Measuring Inconsistency, Indeterminacy, and Error in Adjudication

Joshua B. Fischman[*]

Abstract

Scholars have long debated whether, and to what extent, law constrains judicial decisions. Because law cannot be objectively measured, many believe that judicial decisions cannot be empirically evaluated on grounds internal to the practice of law. This Article demonstrates that empirical analysis of judicial decisions can nevertheless provide objective, albeit limited, conclusions about subjective criteria for evaluating a system of adjudication. It begins by formalizing three criteria: inter-judge inconsistency, legal indeterminacy, and judicial error. It then clarifies what can be learned about these criteria from observational data on single-judge adjudication. The precise level of inconsistency cannot be identified, but it is possible to estimate a range of feasible values. Similarly, rates of indeterminacy and error cannot be estimated in isolation, but it is possible to estimate a curve that identifies feasible combinations of these rates. The methodologies developed in this Article are illustrated using data on immigration adjudication.

[*] Associate Professor, University of Virginia School of Law, jfischman@virginia.edu. I am deeply grateful to Charles Barzun, Mike Gilbert, Fred Schauer, and Micah Schwartzman for many helpful discussions during the development of this paper. I also thank Ivan Canay, Michael Churgin, John Duffy, Paul Edelman, Kent Greenawalt, Rich Hynes, Charles Manski, David Martin, Greg Mitchell, John Monahan, John Pepper, George Rutherglen, Sean Sullivan, Barry Weingast, and audiences at the University of Virginia School of Law, the University of Texas School of Law, Washington University School of Law, Northwestern University Law School, the Conference on Empirical Legal Studies, the University of Toronto Faculty of Law, George Washington University Law School, Harvard Law School, George Mason University School of Law, and Northwestern Department of Economics for helpful comments. I thank Mariusz Kolczykiewicz for excellent research assistance.

I. Introduction

For much of the last century, scholars have debated whether, and to what extent, law constrains judicial decisions. Rule skeptics—including many legal realists, critical legal scholars, and political scientists—have argued that legal rules are seldom determinate and that judges have substantial discretion to decide cases according to extralegal criteria. Legal positivists such as Hart (1961), on the other hand, agree that legal rules are occasionally indeterminate, but assert that the law is clear and binding in most cases. A third position, taken by interpretivists such as Dworkin (1986) and many proponents of natural law, holds that law is always determinate, although it does not always constrain judicial decision making in practice.

Many empirical studies have documented substantial disparities among adjudicators in criminal (e.g., Everson 1919; Gaudet, Harris, and St. John 1933; Waldfogel 1998; Anderson, Kling, and Stith 1999), social security (Mashaw et al. 1978), and immigration cases (Ramji-Nogales, Schoenholz, and Schrag 2007). More recently, studies of circuit courts have also found differences in case outcomes depending on the composition of the panel (e.g., Revesz 1997; Cross and Tiller 1998; Sunstein et al. 2006; Miles and Sunstein 2006, 2008). Such studies provide intuitive confirmation that judicial decisions are not fully constrained by law, but they do not attempt to quantify compliance with rule-of-law values. It is true that the existence of disparities among judges refutes the claim that judging is perfectly objective and predictable, but no one has even taken this claim seriously (Tamanaha 2009: 27–43). The meaningful empirical question is not whether a system of adjudication meets such an ideal, but rather the degree to which it falls short (Ibid.: 145–48).

One reason why answers have been elusive is that inquiries about legal constraint have both conceptual and empirical components, yet there is little engagement between philosophical and quantitative analyses of law (Galligan 2010). The conceptual part of the inquiry addresses the question, "When does law obligate a judge to reach a particular outcome?" The answer to this question depends on which sources of law are considered authoritative and what obligations they establish. Theories of law that include moral principles as valid sources, for example, may determine unique outcomes more frequently than narrow conceptions of law that rely exclusively on legal texts.

The empirical component addresses the question, "How often do judges deviate from their legal obligations?" Although this inquiry is essentially empirical, it is posed from an internal point of view. For this reason, it might seem that objective answers are unattainable. As Barry Friedman observed, as long as “[t]here are deep philosophical debates within the legal academy itself about what law is, … it [will be] difficult to make claims about law’s influence that are readily subject to falsification” (2006: 265–66).

In this Article, I demonstrate that it is in fact possible to make progress on the empirical inquiry without first resolving the conceptual inquiry. I do so by formalizing three measures—inconsistency, indeterminacy, and error—and clarifying what can be revealed about them using empirical analysis of observational data. I use the term “inconsistency” to capture how often the outcome of a case will depend on the identity of the judge selected to decide it; formally, it represents the probability that two judges would differ in their disposition of a randomly selected case. “Indeterminacy” denotes the proportion of cases in which the law fails to require a unique result. “Error” represents the proportion of cases in which the judge’s decision conflicts with the result required by law.

These measures can serve to quantify the gap between how judges *actually* decide cases and how they *ought to* decide cases. Because these measures coincide with widely accepted concepts in legal theory, they can provide credible justification for normative claims about legal rules and the design of adjudicatory institutions. The federal sentencing guidelines, for example, were motivated in large part by empirical studies of inter-judge sentencing disparity (Stith and Cabranes 1998: 104–12) and claims about inconsistency and indeterminacy in pre-guidelines sentencing (Frankel 1974: 1–12). The Social Security Administration similarly justified its vocational grid for disability determinations on the ground that it would reduce inconsistency among adjudicators (43 Fed. Reg. 55,349 [1978]). More recently, some scholars have argued that inconsistency between majority-Democratic and majority-Republican circuit court panels justifies greater judicial deference to administrative agencies (Miles and Sunstein 2006, 2008) or partisan balancing within panels (Cross and Tiller 1999).

Because inconsistency is formulated purely in terms of observable judicial decisions, and not with respect to the contested requirements of law, it would appear to be objectively measureable. For this reason, most empirical studies of judicial decision making present results that can be interpreted in terms of inconsistency. Indeterminacy and error, on the other hand, are only cognizable from an internal point of view, and therefore might appear to be outside the province of empirical inquiry (Kysar 2007; Edwards and Livermore 2009). I show, however, that although indeterminacy and error cannot be measured in isolation, it is possible to derive informative results by examining them jointly. Specifically, for any data set of adjudication outcomes in which cases are assigned randomly, it is possible to estimate a minimum rate of error as a function of the proportion of cases that are assumed to be indeterminate. These estimates can be used to construct an "indeterminacy-error curve" that demarcates a boundary between feasible and infeasible combinations of indeterminacy and error rates.

The empirical methods employed in this Article draw upon recent advances in the estimation of partially identified econometric models (Manski 2003; Tamer 2010). These methods enable statistical inference with minimal assumptions; the tradeoff is that they yield weaker estimates than standard econometric approaches. Even with an infinite amount of data, the methodology developed here could not derive precise estimates of inconsistency, indeterminacy, and error, but could only generate upper and lower bounds on inconsistency and joint bounds on indeterminacy and error. However, the methodology clarifies what additional sources of data can sharpen the estimates and demonstrates the need for strong assumptions in order to justify legal or institutional reforms.

The organization of this Article proceeds as follows. Section II provides the conceptual framework for operationalizing inconsistency for the purpose of empirical analysis. It then addresses the statistical identification of inconsistency, revealing what can be learned about the rate of inconsistency in an idealized setting. Drawing upon the framework used to analyze inconsistency, Section III shows how to construct joint bounds on indeterminacy and error rates. Section IV addresses statistical inference, developing methods for generating confidence intervals for the rate of inconsistency and confidence regions for indeterminacy and error rates using observational data. Section V provides an illustration of the framework developed in this Article using data on administrative asylum adjudication. Section VI concludes. An appendix provides most mathematical derivations as well as a table with descriptions of all variables used in the analysis.

II. Inconsistency

Many empirical studies find disparities among judges' decision rates, and conclude that law fails to fully constrain judges' decisions. Such claims are intuitive: if judges reach various outcomes at significantly different rates in randomly assigned cases, then their decisions cannot be uniquely determined by law. Yet comparisons of judges' decision rates also have important limitations. Although such statistics are useful for testing whether judges are perfectly consistent, they do not by themselves provide enough information to estimate the magnitude of inconsistency.

To illustrate, consider two judges, both of whom would decide in favor of plaintiffs 50% of the time in an identical set of cases. It is possible that both judges would side with plaintiffs in the same 50% of cases, in which case their decisions would be perfectly consistent. On the other hand, it is possible that the second judge would decide every case contrary to how the first judge would decide, so that their decisions would be perfectly inconsistent. Alternatively, both judges could be deciding cases by flipping coins. These scenarios have sharply different normative implications, but the mere lack of disparity between the judges' decision rates does not allow us to distinguish among them. Even if the judges were perfectly consistent, they could be always correct, always incorrect, or anywhere in between.

Even if we knew the precise rate of inconsistency between judges, it would be a misleading indicator of legal constraint, since adjudication can be highly consistent even when law fails to constrain judicial decisions.[1] Consider another hypothetical court with two judges, both of whom decide every case incorrectly, but in exactly the same way. In this hypothetical court, both judges would be perfectly consistent, but all of the decisions would be contrary to law. Now imagine that a third judge is appointed, and that this judge decides every case correctly. Inconsistency would increase, since the third judge would always differ from the first two, but the rate of error would also decrease.

Although inconsistency provides evidence that law does not fully constrain judges, this example demonstrates that it cannot be used to comparing the degree of legal constraint in different courts or different time periods. Nevertheless, inconsistency may be of theoretical interest for two additional reasons. First, it provides a measure of the degree of predictability in adjudication (Coleman and Leiter 1993; Waldron 2007). Most theories of the rule of law require that individuals have notice regarding how the law will be applied and the opportunity to conform their behavior to its requirements. To the extent that the outcomes of potential disputes

[1] For example, many legal realists (e.g., Moore and Hope 1929; Cohen 1935; Llewellyn 2011) and critical legal scholars (e.g., Kairys 1984; Singer 1984; Dalton 1985) argued that legal rules were largely indeterminate but that adjudication nevertheless followed predictable patterns.

would be invariant to the judges selected to adjudicate them, parties could predict how the law will be applied and plan accordingly. This will be true even if consistency stems from extralegal influences. Second, inconsistency provides evidence of comparative injustice, in the sense that like parties are not being treated alike (Waldron 2007). It would be unjust for a defendant facing a harsh judge to receive a longer prison term than an equally culpable defendant who was assigned to a lenient judge. There is substantial disagreement, however, regarding whether the comparative discrepancy is itself a form of injustice; some argue that such discrepancy is merely evidence that one of the sentences was not determined according to law in a non-comparative sense (Westen 1982).

The importance of predictability and comparative justice will necessarily depend on context. In circumstances in which law enables unconnected individuals to coordinate their activities, predictability may be a central concern (Shapiro 2011: 132). As Justice Brandeis observed, it may be "more important that the applicable rule of law be settled than that it be settled right" (*Burnet v. Coronado Oil & Gas Co*., 285 U.S. 393, 406 [1932]). Comparative justice will be of greater concern if outcomes are not uniquely determined by law but legal or moral considerations prescribe equal treatment among certain parties.

A. Empirical Framework

Let C denote a set of legal questions, and let J denote a set of n judges (or administrative decision makers) who could potentially decide the questions in C. In typical studies, C will consist of cases involving a common legal issue, but the validity of the analysis does not require any degree of similarity among the cases. Alternatively, the unit of observation might not be cases but rather particular issues within cases (such as whether a plaintiff has standing), abstract hypotheticals, or administrative determinations. The only restriction placed on C, to avoid the possibility of selection bias, is that the criteria for inclusion of a case in C must be independent of the judge assigned to adjudicate that case.[2] In this Article, I shall ignore any heterogeneity among the cases in C; future research will consider how to incorporate information about particular cases into the framework developed here.

[2] This condition may not be satisfied, for example, in a study that only examines published opinions if the decision to publish is not independent of the judge deciding the case. Similarly, studies that examine cases that decide a particular legal issue might violate the condition if some judges are less likely to reach the merits on this issue.

To operationalize the concept of inconsistency, imagine that we could observe two judges' decisions in a case under idealized conditions. The case would be litigated before both judges, with perfect replication, and the judges would issue their decisions in complete isolation. Let $Y_i(c)$ and $Y_j(c)$ denote the respective decisions of judges i and j in case c in this idealized comparison. In the causal model of Neyman (1935) and Rubin (1974), the cases are the units of observation, judges i and j can be viewed as "treatments," and the decisions $Y_i(c)$ and $Y_j(c)$ are "potential outcomes." For simplicity, assume that these decisions can be coded in a dichotomous manner, so that $Y_i(c)$ takes on the values of zero and one.[3] To maintain generality, we refer to a decision as "positive" when $Y_i(c) = 1$ and "negative" when $Y_i(c) = 0$. Depending on the context, a positive decision might mean that a plaintiff obtained some particular form of relief, or that the judge's decision coincides with a liberal outcome. The results will be valid irrespective of the method for coding outcomes, but some coding choices may generate more informative results than others.

If $Y_i(c) \neq Y_j(c)$—that is, if the outcome of case c would have depended on whether it was assigned to judge i or judge j—then we say that these judges' decisions in case c would be inconsistent. Define "pairwise inconsistency" D_{ij} as the proportion of cases in C that would be decided inconsistently between judges i and j under these idealized conditions, and let "average inconsistency" D be the average rate of pairwise inconsistency among all pairs of judges in J. This can be interpreted as the probability that a randomly selected case would be decided inconsistently between a pair of randomly selected judges under idealized conditions. Note that we are only estimating inconsistency among cases that are litigated; these will not be representative of the universe of potential disputes (Priest and Klein 1984). Whether, and how, estimates of inconsistency can be extrapolated is a subject for future research.

B. Pairwise Inconsistency

In courts where cases are randomly assigned to a single judge, it is never possible to simultaneously observe $Y_i(c)$ and $Y_j(c)$ for two judges i, j. This is due to what Holland (1986)

[3] For simplicity, I shall also assume that each decision is non-stochastic, i.e., that the outcome would always be the same if a case were assigned to the same judge, and that judges do not "drift" ideologically during the period of study. If judicial decisions are stochastic, then the results in this article still hold with respect to *expected* inconsistency and error rates.

calls the “fundamental problem of causal inference.” For any case c, we only observe the outcome for the judge who actually decided the case. The decisions that would have been rendered by the other judges are unobserved “potential outcomes.”

The standard approach in the Neyman-Rubin model is to estimate an average treatment effect $ATE = \mathrm{E}\left[Y_j(c) - Y_i(c)\right]$, representing the average effect of reassigning a case from judge i to judge j. As long as assignment of judges is random, it is possible to derive a valid estimate of the ATE with minimal assumptions.[4] For this reason, many studies of judicial behavior have reported average treatment effects, such as the average effect of replacing a male judge with a female judge (Boyd, Epstein, and Martin 2010).

The problem with this approach is that the ATE does not have any normative significance; it can be directly measured, but it is not a valid metric for evaluating systems of adjudication. In the example discussed at the beginning of this section, in which both judges favor plaintiffs 50% of the time, the ATE would be zero. Yet this result is compatible with highly desirable as well as highly undesirable scenarios.

Although inconsistency cannot be expressed as an ATE, it is possible to derive bounds on inconsistency by exploiting information about the judges’ rates of reaching positive decisions. Let r_j represent the proportion of cases in C in which judge j would reach a positive decision. In practice, we cannot know r_j exactly, since we only observe judge j’s decisions in a subset of the cases. However, if these cases are representative, then we can use statistical methods to estimate the distribution of r_j.

For the purpose of deriving bounds on inconsistency, suppose that we have exact knowledge of r_i and r_j. These rates specify the marginal distributions of Y_i and Y_j, but the inconsistency rate D_{ij} is determined by the joint distribution of Y_i and Y_j, which is not observed. This is illustrated in the following contingency table, in which the joint probabilities are denoted $p_{00}, p_{01}, p_{10}, p_{11}$.

[4] In particular, it is necessary to assume that judges’ decisions (or potential decisions) will not be affected by which judges were assigned other cases. In the Neyman-Rubin framework, this assumption is known as the “stable unit treatment value assumption” (Rubin 1980). This could potentially be violated if judges’ decisions are influenced by the precedential authority of other judges’ decisions. For the remainder of the discussion, I make the simplifying assumption that there is no precedential influence among the cases in C. This may be a reasonable assumption in many systems of administrative adjudication. In addition, when C encompasses a short time period, the impact of prior cases might dominate the precedential effect of cases within C. If the cases span a longer time period, a model with time controls may be adequate to control for the effects of precedent.

TABLE 1
CONTINGENCY TABLE FOR JUDGES' DECISIONS

	j negative ($Y_j = 0$)	j positive ($Y_j = 1$)	Total
i negative ($Y_i = 0$)	p_{00}	p_{01}	$1 - r_i$
i positive ($Y_i = 1$)	p_{10}	p_{11}	r_i
Total	$1 - r_j$	r_j	

The ATE can be represented in terms of the joint probabilities as $p_{01} - p_{10}$ or in terms of the marginal probabilities as $r_j - r_i$; it can be consistently estimated because it can be expressed as a function of the marginal probabilities. The rate of inconsistency between judges i and j is given by $D_{ij} = p_{01} + p_{10}$, the sum of the joint probabilities corresponding to disagreement between i and j. Although these joint probabilities cannot be expressed as a function of marginal probabilities, Fréchet (1951) and Hoeffding (1940) provide bounds on the joint probabilities in terms of the marginals. The Fréchet-Hoeffding bounds for p_{01} and p_{10} are as follows:

$$\begin{aligned} \max\{r_i - r_j, 0\} \le p_{10} \le \min\{r_i, 1 - r_j\} \\ \max\{r_j - r_i, 0\} \le p_{01} \le \min\{r_j, 1 - r_i\} \end{aligned} \tag{1}$$

Summing the two inequalities in (1) yields the following result.

PROPOSITION 1: Let $\underline{D}_{ij} = |r_i - r_j|$ and $\overline{D}_{ij} = \min\{r_i + r_j, 2 - r_i - r_j\}$. Then the rate of pairwise inconsistency D_{ij} between any two judges i and j satisfies

$$\underline{D}_{ij} \le D_{ij} \le \overline{D}_{ij}, \tag{2}$$

and both bounds are sharp.

PROOF: See Appendix.

To provide some intuition for the above result, consider the following comparison between Justices Thomas and Ginsburg. According to the Spaeth Supreme Court Database (2011), Justice Thomas voted in the liberal direction 29% of the time during the 2000–2009

Terms, while Justice Ginsburg voted in the liberal direction 62% of the time.[5] Given only this information, and not information about their specific votes, what could be inferred about their rate of disagreement? The lower bound in Proposition 1 shows that they must disagree at least $|62\% - 29\%| = 33\%$ of the time. This lower bound would occur if *all* of Justice Thomas's liberal votes coincided with liberal votes by Justice Ginsburg. Proposition 1 also shows that they can disagree in at most $\min\{62\% + 29\%, 2 - 62\% - 29\%\} = 91\%$ of the cases. This would occur only if each justice's liberal votes coincided with conservative votes by the other justice.

Without any additional information, the data on the two justices' voting rates can only show that their rate of disagreement lies somewhere between 33% and 91%. If this interval seems surprisingly wide, and if the conditions for achieving the upper bound sound implausible, it is only because the justices' votes are in fact simultaneously observable, their voting patterns are widely known, and many readers have rough sense for what it means for a vote to be "liberal." In this instance, because the justices' votes are simultaneously observable, we can test our intuitions on the data. Table 2 provides a contingency table showing the joint outcomes for the two justices over the 2000–2009 Terms.

TABLE 2
CONTINGENCY TABLE FOR JUSTICES THOMAS AND GINSBURG:
PROPORTION OF LIBERAL VOTES, 2000–2009 TERMS

	Thomas Conservative	Thomas Liberal	Total
Ginsburg Conservative	34%	4%	38%
Ginsburg Liberal	37%	25%	62%
Total	71%	29%	

The true rate of disagreement between Justice Thomas and Justice Ginsburg is 41%, which can be found by adding the entries in off-diagonal cells. This rate of disagreement is eight percentage points higher than the lower bound, but still quite far from the upper bound. Although

[5] The Spaeth Supreme Court Database determines which votes are liberal and conservative based on the issues presented in each case. Although this coding method usually conforms to readers' expectations, various aspects of the coding procedure have been criticized by Shapiro (2009), Landes and Posner (2010), and Harvey and Woodruff (forthcoming).

this result may not seem surprising, intuition alone could not have revealed the precise rate of disagreement. In other applications, where outcomes are not simultaneously observable and judges are less well-known, intuition may be an even weaker guide.

The conditions under which the bounds on inconsistency are achieved can be formulated precisely using the following definitions.

DEFINITION: Given judges i and j, we say that judges i and j are *monotonic* if either $Y_i(c) \geq Y_j(c)$ for all c or $Y_i(c) \leq Y_j(c)$ for all c.

DEFINITION: Given judges i and j, we say that judges i and j are *countermonotonic* if either $Y_i(c) \leq 1 - Y_j(c)$ for all c or $Y_i(c) \geq 1 - Y_j(c)$ for all c.

Two judges are monotonic if the decisions of one judge are always at least as positive as the decisions of the other judge; this corresponds to the concept of monotonicity used in the literature on treatment effects (e.g., Imbens and Angrist 1994; Manski 1997). Similarly, judges i and j are countermononotic if judge i would be monotonic with a judge who always disagrees with judge j. In the contingency table in Table 1, monotonicity corresponds to the scenario where either $p_{01} = 0$ or $p_{10} = 0$, while countermonotonicity corresponds to the situation where either $p_{00} = 0$ or $p_{11} = 0$.

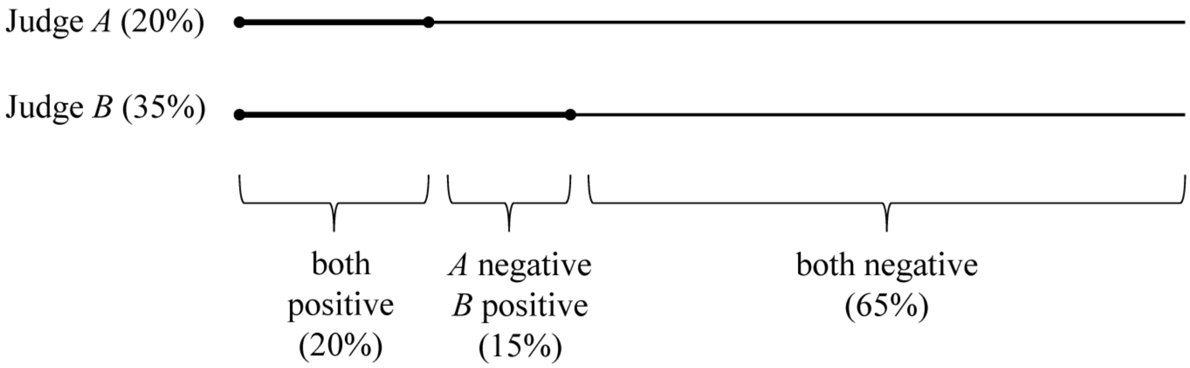
PROPOSITION 2: If judges i and j are monotonic, then the rate of pairwise inconsistency D_{ij} between i and j achieves the lower bound in Proposition 1 (i.e., $D_{ij} = \underline{D_{ij}}$). If judges i and j are countermonotonic, then the rate of pairwise inconsistency D_{ij} achieves the upper bound in Proposition 1 (i.e., $D_{ij} = \overline{D}_{ij}$). Under either of these assumptions, D_{ij} will be point-identified.

PROOF: See Appendix.

To illustrate how the bounds are achieved, consider two judges A and B with decision rates of 20% and 35%, respectively. Figure 1 illustrates a monotonic voting pattern in which the lower bound is achieved. The lines represent an unordered space of cases, where the shaded bars represent positive votes and the unshaded bars represent negative votes. In this illustration, judge B always reaches a positive decision whenever judge A does, so that $Y_A(c) \leq Y_B(c)$ for all c. Both judges reach a positive decision in 20% of the cases, and both reach a negative decision in

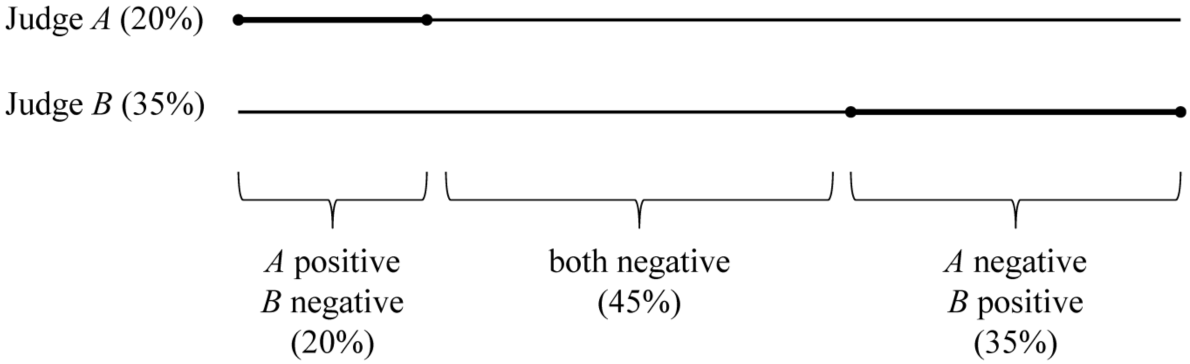
65% of the cases. Disagreement between judges A and B occurs only in the remaining 15% of cases in which judge B reaches a positive decision but judge A reaches a negative decision.

FIGURE 1
LOWER BOUND ON PAIRWISE INCONSISTENCY: MONOTONIC JUDGES



The upper bound in Proposition 1 is achieved when the judges are countermonotonic, as illustrated in Figure 2. Here, there is no overlap between the positive votes of judge A and the positive votes of judge B (i.e., $p_{11} = 0$), and the two judges disagree in 45% of the cases.

FIGURE 2
UPPER BOUND ON PAIRWISE INCONSISTENCY: COUNTERMONOTONIC JUDGES


Judge *A* (20%)
Judge *B* (35%)
A positive *B* negative (20%)
both negative (45%)
A negative *B* positive (35%)

To many readers, the alignment represented in Figure 2 may seem less plausible than the alignment represented in Figure 1. This intuition is not empirically testable, however, unless there exists detailed information about the cases or a representative sample of cases in which the judges decisions are simultaneously observable. Intuition about judicial behavior may help to narrow the range of inconsistency, but it is at best an imperfect guide. Although it may be reasonable to assume some degree of positive association between the judges' decisions, the judges will be monotonic only if all they can be mapped perfectly onto a left-right spectrum, a notion that is typically dismissed as "absurd" (Edwards and Livermore 2010: 1916). Once it is acknowledged that monotonicity is implausible, intuition cannot tell us exactly how close or far the level of inconsistency will be from the lower bound.

Whether the actual level of inconsistency will be close to the lower bound will depend on how the set C of cases is defined. Intuition and experience suggest that the judges' will be more closer to monotonicity when C is defined narrowly to include cases involving a single area of law, so that judicial preferences will be more likely to be unidimensional (Fischman and Law 2009). On the other hand, the judges may be further from monotonicity when C is defined broadly to encompass cases involving many types of issues.

Similarly, how close the level of inconsistency is to the lower bound will depend on how the outcomes are coded. To illustrate, suppose that C consists of discrimination cases and that these cases are coded positively if a judge provides some relief to an employee plaintiff. Although judges may disagree about the threshold for providing relief, it may seem plausible that the judges' would be reasonably close to monotonic in typical cases. If C includes reverse discrimination cases, however, then those least likely to provide relief in typical cases might be most sympathetic to plaintiffs with reverse discrimination claims. If this is true, then coding decisions for plaintiffs in reverse discrimination cases as negative decisions would generate more informative bounds.

C. Average Inconsistency

The concept of pairwise inconsistency can be generalized from the two-judge example to the case of n judges by averaging the measures of pairwise inconsistency over all pairs of judges:

$$D = \frac{2}{n(n-1)} \sum_{i<j} D_{ij}.$$

This can be interpreted as the proportion of cases in which two randomly selected judges would reach different outcomes. If every case in C were reassigned to a different judge, D represents the expected proportion of cases that would result in different outcomes.

The following result establishes bounds on average inconsistency in terms of the judges' decision rates.

PROPOSITION 3: Let $R = \sum r_i$, and let

$$\underline{D} = \frac{2}{n(n-1)} \sum_{i<j} \left| r_i - r_j \right| \text{ and } \overline{D} = \frac{2R}{n} + \frac{2\lfloor R \rfloor (\lfloor R \rfloor + 1 - 2R)}{n(n-1)},$$

where $\lfloor R \rfloor$ denotes the largest integer less than or equal to R. Then the average inconsistency rate D satisfies $\underline{D} \leq D \leq \overline{D}$, and both bounds can be achieved.

PROOF: See Appendix.

Note that the lower bound on average inconsistency is simply the average of the pairwise lower bounds. This bound will be achieved when all pairs of judges are monotonic. The upper bound on average inconsistency is achieved when the positive decisions are spread out as evenly as possible among the cases, so that every case receives a proportion of positive decisions that is as close as possible to the average. The upper bound on average inconsistency, however, will generally not be the average of the upper bounds on pairwise inconsistency, because it may not be possible for all judges to be countermonotonic with each other.[6] For example, suppose that one judge reaches positive decisions in half of the cases, and a second judge reaches positive decisions in the other half of the cases. These judges would be countermonotonic with each other, but it would be impossible for a third judge to be countermonotonic with both of them.

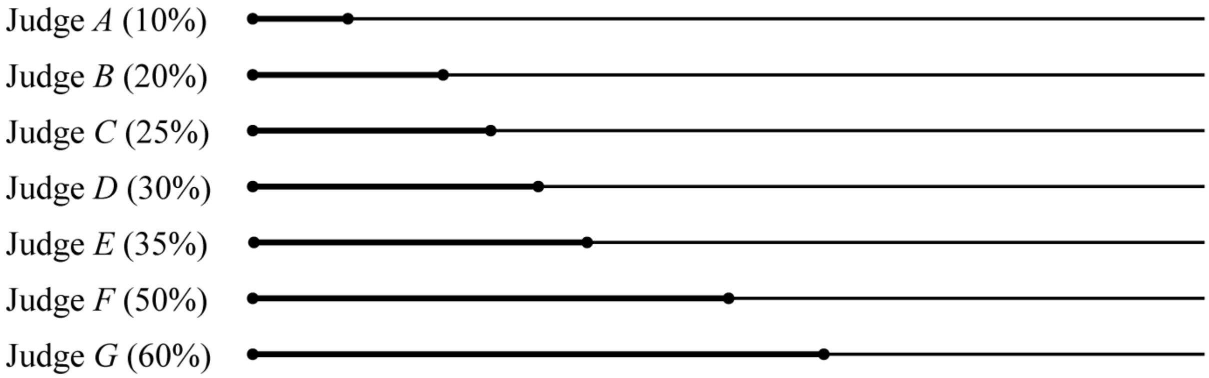
To provide an example of the average inconsistency bounds, consider a hypothetical court with seven judges. Suppose that any case would be equally likely to be assigned to any of these judges, and that they would decide in the positive direction at rates of 10%, 20%, 25%, 30%, 35%, 50%, and 60%, respectively. The lower bound on average inconsistency for these judges is 21%. The alignment that achieves this bound, which is depicted in Figure 3, has two features worth noting. First, every pair of judges is monotonic, so that pairwise inconsistency achieves the lower bound for every pair. Second, the judges would agree on a relative ordering of

[6] The upper bound on inconsistency will be the average of the pairwise upper bounds only if $\sum r_i \leq 1$ or $\sum r_i \geq n-1$. This result follows from Theorem 3.7 in Joe (1997: 61–63).

the cases: all seven judges would reach a positive decision in the cases depicted at the left end of the spectrum, and none would do so in the cases depicted on the right.

FIGURE 3

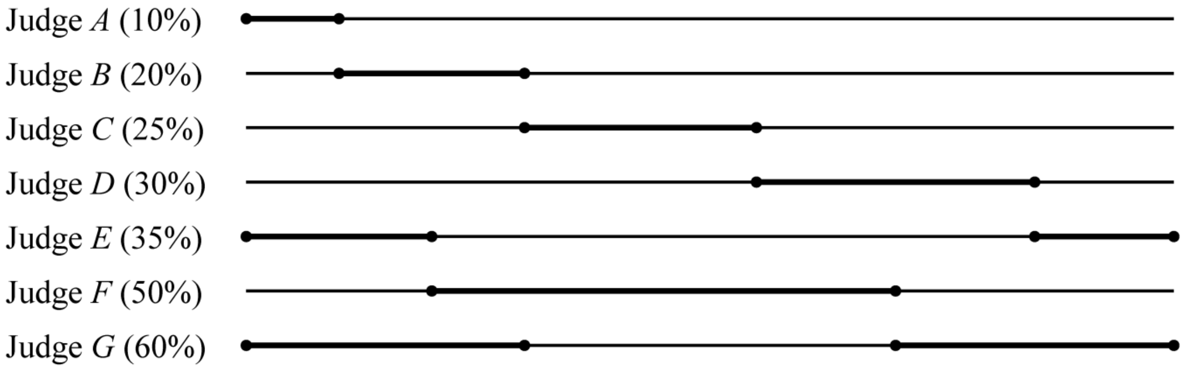
LOWER BOUND ON AVERAGE INCONSISTENCY: MONOTONIC JUDGES



The upper bound on average inconsistency for these judges is 50.5%. An alignment that achieves this upper bound in depicted in Figure 4. Note that some, but not all, pairs of judges are countermonotonic; given the hypothesized decision rates, it would be impossible for all pairs to be countermonotonic. The positive decisions are distributed as evenly as possible among the cases, so that all cases would result in positive decisions by either two or three judges. Thus, the judges would not agree on a common ordering of the cases, as they would when the lower bound is achieved.

Figure 4

Upper Bound on Average Inconsistency: Evenly Distributed Decisions



The alignment in Figure 4 may seem far-fetched; it is widely believed that judges do share a common understanding regarding what constitutes a strong or weak case. However, it is possible that judges would have a similar rank ordering among the universe of potential disputes but no common ordering within the set of litigated cases. In models of settlement such as Priest and Klein (1984), for example, the strongest and weakest cases will be settled, and probability of a plaintiff victory will be roughly equal among the remaining cases. If settlement typically occurs before the judge is announced, and if all remaining litigated cases have an equal probability of a positive decision, then the positive votes would be evenly distributed, as in Figure 4. Thus, the selection of disputes for litigation could result in a level of inconsistency that coincides with the upper bound, even if the judges have a similar rank ordering among litigated disputes.

D. Improving the Bounds

Subsections II.B and II.C demonstrated that inconsistency is only partially identified when judges' decisions in the same cases are not simultaneously observable. In practice, such bounds will be of limited use in policymaking. If one goal of an institutional reform is to reduce inconsistency, it will be difficult to assess whether the reform is successful. Estimates of inconsistency pre- and post-intervention will be interval-identified, and these intervals will typically overlap. When this occurs, it will be impossible to determine whether the intervention increased or decreased inconsistency. Thus, evaluation of institutional changes will typically require additional data and additional assumptions. This subsection briefly discusses several strategies for improving the bounds, which will developed in more detail in future research.

The discussion in the previous sections ignored any heterogeneity among the cases. An alternative approach is to exploit data on case characteristics. Suppose that the case space C can be divided according to case attributes into m partitions. For example, if C consists of immigration cases, it could be partitioned by country of origin or legal claim. One could estimate inconsistency on each partition, and then take a weighted average of these estimates to derive an estimate of inconsistency on C. Because the lower bound functions in Propositions 1 and 3 are globally convex and the upper bound functions are globally concave, Jensen's Inequality ensures that this approach will always generate bounds that are at least as informative as estimating inconsistency on the entire case space. However, because the upper and lower bound functions are both piecewise linear, it is possible that exploiting case characteristics will have no effect on the bounds.

Another possibility is to augment the preceding analysis with data that can shed light on the joint distribution of judges' decisions. One approach would entail surveying the judges on how they would decide a sample of the cases in C, as in Partridge and Eldridge (1974). Because the judges' decisions would be simultaneously observable for the surveyed cases, it would be possible to estimate measures of association among the judges' decisions, which could then provide estimates of the joint densities in Table 1.

The common criticism of surveys is that they lack external validity (Diamond and Zeisel 1975; Conley and O'Barr 1988; Sisk, Heise and Morriss 1998; Stith and Cabranes 1998: 109–10; Anderson, Kling, and Stith 1999). Simplified scenarios in written questionnaires may not present the same stimuli as actual cases: judges are not exposed to advocacy from both sides, they are not required to write opinions justifying their decisions, and they do not need to consider the impact of their judgments on actual parties. In some circumstances, judges may be loath to cooperate, especially if they are concerned that the research could support reforms that they oppose. A more credible approach would be to use sentencing councils, in which judges independently recommend sentences for actual offenders and discuss their recommendations with colleagues (Diamond and Zeisel 1975).

Another possibility is to exploit data from appealed cases. Because appeals are typically decided by multimember courts, the appellate judges' decisions will be simultaneously observable, and measures of association among appellate judges could conceivably be extrapolated to estimate the joint distributions among the judges in J. Decisions in appealed cases will have authenticity that survey responses lack, but this approach nevertheless raises concerns about external validity. The appealed cases may not be representative of the cases in C and appellate judges may behave differently from the judges in J. Furthermore, judges on multimember courts typically deliberate and may be strongly influenced by consensus norms (Fischman 2011, forthcoming), so their decisions will not be independent. Whether and how estimates from appellate cases can be extrapolated to improve estimates of inconsistency will require further testing.

III. Indeterminacy and Error

Inconsistency measures how often judges would agree with each other, but not how often their decisions coincide with results required by law. This section demonstrates that judicial compliance with legal obligation can be measured in terms of indeterminacy and error. The indeterminacy rate refers to the proportion of cases in which the law does not compel a unique outcome. The error rate represents the proportion of cases in which the judge's decision is incompatible with the requirements of law.

There have been extensive debates about the extent of legal indeterminacy, as well as its normative implications (e.g., Coleman and Leiter 1993; Kress 1989; Leiter 1995; Singer 1984; Solum 1987). The primary normative concern is that judicial decisions are viewed as less legitimate if they are merely the product of judges' discretion and not determined uniquely by legal rules. Most of these discussions focus on metaphysical indeterminacy (whether legal questions have correct answers), although some, such as Kress (1990), consider epistemic indeterminacy (whether the answers to legal questions are knowable). Both conceptions of indeterminacy are compatible with the framework developed in this section.

A. Empirical Framework

The approach used to derive bounds on the rate of inconsistency can be modified to derive bounds on the rates of indeterminacy and error. Although these rates cannot be measured in isolation—at least without making strong assumptions about what results the law requires—it is possible to derive joint bounds on these rates. As in the previous section, we assume that we have perfect knowledge of each judge's decision rate r_i among all cases in C.

Under any theory of law, we can define a fraction z_0 of cases in C for which a negative outcome is the only legally justifiable result, a fraction z_1 of cases for which a positive outcome is the only justifiable result, and a fraction I of cases in which the law is indeterminate. The exclusivity of these three categories requires that $z_0 + z_1 + I = 1$. For any judge j, consider the following contingency table:

TABLE 3

CONTINGENCY TABLE FOR JUDGE j'S DECISIONS WITH RESPECT TO CORRECT OUTCOME

	Law Requires Negative Outcome	Law Requires Positive Outcome	Law is Indeterminate	Total
j negative ($Y_j = 0$)	p_{00}	p_{01}	p_{0I}	$1 - r_j$
j positive ($Y_j = 1$)	p_{10}	p_{11}	p_{1I}	r_j
Total	z_0	z_1	I	

As in the previous section, we cannot observe the joint probabilities $p_{00}, p_{01}, p_{10}, p_{11}$. The challenge in analyzing indeterminacy and error is that we also cannot observe the marginal probabilities z_0, z_1, and I, since the correctness of outcomes cannot be objectively measured. Let E_j represent the proportion of the cases in C that would be decided erroneously by judge j. Then $E_j = p_{10} + p_{01}$, since these probabilities correspond to the situations in which judge j's decision conflicts with the result required by law.

The Fréchet-Hoeffding bounds for p_{01} and p_{10} are as follows:

$$\max\{r_j + z_0 - 1{,}0\} \leq p_{10} \leq \min\{r_j, z_0\}$$
$$\max\{z_1 - r_j, 0\} \leq p_{01} \leq \min\{1 - r_j, z_1\}$$

Adding these inequalities and substituting $z_0 = 1 - I - z_1$ yields

$$\underline{E}_j(z_1, I) \leq E_j \leq \overline{E}_j(z_1, I), \tag{3}$$

where $\underline{E}_j(z_1, I) = \max\{r_j - z_1 - I, 0\} + \max\{z_1 - r_j, 0\}$ and $\overline{E}_j(z_1, I) = \min\{r_j, 1 - I - z_1\} + \min\{1 - r_j, z_1\}$.

The expected error rate $E = \sum w_j E_j$ is the weighted sum of the individual judges' error rates, where w_j is the proportion of cases decided by judge j. Inequality (3) provides a lower bound on E in terms of z_1 and I:

$$E \geq \sum_j w_j \, \underline{E}_j(z_1, I).$$

We can now construct a lower bound on the expected error rate in terms of the indeterminacy rate.

PROPOSITION 4: Let $\underline{E}(I)$ denote the lower bound on the expected error rate, given a rate of indeterminacy I. Then

$$\underline{E}(I) = \min_{0 \leq z_1 \leq 1 - I} \sum_j w_j \, \underline{E}_j(z_1, I), \tag{4}$$

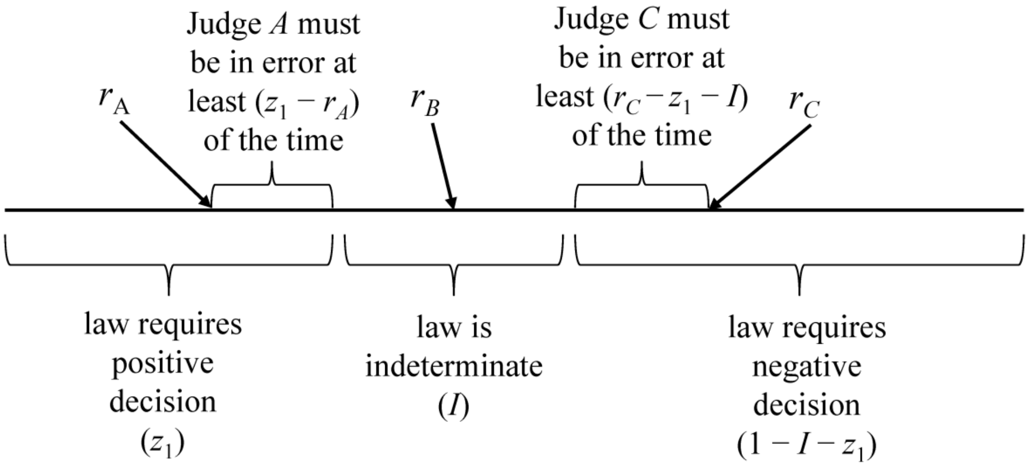
and this lower bound can be achieved for some combination of judicial votes and correct legal outcomes. The function $\underline{E}(I)$ will be nonincreasing in I.

PROOF: See Appendix.

For any given values of $r_1, \cdots, r_n$ and any hypothesized rate of indeterminacy I, this bound can be calculated by evaluating the above expression for all values of z_1 in the range $0 \leq z_1 \leq 1 - I$. The fact that $\underline{E}(I)$ is nonincreasing means that there is an explicit tradeoff between legal indeterminacy and judicial error when interpreting disparity.

The intuition for Proposition 4 is illustrated in Figure 5. Assume that we know the true rate of indeterminacy I and the proportion of cases z_1 in which the only correct outcome is the positive decision. Under these assumptions, a judge who is never in error must have a decision rate satisfying $z_1 < r_j < z_1 + I$. Now suppose that we have three judges i, j, and k as depicted in Figure 3 with $r_i < z_1 < r_j < z_1 + I < r_k$. Under these assumptions, judge i would be wrong in at least a proportion $(z_1 - r_i)$ of the cases in C, corresponding to the second term of $\underline{E}_j(z_1, I)$. Similarly, judge k would be wrong in at least a proportion $(r_k - z_1 - I)$ of the cases, corresponding to the first term of $\underline{E}_j(z_1, I)$. Because judge j's rate is within the permissible range, it is conceivable that judge j is never wrong. In practice, of course, we cannot know I and z_1. But by keeping I fixed and allowing z_1 to vary within the possible range of values, we can derive the lower bound for the expected error rate in terms of the indeterminacy rate.

FIGURE 5
LOWER BOUND ON ERROR FOR INDIVIDUAL JUDGES


r_A
Judge A must be in error at least $(z_1 - r_A)$ of the time
r_B
Judge C must be in error at least $(r_C - z_1 - I)$ of the time
r_C
law requires positive decision (z_1)
law is indeterminate (I)
law requires negative decision $(1 - I - z_1)$

Two particular consequences of Proposition 4 are worth noting. First, consider the Dworkinian thesis that every case has a correct answer. Under this conception, $I = 0$, so equation (4) reduces to

$$\underline{E}(0) = \min_{0 \le z_1 \le 1} \sum_j w_j \left| r_j - z_1 \right|,$$

where z_1 would correspond to the decision rate of a Herculean judge. The above expression is minimized when $z_1 = r^{med}$, the weighted median of the r_j's (Wooldridge 2002: 348). Thus, if $I = 0$,

$$\underline{E}(0) = \sum_j w_j \left| r_j - r^{med} \right| \tag{5}$$

If the decision rate of a Herculean judge deviates from the decision rate of the median judge, then the expected error rate will exceed the lower bound.

Second, consider a skeptical view of law, in which law consists merely of "prophecies of what the courts will do in fact" (Holmes 1897: 994) or "specific past decisions, and guesses as to actual specific future decisions" (Frank 1930: 47). Since such a view of law cannot accommodate the concept of legal error (Bix 2009), it must hold that $E = 0$. This means that there must exist a value of z_1 for which the summation on the right-hand side of equation (4) is always zero. This requires $z_1 \leq r_j \leq z_1 + I$ for every j, which implies that $I \geq r^{max} - r^{min}$, the difference between the maximum and minimum decision rates. Thus, if judges are cannot err, then at least $\left(r^{max} - r^{min}\right)$ of the cases must be indeterminate. This quantity can also be interpreted as a measure of unpredictability, representing the proportion of cases in which at least some judges will disagree as to the outcome.

It is also possible to construct an upper bound on error, which is given by Proposition 5.

Proposition 5: Let $\overline{E}(I) = 1 - I - \underline{E}(I)$. Then the expected rate of error as to result satisfies $\underline{E}(I) \leq E \leq \overline{E}(I)$, and both bounds can be achieved.

Proof: See Appendix.

Note that the error rate discussed here refers only to errors as to result, where the case outcomes can be coded dichotomously. If there are multiple results corresponding to positive and negative decisions—for example, if a court provides relief to a plaintiff, but has a choice regarding remedies—then there could be an error as to the remedy even if the decision corresponds to the correct dichotomous outcome. Thus, the rate of error as to remedy will always be at least as large as the rate of error as to result. Similarly, although an erroneous result necessarily implies incorrect justification, the converse does not hold; an incorrect justification can still lead to a correct result. Thus, the lower bound on the rate of error estimated here will also be a lower bound on the rate of error as to remedy or justification. However, the upper bound given in Proposition 5 will not be an upper bound on the rate of error as to remedy or justification. It is conceivable that *every* decision could be incorrectly justified, even if some fortuitously reached the correct result. The bounds on error as to remedy or justification are summarized in the following corollary.

Corollary 1: The expected rates of error as to remedy or as to justification must satisfy $\underline{E}(I) \leq E \leq 1$, and any rate within this range is possible.

This upper bound is highly conservative, and can only be achieved when the correct answer in each determinate case is contrary to the outcome that a majority of judges would have reached. This will be implausible in most applications but could occur when the law is out of date or when a legal system is dysfunctional or unjust. For the remainder of the discussion, I ignore the upper bound on error, since it only applies as to result and is unlikely to be binding in typical applications.

By evaluating the lower bound $\underline{E}(I)$ on the expected error rate for all values of I between 0 and 1, we can construct an "indeterminacy-error curve." This curve visually depicts how inter-judge disparity can be decomposed into indeterminacy and error. Combinations of indeterminacy and error rates that lie above this curve would be feasible, while those combinations below the curve would not.[7]

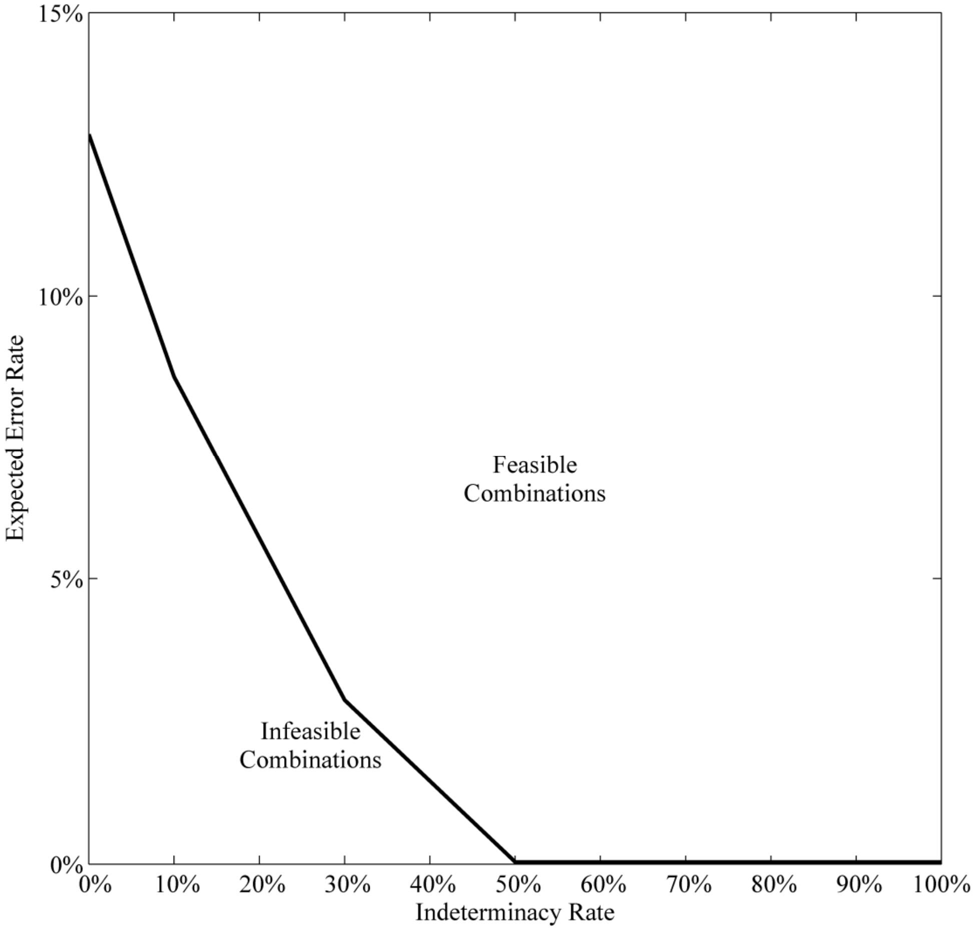
To illustrate, consider the hypothetical court discussed in the previous section, in which the judges have decision rates of 10%, 20%, 25%, 30%, 35%, 50%, and 60%, respectively. Figure 6 depicts an indeterminacy-error curve for this hypothetical court. The region below the curve represents combinations of indeterminacy and error rates that are infeasible given the judges' voting rates, while the region above the curve represents feasible combinations.

If we were to assume that every case has a unique correct outcome, what could we say about the error rate? Clearly, given the disparate decision rates, the judges cannot all be correct in every case. The rate of error will be minimized when the legally correct rate of positive decisions coincides with the decision rate of the median judge. If this occurs—if exactly 30% of the cases require positive decisions—then the error rate will be minimized at 12.9%, the point at which the curve intersects the vertical axis. The error rate, of course, could be much higher: the lower bound coincides with the true error rate only if all judges are monotonic with respect to each other and the median judge is always correct.

Under an assumption of judicial infallibility, the indeterminacy rate must be at least 50%, as shown by the intersection of the curve with the horizontal axis. This bound is determined by the extreme judges—the ones with 10% and 60% decision rates. Since these judges would reach different outcomes 50% of the time, we must acknowledge at least this rate of indeterminacy if neither of these judges is ever wrong.

[7] If we are interested in errors as to result, then it would also be necessary to verify that the error rate does not exceed the upper bound $\overline{E}(I)$.

FIGURE 6
INDETERMINACY-ERROR CURVE FOR HYPOTHETICAL COURT IN WHICH JUDGES HAVE DECISION RATES OF 10%, 20%, 25%, 30%, 35%, 50%, 60%



Since the lower bound $\underline{E}(I)$ is achievable for every value of I, this curve will capture all the information that observational data on judges' decision rates can reveal about rates of indeterminacy and error, in the absence of further assumptions. The methodology developed here can identify which combinations of indeterminacy and error rates are compatible with the data, but cannot say whether any such combination is more plausible than any other. That the methodology does not rely on contestable assumptions about law is therefore both a strength and a weakness. Because it can present empirical conclusions regarding indeterminacy and error in a manner that is uncontroversial, it can establish a "domain of consensus" (Manski 2003: 3) among scholars with sharply divergent views about the nature of law. However, these empirical conclusions must necessarily be weak, since no empirical study can resolve philosophical debates about legal indeterminacy or normative debates about the content of law. Once a domain of consensus is established, observers can supplement the empirical results with subjective assumptions. Proceeding in this manner helps to clarify which claims are based on empirical analysis and which are subjective.

B. Improving the Bounds

As with inconsistency, the joint bounds on indeterminacy and error will typically not be precise enough on their own to justify policy interventions. The estimates can conceivably be improved by using additional data, as in Section II.D. But with indeterminacy and error, it will invariably be necessary to impose additional untestable assumptions about law.

For example, an observer who subscribes to the "right answer thesis" could impose the additional assumption that $I = 0$, yielding the lower bound on error in equation (5). One who believes that legal indeterminacy is a marginal phenomenon could impose an upper bound on the indeterminacy rate. It is also possible to impose bounds on z_0 and z_1, the proportion of cases for which the law requires negative and positive outcome, respectively. Justification for such bounds would necessarily rely on a qualitative analysis of the cases in C.

It may be possible to generate a more useful upper bound on error by incorporating an assumption that if most or all judges would decide a case in the same way, then that decision must be the legally correct one (Mashaw et al. 1978; Greenawalt 1990). Bix (2009) provides a justification for this claim from a positivist perspective: if law is a matter of social fact, then consensus among judges regarding the disposition of cases would constitute the meaning of the law. Thus, if the judges in J are representative of the interpretive community, it would be impossible for all or nearly all of the judges to be in error.

IV. Inference with Observational Data

[NOTE TO NBER PARTICIPANTS: It has come to my attention that there are problems with the way I am using the bootstrap, but I haven't had time to correct it. I suspect the results will not change much. In any event, the sampling variation appears to be quite small relative to the identification uncertainty.]

The discussion in the previous two sections proceeded under a simplifying assumption: that the decision rates r_j of the judges could be known with certainty. Abstracting away all problems of statistical inference, these sections showed what could be identified about the quantities of interest. In practice, of course, the decision rates will not be known precisely; they can only be estimated from observational data. This section discusses how to construct confidence intervals for inconsistency and how to derive an indeterminacy-error curve that accounts for statistical uncertainty about judges' decision rates.

There will typically be a variety of ways of deriving estimates of the r_j's from observational data. If the assignment of cases to judges is fully random, in the sense that each case in C is equally likely to be assigned to each judge in J, then we may simply use the sample decision rate (the proportion of positive decisions) to derive an estimate $\hat{r}_j$ for each judge j. Otherwise, we can use regression models that include variables that explain the likelihood of assignment, such as time periods and districts, to derive an estimate $\hat{r}_j$ for each judge. If assignment is random conditional on these covariates, then it is not necessary to include case characteristics in a regression, although they could potentially increase the efficiency of the estimates. We can then use the estimates $\hat{r}_j$ to construct consistent estimates of $\widehat{\underline{D}}$, $\widehat{\overline{D}}$, $\underline{\hat{E}}(I)$ and $\widehat{\overline{E}}(I)$.

Asymptotically valid confidence intervals can be derived using the bootstrap, as in Horowitz and Manski (2000).[8] Let r denote a column vector of the r_j's. Construct a series of T simulated data sets, drawn from the original data set with replacement, and let r_t^* denote the estimate of r from the t^{th} bootstrap sample. For each r_t^*, we can construct bootstrap estimates of

[8] The confidence intervals are constructed to cover the entire identified region with specified probability, not merely the true parameter. In the case of inconsistency, the confidence intervals will be conservative with regard to the true parameter. There may not exist "true" parameters for indeterminacy and error in an objective sense. Imbens and Manski (2004) provide a method for constructing tighter confidence intervals that cover the true parameter with the specified probability, but their method requires that the estimators for the bounds be asymptotically normal, which will not necessarily be satisfied.

The upper and lower bounds on inconsistency will be asymptotically normal only if all of the r_j's are distinct and $r_i + r_j \neq 1$ for all i, j. If these conditions hold, then the upper and lower bounds on each D_{ij} will be asymptotically normal; see part (i) of Appendix E in Heckman, Smith, and Clements (1997). The bounds on error $\underline{E}(I)$ and $\overline{E}(I)$ will fail to be asymptotically normal for some values of I; this will occur whenever any of the maximization constraints inside the summation in equation (6) are binding. This is most easily seen in the case where $I = r^{max} - r^{min}$, so that $\underline{E}(I) = 0$ and the distribution of $\underline{\hat{E}}(I)$ converges to the distribution of $\max\{(\hat{r}^{max} - \hat{r}^{min}) - (r^{max} - r^{min}), 0\}$. The first term inside the maximization will be asymptotically normal but the asymptotic distribution of $\underline{\hat{E}}(I)$ will be truncated normal.

the bounds $\underline{D}_t^*$, $\overline{D}_t^*$, $\underline{E}_t^*(I)$, and $\overline{E}_t^*(I)$. By repeated bootstrap sampling, we can estimate the distribution of these bounds conditional on the data.

For any parameter of interest, we can generate an interval determined by its upper and lower bound for each bootstrap sample. To construct a $(1-\alpha)$ confidence interval, we then find the smallest interval that fits a proportion $(1-\alpha)$ of these intervals. For example, to generate a 99% confidence interval for D, we generate a series of bootstrap intervals $\left[\underline{D}_t^*, \overline{D}_t^*\right]$, and find the smallest interval that encompasses 99% of these intervals.

Although this methodology provides consistent estimators of the true bounds, there may be substantial finite-sample bias in some applications. This will be especially relevant for the estimates of the lower bounds on inconsistency and error, both of which will typically overstate the true bounds.[9] This can be seen most easily by considering the case in which all judges' true decision rates are exactly equal. The lower bounds on inconsistency and error must be zero, but estimates derived from finite samples will almost always be positive.

The bias will be largest when many judges have similar decision rates and the number of observations is small. To correct for the bias, I adjust all estimates by the bootstrap estimate of bias provided by Efron and Tibshirani (1993: 125).[10] This adjustment will provide valid confidence intervals in most applications, however, the adjustment may still be inadequate in small samples when judges' decision rates are exactly equal or nearly so. More sophisticated adjustments may be necessary in such circumstances.

V. Illustration: Immigration Adjudication

To illustrate how the methodology developed here can be applied to observational data, I derive estimates of inconsistency and construct an indeterminacy-error curve using data involving asylum adjudication in administrative immigration courts. These courts provide a natural application because they hear a large volume of cases—more than 300,000 per year—and

[9] If $\hat{r}$ is an unbiased estimate of r, then the bias on the inconsistency lower bound is positive because $\underline{D}$ is a convex function of r. Thus, by Jensen's inequality, $\mathrm{E}\left[\underline{D}(\hat{r})\right] > \underline{D}(r)$. The lower bound on error is not a globally convex function of r but has many local convexities, so bias will typically be positive in practice.

[10] For example, the estimate of bias for the inconsistency lower bound would take the form $\frac{1}{T}\sum \underline{D}_t^* - \underline{\widehat{D}}$.

a prominent study by Ramji-Nogales, Schoenholz, and Schrag (2007) documented large inter-judge disparities in the resolution of these claims.

Any alien who is physically present in the United States may petition for asylum. In order to meet the statutory requirements for asylum, petitioners must demonstrate that they are "unable or unwilling to return to" their home countries due to "persecution or a well-founded fear of persecution on account of race, religion, nationality, membership in a particular social group, or political opinion" (8 U.S.C. §§ 1101(a)(42)(A)). A grant of asylum permits petitioners to remain in the United States, seek employment, bring certain family members to the United States, and potentially seek permanent residence.

Data on adjudication outcomes from 1996–2004 were obtained from the Executive Office of Immigration Review through a Freedom of Information Act request, and are made available by the organization asylumlaw.org.[11] With a few exceptions,[12] cases are randomly assigned to judges (Ramji-Nogales et al. 2007), however, this applies only within a particular court and time period.[13] Although the validity of the methodology does not require that the cases be similar to each other, I restrict the data in this illustration to a set of homogeneous cases in order to best ensure that each judge hears a comparable mix of cases. First, I examine only cases adjudicated in 2003, the most recent full year for which data are available. Second, I restrict analysis to the New York Immigration Court, which has the highest case volume among the 53 immigration courts. Third, I focus only on petitioners of Chinese origin, who comprise 52% of the claims filed in New York in 2003.[14] Finally, I excluded defensive asylum claims, which are raised during the course of deportation proceedings. Most of these claims involved detained aliens, for which the assignment of judges may be non-random. The remaining affirmative asylum claims comprise 48% of all claims. These cases are referred to immigration court when an affirmative application for asylum is denied and the applicant does not have lawful immigration status.

[11] The data are available at http://www.asylumlaw.org/legal_tools/index.cfm?fuseaction=showJudges2004.

[12] One exception is that one judge within each court may be designated to hear claims involving unaccompanied juveniles (Ramji-Nogales et al. 2007).

[13] This limitation could be addressed in future research by including court and time controls.

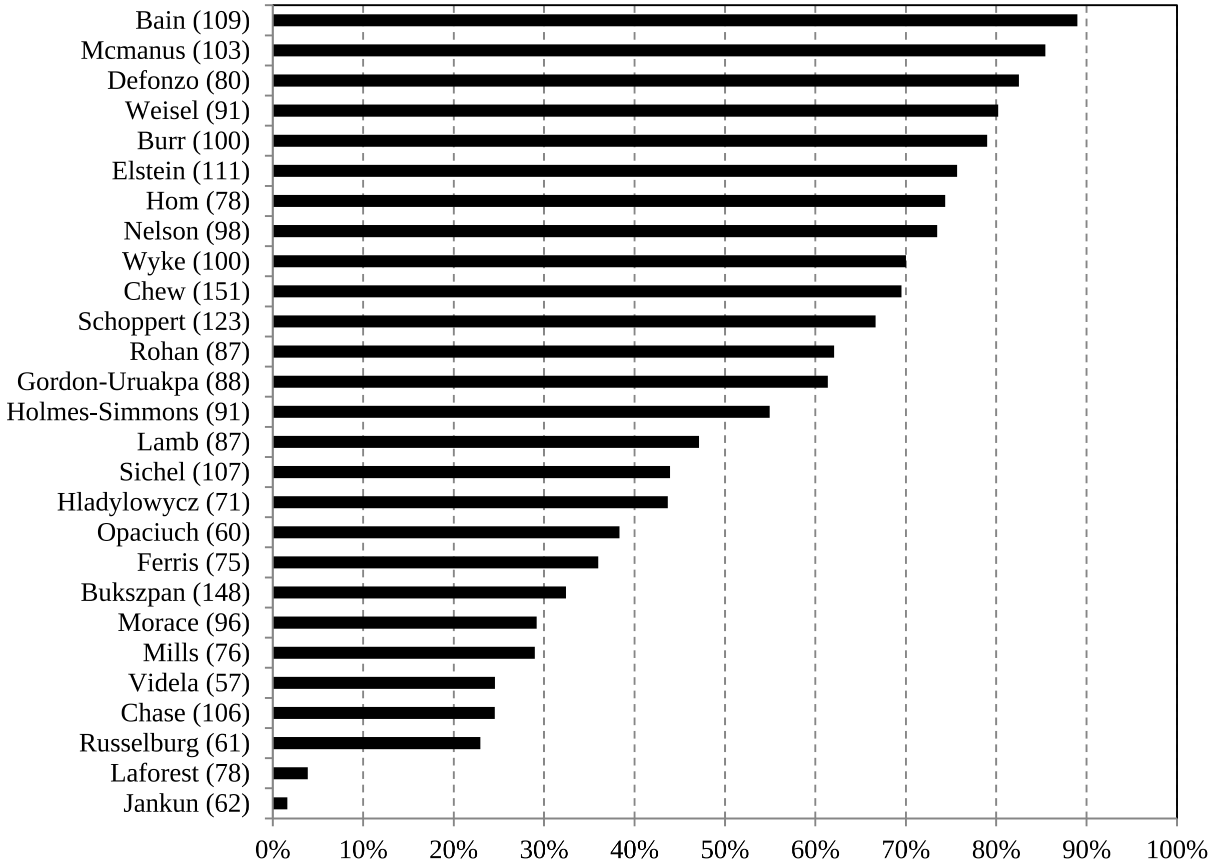
[14] A chi-square test that cases involving claimants of different origins are randomly distributed among judges weakly rejects ($p = 0.06$). This may be due to the fact that the judges were not uniformly active throughout 2003, and that subtle trends in application rates by country of origin could interact with the judges varying activity rates throughout the year. Although deviations from randomness appear to be minor, the assumption that judges' caseloads are comparable is more justified if analysis is restricted to a single country of origin.

There are 2593 cases remaining once the data is restricted to cases involving affirmative claims involving Chinese asylum-seekers in the New York Immigration Court in 2003. The adjudication outcomes were coded in the data in six ways. I classify the outcomes "grant" and "conditional grant" as positive decisions and the outcomes "denied," "abandoned," "withdrawn," and "other" as negative decisions.[15] Figure 7 provides rates of positive decisions ("grant" or "conditional grant") for each of the immigrations judges who decided at least 50 cases in the data. The inter-judge disparities are striking: one judge reaches a positive outcome in 89% of the cases, while another does so in only 2% of the cases. The decision rates, moreover, are spread uniformly throughout this range; the disparities are not merely the result of a few outlier judges. These disparities may be caused by a variety of factors. One such factor may be judges' differing interpretations of the statutory term "well-founded fear of persecution." Another may be that some judges have a greater tendency to find alien's accounts of persecution credible, while others are far more skeptical. Because these asylum cases involve fact-finding as well as legal interpretation, the concepts of indeterminacy and error in this context must be understood to encompass factual as well as legal determinations.

[15] "Conditional grant" can be awarded to aliens raising claims involving coercive population control measures. These grants were conditional due to a quota that limited the number of such claims that could be granted in a given year. A claim is "abandoned" if the applicant fails to appear for a scheduled hearing, whereas "withdrawal" requires an affirmative step by the claimant. The "other" category includes changes of venue as well as other forms of relief besides asylum, such as cancellation of removal or voluntary departure. Many studies drop cases with outcomes classified as "abandoned," "withdrawn," or "other," however these outcomes are not independent of the judges assigned to decide the claims ($p < 0.001$). Claims are more likely to be abandoned or withdrawn when a petitioner is assigned to a judge who is less likely to grant asylum. Thus, dropping these cases would introduce selection bias.

FIGURE 7

RATES OF "POSITIVE DECISIONS" (GRANT OR CONDITIONAL GRANT) FOR NEW YORK IMMIGRATION JUDGES IN AFFIRMATIVE CASES INVOLVING CHINESE ALIENS, 2003



Note: The number of cases decided by each judge is indicated in parentheses. Only judges with at least 50 cases are displayed.

Clearly, the disposition of some of these claims would have depended on the judges to which the claims were assigned. If we imagine a counterfactual in which each of these claims would be assigned randomly to a different judge, how many of them would turn out differently? To answer this question, we need estimates of average inconsistency, which are provided in Table 4.

TABLE 4
BOUNDS FOR AVERAGE INCONSISTENCY: AFFIRMATIVE ASYLUM CASES INVOLVING CHINESE ALIENS, NEW YORK IMMIGRATION COURT, 2003

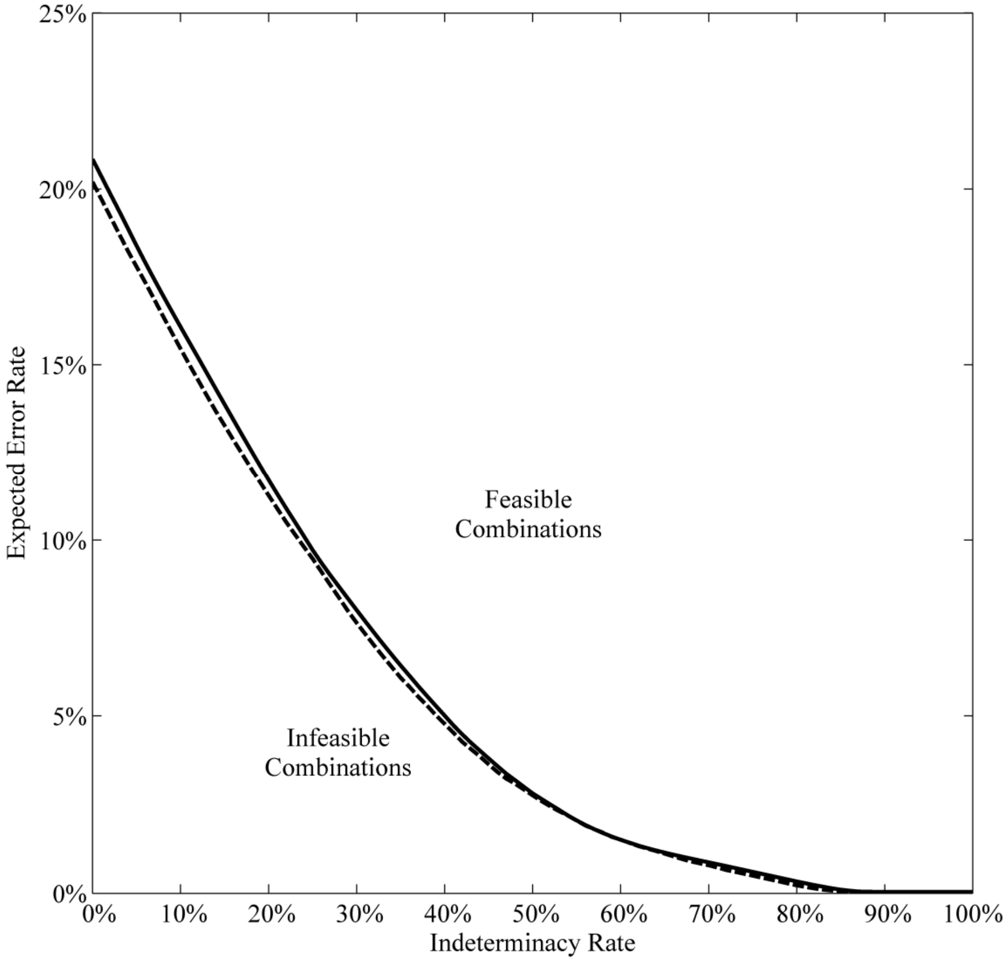
	Lower Bound	Upper Bound
Point Estimates	28.8%	51.7%
99% Confidence Interval	26.6%	51.7%

For the purpose of illustration, I report bounds derived from point estimates as well as a 99% confidence interval. The former are derived under the simplifying assumption that the judges' sample decision rates are their true decision rates. The confidence intervals reported on the right side are computed according to the procedure outlined in Section IV. Due to the large amount of data, the two sets of bounds are similar, but as a general matter, confidence intervals will be wider than bounds derived from point estimates. The results reveal an average inconsistency measure between 26.6% and 51.7%. Thus, a randomly selected pair of judges would disagree about the disposition of a randomly selected case at least one-quarter of the time, and perhaps as often as one-half of the time.

The inconsistency results suggest that some proportion of the cases must either be indeterminate or wrongly decided. This can be seen more clearly in the indeterminacy-error curve, which is provided in Figure 8. As with inconsistency, I report a curve derived from the sample decision rates (the solid line) as well as a 99% confidence curve (the dashed line), derived from simulations using the methodology in Section IV. Due to the large amount of data, the difference is slight, but the confidence curve will always be more conservative.

FIGURE 8

INDETERMINACY-ERROR CURVE: AFFIRMATIVE ASYLUM CASES INVOLVING CHINESE ALIENS, NEW YORK IMMIGRATION COURT, 2003



Note: Solid line is based on sample decision rates. Dashed line is a 99% confidence curve.

The curve illustrates how the disparities among the judges' decision rates can be decomposed into combinations of indeterminacy and error. If we assume that each case has only one legally correct result, then we should expect that at least 20% of the cases will be wrongly decided. If we assume, on the other hand, that 20% of the cases are indeterminate, then we should expect that at least 10% of decisions are incorrect. The expected error rate will always be positive unless at least 84% of the cases are indeterminate. This implies that all judges in the sample would agree on the same disposition in no more that 16% of the cases.

Empirical studies of immigration adjudication have focused on inconsistency rather than indeterminacy or error, in large part because of the difficulty in relating the latter concepts to empirical findings. Yet normative claims about how immigration adjudication ought to be reformed are necessarily informed by concerns about all three concepts. For example, Ramji-Nogales et al. (2007) and Legomsky (2007) acknowledge that bureaucratic controls on adjudicators would increase consistency, but oppose them because they might increase error rates and threaten process values such as adjudicator independence. Ramji-Nogales et al. advocate better screening of immigration judges, more training and resources, and closer appellate scrutiny of asylum determinations, all of which would likely decrease the error rate among determinate cases. But to the extent that cases are indeterminate and decisions are purely a product of judicial discretion, it is less obvious why that discretion should be exercised by an appellate court or an adjudicator with better legal training. Legomsky argues that such proposals would only result in "marginal improvements" (445), in part because he takes the position that many of the cases are indeterminate (425).

Of course, assessing the proportion of cases that are indeterminate and setting tolerable rates of inconsistency and error will necessarily be subjective (Mashaw 1983: 149–51). Such determinations cannot be made in the abstract; they must rely on knowledge of the law, experience with the types of cases under examination, and recognition of the capacities of adjudicators. Nevertheless, it may be easier to reach agreement on these determinations—or at least to clarify grounds for disagreement—than it would be to specify a standard for correctness.

VI. Conclusion

In recent years, the empirical scholarship on judicial behavior has generated numerous claims that case dispositions depend, to some degree at least, on the identities of the adjudicators. Some studies have compared individual judges; others have found significant differences along party, gender, or racial lines or by composition of circuit court panels. These studies typically estimate and compare judges' decision rates, but require readers to assess whether the reported disparities are substantively meaningful and what the normative implications might be.

I have argued that such comparisons of judicial decision rates alone are not sufficient to answer many important questions about law and judicial behavior. In contexts in which legal predictability and comparative justice are important normative values, studies of judicial decision making should report bounds on the rate of inconsistency. On the other hand, empirical claims about the extent to which judicial decisions are justified by legal rules should be characterized in terms of indeterminacy and error. The relationship between judicial voting rates and the feasible rates of inconsistency, indeterminacy, and error is sufficiently complex that readers of empirical studies cannot be expected to reach intuitive conclusions from summary statistics.

It may seem unfortunate that these rates cannot be precisely estimated and that bounds may be quite wide in practice. Further developments may enable stronger inferences by exploiting additional sources of data or by incorporating defensible assumptions about law and judicial behavior. Nevertheless, it is important to recognize the limits of empirical analysis. Much can be gained by distinguishing claims based on credible empirical inference from those based on untestable behavioral or jurisprudential assumptions.

Appendix

PROOF OF PROPOSITION 1: It was demonstrated in the text that equation (2) is always satisfied. It remains to be shown that both bounds can be achieved. For any rate of inconsistency $D_{ij} \in [\underline{D}_{ij}, \overline{D}_{ij}]$, let $p_{00} = 1 - \frac{1}{2}(r_i + r_j + D_{ij}), p_{01} = \frac{1}{2}(r_j - r_i + D_{ij}), p_{10} = \frac{1}{2}(r_i - r_j + D_{ij})$, and $p_{11} = \frac{1}{2}(r_i + r_j - D_{ij})$. Then it can be verified that $p_{01} + p_{10} = D_{ij}$, all of the joint probabilities are bounded between 0 and 1, and the joint probabilities sum to the correct marginal probabilities.

PROOF OF PROPOSITION 2: When $D_{ij} = \underline{D}_{ij}$, it follows from above that either $p_{01} = 0$ or $p_{10} = 0$. When $D_{ij} = \overline{D}_{ij}$, it follows from above that either $p_{00} = 0$ or $p_{11} = 0$.

PROOF OF PROPOSITION 3: Represent C by the unit interval and let $C_j = \{c \mid Y_i(c) = 1\}$, so that $\Pr(c \in C_j) = r_j$. The lower bound on average inconsistency can be derived by averaging the lower bound in equation (2) over all pairs of judges. The lower bound can be achieved by setting $C_j = [0, r_j]$.

Deriving the upper bound is more complicated, since the average of the pairwise upper bounds is not necessarily achievable. Let $R = \sum r_j$ and let $q_k = \Pr(\sum_{i=1}^{n} Y_i = k)$ denote the proportion of cases with exactly k positive votes. Then $D_{ij} = E\left[(Y_i - Y_j)^2\right]$ and $D = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} E\left[(Y_i - Y_j)^2\right] = \frac{2}{n(n-1)}\left(nR - \sum_{i=1}^{n} \sum_{j=1}^{n} E[Y_i Y_j]\right)$.
Now $\sum_{i=1}^{n} \sum_{j=1}^{n} E[Y_i Y_j] = E[\sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j] = E[(\sum_{i=1}^{n} Y_i)^2] = \sum_{k=1}^{n} q_k k^2$, so

$$D = \frac{2}{n(n-1)}\left(nR - \sum_{k=1}^{n} q_k k^2\right). \qquad \text{(A1)}$$

Thus, maximizing D is equivalent to minimizing $\sum_{k=1}^{n} q_k k^2$. Constraints on the q_k terms are obtained by observing that $\sum_{k=1}^{n} q_k = 1$ and $\sum_{k=1}^{n} q_k k = R$. This leads to the constrained optimization problem[16]

$$\min_{q_1,\cdots,q_n} \sum_{k=1}^{n} q_k k^2 \text{ such that } \sum_{k=1}^{n} q_k k = R, \sum_{k=1}^{n} q_k = 1, \text{and } q_k \geq 0 \text{ for all } k.$$

The Lagrangian takes the form

$$L = \sum_{k=1}^{n} q_k k^2 - \mu_1 \left(\sum_{k=1}^{n} q_k k - R \right) - \mu_2 \left(\sum_{k=1}^{n} q_k - 1 \right) - \sum_{k=1}^{n} \lambda_k q_k,$$

and the first-order conditions on the q_k terms are

$$k^2 - \mu_1 k - \mu_2 = \lambda_k. \tag{A2}$$

This means that $\lambda_k = 0$ for at most two distinct values of k, since the expression on the left side of (A2) has at most two distinct roots. It follows from the complementary slackness conditions that $q_k > 0$ for at most two distinct values of k.

Suppose $q_i, q_j > 0$ with $i > j$. Then $q_i + q_j = 1$ and $q_i i + q_j j = R$ implies that $q_i = \frac{R-j}{i-j}$ and $q_j = \frac{i-R}{i-j}$ and $j < R < i$. Substituting the above expressions for q_i, q_j into the minimand yields the expression $R(i+j) - ij$. This expression is strictly increasing in i and decreasing in j, so the constraints $j < R < i$ must be binding. For non-integer R, it follows that $i = \lfloor R \rfloor + 1$ and $j = \lfloor R \rfloor$. When R is an integer, a similar approach shows that $q_R = 1$ and $q_k = 0$ for $k \neq R$. In either case, the minimand reduces to $R(2\lfloor R \rfloor + 1) - \lfloor R \rfloor(\lfloor R \rfloor + 1)$. Substituting this into the right-hand side of (A1) yields the upper bound in Proposition 3.

It only remains to be shown that there exist $C_1, \cdots, C_n$ that achieve the minimizing values of $q_1, \cdots, q_n$. Define $f(x) = x - \lfloor x \rfloor$, so that the range of $f(x)$ is the unit interval. Let $B_j = \left[\sum_{i=1}^{j-1} r_i, \sum_{i=1}^{j} r_i\right)$ and let $C_j = f(B_j)$, so that $\mu(C_j) = r_j$. Then any $x \in [0, R - \lfloor R \rfloor)$ will be included in exactly $\lfloor R \rfloor + 1$ of the C_j's, and any $x \in [R - \lfloor R \rfloor, 1]$ will be included in exactly $\lfloor R \rfloor$ of the C_j's. Thus $q_{\lfloor R \rfloor} = \lfloor R \rfloor + 1 - R$, $q_{\lfloor R \rfloor + 1} = R - \lfloor R \rfloor$, and $q_k = 0$ for all other k, as required.

PROOF OF PROPOSITION 4: It was demonstrated in the text that the bound always holds. To show that it can be achieved, let $C_j = [0, r_j]$. Let Z_0, Z_1 denote the sets of cases in which the law

[16] Note that there are additional upper bounds on the q_k terms, which are difficult to characterize. To provide one example, it must always hold that $q_n \leq r^{min}$. I proceed by ignoring these constraints in the optimization and then demonstrating that the minimum can be achieved.

requires a negative and positive decision, respectively, and let Z_I denote the set of cases in which the law is indeterminate. If $Z_1 = [0, z_1^*]$, $Z_I = (z_1^*, z_1^* + I]$, and $Z_0 = (z_1^* + I, 1]$, where z_1^* is the value of z_1 that minimizes the expression in equation (4), then the lower bound will be achieved. Thus, over the subset of determinate cases, all judges must be monotonic with respect to each other and with a hypothetical judge who is always correct.

To show that $\underline{E}(I)$ is nonincreasing, suppose that z_1^* minimizes the expression in (4) for some I. Then for $I' > I$,

$$\underline{E}(I') \leq \sum_j w_j \left[\max\left\{r_j - z_1^* - I', 0\right\} + \max\left\{z_1^* - r_j, 0\right\}\right] \leq \underline{E}(I).$$

Proof of Proposition 5: Following the same procedure used to derive the lower bound on error yields $E \leq \sum_j w_j \overline{E}_j(z_1, I)$, and $\overline{E}(I) = \max_{0 \leq z_1 \leq 1-I} \sum_j w_j \overline{E}_j(z_1, I)$.

Note that

$$\begin{aligned}
&\underline{E_j}(z_1, I) + \overline{E}_j(1 - I - z_1, I) \\
&\qquad = \max\left\{r_j - z_1 - I, 0\right\} + \max\left\{z_1 - r_j, 0\right\} + \min\left\{r_j, z_1\right\} \\
&\qquad + \min\left\{1 - r_j, 1 - I - z_1\right\} \\
&\qquad = \left[\max\left\{r_j - z_1 - I, 0\right\} + \min\left\{r_j - z_1 - I, 0\right\} + 1 - r_j\right] \\
&\qquad + \left[\max\left\{z_1 - r_j, 0\right\} + \min\left\{z_1 - r_j, 0\right\} + r_j\right] = [1 - z_1 - I] + z_1 = 1 - I.
\end{aligned}$$

It follows that $\sum_j w_j \underline{E_j}(z_1, I) + \sum_j w_j \overline{E}_j(1 - I - z_1, I) = 1 - I$. Thus, if z_1^* minimizes $\sum_j w_j \underline{E_j}(z_1, I)$, then $1 - I - z_1^*$ will maximize $\sum_j w_j \overline{E}_j(z_1, I)$, and $\sum_j w_j \underline{E_j}(z_1^*, I) + \sum_j w_j \overline{E}_j(1 - I - z_1^*, I) = 1 - I$. Hence $\underline{E}(I) + \overline{E}(I) = 1 - I$.

Appendix Table: Description of Variables

C	Set of cases or legal questions to be analyzed
J	Set of judges who could decide the cases in C
n	Number of judges
$Y_j(c)$	Decision judge j would reach in case c (equals zero or one)
D_{ij}	Proportion of cases in C in which judges i and j would reach different outcomes ("pairwise inconsistency")
D	Average rate of pairwise inconsistency among all pairs of judges in J ("average inconsistency")
r_j	Judge j's hypothetical rate of positive decisions among the cases in C
$p_{00}, p_{01}, p_{10}, p_{11}$	Joint probabilities of outcomes for a pair of judges
$\underline{D}_{ij}, \overline{D}_{ij}$	Lower and upper bounds on pairwise inconsistency
$\underline{D}, \overline{D}$	Lower and upper bounds on average inconsistency
R	Sum of the judges' decision rates r_j
z_0	Proportion of cases in C in which a negative outcome is the only correct answer
z_1	Proportion of cases in C in which a positive outcome is the only correct answer
I	Proportion of cases in C in which the outcome is indeterminate
w_j	Proportion of cases in C decided by judge j
E_j	Judge j's error rate of among the cases in C
$\underline{E}_j, \overline{E}_j$	Lower and upper bounds on judge j's error rate
E	Expected proportion of cases in C that were wrongly decided
$\underline{E}(I), \overline{E}(I)$	Lower and upper bounds on the error rate (as a function of the indeterminacy rate)
$\hat{r}_j$	Estimate of judge j's true decision rate r_j, derived from decisions observed in the data
$\widehat{\underline{D}}, \widehat{\overline{D}}, \hat{\underline{E}}(I), \widehat{\overline{E}}(I)$	Estimates of the various lower and upper bounds, derived from the estimated decision rates $\hat{r}_j$
$\underline{D}_t^*, \overline{D}_t^*, \underline{E}_t^*(I), \overline{E}_t^*(I)$	Estimates of the various lower and upper bounds, derived from the tth bootstrap sample

Bibliography

Anderson, James M., Jeffrey R. Kling, and Kate Stith. 1999. Measuring Interjudge Disparity: Before and After the Federal Sentencing Guidelines. *Journal of Law and Economics* 42:271–307.

Bix, Brian H. 2009. Global Error and Legal Truth. *Oxford Journal of Legal Studies* 29(3):535–547.

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. Untangling the Causal Effects of Sex on Judging. *American Journal of Political Science* 54:389–411.

Cohen, Felix. 1935. Transcendental Nonsense and the Functional Approach. *Columbia Law Review* 35:809–49.

Coleman, Jules L. and Brian Leiter. 1993. Determinacy, Objectivity, and Authority. *University of Pennsylvania Law Review* 142:549–637.

Conley, John M. and William M. O'Barr. 1988. Fundamentals of Jurisprudence: An Ethnography of Judicial Decision Making in Informal Courts. *North Carolina Law Review* 66:467–507.

Cross, Frank B. and Emerson H. Tiller. 1998. Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals. *Yale Law Journal* 107:2155–76.

Dalton, Clare. 1985. An Essay in the Deconstruction of Contract Doctrine. *Yale Law Journal* 94(5):997–1114.

Diamond, Shari Seidman and Hans Zeisel. 1975. Sentencing Councils: A Study of Sentence Disparity and its Reduction. University of Chicago Law Review 43:109–149.

Dworkin, Ronald. 1986. *Law's Empire*. Cambridge, Mass.: Harvard University Press.

Edwards, Harry T. and Michael A. Livermore. 2009. Pitfalls of Empirical Studies that Attempt to Understand the Factors Affecting Appellate Decisionmaking. *Duke Law Review* 58:1895–1989.

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Everson, George. 1919. The Human Element in Justice. *Journal of the American Institute of Criminal Law and Criminology* 10:90–99.

Feinberg, Joel. 1974. Noncomparative Justice. *Philosophical Review* 83(3):297–338.

Fischman, Joshua B. and David S. Law. 2009. What Is Judicial Ideology, and How Should We Measure It? *Washington University Journal of Law and Policy* 29:133–214.

Fischman, Joshua B. 2011. Estimating Preferences of Circuit Judges: A Model of Consensus Voting. *Journal of Law and Economics* 54:781–809.

Fischman, Joshua B. Forthcoming. Interpreting Circuit Court Voting Patterns: A Social Interactions Framework. *Journal of Law, Economics and Organization.*

Frank, Jerome. 1930. *Law and the Modern Mind*. New York: Brentano's.

Frankel, Marvin E. 1973. *Criminal Sentences: Law Without Order*. New York: Hill and Wang.

Fréchet, M. 1951. Sur les Tableux de Corrélation Dont les Marges sont Données. *Annales de l'Universite de Lyon A*, Series 3, 14:53–77.

Friedman, Barry. 2006. Taking Law Seriously. *Perspectives on Politics* 4(2):261–76.

Galligan, D.J. 2010. Legal Theory and Empirical Research, in Peter Cane and Herbert M. Kritzer, eds., *The Oxford Handbook of Empirical Legal Research*. Oxford: Oxford University Press.

Gaudet, Frederick J., George S. Harris and Charles W. St. John. 1933. Individual Differences in the Sentencing Tendencies of Judges. *Journal of Criminal Law and Criminology* 23(5):811–818.

Greenawalt, Kent. 1990. How Law Can Be Determinate. *UCLA Law Review* 38:1–86.

Hart, H.L.A. 1961. *The Concept of Law*. Oxford: Clarendon Press.

Harvey, Anna and Michael J. Woodruff. Forthcoming. Confirmation Bias in the United States Supreme Court Judicial Database. *Journal of Law, Economics, and Organization.*

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. *Review of Economic Studies* 64(4):487–535.

Hoeffding, W. 1940. Masstabinvariate Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, 5(3), 179–233. Reprinted as "Scale-Invariant Correlation Theory," *The Collected Works of Wassily Hoeffding* (1994), edited by Fisher, N.I. and P.K. Sen, pp. 57–107, New York: Springer.

Holland, Paul W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81:945–60.

Holmes, Jr., Oliver Wendell. 1897. The Path of the Law. *Harvard Law Review* 10:457–78.

Horowitz, Joel L. and Charles F. Manski. 2000. Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data. *Journal of the American Statistical Association* 95:77–84.

Imbens, Guido W. and Joshua D. Angrist. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62(2): 467−475.

Imbens, Guido W. and Charles F. Manski. 2004. Confidence Intervals for Partially Identified Parameters. *Econometrica* 72(6):1845–57.

Joe, Harry. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman-Hall.

Kairys, David. 1984. Law and Politics. *George Washington Law Review* 52(2):243–62.

Kress, Ken. 1989. Legal Indeterminacy. *California Law Review* 77:283–337.

Kress, Ken. 1990. A Preface to Epistemological Indeterminacy. *Northwestern University Law Review* 85(1):134–47.

Kruskal, William H. 1958. Ordinal Measures of Association. *American Statistical Association Journal* 53:814–51.

Kysar, Douglas A. 2007. The Jurisprudence of Experimental Law and Economics. *Journal of Institutional and Theoretical Economics* 163:187–98.

Landes, William M. and Richard A. Posner. 2009. Rational Judicial Behavior: A Statistical Study. *Journal of Legal Analysis* 1:775–831.

Legomsky, Stephen H. 2007. Learning to Live with Unequal Justice: Asylum and the Limits to Consistency. *Stanford Law Review* 60:413–74.

Leiter, Brian. 1995. Legal Indeterminacy. *Legal Theory* 1:481–492.

Llewellyn, Karl. 2011. *The Theory of Rules*. Chicago: University of Chicago Press.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer.

Mashaw, Jerry L. 1983. *Bureaucratic Justice: Managing Social Security Disability Claims*. New Haven, Conn.: Yale University Press.

Mashaw, Jerry L., Charles J. Goetz, Frank I. Goodman, Warren F. Schwartz, Paul R. Verkuil, and Milton M. Carrow. 1978. *Social Security Hearings and Appeals: A Study of the Social Security Administration Hearing System*. Lexington, Mass.: Lexington Books.

Miles, Thomas J. and Cass R. Sunstein. 2006. Do Judges Make Regulatory Policy? An Empirical Investigation of Chevron. *University of Chicago Law Review* 73:823–82.

Miles, Thomas J. and Cass R. Sunstein. 2008. The Real World of Arbitrariness Review, 75 *University of Chicago Law Review* 761.

Moore, Underhill, and Theodore S. Hope, Jr. 1929. An Institutional Approach to the Law of Commercial Banking. *Yale Law Journal* 38(6):703–19.

Neyman, J., with K. Iwaszkiewicz and S. Kolodziejczyk. 1935. Statistical Problems in Agricultural Experimentation. *Supplement to the Journal of the Royal Statistical Society* 2(2):107–54.

Partridge, Anthony, and Eldridge, William B. 1974. *The Second Circuit Sentencing Study: A Report to the Judges*. Washington, D.C.: Federal Judicial Center.

Priest, George L. and Benjamin Klein. 1984. The Selection of Disputes for Litigation. *Journal of Legal Studies* 13(1):1–55.

Ramji-Nogales, Jaya, Andrew Schoenholtz, and Philip G. Schrag. 2007. Refugee Roulette: Disparities in Asylum Adjudication. *Stanford Law Review* 60:295–412.

Revesz, Richard L. 1997. Environmental Regulation, Ideology, and the D.C. Circuit. *Virginia Law Review* 83:1717–72.

Rubin, Donald B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 6(5): 688–701.

Rubin, Donald B. 1980. Discussion of “Randomization Analysis of Experimental Data in the Fisher Randomization Test” by Basu. *Journal of the American Statistical Association* 75:591−593.

Shapiro, Carolyn. 2009. Coding Complexity: Bringing Law to the Empirical Analysis of the Supreme Court. *Hastings Law Journal* 60:477–540.

Shapiro, Scott J. 2011. *Legality*. Cambridge, Mass.: The Belknap Press of Harvard University Press.

Singer, Joseph William. 1984. The Player and the Cards: Nihilism and Legal Theory. *Yale Law Journal* 94(1):1–70.

Sisk, Gregory C., Michael Heise, and Andrew P. Morriss. 1998. Charting the Influences on the Judicial Mind: An Empirical Study of Judicial Reasoning. *New York University Law Review* 73:1377–1500.

Solum, Lawrence B. 1987. On the Indeterminacy Crisis: Critiquing Critical Dogma. *University of Chicago Law Review* 54:462–503.

Spaeth, Harold. 2011. *The Supreme Court Database*. Available at http://scdb.wustl.edu/data.php (accessed July 1, 2011).

Stith, Kate and José A. Cabranes. 1998. *Fear of Judging: Sentencing Guidelines in the Federal Courts*. Chicago: University of Chicago Press.

Sunstein, Cass R., David Schkade, Lisa M. Ellman, and Andres Sawicki. 2006. *Are Judges Political?: An Empirical Analysis of the Federal Judiciary*. Washington, DC: Brookings Institution Press.

Tamanaha, Brian Z. 2009. *Beyond the Formalist-Realist Divide: The Role of Politics in Judging*. Princeton, N.J.: Princeton University Press.

Tamer, Elie. 2010. Partial Identification in Econometrics. *Annual Review of Economics* 2:167–95.

Tiller, Emerson H. and Frank B. Cross. 1999. A Modest Proposal for Improving American Justice. *Columbia Law Review* 99:215–34.

Waldfogel, Joel. 1998. Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines? *International Review of Law and Economics* 18:293−304.

Waldron, Jeremy. 2007. Lucky in Your Judge. *Theoretical Inquiries in Law* 9:185–216.

Westen, Peter. 1982. The Empty Idea of Equality. *Harvard Law Review* 95(3):537–96.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press.