights to generate a predicted outcome in any case, given a set of facts. Where such fact weights are those taken from a logit analysis of Supreme Court decisions, we can then predict

[4] As Friedman (2006) argues, the term "case facts" can obscure their often interpretative nature and mask the distinction between fact and law. Richards and Kritzer (2002) use the term "case factors" instead. For clarity's sake, we employ the more common term "facts," while acknowledging that the presence or absence of a particular case element is not always so clear-cut as many political science accounts would suggest. Indeed, some of the facts we use in our analysis, such as whether probable cause existed for a search, might more properly be called "legal judgments." Doing so, however, would not change our substantive conclusions.

[5] Since Segal's (1984) path-breaking work on Fourth Amendment cases (updated in Segal & Spaeth (1993, 2002) and extended to the federal courts of appeals in Songer et al. (1994), fact-pattern analyses have been applied to several other areas of the law, including sex discrimination (Segal & Reedy 1988); obscenity (McGuire 1990; Hagle 1991; Songer & Haire 1992); the death penalty (George & Epstein 1992; Hall & Brace 1996); the Establishment Clause (Ignagni 1994; Kritzer & Richards 2003); the law of confessions (Benesh 2002); freedom of expression (Richards & Kritzer 2002); judicial review (Emmert 1992); and school financing (Swinford 1991). Segal also uses fact patterns to explore legal change over time and differences between Justices (Segal 1985, 1986).

[6] See Wahlbeck (1997) for an important exception, however.

what the Supreme Court would do in any case, even one it does not hear, so long as we can measure the relevant fact variables.

Accordingly, our analysis is designed to test the validity of inferences drawn both with respect to the influence of case facts on outcomes and to outcomes directly. It could be that the fact weights we observe are affected by which cases the Supreme Court chooses to hear. That is, if the Supreme Court selected cases differently, we might observe very different fact weights and draw different conclusions as to the state of the law, jurisprudential changes, and the like. In turn, we would also predict different case outcomes.

In addition to drawing general conclusions about the inferences in judicial politics, we connect our results back to two specific lines of inquiry, one of each type: "jurisprudential regime" theory, which focuses primarily on case facts, and the judicial compliance literature, which focuses primarily on case outcomes. Our results undercut the key conclusions drawn in these two research areas. To be clear, our purpose is not to single out these studies—both are highly innovative and, despite our concerns, the arguments they make are ones to which scholars should pay close attention. Rather, we highlight them to demonstrate the depth of the problem using specific examples. Indeed, our results raise questions for a wide range of research, as we make clear when discussing our findings.

A. Jurisprudential Regimes

In an important series of papers, Kritzer and Richards use fact weights to test their theory that the influence of law within the Supreme Court can be found in the form of "jurisprudential regimes" (Richards & Kritzer 2002; Kritzer & Richards 2003, 2005). These regimes "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 & Kritzer 2002:305). Their research design is based on finding statistically significant differences in fact weights before and after a specific break point in time determined by an important decision by the Court. This Supreme Court decision is said to restructure legal doctrine, thus establishing a new jurisprudential regime, which changes the way the Court itself will decide future cases.

If the jurisprudential regimes theory is right, these conclusions would mark considerable progress in the elusive quest to demonstrate empirically that "law matters" in Supreme Court decision making. Tests of the theory depend on reliable estimation of the magnitude and significance of fact weights. Unfortunately, our results show that case selection has dramatic effects on such estimation. If we cannot reliably measure fact weights within a single period, we almost certainly cannot reliably test differences across periods.

B. Compliance and the Noncompliance Puzzle

In a provocative and innovative paper, Klein and Hume (2003) set out to test whether fear of reversal can explain the high rate of compliance generally observed in studies of lower federal court interpretations of Supreme Court doctrine. To do so, they examine search and seizure cases and develop a straightforward measure of compliance for use as a dependent variable: Did the circuit court decide the case in the same direction as we would expect the Supreme Court to, if it were hearing the same case? If it does, the lower court is said to have complied (Klein & Hume 2003:586). Klein and Hume make such predictions using estimates of the Supreme Court's fact weights.

To foreshadow our concern, note that using this measure of compliance requires estimates of the Supreme Court's preferred fact weights to be reliable. Based on this measure, Klein and Hume find noncompliance in 21 percent of the lower court cases in their sample, a percentage that conforms to the repeated finding of widespread compliance among the lower federal courts (Gruhl 1980; Songer & Sheehan 1990).

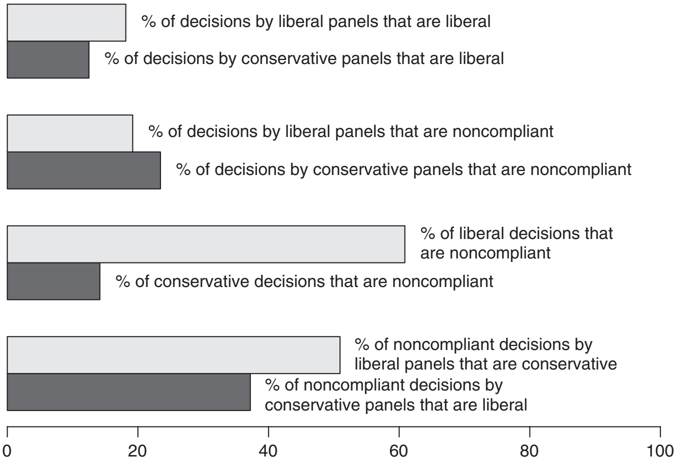
However, when we look closely at the particular decisions that are labeled as compliant or noncompliant by this method, the face validity of this measure is called into question. Although Klein and Hume do not break down compliance rates by ideological direction, we first divide the sample into liberal and conservative lower court panels (based on whether the median ideology score in the panel was less than or greater than the median ideology of all the panels combined).[7] For search and seizure cases, we would expect liberal judges to want to strike down searches (the liberal outcome) more often than conservative judges, who will want to uphold searches (the conservative outcome) more often, though they are likely to agree on the outcome in a significant percentage of cases. Even when they disagree, the tendency to comply with the Supreme Court's doctrine will further mediate differences in behavior.

[7] We use the same ideology scores as Klein and Hume (2003:587). These scores are based on the characteristics of the judges, such as their legal experience and the party of their appointing president.

Indeed, the first set of bars in Figure 1 shows that the direction of case outcomes varies little by panel ideology—liberal panels decide liberally 18 percent of the time; conservative panels decide liberally 12 percent of the time. The second set of bars in Figure 1 shows that panels also do not differ much in terms of their percentage of decisions measured as compliant—liberal panels comply 77 percent of the time while conservative panels comply 81 percent of the time.

However, when we tabulate noncompliance by the direction of the lower court outcome, a more one-sided picture emerges. The third set of bars in Figure 1 shows that a majority of liberal decisions—61 percent—are measured as noncompliant, compared to 14 percent of conservative decisions. Although one might seek to explain this discrepancy by assuming that liberal panels are simply more likely to disobey the Supreme Court, breaking

Figure 1: The noncompliance puzzle: Why do appeals court judges seem to act against both their own preferences and the Supreme Court's?



Note: Percentages are based on data analyzed in Klein and Hume (2003). The first set of bars shows that the direction of case outcomes varies little by panel ideology. The second set shows that panels also do not differ much in terms of their percentage of decisions measured as compliant. Both results make sense. However, the third set shows that a majority of liberal decisions are measured as noncompliant, compared to only 14 percent of conservative decisions. The final set of bars shows that counter to expectations, we do not see liberal panels being noncompliant solely in pursuit of liberal outcomes or conservative panels being noncompliant solely in pursuit of conservative outcomes. Instead, in half the decisions where they were measured as noncompliant, liberal panels actually found the search reasonable (the conservative outcome), while conservative panels found the search unreasonable (the liberal outcome) in more than one-third of the decisions in which they were measured as noncompliant.

the data down still further reveals an even stranger pattern that belies this possibility.

If the compliance measure is accurately picking up noncompliance, we should see liberal panels being noncompliant almost always in pursuit of liberal outcomes and conservative panels being noncompliant almost always in pursuit of conservative outcomes. The final set of bars in Figure 1 shows otherwise. In 51 percent of the decisions where they were measured as noncompliant, liberal panels actually found the search reasonable (the conservative outcome, in cases where a liberal outcome would have meant compliance). Similarly, conservative panels found the search unreasonable—the liberal outcome—in 37 percent of the decisions in which they were measured as noncompliant (where following their own ideological preference would have actually meant compliance). Taken together, in 42 percent of the decisions where lower courts were measured as noncompliant, they seemingly disobeyed the Supreme Court just to act against their own interests. No theory of judicial behavior or of jurisprudence would suggest this pattern.

The stakes in this debate are quite high. Klein and Hume use their compliance measure to explore whether it is the fear of reversal that drives compliance and conclude that it is not. If this conclusion is correct, it would call into question a wide range of formal and empirical models that assume lower courts seek to avoid reversal (McNollgast 1995; Cameron et al. 2000; Lax 2003; Kastellec 2007). The puzzle above, however, calls into question the face validity of the measure of compliance they use to draw this conclusion. Our results that follow show that the measure is severely biased by the Court's case-selection strategy. Although measurement error may explain some of this deviance, we believe that the heart of this puzzle lies in case selection. (Indeed, in the Appendix, we present a "solution" to the noncompliance puzzle that is grounded in case selection.)

III. Data and Methods

How might selection bias affect standard research designs? Normally, a researcher will only be able to estimate fact weights based on the small number of cases the Supreme Court actually hears. Call these the "in-sample" cases and the remainder the "out-of-sample" cases, which are analogous to the large number of cases heard by trial courts and appeals courts that are not heard by the Supreme Court. Using the estimates of fact weights from the in-sample cases, the researcher can draw inferences about (1) the significance of particular case facts and (2) the correct classification of out-of-sample case outcomes (liberal or conservative).

How can we tell whether these inferences are valid? We do so using a series of simulations, employing lower court data as the population of cases from which the Supreme Court might choose. The unique benefit of a simulation is that we will know the truth by construction, and so can evaluate estimates and inferences against the truth.[8]

Our data are drawn from a random sample of published courts of appeals search and seizure cases from 1960 to 1990.[9] Each case is coded for the presence or absence of the relevant case facts or circumstances.[10] Note that our results do not depend on any meaningful substantive interpretation of these variables (see Friedman 2006) and, indeed, we could have chosen to use simulated data to the same effect. Rather, we use actual data to ground our analysis in the context of a particular area of the law, and

[8] Simulations are often used by statisticians and social scientists to check the bias and accuracy of estimates from empirical models or to construct certain estimates. Simulation techniques generally have two features. The first is setting up a replica or model capturing how we think the real world works, reducing it to its key moving parts. We may be able to manipulate various moving parts of this replica in ways we could not in the "real world" (or outside an experimental setting). The second feature is repetition, particularly for models with a random element, so we can know how variable or reliable our observations are. To give a basic example, consider the problem of estimating the probability of the result of rolling a 7 with two, fair six-sided dice. One way would be to calculate the probability analytically, taking the possible outcomes that result in 7 divided by the total number of outcomes. Another way would be to simulate the rolling of two dice, perhaps at a craps table or on a computer, and simply count the proportion of rolls that sum to 7. Given enough simulations, this method produces an unbiased estimate of the probability. Or, suppose we wanted to know whether a pair of dice are weighted. We could roll them many times to compare the results to what we know should be true analytically or to our observations of rolling dice we know to be fair. The latter is, in effect, what we are doing in this article. See Robert and Casella (2005) for a general overview of simulation-based methods.

[9] These data were originally collected and presented Songer et al., and were subsequently analyzed in Songer et al. (1995), Cameron et al. (2000), and Klein and Hume (2003). See Segal (1984) for a complete description of the variables we employ.

[10] As noted above, some of the case variables might properly be labeled "legal judgments." In addition, one of the variables we use, MEMBERSHIP CHANGE, is not a case fact at all but a measure of judicial ideology.

in known case data, to avoid any concerns that we have cherry-picked cases or rigged any particular correlations between variables in order to produce our results.

Since this is only a random subset of a larger pool from which the Supreme Court might draw in reality, we append 10 copies of this random sample together. This leaves us with a "universe" of 12,610 cases from which to sample—a procedure that means we are effectively sampling with replacement. If we did not increase the base set of cases (i.e., if we sampled *without* replacement), the selected cases in our analyses (see below) would have constituted up to 25 percent of the universe, which, of course, is a much larger percentage than the Supreme Court hears in *any* issue area. Thus, using a larger data set more closely parallels actual research designs based on the Court's decisions. This procedure does not change any substantive findings, since these cases are already a random sample, and this preserves all correlations between variables that already existed in the sample.

We proceed as follows. We first assign the Supreme Court's "true" fact weights, which capture how the Supreme Court wants these facts treated, in its own decisions and in lower court decisions. Although any set of weights (within reason) would be sufficient for our purposes, we connect our simulation to actual Fourth Amendment cases by applying (with two exceptions) the weights obtained from a logit analysis of the Supreme Court's search and seizure decisions from the 1962 to 1998 terms.[11] Note that we strongly suspect these are not accurate measures of the Court's true weights because they were affected by selection bias. We use these particular weights, however, so we can ask the question: If they *were* the true weights, would we detect them correctly? An alternative strategy would be to assume a different set (or sets) of weights, but these are thought to be plausible and so seem a good starting point.

These fact weights are then applied to the universe of cases to define the true probability of a conservative outcome in each case. This probability

[11] The weights employed were: 1.97 – 2.77 × home – 2.34 × business – 1.94 × person – 1.94 × car 1.53 × extent of search +1.73 × warrant +2.89 × incident to arrest +1.03 × after lawful arrest +1.37 × exception to warrant requirement +0.29 × membership change. For descriptions of each, see Segal and Spaeth (2002:Ch. 8). The two exceptions are that, for simplicity, we set the weights for "after unlawful arrest" and "probable cause" to 0 (these were insignificant in Segal and Spaeth's analysis).

is obtained by taking the inverse logit of the result of the fact weights applied to each case fact.[12]

For each case, we draw an outcome based on its probability (e.g., a case with a predicted probability of 73 percent still had a 27 percent chance of receiving a classification of 0, where 0 designates a liberal classification and 1 a conservative classification). These outcomes define the true outcomes.[13]

We then devised a series of case-selection rules, which are described in detail below. The resulting set of cases—the in-sample cases—is the very set from which the researcher would seek to draw general inferences about fact weights and about out-of-sample cases. As such, the next step was to investigate how the selection of the Supreme Court's docket affects our inferences. For each simulation, we ran a logit of the in-sample case outcomes on the case facts and recorded the estimated fact weights and other results. We then compare the estimated influence of case facts to the true influences. Finally, we use the estimated fact weights from the in-sample cases to generate out-of-sample predictions, which are compared to the true outcomes for those cases.

A. Simulation Summary

There are three main steps to our simulations.

1. Define the true parameters and outcomes.
   a. Define a universe of cases and assume a set of fact weights. These are the true weights.
   b. Apply these assumed weights to define the true probability of a conservative outcome for each case in the universe.
   c. For each case, randomly draw the true outcome given this probability.[14]

[12] The inverse logit function $\left(\log\mathrm{it}^{-1}(x)=\frac{e(x)}{1+e(x)}\right)$ transforms continuous values to the (0,1) range (Gelman & Hill 2007:80). For example, if the fact weights applied to the facts in a particular case sum to 2.3, the probability of a conservative outcome in that case = $\mathrm{logit}^{-1}(2.3) = 0.9$.

[13] This procedure resulted in conservative outcomes roughly 75 percent of the time, which is similar to the proportion of conservative outcomes in lower court cases in Songer et al. (1994).

[14] Simply converting, for example, each case with an 80 percent probability of a 1 to a 1 would be incorrect, as a probability of 80 percent means that there is a 20 percent chance the case would receive a 0.

2. Estimate parameters and outcomes from a case sample.
    a. Take a sample of cases, based on specified selection criteria (discussed in detail below).
    b. Estimate fact weights, using only this sample, by running a logit of the true outcome (the dependent variable) on the case facts (the independent variables).
    c. Use these in-sample estimated weights to classify the outcomes of all out-of-sample cases. Call these the "estimated" outcomes.[15]
3. Compare our estimations to the truth.
    a. Compare the estimated fact weights to true weights.
    b. Compare the estimated outcomes to the true outcomes. (If these match, then using the in-sample cases to form weights did not lead to a classification error. If they do not match, then using the in-sample cases biased our estimation of the preferred case outcome.)

For each selection strategy, we repeat each simulation 1,000 times.

B. Selection Strategies

The strategies that we study are not intended to exhaust all possible ways the Supreme Court may choose cases. However, they do represent a variety of strategy types, from error correction to ideological behavior to legal development.

We consider three main types of strategies. First, we look at "close cases." These are cases that could go either way, defined as cases where the latent probability of being conservative, according to the true weights, is close to 50 percent. These can be thought of as hard cases, given the close calls involved. They might be more prone to error or even manipulation by lower courts. Such cases might also lead to circuit splits, thereby increasing the probability that the Supreme Court would grant cert (Perry 1991). We break down this analysis further by whether the Supreme Court also selects on the basis of the direction of the outcome, taking only those close cases in which the desired outcome is liberal (as a liberal Court might) or only those with conservative outcomes.

[15] We maximize prediction success by following the standard convention of assigning conservative outcomes to all cases with predicted probabilities greater than or equal to 0.5, and liberal outcomes otherwise.

Second, we look at "anomalous cases," those that look like they should be decided in the conservative direction, but in which the true decision is nonetheless liberal, and vice versa. In contrast to close cases, these are cases that seem clear-cut because many factors in the case push in one direction, so that the latent probability is very high or very low. The catch is that the outcome does not match the apparent predisposition of the case. That is, cases are taken where the latent probability (given the standard case facts) is low, suggesting a clear liberal outcome, but the true outcome is nevertheless conservative (as will happen a nonnegligible percentage of the time), or vice versa (where the latent probability is high but the true outcome is nonetheless liberal). These are cases not covered well by existing doctrine and the standard facts. Therefore, they are unlikely to be decided correctly by lower courts and are perhaps the very cases the Supreme Court is most likely to take (to address new issues not covered by existing doctrine). Like close cases, we break this strategy down by whether the Court seeks such cases with liberal or conservative outcomes. We would expect the selection of anomalous cases to yield worse inferences than close cases.

Finally, we present an example of a different type of strategy, where the Court selects on the basis of specific case facts. This selection strategy oversamples cases in which the search took place in a home, no exceptions were present, and the true outcome was conservative.

In each simulation, we draw 300 cases into the in-sample. Of these, roughly 40 percent (or 120) are drawn according to the given selection strategy, and roughly 60 percent (or 180) are drawn from a random sample of the universe of cases, for a total of 300. Including a subset of random cases has several benefits. First, it biases our results *against* showing selection effects. Second, it also mitigates the problem of complete or quasi-complete separation in the logit estimations due to perfect or near-perfect prediction among certain independent variables and the dependent variable (Zorn 2005).[16] Finally, it reflects the fact that not all cases will be chosen using a single selection strategy or for the same reason. Including a significant

[16] Separation nevertheless occurs in a small percentage of simulations, resulting in those observations being dropped from the logit estimation. Table 1 presents information on the distribution of random cases and selected cases across selection strategies that remain in the logit estimations. As Table 1 demonstrates, most of the dropped observations are selected cases, which increases the percentage of random cases in the in-sample, thereby biasing against finding selection effects. Our results are fully robust to including only simulations where all 300 observations enter the logit estimation.

number of random cases is similar to assuming the Court is using a mixture of strategies for some portion of its docket.[17]

In addition to these nonrandom strategies, we also provide two baselines for judging the impact of case selection, running our simulations on the full universe of cases ("all cases") and on a random selection of cases ("random selection"). As Priest and Klein (1984:1) note, "if all legal disputes, or even a random sample of these disputes, were tried to judgment and then appealed, the inferences from legal rules to social behavior would be straightforward." Altogether, we compare the results of the "all cases" base line and eight case-selection methods, including "random selection." Table 1 provides detailed information on each selection strategy.

We can now turn to our analysis of substantive issues using these data and methods, which we break down into inferences about the effects of fact weights on case outcomes and about case outcomes directly.

IV. Results and Discussion

A. Inferences About Case Facts

1. Will We Correctly Identify Significant Case Facts?

That is, using the logit results from only the in-sample cases, are the significant case facts actually found to be significant? For each of the selection strategies and for nine case facts, Figure 2 shows the percentage of simulations in which we would conclude that a given case fact indeed has a statistically significant coefficient (at the 95 percent confidence level) and is in the same direction as the true value.[18] As the top bar for each fact shows, when all cases are included, each fact is significant in nearly every simulation. Thus, if the in-sample cases were leading to valid inferences and our conclusions as to case facts were not affected by selection bias, each fact should be statistically significant in nearly 100 percent of simulations.

[17] One way to extend our analysis would be to see how the problems induced by selection bias increase as the percent of selected cases in the in-sample increases.

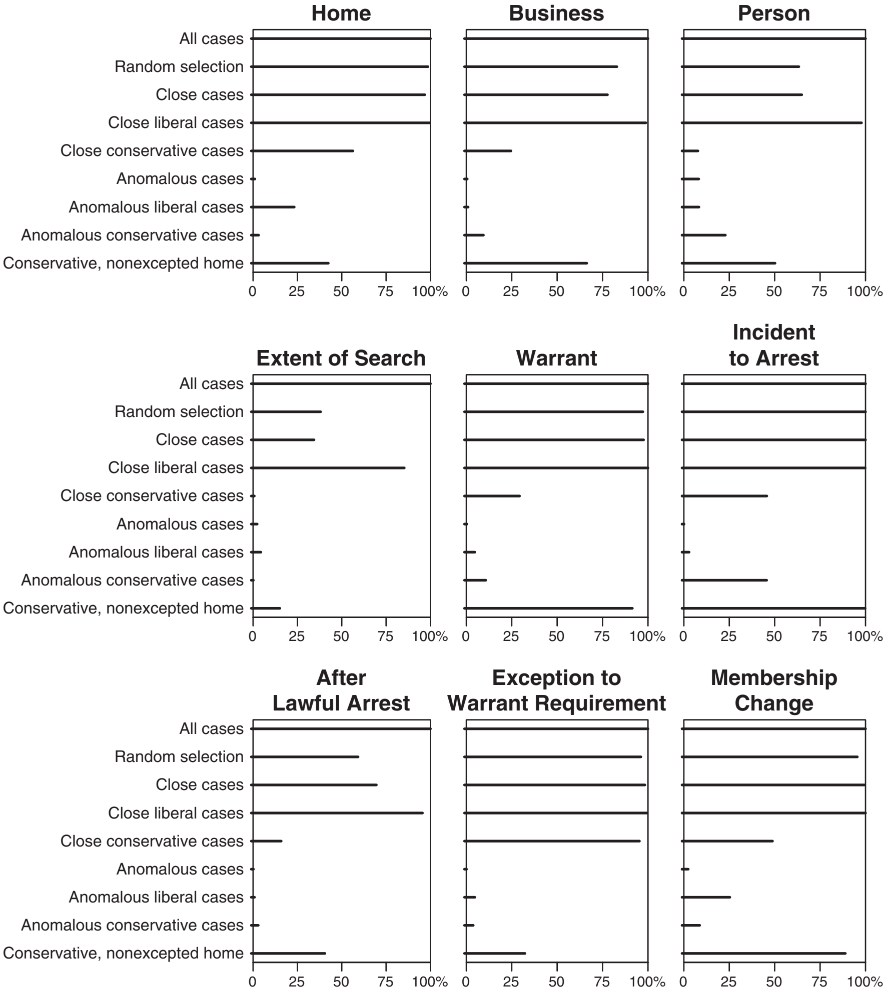
[18] Following the advice of Gelman et al. (2002), Epstein et al. (2006, 2007), and Kastellec and Leoni (2007), we present the majority of results graphically rather than in tabular form.

Table 1: Selection Criteria for Case-Selection Strategies

Selection Strategy	Selection Criteria (Latent probability range)	Outcome	Random Cases				Selected Cases				Total Cases			
			Min	Mean	Median	Max	Min	Mean	Median	Max	Min	Mean	Median	Max
Random	—	—	273	298	300	300	—	—		—	273	298	300	300
Close	0.45–0.55	—	120	179	180	180	112	118	118	120	240	297	298	300
Close liberal	0.45–0.55	Liberal	106	178	180	180	104	117	118	120	221	295	298	300
Close conservative	0.45–0.55	Conservative	100	176	180	180	75	116	117	120	211	292	297	300
Anomalous	0–0.36, 0.87–1	—	173	180	180	180	109	118	118	120	287	298	298	300
Anomalously liberal	0.87–1	Liberal	175	179	180	180	113	118	118	120	291	298	298	300
Anomalously conservative	0–0.36	Conservative	107	175	180	180	104	117	118	120	212	292	297	300
Conservative, non-excepted home	—	Conservative	113	176	180	180	63	117	118	120	183	292	297	300

Note: For each selection strategy, the table lists the selection criteria for the nonrandomly drawn cases (where applicable); the outcome selection criteria (where applicable); and summary statistics on the number of random cases and selected cases in each simulation that enter the logit estimation, as well as the sum of these two (total cases). The probability of selection is uniform across the latent probability range. The selection criteria for anomalous cases are not symmetric because the distribution of the latent probability range is skewed toward 1. Hence, it is necessary to expand the range to get a sufficient number of anomalous conservative cases. Finally, the number of cases varies across simulations due to some observations not being included in the logit estimation due to complete or quasi-complete separation. Our results are fully robust to including only simulations where all 300 observations enter the logit estimation.

Figure 2: Significance of case facts across selection strategies.



NOTE: For each fact and each selection strategy, the graphs show the percentage of simulations in which the logit of the in-sample cases correctly reveals the case fact to be statistically significant and in the same direction as the true value. Ideally, the bars would go all the way across the plots to 100 percent, as they do for all case facts under "all cases" and for most case facts under "random selection."

Figure 2 instead reveals wide variation both across case facts and selection strategies in the percent of simulations in which one would make the correct inferences about the statistical significance of case facts. Not surprisingly, the "random selection" strategy does very well, as five facts—HOME, WARRANT, INCIDENT TO ARREST, EXCEPTION TO WARRANT REQUIREMENT, and MEMBERSHIP CHANGE—are significant in nearly 100 percent of simulations. Even in random samples, however, the remaining facts are often found to be insignificant, suggesting that the best-case scenario for drawing inferences may nevertheless prove problematic. This suggests that statistical fact-pattern analyses reduce doctrine to the facts with the largest impact—which is fine so long as we recognize that we are constructing an incomplete model of the more nuanced doctrine the Court may be applying.

Once we move to the nonrandom selection strategies, the inferential problems become stark. For EXTENT, for example, significance is never found more than 85 percent of the time, across selection strategies; excluding "close liberal cases," significance is found in no more than 38 percent of simulations. Moreover, for the last five selection strategies, significance is almost never found. Note that certain selection strategies are more likely to lead to incorrect inferences. "Anomalous liberal cases" is a prime example of this. "Close cases" leads to fewer inferential errors, but still remains far from the ideal.

Thus, we conclude that correctly identifying significant case facts depends heavily on how cases are selected.[19] Solely because of the Supreme Court's selection strategy, an analyst would reject a case fact as insignificant that is strongly signifiicant in terms of the Court's preferred fact weights. Indeed, one might even falsely conclude that a given area of law is chaotic or unprincipled as a whole.

We also note that the inferential problem extends to analyzing ideological change on the Court. For several selection strategies, the coefficient on MEMBERSHIP CHANGE is insignificant in a large percentage of simulations. Thus, it is possible that selection could mask the extent or direction of ideological change on the Supreme Court, or the role that ideology plays. (We explore ideology in greater detail in the case outcome section.)

[19] In addition, just as facts we know to be significant may appear insignificant due to case selection, it is possible that insignificant facts could appear significant if correlated with the probability of being selected or with other facts that are significant yet omitted.

2. How Often Will We Falsely Reject the True Effect of a Case Fact?

Another question of interest is how close our estimates of fact weights (from the in-sample cases) will be to the true weights. More precisely, given an estimated fact weight and its standard error, how often would one reject the true weight (at the 95 percent confidence level)?[20]

Figure 3 answers this question by displaying the percentage of simulations in which one would falsely reject the true weight, across facts and across strategies. Ideally, we should see this happening only rarely—and as Figure 3 shows, when all cases are used in the estimation the truth is falsely rejected in only about 5 percent of simulations. However, when we turn to the various case-selection strategies, the true fact weights are falsely rejected across a large percentage of simulations, particularly for the "anomalous" and "anomalous liberal" cases. For INCIDENT TO ARREST, for example, if the Court were employing either of these two selection strategies we would falsely reject the true weight in 100 percent of simulations. Similar patterns emerge for the other eight case facts. Thus, case selection severely complicates efforts to infer true fact weights.

3. How Much Will the Estimated Impact of a Case Fact be Affected by Case Selection?

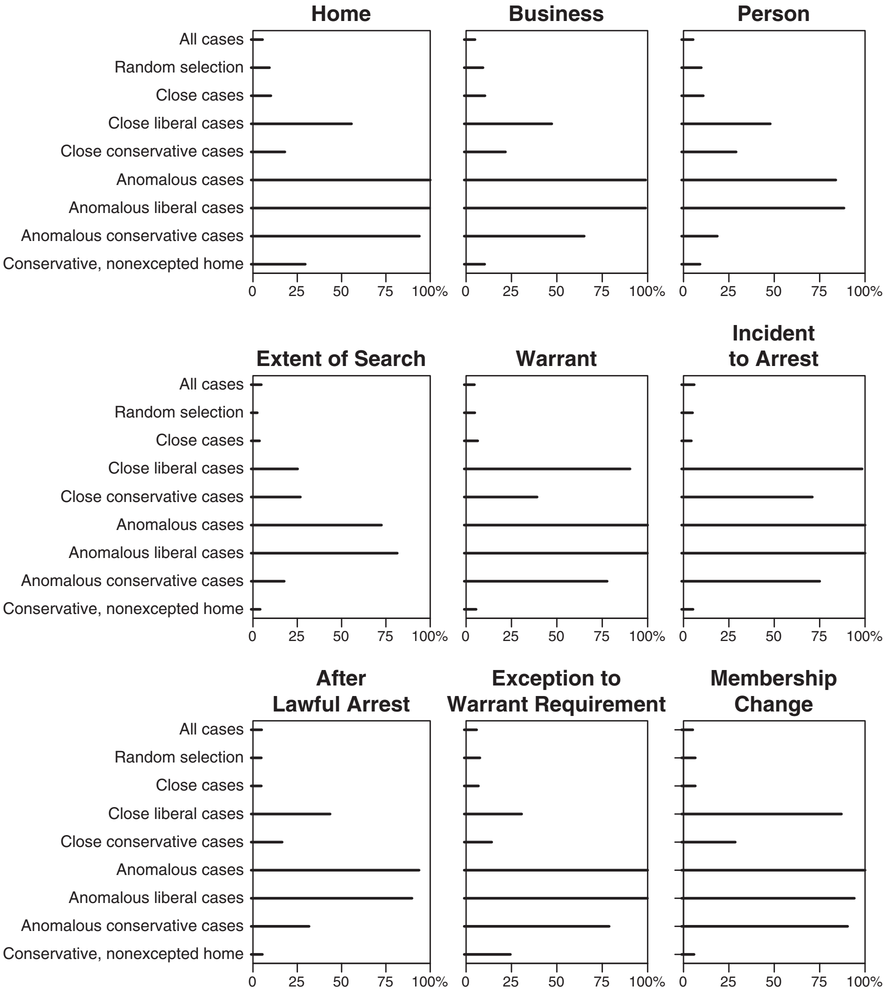
In addition to correctly identifying the true fact weights and their significance, researchers are also (if not more) concerned with estimating substantive significance: how much their presence or absence increases the probability of a conservative decision. Although it is difficult to directly interpret logit coefficients, their magnitude gives a rough sense of how influential each fact weight can be.[21]

Figure 4 illustrates how different selection strategies can lead to wide variation in coefficients. For all 12 case facts and selection strategies, each point depicts the median coefficient across the set of simulations, while the horizontal lines depict the median coefficient plus and minus the median

[20] Note that we might not reject the true weight for two reasons: because we found the right weight or because the variance is simply too high to reject weights (creating so large a confidence interval that it includes the truth). Thus, failing to falsely reject the true fact weight can be a positive or negative event; falsely rejecting truth, however, is clearly problematic.

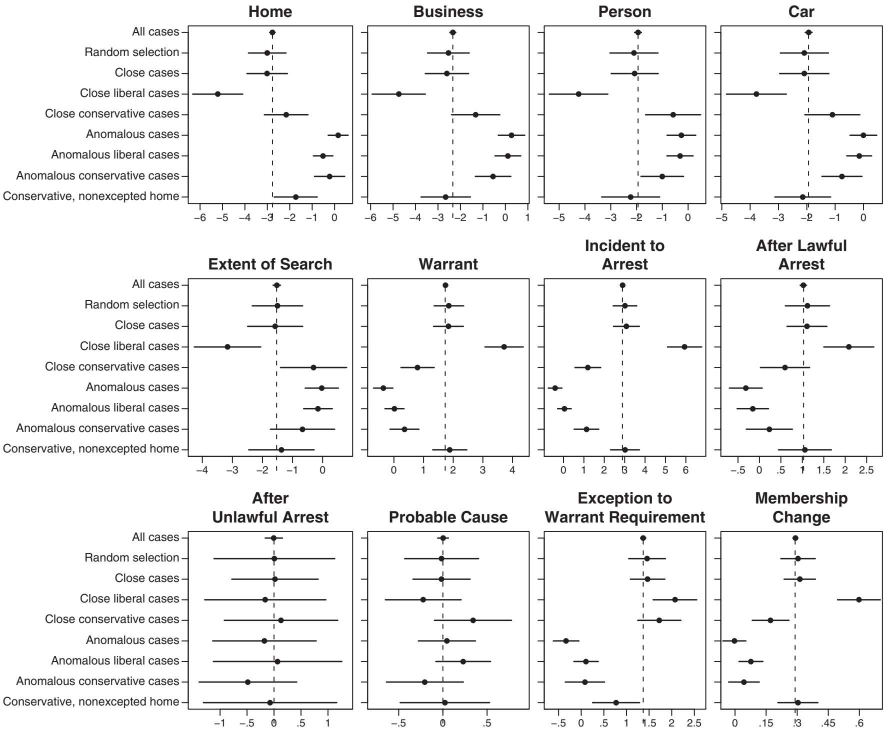
[21] For instance, one can use the "divide-by-four" rule of thumb: dividing a logit coefficient by four gives an upper bound on its probabilistic effect on the dependent variable given a one-unit change in the relevant independent variable (Gelman & Hill 2007:82).

Figure 3: False rejection of true fact weights across selection strategies.



Note: For each fact and each selection strategy, the graphs show the percentage of simulations in which the logit of the in-sample cases would lead to a false rejection of the true fact weight. Ideally, and in contrast to Figure 2, the bars would all be close to 0 percent.

Figure 4: Distribution of fact weights across selection strategies.



NOTE: For each fact and each selection strategy, the points show the median coefficient from each set of simulations, while the horizontal lines depict the median coefficient plus and minus the median standard error from each set of simulations (note that these are not confidence intervals). The vertical dotted lines depict the median coefficient from "all cases" for comparison.

standard error from each set of simulations. Note that these are *not* confidence intervals, and that the goal is not to evaluate when fact weights will be significantly different from "all cases" or "random selection," as we did in Figure 3, but to examine the spread of coefficients across selection strategies.

Whereas our estimates for "all cases" are quite precise, the estimates vary quite a bit within and across selection strategies. Looking at HOME, for example, the median coefficient ranges from a minimum of –5.2 for "close liberal cases" to a maximum of 0.2 for "anomalous cases." Likewise, for EXCEPTION TO WARRANT REQUIREMENT, the median coefficient ranges from a minimum of –0.3 for "anomalous cases" to a maximum of 2.1 for "close liberal cases." Clearly, selection bias will affect our estimates of fact weights.

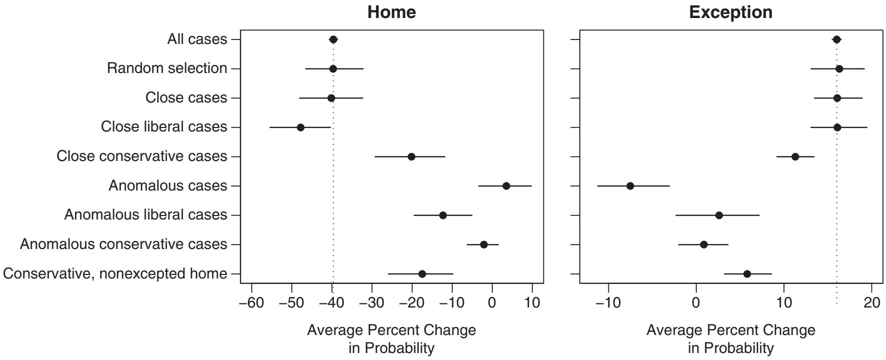
Examining the standard errors across facts and strategies also reveals a disturbing picture. The very selection strategies that yield more precise estimates are the ones for which we obtained (on average) more misleading values of the coefficients. Examining WARRANT, for instance, we see that the smallest standard errors are for the "anomalous" and "anomalous liberal cases," for which the coefficients are farthest from the true fact weight. Thus, it appears that selection bias not only can give misleading estimates, but it may also give one greater false confidence in such estimates.

Another way to explore these issues, given the difficulty of interpreting logit coefficients directly, is to look at the probabilistic effect on outcomes of changing one variable, holding the others constant (see, e.g., Segal & Spaeth 2002:312). We do this for two variables. For each case, we calculate the average predictive change in the predicted probability of a conservative outcome if HOME is set to 0, as compared to the predicted probability if HOME is set to 1, with all other case facts held at their actual values within each case (Gelman & Pardoe 2007). We then average these changes over the entire set of cases to provide an intuitive measure of the average effect of HOME. We then explore how this substantive effect varies within and across strategies. We do the same for EXCEPTION TO WARRANT REQUIREMENT. Figure 5 depicts the results of this procedure. The points show the median change in probability from each set of simulations. The horizontal lines show the range from the 25th percentile value to the 75th percentile value.

From "all cases," we learn that the change in probability for HOME should be roughly –40 percent; that is, a search that takes place in the home on average decreases the probability of upholding a search by 40 percent. Case selection, however, would cause us to infer very different effects. Furthermore, there is wide variation within and across selection strategies. For HOME, the median change in probability ranges from –49 percent for "close liberal cases" to 3.5 percent for the strategy of "anomalous cases."

The results are similar for EXCEPTION TO WARRANT REQUIREMENT. For each nonrandom selection strategy except "close cases" and "close liberal cases," the estimated change in probability is far removed from the actual change in probability seen in "all cases" of 16 percent. In fact, for "anomalous cases," a researcher would actually estimate the change in probability to be negative. Thus, we conclude that selection bias can greatly influence our estimates of the substantive impact of case facts.

Figure 5: Impact of fact weights across selection strategies.



NOTE: The left-hand graph displays the average change in predicted probability of a conservative outcome for a search conducted in a home compared to one not in a home; the right-hand graph displays the average change in predicted probability for a search conducted in which an exception to the warrant requirement was present compared to one with no exception present. The points show the median average change in probability from each set of simulations, while the horizontal lines connect the 25th percentile value of the average change to the 75th percentile. The vertical dotted lines depict the median average change from "all cases" for comparison.

4. Can We Compare Fact Weights Across Periods of Time?

Recall that the jurisprudential regime research design relies on (1) whether specific facts are significant before and after a regime break and (2) statistically significant changes in fact coefficients before and after this regime break. Richards and Kritzer attribute these differences to the effect of the new regime on the Justices' decision making. If a fact weight is insignificant before a regime break but significant after, that is said to demonstrate the influence of precedent.

This research design depends on the ability to estimate fact weights reliably both within and across two different periods of time. However, as our results demonstrate, it is very difficult to trust either the substantive or statistical significance of fact weight estimations in a *single period* of Supreme Court decision making. It is thus difficult to be confident in a delicate comparison of estimates *across* two periods (or two samples), as is required by the jurisprudential regimes design. Even if we could reliably estimate fact weights within a single period, to compare them across time we would have to assume that the selection strategy employed by the Justices remained precisely constant as to all case facts over time. There is little reason to believe that this assumption will be met.[22] Since selection bias would lead to differences in fact weights, findings of such differences are called into question.

5. Can We Compare Fact Weights Across Courts?

Another study that our findings potentially implicate is Songer et al. (1994), which compares estimates of fact weights from the courts of appeals and from the Supreme Court to assess how closely the former set of weights matches the latter set of weights (i.e., how compliant the lower courts are). Finding that most weights are statistically significant in the same direction at both levels of the judicial hierarchy, the authors conclude that the judges on the courts of appeals are generally compliant with the Supreme Court. Note first that if we are concerned solely with statistical significance, and given that they find similarities, then selection bias might not undercut the validity of their findings. One might consider it unlikely that selection bias would cause observed fact weights to be significant in the *same* direction when the true fact weights are actually different across levels. If they had found that the fact weights *differed* in significance across courts, then we might be more concerned that selection bias contributed to such differences.

On the other hand, the authors give little attention to the fact that many of the estimated fact weights differ greatly in *magnitude* across the two levels, with larger coefficients seen for the Supreme Court. If we are to take estimates of fact weights seriously, then indeed these differences call into question a conclusion of compliance—compliance requires not only that a case fact significantly changes the probability of a liberal outcome, but that the change in probability is roughly comparable across courts. For example, based on the estimates presented in Songer et al. (1994:682–83), at the Supreme Court level, searching a person's home increases the probability of a liberal decision in a borderline case by roughly 88 percent, compared to only 22 percent at the court of appeals level—a rather large difference in the treatment of this case fact. Indeed, Table 2 in their paper shows rather disparate treatments of similar cases across levels of the judicial hierarchy. Of course, since our results strongly suggest that selection bias taints estimates at

[22] Indeed, if a new jurisprudential regime has been implemented, we would *expect* case selection to change as well. Moreover, if litigants respond to key Supreme Court decisions and change their decisions on whether to appeal accordingly, the sample of cases will likely change over time, even if the selection strategy does not. This would have the same effect on inference.

the Supreme Court level, it is difficult to draw firm conclusions from this research design.

The findings of a later work, Cameron et al. (2000), rely on ordering lower court cases by their overall "intrusiveness," based on the Supreme Court's estimated fact weights. It is then shown that the Court's cert decisions are based on this intrusiveness measure and the interaction between lower court ideology and the direction of the lower court's decision. Thus, one of their key independent variables likely suffers from measurement error induced by selection bias, which might cast doubt on their causal claims (Greene 2003:84–86).

B. Inferences About Case Outcomes

The analysis so far has focused on the impact of specific case facts on case outcomes. We now focus on case outcomes directly. We start by looking at aggregated measures of case outcomes and then turn to a case-by-case approach.

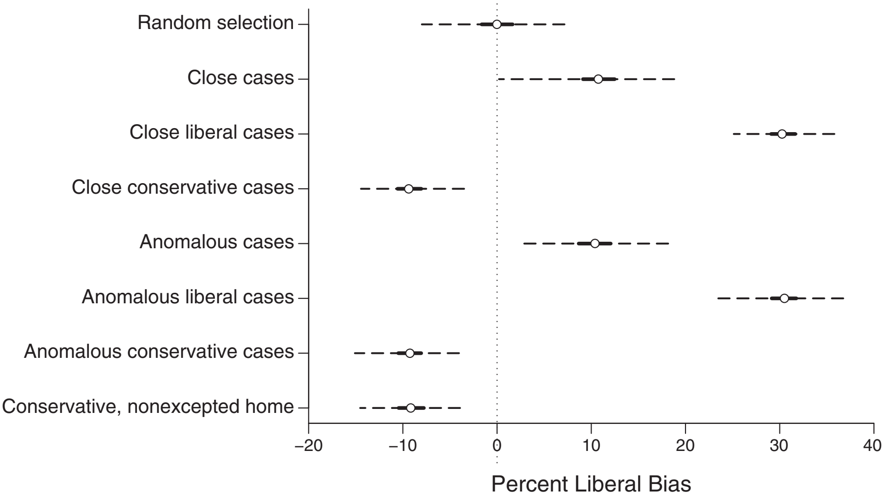
1. Measuring Aggregate Ideology

One nonobvious implication of our results deals with measures of ideological behavior in the Supreme Court. Simple aggregate liberalism scores (e.g., Segal & Spaeth 2002) and more advanced estimates of the ideal points of the Justices (e.g., Martin & Quinn 2002) are based on observed cases. Such scores are used to compare the Court over time, to compare Justices within a Court and over time, and to compare issue areas. How are such measures biased by the particular sample of cases chosen by the Court, given that the measures actually capture an unknown interaction between the cases selected and the decision-making process in those cases?

Consider aggregate liberalism scores, which serve as the foundation for testing the predictions for the attitudinal model (Segal & Spaeth 2002:320–24). These had been used as measures of judicial preferences themselves (e.g., Rohde & Spaeth 1976), and are still used as measures of behavior to be explained by ideology scores (e.g., Segal & Cover 1989). How much is aggregate liberalism biased by the fact that we only observe a nonrandom sample of cases? That is, how much higher or lower is the percentage of liberal decisions in the chosen sample of cases than in the universe of cases?

Figure 6 shows the answer. For each selection strategy, we plot the median "percent liberal bias"—the difference between the percentage of liberal outcomes in the in-sample cases and the percentage of liberal

Figure 6: Bias in measuring aggregate ideology.



NOTE: The graph summarizes the distribution of bias created by observing only the in-sample case outcomes. "Percent liberal bias" is equal to the percentage of liberal outcomes in the in-sample cases minus the percentage of liberal outcomes in the entire universe of cases, that is, in-sample and out-of-sample cases combined. For each selection strategy, the points depict the median bias within each set of simulations; the solid horizontal lines connect the 25th percentile of bias to the 75th percentile, while the dotted horizontal lines connect the minimum bias to the maximum.

outcomes in the entire universe of cases. (We also plot the 25th and 75th percentile biases, and the minimum and maximum biases.) Obviously, a random sample does well—the median bias is 0, as one would expect. Other selection strategies, however, lead to liberalism rates that differ a great deal from the true liberalism of the Court. For "close cases," which does well on other criteria, the median estimated liberalism rating is 11 percentage points higher than the true percentage of roughly 75 percent (the maximum bias is +21 percent). If the Court takes either "close liberal cases," or "anomalous liberal" cases, the median bias is even higher, at +30 percent.

For three of the nonrandom selection strategies, the median bias is negative; for the remainder it is positive. For the liberal and conservative selection strategies, the bias is in the direction one might expect—but for the "close all" or "anomalous all" strategies, it also yields a median liberal bias of about 10 percent, even though these strategies are not biased in terms of case outcomes. In addition, we have no guarantee that the same selection strategy is being used across issue areas or over time, so that unless we are willing to assume these are constant, we cannot safely make such comparisons.

It is unclear how much selection bias affects ideal point estimates such as the commonly used Martin-Quinn scores, which are aggregate measures of Justice ideology, given that they allow ideal points and case-location estimates to vary over time (Martin & Quinn 2002; Epstein et al. 2007b, 2007c). To the extent they do not completely account for such dynamics, changes in case selection might bias ideal point estimates or make it seem like preferences had changed over time (Farnsworth 2007).

2. Measuring Compliance: How Often Do Misclassifications Occur?

When we generalize from the Court's decisions to study its relationships with other courts and actors in the legal system, can we have confidence in our conclusions if we ignore case selection? In particular, our research strategy allows us to assess whether we can use the cases the Court hears to make claims about cases it does not hear, specifically to study compliance in such cases. As noted above, in each simulation, we generated a predicted outcome in the out-of-sample cases based on a logit analysis of the in-sample cases. Of course, we know the true outcome by construction.

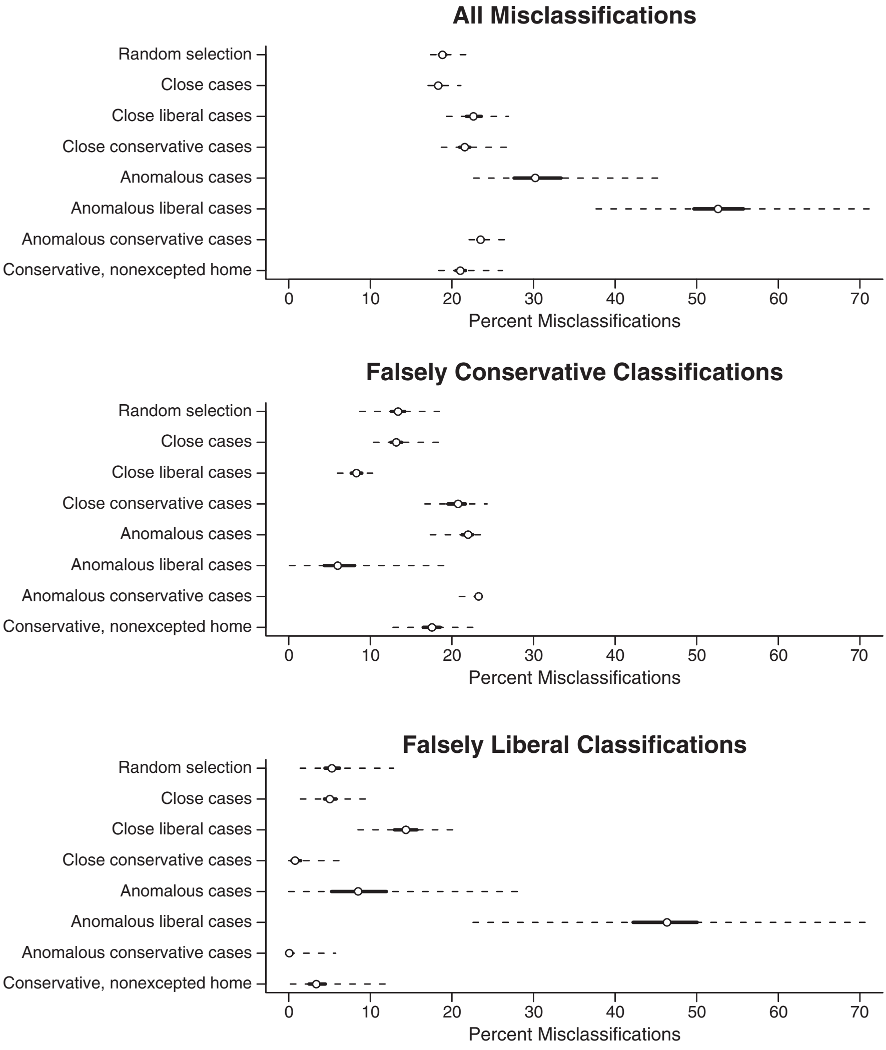
We begin our analysis in this section by analyzing the percentage of out-of-sample cases that are incorrectly classified according to the fact weights generated by the logit analyses of the in-sample cases. That is, we assume that lower courts are perfectly compliant with those outcomes truly preferred by the Supreme Court, and ask how many decisions we would label as noncompliant due to measurement error or selection bias.

The top graph in Figure 7 depicts the distribution of out-of-sample misclassifications across selection strategies.[23] The points depict the median percentage of cases that are misclassified within each set of simulations; the solid horizontal lines connect the 25th percentile of misclassifications to the 75th percentile, while the dotted horizontal lines connect the minimum percent of misclassifications to the maximum.

The percent of misclassifications in the "random selection" simulations provides a good baseline for assessing the other strategies. Although under this selection strategy we were generally able to recover the true fact weights, it nevertheless generates classification errors, on average, in about 19 percent of cases. The relatively high rate of misclassifications here is perhaps initially surprising, but it is less so upon closer reflection. Each case has only an underlying *probability* of a liberal or conservative outcome, so that a case,

[23] In this section, we exclude "all cases," since this leaves no cases for out-of-sample prediction.

Figure 7: Classification of out-of-sample cases based on in-sample logit analyses.



NOTE: The graphs respectively summarize the percentage of cases that are incorrectly classified in either the liberal or conservative direction, the percentage of truly liberal cases that are classified as conservative, and the percentage of truly conservative cases that are classified as liberal. For each, the points depict the median percent of cases that are misclassified within each set of simulations; the solid horizontal lines connect the 25th percentile of misclassifications to the 75th percentile, while the dotted horizontal lines connect the minimum percent of misclassifications to the maximum.

for example, that has an 80 percent probability of being conservative still has a 20 percent chance of being liberal. Thus, while the modal prediction of "conservative" remains the best prediction in each case, it will still lead you astray 20 percent of the time. That is, even for random selection, there will be a nonnegligible degree of measurement error, which should be taken into account in drawing inferences from such compliance studies. It is worth noting that the degree of misclassification is roughly equal to the noncompliance level found by Klein and Hume (2003) and is perhaps greater than the overall amount of noncompliance that previous studies have found to exist among lower federal courts (e.g., Gruhl 1980; Songer & Sheehan 1990).

Moving to the nonrandom strategies, we see that selection bias compounds this problem—the degree of misclassification now ranges from significant to severe. The "close cases" strategy results on average in about an 18 percent misclassification rate; that figure increases slightly when the Court keys on a particular type of case. A severe number of misclassifications occur when we turn to "anomalous cases" and "anomalous liberal cases." In the latter, about 53 percent of cases are incorrectly classified on average; in the former, approximately 30 percent are classified incorrectly.

In sum, depending on the Court's selection strategy, our inferences about the true decision in the out-of-sample cases, according to our estimations from the in-sample cases, are likely to be off in anywhere from 18 to 72 percent of cases. This is all the more troubling in that simply guessing a conservative outcome for each case would itself lead to only a 25 percent misclassification rate. Even if lower courts were perfectly compliant with Supreme Court preferences, we could still find high rates of noncompliance.

If lower courts were noncompliant in some of their decisions, we might falsely label many of these as compliant. Even if we ignore the misclassification of individual cases, we would still not know how these two types of error would balance out, and it would seem coincidental at best if they exactly cancelled out in aggregate compliance measures.

3. The Direction of Misclassifications

We can further break down these results by examining the direction of misclassification. The second graph in Figure 7 depicts the percentage of out-of-sample outcomes incorrectly classified as conservative (i.e., predicted conservative outcomes when the true outcome is liberal), while the third graph in Figure 7 depicts the percentage of out-of-sample outcomes incorrectly classified as liberal (i.e., predicted liberal outcomes when the true outcome is conservative).[24] Given that the true outcome is liberal in only about 25 percent of all cases (on average), the median rate of cases incorrectly classified as conservative is rather high, 6 to 23 percent across selection strategies. When we turn to the rate of cases incorrectly classified as liberal, the error rate can be as high as 71 percent; the median rate ranges from 0 to 46 percent. We emphasize that random measurement error (the occurrence of which might not create selection bias, depending on a given analysis) would not create an error rate of this magnitude, given our findings.

Once again, the larger point is clear: the Court's nonrandom docket significantly interferes with our ability to classify lower court decisions as compliant or noncompliant, based on the Court's estimated fact weights, where the estimation comes solely from the nonrandom cases that it chooses to hear. Moreover, the errors are not randomly distributed across the ideological spectrum, but can be biased in one particular direction depending on the Court's specific selection strategy. This suggests that in an area of the law where most outcomes are properly in the conservative (liberal) direction, when the Supreme Court uses its docket to carve out exceptions, our estimates will be biased in the liberal (conservative) direction.

V. Conclusion

A discretionary docket may be really good for the Supreme Court, but our results show that it can be really bad for scholars of judicial politics. To be sure, scholars *can* reliably draw conclusions within a single set of cases, so long as it is recognized that any effects of case factors so measured cannot be said to represent general preferences of the judges over all cases or as applied to other sets of cases. In addition, research designs in which selection bias cuts *against* finding particular results may also be acceptable, so long as we do not overstate null findings (Gill 1999). We demonstrate, however, that we cannot simply take the Supreme Court's docket of cases, apply fact-pattern analysis, and draw fully reliable conclusions about compliance, doctrinal substance, changes in the law, or the effects of ideology. Thus, our findings have implications not only for the study of compliance and jurisprudential regimes, which we have addressed in detail here, but for

[24] These are analogous to "false positive" and "false negative" errors.

numerous lines of inquiry in the study of law and courts. We surely are not attacking any scholars in particular: given that case selection is almost always set aside in studying courts, they are in good company. Nor do we want to suggest that our results are the last word as to any debates in these research areas; clearly, the questions of when and why lower courts comply with higher courts and whether and how the law constrains judges are deeply important questions worth further in-depth examination.

Rather, in order to examine selection effects in detail, we have highlighted two specific research designs of growing prominence. Both research agendas—compliance in lower courts and changes in the Supreme Court preferences over time—rely on inferences as to the effects of case facts or findings. Replicating the very research designs used in such work, we show that unless selection is taken into account, comparisons of the effects of case factors over time or over sets of cases may not be measured properly even under optimal conditions, and so inferences as to these effects may not be drawn in practice. As such, there is a large degree of uncertainty around the conclusions drawn from such research designs.

Our hope, however, is that this article represents a first word on this problem rather than the last. Our results are strongly suggestive, but not conclusive, as to the exact magnitude of the problem. We have not tried to give an exact point prediction as to this magnitude but, rather, to show how bad it might be under reasonable assumptions. Future work might explore the contours of selection bias further. For example, whereas we assumed a set of true Supreme Court fact weights and simulated draws of cases, others might simulate varying sets of true fact weights to see what range of true weights are compatible with a given set of observed weights.

We see four potential ways of dealing with the issues we have raised. Perhaps the most promising option would be to model the Court's case selection directly, as studies on certiorari have done (see, e.g., Caldeira & Wright 1988; Caldeira et al. 1999; Brenner et al. 2007). These models could then be incorporated into analyses of the Court's decisions, perhaps using a Heckman model (but note that many scholars have expressed concerns about Heckman models, e.g., Winship & Ware 1992; Liao 1995; Sartori 2003; Simmons & Hopkins 2005).

Given the fact that the Court receives thousands of cert petitions each year, such an endeavor seems daunting. However, if a researcher wants to incorporate the cases the Court hears to study the Court's actual decisions, rather than to study the selection process itself, then for many research designs it will suffice to analyze a random sample of nonselected cases (see, e.g., Owens 2007).[25] For some research questions, choice-based sampling techniques may be used to overcome concerns raised by selecting on the dependent variable (see, e.g., Cameron et al. 2000; George & Solimine 2001).[26]

A second possibility is to change the unit of analysis to avoid the selection bias problem entirely. For example, in their study of whether the Supreme Court is constrained by Congress, Harvey and Friedman (2006) look not only at cases in which the Court decides whether to strike down a law, but at the entire universe of statutes that might potentially be struck down. This cleverly sidesteps the case-selection issue.

Third, where no such options are available, scholars might use our findings to guide their expectations as to the likely direction and magnitude of selection bias. Research designs in which selection bias might undercut a finding will have greater validity than those in which selection bias could lead to a false positive (of course, selection bias can still reduce the power of such tests).

Finally, one way of thinking about the selection bias problem is to note that analyses based only on case outcomes do not make use of all information observable from a sample of cases. Studies that focus more closely on substantive policy content might avoid selection issues. To be sure, this would require reading opinions more closely (as, indeed, lower court judges do when dealing with Supreme Court decisions). Wahlbeck (1997, 1998), for example, studies the expansion and contraction of legal rules, rather than just case outcome patterns. Another example is Songer and Sheehan (1990), who conduct a more nuanced interpretive analysis in their study of lower courts' reactions to *New York Times v. Sullivan* (376 U.S. 254, 1964) and *Miranda v. Arizona* (384 U.S. 386, 1966). A promising new approach is that of McGuire and Vanberg (2005), who score the policy positions of written opinions using the "wordscore" method (Laver et al. 2003).

To be sure, any of these options will require more work than simply setting aside selection issues (and each may raise other issues or concerns).

[25] Indeed, one might do better assuming purposeful noncompliance to be rare and using lower court cases to assess preferred Supreme Court doctrine. Note, however, that we would still want to consider the selection issues raised by studying only those cases that reach the courts of appeals.

[26] For a general review of choice-based sampling, see Manski and Lerman (1977) and King and Zeng (2001).

However, this article has shown that such additional effort may be worthwhile for those who study judicial politics. Indeed, it may be unavoidable.

References

Allee, Todd L., & Paul K. Huth (2006) "Legitimizing Dispute Settlement: International Legal Rulings as Domestic Political Cover," 100(2) *American Political Science Rev.* 219.

Ashworth, Scott, Joshua D. Clinton, Adam Meirowitz, & Kris Ramsay (2008) "Design, Inference, and the Strategic Logic of Suicide Terrorism," 102(2) *American Political Science Rev.* 269.

Benesh, Sara C. (2002) *The U.S. Court of Appeals and the Law of Confessions: Perspectives on the Hierarchy of Justice.* New York: LFB Scholarly Publishing.

Bergara, Mario, Barak Richman, & Pablo T. Spiller (2003) "Modeling Supreme Court Strategic Decision Making: The Congressional Constraint, Volume 28, Number 2, May 2003, pp. 247–280(34)," 28(2) *Legislative Studies Q.* 247.

Brenner, Saul, Joseph Whitmeyer, & Harold Spaeth (2007) "The Outcome-Prediction Strategy in Cases Denied Certiorari by the U.S. Supreme Court," 130(1) *Public Choice* 125.

Caldeira, Gregory A., & John R. Wright (1988) "Organized Interests and Agenda Setting in the U.S. Supreme Court," 82(4) *American Political Science Rev.* 1109.

Caldeira, Gregory A., John R. Wright, & Christopher Zorn (1999) "Sophisticated Voting and Gate-Keeping in the Supreme Court," 15(3) *J. of Law Economics & Organization* 549.

Cameron, Charles M., Jeffrey A. Segal, & Donald R. Songer (2000) "Strategic Auditing in a Political Hierarchy: An Informational Model of the Supreme Court's Certiorari Decisions," 94(1) *American Political Science Rev.* 101.

Clermont, Kevin M., & Theodore Eisenberg (2001) "Appeal from Jury or Judge Trial: Defendants' Advantage," 3(1) *American Law & Economics Rev.* 125.

Cross, Frank B. (1997) "Political Science and the New Legal Realism: A Case of Unfortunate Interdiscplinary Ignorance," 92(1) *Northwestern Univ. Law Rev.* 251.

Easterbrook, Frank H. (1982) "Ways of Criticizing the Court," 95(4) *Harvard Law Rev.* 802.

Edwards, Harry T. (1985) "Public Misperceptions Concerning the 'Politics of Judging': Dispelling Some Myths About the D.C. Circuit," 56 *Univ. of Colorado Law Rev.* 619.

Eisenberg, Theodore (1990) "Testing the Selection Effect: A New Theoretical Framework with Empirical Tests," 19(2) *J. of Legal Studies* 337.

Emmert, Craig F. (1992) "An Integrated Case-Related Model of Judicial Decision Making: Explaining State Supreme Court Decisions in Judicial Review Cases," 54(2) *J. of Politics* 543.

Epstein, Lee, Valerie Hoekstra, Jeffrey A. Segal, & Harold J. Spaeth (1998) "Do Political Preferences Change? A Longitudinal Study of U.S. Supreme Court Justices," 60(3) *J. of Politics* 801.

Epstein, Lee, Andrew D. Martin, & Christina L. Boyd (2007a) "On the Effective Communication of the Results of Empirical Studies, Part II," 60 *Vanderbilt Law Rev.* 101.

Epstein, Lee, Andrew D. Martin, Kevin M. Quinn, & Jeffrey A. Segal (2007b) "Ideological Drift Among Supreme Court Judges: Who, When and How Important?" 101 *Northwestern Univ. Law Rev.* 127.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal, & Chad Westerland (2007c) "The Judicial Common Space," 23(2) *J. of Law Economics & Organization* 303.

Epstein, Lee, Andrew D. Martin, & Matthew M. Schneider (2006) "On the Effective Communication of the Results of Empirical Studies, Part 1," 59 *Vanderbilt Law Rev.* 1181.

Farnsworth, Ward (2007) "The Use and Limits of Martin-Quinn Scores to Assess Supreme Court Justices, with Special Attention to the Problem of Ideological Drift," 101 *Northwestern Univ. Law Rev.* 143.

Friedman, Barry (2006) "Taking Law Seriously," 4(2) *Perspectives on Politics* 261.

Gelman, Andrew, & Jennifer Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew, & Iain Pardoe (2007) "Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components," 37(1) *Sociological Methodology* 23.

Gelman, Andrew, Christian Pasarica, & Rahul Dodhia (2002) "Let's Practice What We Preach: Turning Tables into Graphs," 56(2) *American Statistician* 121.

George, Tracey, & Lee Epstein (1992) "On the Nature of Supreme Court Decision Making," 86(2) *American Political Science Rev.* 323.

George, Tracey, & Michael E. Solimine (2001) "Supreme Court Monitoring of Courts of Appeals En Banc," 9 *Supreme Court Economic Rev.* 171.

Gill, Jeff (1999) "The Insignificance of Null Hypothesis Significance Testing," 52(3) *Political Research Q.* 647.

Greene, William H. (2003) *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.

Grier, Kevin B., Michael C. Munger, & Brian E. Roberts (1994) "The Determinants of Industry Political Activity, 1978–1986," 88(4) *American Political Science Rev.* 911.

Grofman, Bernard, & Timothy Brazill (2002) "Identifying the Median Justice on the Supreme Court Through Multidimensional Scaling: Analysis of 'Natural Courts' 1953–1991," 112(3–4) *Public Choice* 55.

Gruhl, John (1980) "The Supreme Court's Impact on the Law of Libel: Compliance by Lower Federal Courts," 33(4) *Western Political Q.* 502.

Hagle, Timothy M. (1991) "But Do They Have to See It to Know It: The Supreme Court's Obscenity and Pornography Decisions," 44(4) *Western Political Q.* 1039.

Hall, Melinda G., & Paul Brace (1996) "Justices' Response to Case Facts: An Interactive Model," 24(2) *American Politics Q.* 237.

Hart, David M. (2001) "Why Do Some Firms Give? Why Do Some Give a Lot?: High-Tech PACs, 1977–1996," 63(4) *J. of Politics* 1230.

Hartnett, Edward A. (2000) "Questioning Certiorari: Some Reflections Seventy-Five Years After the Judges' Bill," 100(7) *Columbia Law Rev.* 1643.

Harvey, Anna, & Barry Friedman (2006) "Pulling Punches: Congressional Constraints on the Supreme Courts Constitutional Rulings, 1987–2000," 31(4) *Legislative Studies Q.* 533.

Heckman, James J. (1979) "Sample Selection Bias as a Specification Error," 47(1) *Econometrica* 153.

Ignagni, Joseph A. (1994) "Explaining and Predicting Supreme Court Decision Making: The Burger Court's Establishment Clause Decisions," 36(2) *J. of Church & State* 301.

Kastellec, Jonathan P. (2007) "Panel Composition and Judicial Compliance on the United States Courts of Appeals," 23(2) *J. of Law Economics & Organization* 421.

Kastellec, Jonathan P., & Eduardo L. Leoni (2007) "Using Graphs Instead of Tables in Political Science," 5(4) *Perspectives on Politics* 755.

King, Gary, & Langche Zeng (2001) "Logistic Regression in Rare Events Data," 9(2) *Political Analysis* 137.

Klein, David E., & Robert J. Hume (2003) "Fear of Reversal as an Explanation of Lower Court Compliance," 37(3) *Law & Society Rev.* 579.

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

Kritzer, Herbert M., & Mark J. Richards (2002) "Deciding the Supreme Court's Administrative Law Cases: Does *Chevron* Matter?" paper prepared for the Annual Meeting of the American Political Science Association. Boston, MA.

—— (2003) "Jurisprudential Regimes and Supreme Court Decisionmaking: The Lemon Regime and Establishment Clause Cases," 37(4) *Law & Society Rev.* 827.

—— (2005) "The Influence of Law in the Supreme Court's Search-and-Seizure Jurisprudence," 33(1) *American Politics Research* 33.

Laver, Michael, Kenneth Benoit, & John Garry (2003) "Extracting Policy Positions from Political Texts Using Words as Data," 97(2) *American Political Science Rev.* 311.

Law, David S. (2005) "Strategic Judicial Lawmaking: Ideology, Publication, and Asylum Law in the Ninth Circuit," 73(3) *Univ. of Cincinnati Law Rev.* 817.

Lax, Jeffrey R. (2003) "Certiorari and Compliance in the Judicial Hierarchy: Discretion, Reputation and the Rule of Four," 15(1) *J. of Theoretical Politics* 61.

Liao, Tim Futing (1995) "The Nonrandom Selection of Don't Knows in Binary and Ordinal Responses: Corrections with the Bivariate Probit Model with Sample Selection," 29 *Quality & Quantity* 87.

Manski, Charles F., & Steven R. Lerman (1977) "The Estimation of Choice Probabilities from Choice Based Samples," 45(8) *Econometrica* 1977.

Martin, Andrew D., & Kevin M. Quinn (2002) "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999," 10(2) *Political Analysis* 134.

—— (2007) "Assessing Preference Change on the U.S. Supreme Court," 23(2) *J. of Law, Economics & Organization* 365.

McGuire, Kevin T. (1990) "Obscenity, Libertarian Values, and Decision Making in the Supreme Court," 18(1) *American Politics Q.* 47.

McGuire, Kevin T., & Georg Vanberg (2005) "Mapping the Policies of the U.S. Supreme Court: Data, Opinions, and Constitutional Law," paper presented at the Annual Meeting of the American Political Science Association. Washington, DC.

McNollgast (1995) "Politics and the Courts: A Positive Theory of Judicial Doctrine and the Rule of Law," 68(6) *Southern California Law Rev.* 1631.

Merritt, Deborah Jones, & James J. Brudney (2001) "Stalking Secret Law: What Predicts Publication in the United States Courts of Appeals," 54 *Vanderbilt Law Rev.* 71.

Owens, Ryan J. (2007) "The Separation of Powers, The Supreme Court, and Strategic Agenda Setting," manuscript. Washington University.

Perry, Jr., H. W. (1991) *Deciding to Decide: Agenda Setting in the United States Supreme Court.* Cambridge: Harvard University Press.

Poe, Steven C., & James Meernik (1995) "US Military Aid in the 1980s: A Global Analysis," 32(4) *J. of Peace Research* 399.

Priest, George L., & Benjamin Klein (1984) "The Selection of Disputes for Litigation," 13(1) *J. of Legal Studies* 1.

Richards, Mark J., & Herbert M. Kritzer (2002) "Jurisprudential Regimes in Supreme Court Decision Making," 96(2) *American Political Science Rev.* 305.

Robert, Christian P., & George Casella (2005) *Monte Carlo Statistical Methods.* New York: Springer.

Rohde, David W., & Harold J. Spaeth (1976) *Supreme Court Decision Making.* San Francisco, CA: W.W. Freeman & Co.

Sartori, Anne E. (2003) "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions," 11(2) *Political Analysis* 111.

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

—— (1985) "Measuring Change on the Supreme Court: Examining Alternative Models," 29(3) *American J. of Political Science* 461.

—— (1986) "Supreme Court Justices as Human Decision Makers: An Individual-Level Analysis of the Search and Seizure Cases," 48(4) *J. of Politics* 938.

—— (1997) "Separation of Powers Games in the Positive Theory of Congress and Courts," 91(1) *American Political Science Rev.* 28.

Segal, Jeffrey A., & Albert D. Cover (1989) "Ideological Values and the Votes of U.S. Supreme Court Justices," 83(2) *American Political Science Rev.* 557.

Segal, Jeffrey A., & Cheryl D. Reedy (1988) "The Supreme Court and Sex Discrimination: The Role of the Solicitor General," 41(3) *Western Political Q.* 553.

Segal, Jeffrey. A., & Harold J. Spaeth (1993) *The Supreme Court and the Attitudinal Model.* New York: Cambridge University Press.

—— (2002) *The Supreme Court and the Attitudinal Model Revisited.* New York: Cambridge University Press.

Siegelman, Peter, & John J. Donohue III (1990) "Studying the Iceberg from Its Tip: A Comparison of Published and Unpublished Employment Discrimination Cases," 24(5) *Law & Society Rev.* 1133.

Simmons, Beth, & Daniel Hopkins (2005) "The Constraining Power of International Treaties: Theory and Methods," 99(4) *American Political Science Rev.* 623.

Songer, Donald R., Charles M. Cameron, & Jeffrey A. Segal (1995) "An Empirical-Test of the Rational-Actor Theory of Litigation," 57(4) *J. of Politics* 1119.

Songer, Donald R., & Susan Haire (1992) "Integrating Alternative Approaches to the Study of Judicial Voting: Obscenity Cases in the United-States Courts of Appeals," 36(4) *American J. of Political Science* 963.

Songer, Donald R., Jeffrey A. Segal, & Charles M. Cameron (1994) "The Hierarchy of Justice: Testing a Principal-Agent Model of Supreme-Court Circuit-Court Interactions," 38(3) *American J. of Political Science* 673.

Songer, Donald R., & Reginald S. Sheehan (1990) "Supreme Court Impact on Compliance and Outcomes: *Miranda* and *New York Times* in the United States Courts of Appeals," 43(2) *Western Political Q.* 297.

Songer, Donald R., Reginald S. Sheehan, & Susan B. Haire (2000) *Continuity and Change on the U.S. Courts of Appeals*. Ann Arbor, MI: University of Michigan Press.

Spiller, Pablo T., & Rafael Gely (1992) "Congressional Control or Judicial Independence: The Determinants of US Supreme Court Labor-Relations Decisions 1949–1988," 23(4) *Rand J. of Economics* 463.

Swinford, Bill (1991) "A Predictive Model of Decision Making in State Supreme Courts: The School Financing Cases," 19(3) *American Politics Q.* 336.

Timpone, Richard J. (1998) "Structure, Behavior, and Voter Turnout in the United States," 92(1) *American Political Science Rev.* 145.

Wahlbeck, Paul J. (1997) "The Life of the Law: Judicial Politics and Legal Change," 59(3) *J. of Politics* 778.

—— (1998) "The Development of a Legal Rule: The Federal Common Law of Public Nuisance," 32(3) *Law & Society Rev.* 613.

Winship, Christopher, & Robert D. Ware (1992) "Models for Sample Selection Bias," 18 *Annual Rev. of Sociology* 327.

Zorn, Christopher (2005) "A Solution to Separation in Binary Response Models," 13(2) *Political Analysis* 157.

Appendix: Solving the Noncompliance Puzzle

This appendix presents a possible solution to the noncompliance puzzle, based on our simulations. Recall that there are two possible true outcomes and two possible estimated decisions, creating four categories in which cases might fall. Let the percentage of each category be A, B, C, and D, as shown in Table A1. For example, Category B contains cases in which the true decision is conservative, but we falsely estimate it to be liberal given our sample weights.

Table A1: Inferences About Lower Court Behavior

			If the lower court complies . . .			If the lower court does not comply . . .		
Case Categories	True decision should be	Predicted decision is	The decision is	Is it truly compliant?	Would we think it compliant?	The decision is	Is it truly compliant?	Would we think it compliant?
A%	1	1	1	Y	Y	0	N	N
B%	1	0	1	Y	N	0	N	Y
C%	0	1	0	Y	N	1	N	Y
D%	0	0	0	Y	Y	1	N	N
Total				A + B + C + D	A + D		(0%)	B + C

			If the lower court votes liberally . . .			If the lower court votes conservatively . . .		
Case Categories	True decision should be	Predicted decision is	The decision is	Is it truly compliant?	Would we think it compliant?	The decision is	Is it truly compliant?	Would we think it compliant?
A%	1	1	0	N	N	1	Y	Y
B%	1	0	0	N	Y	1	Y	N
C%	0	1	0	Y	N	1	N	Y
D%	0	0	0	Y	Y	1	N	N
Total				C + D	B + D		A + B	A + C

NOTE: The table examines when out-of-sample conclusions will be accurate, for all four combinations of true outcomes and predicted outcomes. For example, Category B contains cases in which the true decision is conservative, but we falsely estimate it to be liberal given our sample weights. The table reveals the possibility for broad inferential errors: a consistently compliant lower court—one that always decides in favor of the outcome truly preferred by the Supreme Court—will nevertheless be labeled as noncompliant (B + C) percent of the time. A noncompliant lower court, one that always makes the opposite decision, will be labeled as compliant (B + C) percent of the time. A recalcitrant liberal lower court (one that votes liberally no matter what the case or the Supreme Court's preferences) will be compliant, solely due to coincidence, (C + D) percent of the time, but it will be labeled as compliant (B + D) percent of the time. A recalcitrant conservative lower court (one that votes conservatively no matter what) will be compliant (A + B) percent of the time, but will be labeled as compliant (A + C) percent of the time.

A consistently compliant lower court—one that always decides in favor of the outcome truly preferred by the Supreme Court—will nevertheless be labeled as noncompliant (B + C) percent of the time. A noncompliant lower court, one that always makes the opposite decision, will be labeled as compliant (B + C) percent of the time. A recalcitrant liberal lower court (one that votes liberally in every case) will be compliant, solely due to coincidence, (C + D) percent of the time, but it will be labeled as compliant (B + D) percent of the time. A recalcitrant conservative lower court (one that votes conservatively no matter what) will be compliant (A + B) percent of the time, but will be labeled as compliant (A + C) percent of the time. Note that we cannot even tell if our estimates of compliance are biased upward or downward, unless we know the true outcomes to begin with. That is, the error in measuring the recalcitrant conservative lower court's compliance level is (A + C) – (A – B) = (C – B) percentage points, which can be positive or negative.

Table A2 depicts how this actually occurs in our simulations. Consider the mean values for the "close liberal cases" strategy: A = 63 percent, B = 14 percent, C = 8 percent, and D = 14 percent. If all lower court decisions were truly compliant, 22 percent (B + C) would still be falsely labeled as noncompliant; conversely, if all were noncompliant, the same percentage would be falsely labeled as compliant. A recalcitrant liberal lower court would technically still be compliant in 22 percent of cases (C + D), but be measured as complying in 28 percent of cases (B + D), while a recalcitrant conservative lower court will comply by coincidence alone in 77 percent of cases (A + B), but be measured as complying 70 percent of the time (A + C). Finally, of all compliant liberal decisions, 36 percent would be falsely labeled as noncompliant (C/(C + D)), while of all compliant conservative decisions, 18 percent (B/(A + B)) would be labeled as noncompliant.

One further calculation fills in the final piece of the puzzle. We can break down the 22 percent of cases in which a liberal, but perfectly obedient, court would be measured as noncompliant. We would think that 8 percent (C) of the time, the liberal lower court was deciding against higher court doctrine and pursuing its own interests (liberal decisions where conservative decisions are due). However, in 14 percent (B) of cases, we would think this truly compliant liberal lower court was deciding against higher court doctrine *and* counter to its own interests (making conservative decisions where liberal ones are due). That is, 64 percent (B/(B + C)) of the so-called noncompliance from truly compliant liberal lower courts would be, paradoxically, in the conservative direction—and all because of the bias caused by case selection.

Table A2: Out-of Sample Inferences Across Selection Strategies

Case Categories (Mean Values)	*Random*	*Close*	*Close Liberal*	*Close Conservative*	*Anomalous*	*Anomalous Liberal*	*Anomalous Conservative*	*Conservative Nonexcepted Home*
% of true positives (A)	71%	72%	63%	76%	68%	31%	76%	73%
% of false negatives (B) % of false	5	5	14	1	9	46	~0	4
positives (C)	13	13	8	21	22	6	23	18
% of true negatives (D)	10	10	14	3	1	16	~0	5
Puzzling liberal “noncompliance” (B/B + C)	28	28	64	5	29	88	~0	18
Puzzling conservative “noncompliance” (C/B + C)	72	72	36	95	71	12	100	82

NOTE: For each selection strategy, the table presents the percentages of cases that fall into each possible case categories (as described in Table A1), as well as the percent of puzzling liberal noncompliance and puzzling conservative noncompliance—respectively, how often a perfectly compliant liberal lower court would be misclassified as not complying in the conservative direction, and how often a perfectly compliant conservative lower court would be misclassified as not complying in the liberal direction.