Preliminary draft. Do not cite or circulate.

Inconsistency, Indeterminacy, and Error in Adjudication

Joshua B. Fischman

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, is that law is always determinate, although it does not always constrain judicial decisions in practice.

Despite the rapid growth in empirical scholarship on judicial decision making, little progress has been made in measuring the influence of law on judicial decisions. Many empirical studies have demonstrated that case outcomes vary significantly depending on the characteristics of the deciding judges (e.g., Revesz 1997; Cross and Tiller 1998; Sunstein et al. 2006; Cox and Miles 2008; Boyd, Epstein, and Martin 2010), while others have documented disparities in decision rates among individual adjudicators (e.g., Everson 1919; Gaudet, Harris, and St. John 1933; Mashaw et al. 1978 ; Ramji-Nogales, Schoenholz, and Schrag 2007). These findings, however, only address the impact of law in an indirect way. To be sure, the existence of significant behavioral differences among judges repudiates the claim that judging is objective and mechanical, but this claim has never had many prominent adherents (Tamanaha 2010, 27–43). The important empirical questions are not *whether* the identity of the presiding judge may influence the outcome of adjudication or *whether* law may fail to constrain some judicial decisions, but rather *how much* and *how often* (Ibid., 145–48).

One reason why answers have been elusive is that these inquiries 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

will depend on which sources of law are considered authoritative and what obligations they establish. Theories of law that include moral principles as valid sources (e.g., Dworkin 1986) will determine unique outcomes more frequently than narrow conceptions of law that rely exclusively on legal texts.

The second component addresses the question, "How often do judges deviate from their legal obligations?" Although this inquiry is essentially empirical, it presupposes a fully specified theory of law. 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 second component of the above inquiry without resolving the first. I do so by analyzing three qualities of adjudication—*inconsistency*, *indeterminacy*, and *error*—and clarifying what can be revealed about them using empirical analysis of observational data. I use the term *inconsistency* to measure 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*, which addresses the first component discussed above, measures the proportion of cases in which the law fails to require a unique result. *Error* relates to the second component, measuring the proportion of cases in which the judge's decision conflicts with the result required by law.

Because inconsistency is formulated purely in terms of judicial decisions, and not with respect to legal obligation, it would appear to be naturally suited to empirical analysis. 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, might seem to be entirely outside the province of empirical inquiry (Edwards and Livermore 2009; cf. Kysar 2007). I show, however, that although these qualities cannot be empirically analyzed 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 intuition for the methodology runs as follows: if two judges disagree about a legal question, then either the law is indeterminate, at least one of the judges is wrong, or both. Disagreement cannot occur if the law is determinate and both judges are correct. Similarly, if two judges exhibit significant behavioral disparities in randomly assigned cases, then this must point to some combination of indeterminacy and error. The methodology developed in this paper determines precisely which combinations are feasible.

Although inconsistency is often used as a measure the effect of legal constraint, I argue that it can be misleading in this regard, since judges may exhibit similar behavior due to legal constraint or extralegal influences. If adjudication is inconsistent, then law cannot fully constrain judicial decisions, but the converse does not hold. Nevertheless, inconsistency is of interest for two additional reasons: it provides a measure of predictability in adjudication and the degree to which like cases are treated alike.

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.

The organization of this article proceeds as follows. Section II begins with the meaning and implications of inconsistency, and provides the conceptual framework necessary to operationalize 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 a simplified setting in which cases are undifferentiated and data are unlimited. Drawing upon the framework used to analyze inconsistency, and proceeding under the same assumptions, 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 inconsistency,

indeterminacy, and error rates from actual observational data. Section V provides an illustration of the framework developed in this article using data on administrative asylum adjudication. Section VI concludes. Most mathematical derivations are in a separate appendix.

II. Inconsistency

Many empirical studies find disparities among judges' decision rates, and conclude that law fails to fully constrain judges' decisions. Such claims are easy to understand: if judges reach significantly different rates of outcomes in randomly assigned cases, then their decisions cannot be consistent with each other, and also cannot be uniquely determined by legal reasons. Nevertheless, I argue that statistics on inter-judge disparities can provide misleading measures of inconsistency, since two judges could have identical decision rates yet frequently disagree with each other.

Furthermore, inconsistency can be misleading indicator of legal constraint, since adjudication could be highly consistent even when law fails to constrain judicial decisions.[1] To illustrate, imagine a hypothetical court with two judges, both of whom always decided 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 proportion of correct decisions would also increase.

Although inconsistency provides evidence that law does not fully constrain judges, this example demonstrates that it cannot be used as a basis for comparing different courts or examining changes in the same court over time. 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

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

opportunity to conform their behavior to its requirements. To the extent that the outcomes of potential disputes 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). If one defendant receives a longer prison term than another equally culpable defendant, only because the first defendant faced a harsher judge, then most observers would perceive such a disparity to be unjust. There is substantial disagreement, however, regarding whether the comparative discrepancy is itself a form of injustice, or merely evidence that one of the sentences was not determined according to law.[2]

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 require equal treatment among certain parties.

a. Empirical Framework

In order to derive empirical estimates of inconsistency, it is first necessary to specify the context in which it will be measured. 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 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

[2] For example. some scholars (e.g., Westen 1982) have argued that comparative justice is a superfluous concept. According to this view, the two defendants can be deemed to be equally culpable only if the law prescribes the same sentence for both; if their sentences differ, it is sufficient to note that one of them has not been treated according to the requirements of law. If the law governing these sentences is indeterminate, however, or if both sentences are wrongly determined, unequal sentences for comparable offenders could still give rise to claims of comparative injustice.

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.[3] In the current article, I shall treat the cases in C as undifferentiated; future work may consider how to incorporate information about particular cases into the framework developed here.

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. 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.[4] To maintain generality, we refer to decisions as "positive" or "negative." Depending on the context, a positive decision might mean that a certain type of plaintiff obtained some kind of relief, or that the judge's decision coincides with the result that political liberals would prefer. The validity of the analysis does not depend on the choice of coding method, although 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. Define *average inconsistency* D as 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.

The idealized conditions under which we conceptualize inconsistency are meant to capture three essential properties. The first property, which I call *joint observability*, holds whenever we can simultaneously observe the decisions that two judges would reach in the same

[3] 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.

[4] I shall also make the simplifying assumptions 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.

case. The second property, which I call *autonomy*, requires that one judge's decision will not influence, or be influenced by, any other judge's decisions. The third property, which I refer to as *authenticity*, requires that the decisions should correspond to those that would be rendered in a real-life setting.

In practice, of course, it is infeasible to conduct an ideal experiment that satisfies all three of these properties. It is important, therefore, to consider what sort of inference can be made about inconsistency in non-ideal settings. For the purpose of this analysis, most empirical studies of judicial behavior can be grouped into three broad categories on the basis of which of the three above properties is omitted.

The first category consists of experimental studies that use questionnaires to elicit how judges would adjudicate simulated cases (e.g., Austin and Williams 1977; Clancy et al. 1981; Guthrie, Rachlinski, and Wistrich 2007; Kapardis and Farrington 1981; Partridge and Eldrige 1974; Van Koppen and Ten Kate 1984; Wistrich, Guthrie, and Rachlinski 2005). If the questionnaires are properly administered and each judge responds to each question, then the autonomy and joint observability properties will both be satisfied. For the purpose of measuring inconsistency, such studies are ideal: one need only compare any two judges' survey responses and see how often they disagree.[5]

The common criticism of experimental studies is that they lack authenticity (Anderson, Kling, and Stith 1999; Conley and O'Barr 1988; Sisk, Heise and Morriss 1998; Stith and Cabranes 1998, 109–10). Highly simplified scenarios presented in written questionnaires may not present the same stimuli as actual cases. In a laboratory setting, judges are not exposed to advocacy from both sides, are not required to write opinions justifying their decisions, and do not need to consider the impact of their judgments on actual parties. Further concerns about authenticity arise when surveys are conducted on students rather than actual judges.

The second category of studies, which will be the focus of this article, consists of observational studies of single-judge adjudication. These studies typically involve single judges

[5] There is a vast literature on measures of interrater agreement that are applicable when decisions are jointly observable. For a survey, see Banerjee et al. (1999).

deciding non-identical but randomly assigned cases. Such studies will satisfy the autonomy[6] and authenticity properties, but not joint observability: for any case c, we will only observe $Y_j(c)$ for a single judge j. I show that it is possible to make inferences about how often judges i and j *would* disagree about the correct disposition of case c in this category of studies, even though we can never simultaneously observe $Y_i(c)$ and $Y_j(c)$. However, it is not possible to know when the two judges would disagree or what their grounds of disagreement would be.

The final category of studies consists of observational studies of multi-member courts that decide cases en banc, such as the U.S. Supreme Court (Martin and Quinn 2002; Bailey 2007), federal circuit courts sitting en banc (George 1998), or the Canadian Supreme Court (Alarie and Green 2007). Authenticity and joint observability will be satisfied in these studies, since each judge takes a public position in every case. The autonomy condition will be violated, however, since such courts decide cases through deliberation. Evidence suggests that this deliberative influence will lead to greater conformity among judges' observed votes than would occur under autonomous decision making (Fischman 2011a, b).

In this final category, observed disagreement will provide a lower bound for inconsistency. This lower bound might only be weakly informative, especially if decisions are announced under a strong norm of consensus. It may be possible to provide stronger bounds on inconsistency if data on interim decisions are available, such as conference votes of Supreme Court justices (Brenner 1980; Epstein, Segal, and Spaeth 2001; Post 2001).

b. Pairwise Inconsistency

In studies lacking joint observability, we can never simultaneously observe $Y_i(c)$ and $Y_j(c)$ for two judges i, j. The challenge of making inference about inconsistency in such studies can be understood in terms of the causal model of Neyman (1935) and Rubin (1974), where

[6] Although judges will not be influenced by other judges' hypothetical decisions in the same case, autonomy could potentially be violated if judges decisions are influenced by the precedential authority of other judges' decision in other cases in C. 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 regression with time controls may be adequate to control for the effects of precedent.

cases are the units of observation and each judge is a distinct "treatment." 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 "potential outcomes." This leads to what Holland (1986) refers to as the "fundamental problem of causal inference"; it is impossible to observe the effect of assigning a case to judge i as opposed to judge j if we cannot observe both judges' decisions.

Although the effect of judicial assignment cannot be observed in this context, it is possible to use statistical inference when the assignment of cases is independent of the potential outcomes, as is typical when cases are randomly assigned.[7] In particular, it is possible to estimate an *average treatment effect* $\mathrm{E}\left[Y_j(c) - Y_i(c)\right]$, representing the average effect of reassigning a case from judge i to judge j, by comparing the two judges' rates of reaching positive decisions. 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).

Unfortunately, inconsistency cannot be estimated in this manner, since it cannot be expressed as an average treatment effect. However, it is possible to derive bounds on inconsistency by exploiting information about the judges' rates of reaching positive decisions. For each judge j, let r_j represent the proportion of cases in C in which the judge 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.

To clarify what can be identified about inconsistency between judges i and j from the data lacking joint observability, it is helpful to suppose that we have *exact* knowledge of r_i and r_j. These rates specify the *marginal densities* of Y_i and Y_j, but the inconsistency rate D_{ij} is determined by the *joint density* of Y_i and Y_j, which is not observed. This is illustrated in the following contingency table, in which the joint densities are denoted $p_{00}, p_{01}, p_{10}, p_{11}$.

[7] When case assignment is non-random but unconfounded (Rubin 1990), it may also be possible to employ matching methods, as in Boyd, Epstein, and Martin (2010).

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

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

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

PROPOSITION 1: Define $\underline{D}_{ij}$, the lower bound on pairwise inconsistency, and $\overline{D}_{ij}$, the upper bound on pairwise inconsistency, as follows:

$$\begin{aligned} \underline{D}_{ij} &= |r_i - r_j| \\ \overline{D}_{ij} &= \min\{r_i + r_j, 2 - r_i - r_j\}. \end{aligned} \tag{2}$$

Then the rate of pairwise inconsistency D_{ij} between any two judges i and j satisfies

$$\underline{D}_{ij} \leq D_{ij} \leq \overline{D}_{ij}, \tag{3}$$

and both bounds can be achieved.

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. 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 jointly observable, their voting patterns are widely known, and most readers have an intuitive sense for what it means for a vote to be "liberal."[8] In this instance, because the justices' votes are jointly 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%	

[8] In some instances, intuition could be misleading; the directional coding of case outcomes in the Spaeth Supreme Court Database does not always conform to readers' expectations. For criticism of the coding methods used in the database, see Harvey and Woodruff (2011); Landes and Posner (2010); Shapiro (2009).

The true rate of disagreement between Justice Thomas and Justice Ginsburg is 41%, which can be found by adding the entries in which their decisions differ. This rate of disagreement is eight percentage points higher than the lower bound, but still quite far from the upper bound. Although this result may not seem surprising, intuition alone could not have revealed the precise rate of disagreement. In other applications, where outcomes are not jointly observable and judges are less well-known, intuition may be an even weaker guide.

The conditions under which the true degree of inconsistency will achieve one of the bounds can be most easily formulated using the concepts of *concordance* and *discordance*. Two judges i and j are *concordant* if judge i reaches positive decisions more frequently in cases in which judge j reaches positive decisions; they are *discordant* if judge i reaches positive decisions more frequently in cases in which judge j reaches negative decisions (Kruskal 1958). Note that concordance differs from agreement; if two judges would rank the cases in similar order but have different thresholds for reaching a positive decision, then they could be highly concordant and yet disagree most of the time.

Two judges are *perfectly concordant* if the decisions of one judge are always at least as positive as the decisions of the other judge. Judges i and j are *perfectly discordant* if judge i would be perfectly concordant with a judge who always disagrees with judge j. In the contingency table in Table 1, perfect concordance corresponds to the scenario where either $p_{01} = 0$ or $p_{10} = 0$. Perfect discordance corresponds to the situation where either $p_{00} = 0$ or $p_{11} = 0$, i.e., one of the probabilities on the diagonal is zero. These definitions are formalized below.

DEFINITION: Given judges i and j, we say that judges i and j are *perfectly concordant* 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 *perfectly discordant* 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.

PROPOSITION 2: If judges i and j are perfectly concordant, then the rate of 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

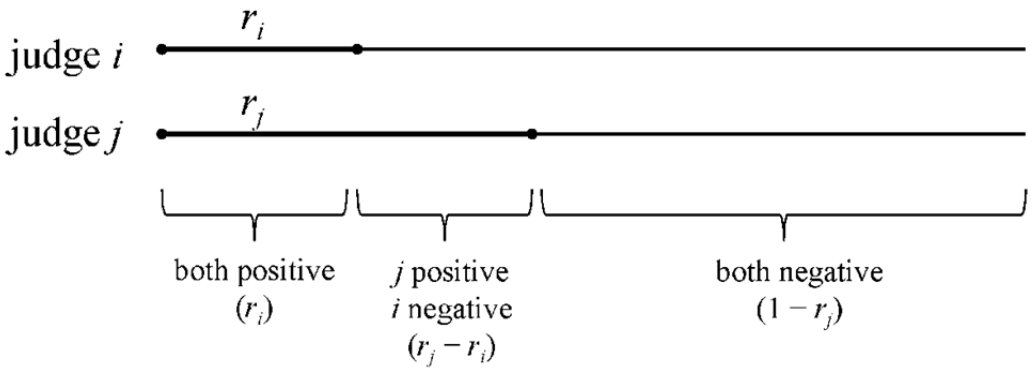
perfectly discordant, then the rate of inconsistency D_{ij} between i and j achieves the upper bound in Proposition 1 (i.e., $D_{ij} = \overline{D}_{ij}$). Under either of these assumptions, D_{ij} will be point-identified.

PROOF: See Appendix.

Figure 1 illustrates perfectly concordant voting patterns in which the lower bound is achieved. The lines represent the space of cases, with the heavily shaded bars representing each judge's positive votes. In this illustration, judge j always reaches a positive decision whenever judge i does; thus $Y_i(c) \leq Y_j(c)$ for all c. In a proportion r_i of the cases, both judges reach a positive decision, and in a proportion $1 - r_i$ of the cases, both reach a negative decision. Disagreement between judges i and j occurs only in the proportion $r_j - r_i$ of cases in which judge j reaches a positive decision and judge i reaches a negative decision.

FIGURE 1

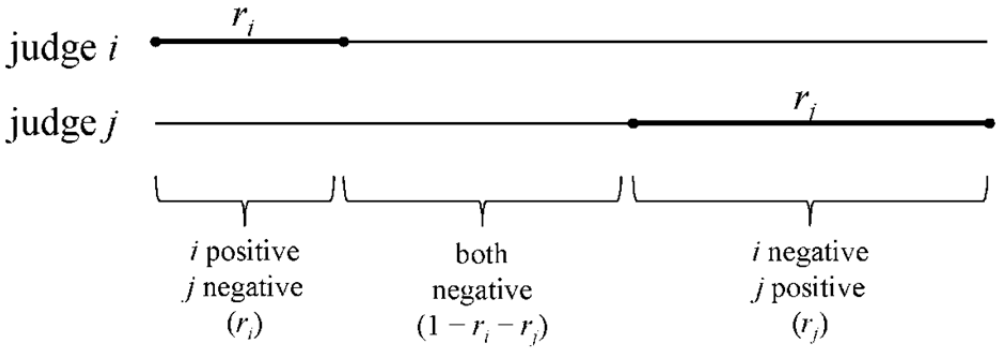
LOWER BOUND ON PAIRWISE INCONSISTENCY: PERFECT CONCORDANCE



The upper bound in Proposition 1 is achieved when the judges are perfectly discordant. Figure 2 provides one such representation of perfectly discordant voting patterns. Here, there is no overlap between the positive votes of judge i and the positive votes of judge j (i.e., $p_{11} = 0$), and the two judges disagree in a proportion $r_i + r_j$ of the cases.

FIGURE 2

UPPER BOUND ON PAIRWISE INCONSISTENCY: PERFECT DISCORDANCE



To many readers, the alignment represented in Figure 2 may seem less plausible than the alignment represented in Figure 1. Any claim about concordance between judges, however, is not empirically testable unless we either have more detailed information about the cases or a representative sample of cases in which decisions are jointly observable. Although intuition about judicial behavior may help to narrow the range of inconsistency, it is at best an imperfect guide. Although it may be reasonable to presume some degree of concordance among judges, it is rare that judges will be perfectly concordant. This would require that all judges could be mapped perfectly onto a left-right spectrum, a notion that Edwards and Livermore (2010, 1916) dismiss as "absurd." Once we acknowledge that perfect concordance is unlikely, intuition cannot tell us exactly how close or far the degree of inconsistency will be from the lower bound.

The degree of concordance among judges, moreover, will be contingent on how the set C of cases is defined and how the outcomes Y_i are coded. Intuition and experience suggest that ordering should be more concordant 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, concordance may be weaker when C is defined broadly to encompass cases involving many types of issues.

Similarly, the degree of concordance will depend on how the outcomes Y_i are coded. To illustrate, suppose that C consists of discrimination cases and that 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 judges' decisions would be reasonably concordant in typical cases. If C includes reverse discrimination cases, however, there may be

some degree of discordance among the judges' decisions: those least likely to provide relief in typical cases may be more sympathetic to plaintiffs with reverse discrimination claims. It might be possible to increase concordance by coding decisions for plaintiffs in reverse discrimination cases as negative decisions, so that the coding aligns with commonly understood liberal/conservative distinctions.

In this example, the degree of concordance depends upon how C is defined, what proportion of C consists of reverse discrimination cases, and how the outcomes are coded. Intuition about judicial behavior and estimates of concordance from other contexts may only be of limited use. Even if intuition suggests that the degree of inconsistency is close to the lower bound, it may be impossible to know precisely how close it is.

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

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

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

where $\lfloor R \rfloor$ denotes the largest integer less than or equal to R. Then the rate of average inconsistency 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 perfectly concordant.

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 perfectly discordant with each other.[9] 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 perfectly discordant with each other, but it would be impossible for a third judge to be perfectly discordant 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. Then it follows from Proposition 3 that average inconsistency must range between 21.0% and 50.5%.

III. Indeterminacy and Error

The same approach used to derive bounds on the rate of inconsistency can also be used 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 without making strong assumptions. 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, there must exist a fraction z_0 of cases in C for which a negative outcome is the only legally justifiable result, a fraction z_1 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:

[9] 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).

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 densities $p_{00}, p_{01}, p_{10}, p_{11}$. In this case, we also cannot observe the marginal densities z_0, z_1, and I, since legal justification 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 densities correspond to the situations in which judge j's decision conflicts with what the law requires.

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

$$\underline{E_j}(z_1, I) \leq E_j \leq \underline{E_j}(z_1, I), \tag{4}$$

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 (4) 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

$$\underline{E}(I) = \min_{0 \le z_1 \le 1-I} \sum_j w_j \, \underline{E}_j(z_1, I), \tag{5}$$

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.

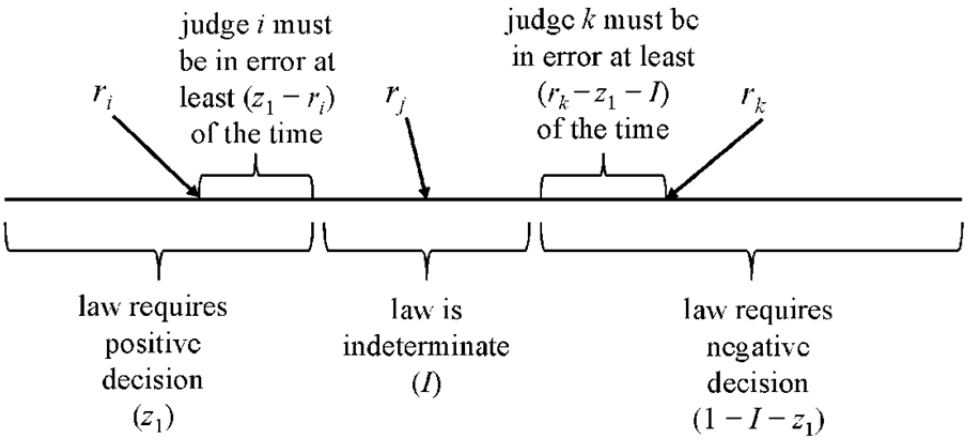
PROOF: See Appendix.

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 $0 \le z_1 \le 1 - I$. 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 3. 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 inside the summation in inequality (5). Similarly, judge k would be wrong in at least a proportion $(r_k - z_1 - I)$ of the cases, corresponding to the first term inside the summation. 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 3

ILLUSTRATION OF THE LOWER BOUND ON ERROR



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 that equation (5) 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,

$$\underline{E}(0) = \sum_j w_j \left| r_j - r^{med} \right| \text{ if } I = 0. \tag{6}$$

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 (5) 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 infallible, 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. However, this upper bound only applies to errors as to result, and not as to remedy or justification.

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. According to some points of view, *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.

In most applications, the upper bound will not be interest, since it 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 unless a legal system is deeply dysfunctional or grossly unjust.[10] 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,[11] while those combinations below the curve would not be.

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 4 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 had a unique correct outcome, what could we say about the error rate? Clearly, given their disparate decision rates, these 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 perfectly concordant 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

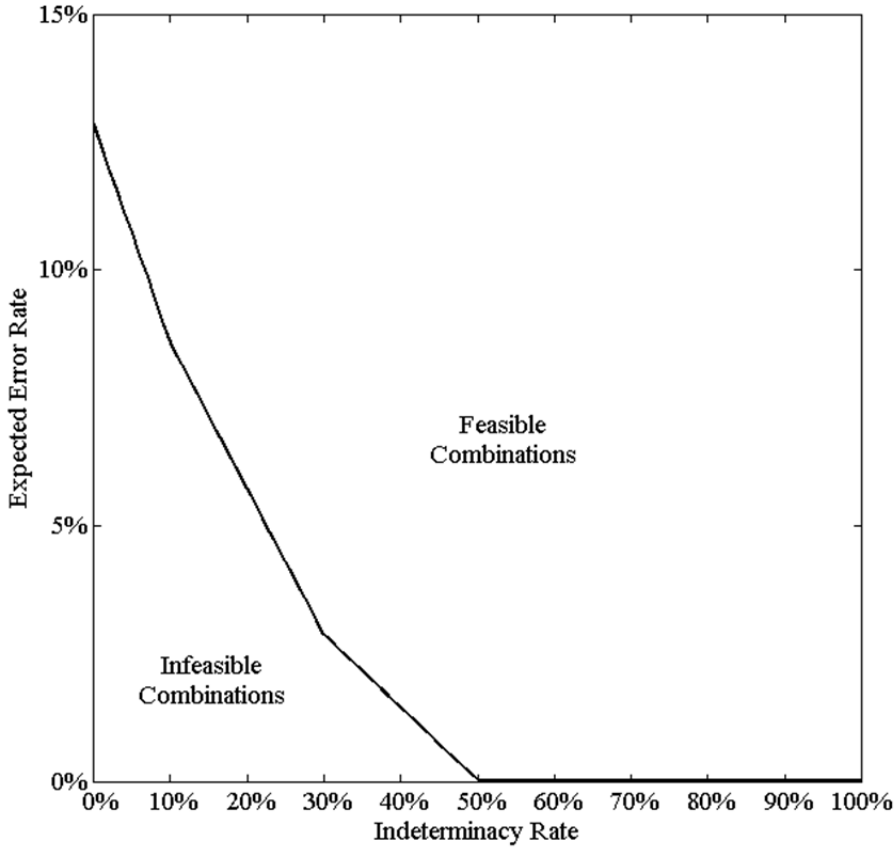
[10] It may be possible to generate a more useful upper bound on error by establishing a presumption that if most judges would decide a case in the same way, then that decision must be the legally correct one, as in Mashaw et al. (1978) or Greenawalt (1990).

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

different outcomes 50% of the time, we must acknowledge at least this rate of indeterminacy if neither of these judges is ever wrong.

FIGURE 4

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



This curve will capture all the information that empirical data on judges' decision rates can reveal about rates of indeterminacy and error, since the lower bound $\underline{E}(I)$ is achievable for every value of I. 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 constestable assumptions about law is 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 limited,

since no empirical study can resolve philosophical debates about legal indeterminacy or normative debates about the content of law.

Once this domain of consensus is established, observers can supplement the empirical results with subjective assumptions or qualitative observations. Proceeding in this manner helps to clarify which claims are based on empirical analysis and which are subjective. 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 (6). One who believes that legal indeterminacy is a marginal phenomenon could impose an upper bound on the indeterminacy rate. A lawyer who is familiar with the cases analyzed in the data might propose a minimum or maximum permissible rate of positive decisions.

IV. Inference with Observational Data

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 $\underline{\hat{D}}$, $\widehat{\overline{D}}$, $\underline{\hat{E}}(I)$ and $\widehat{\overline{E}}(I)$.

Asymptotically valid confidence intervals can be derived using the bootstrap, as in Horowitz and Manski (2000).[12] 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 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.[13] This can be seen most easily by considering the case in which all judges true

[12] 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 always 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.

[13] 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 that bias will typically be positive in practice.

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).[14] 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 because recent empirical scholarship (Ramji-Nogales, Schoenholz, and Schrag 2007) has 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 standard 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

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

by the organization asylumlaw.org.[15] With a few exceptions,[16] cases are randomly assigned to judges within each court (Ramji-Nogales et al. 2007). I restrict the data set in several ways in order to focus on a homogeneous set of cases and to avoid violations of random assignment. 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. Both of these restrictions are essential, since cases are only randomly assigned within a particular court and time period.[17] Third, I focus only on petitioners of Chinese origin, who comprise 52% of the claims filed in New York in 2003.[18] Finally, I excluded "defensive" asylum claims, which are raised during the course of deportation proceedings. Some of these claims involved detained aliens, for which assignment may be non-random. The remaining "affirmative" asylum claims comprise 48% of the claims.

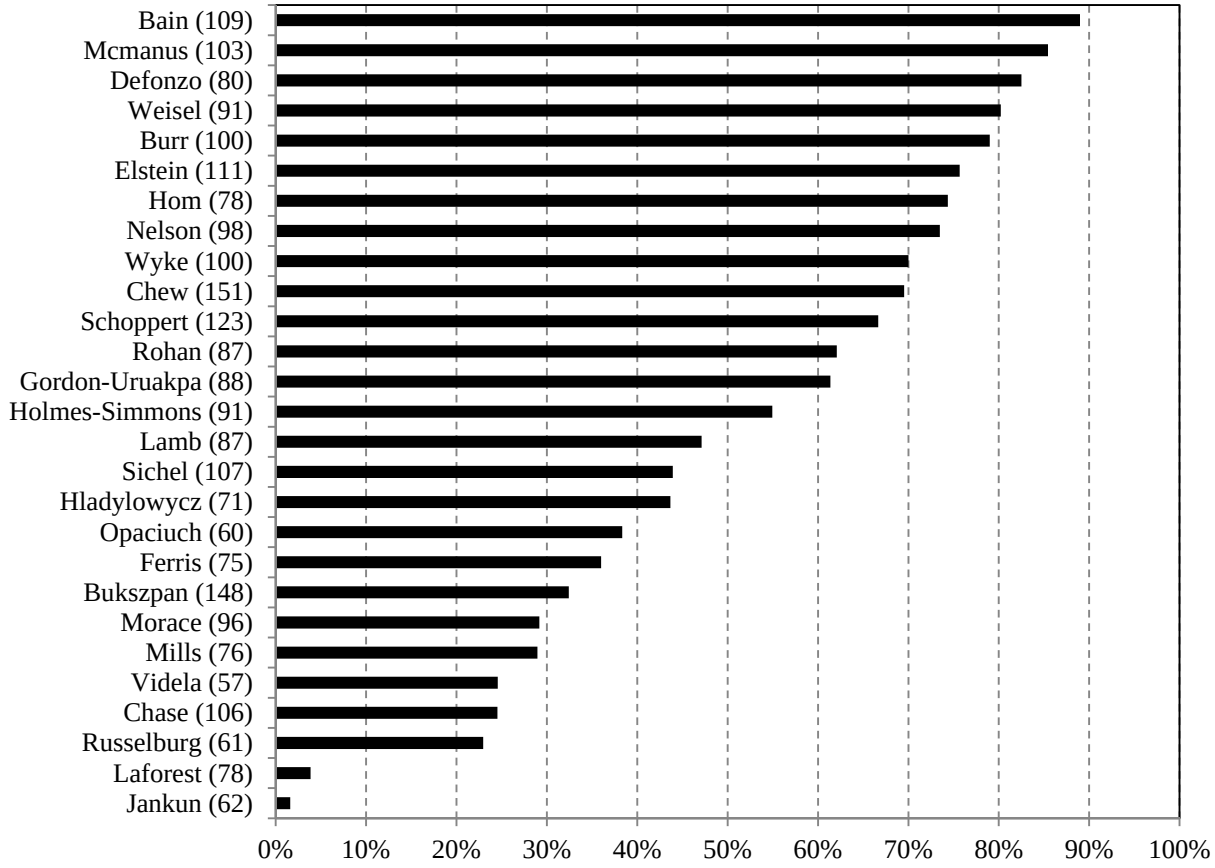
There are 2593 cases remaining once the data is restricted to cases involving affirmative claims involving Chinese asylum-seekers in the New York immigration court in 2003. The adjudication outcomes were coded in the data in six ways. I classify the outcomes "grant" and "conditional grant" as positive decisions and the outcomes "denied," "abandoned," "withdrawn," and "other" as negative decisions.[19] Figure 5 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 in only 2%. The decision rates, moreover, are spread uniformly throughout this range; the disparities are not merely the consequence of a few outlier judges.

[15] The data set can be found at http://www.asylumlaw.org/legal_tools/index.cfm?fuseaction=showJudges2004.
[16] One exception is that one judge within each court may be designated to hear claims involving unaccompanied juveniles (Ramji-Nogales et al. 2007).
[17] The issues could be addressed with court and time controls, but at the expense of additional complexity.
[18] 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.
[19] "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 in venue as well as other forms of relief besides asylum, such as cancellation of removal. 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.

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.

FIGURE 5

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.

It is clear from Figure 5 that the disposition of many of these claims would have depended on the judge to which the claim was 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 inconsistency, which are provided in Table 4.

TABLE 4

ESTIMATES OF AVERAGE INCONSISTENCY:
AFFIRMATIVE ASYLUM CASES INVOLVING CHINESE ALIENS
NEW YORK IMMIGRATION COURT, 2003

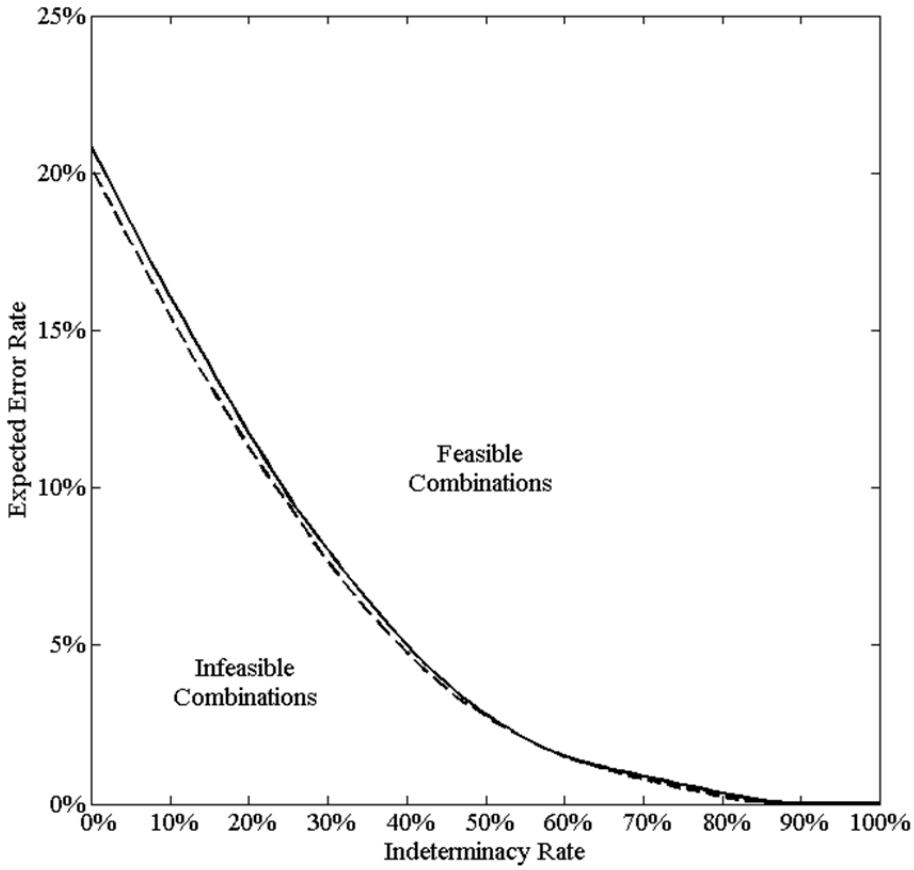
	Lower Bound	Upper Bound
Point Estimates	28.8%	51.7%
99% Confidence Interval	26.6%	51.7%

For the purpose of illustration, I report bounds based on point estimates as well as a true 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 V. 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 6. 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 6

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. Under an assumption that the law is fully determinate, for example, it is expected that at least 20% of the decisions will be wrong. If we assume that at most 20% of the cases are indeterminate, then at least 10% of cases are expected to be wrongly decided. The expected error rate will always be positive unless at least 84% of the cases are indeterminate. Thus, there are at most 16% of the cases in which all of the judges would agree on the same disposition.

To a large degree, the recent debate on reforming immigration adjudication has focused on inconsistency rather than accuracy, and commentators have pointedly refrained from drawing conclusions about error from empirical findings of inter-judge disparity (e.g., Legomsky 2007; Ramji-Nogales et al. 2007). The indeterminacy-error curve can enable this debate to address questions of accuracy, while forcing observers to clarify their beliefs about indeterminacy and tolerable rates of error.

VI. Conclusion

The methodology developed in this article demonstrates that it is possible to use empirical data on judicial decision to generate estimates of about inconsistency, indeterminacy, and error that are suitable for normative evaluation. In particular, it is possible to make rigorous inferences about these rates without making assumptions about judicial behavior and without specifying a theory of law. On the other hand, this article also clarifies the limits of empirical analysis. Although the framework developed here can estimate bounds for inconsistency and joint bounds for indeterminacy and error, it also demonstrates that it is impossible to distinguish among values within the feasible range with typical observational data.

One final limitation is worth emphasizing: this article only addresses estimation of inconsistency, indeterminacy, and error within the set of cases under examination. Of course, cases that are litigated or appealed will not be representative of the universe of potential disputes (Priest and Klein 1984). Whether, and how, these estimates can be extrapolated is a challenging problem that will require careful attention to the process by which disputes are selected for litigation.

Bibliography

Alarie, Benjamin R.D. and Andrew Green. 2007. "The Reasonable Justice: An Empirical Analysis of Frank Iacobucci's Career on the Supreme Court of Canada." *University of Toronto Law Journal* 57:195–226.

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.

Austin, William and Thomas A. Williams III. 1977. "A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity." *Journal of Criminal Law and Criminology* 68(2):306–10.

Bailey, Michael A. 2007. "Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3):433–48.

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeny, and Debajyoti Sinha. 1999. "Beyond Kappa: A Review of Interrater Agreement Measures". *Canadian Journal of Statistics* 27(1):3–23.

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.

Brenner, Saul. 1980. "Fluidity on the United States Supreme Court: A Reexamination." *American Journal of Political Science* 24(3):526–35.

Clancy, Kevin, John Bartolomeo, David Richardson, and Charles Wellford. 1981. "Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity." *Journal of Criminal Law and Criminology* 72(2):524–54.

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.

Cox, Adam B. and Thomas J. Miles. 2008. "Judging the Voting Rights Act." *Columbia Law Review* 108:1–54.

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.

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.

Epstein, Lee, Jeffrey A. Segal, and Harold J. Spaeth. 2001. "The Norm of Consensus on the U.S. Supreme Court." *American Journal of Political Science* 45:362–77.

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. Forthcoming. "Estimating Preferences of Circuit Judges: A Model of 'Consensus Voting.'" *Journal of Law and Economics*.

Fischman, Joshua B. 2011. "Interpreting Circuit Court Voting Patterns: A 'Social Interactions' Framework," working paper available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1636002.

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

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. *Journal of Criminal Law and Criminology* 23(5):811–818.

George, Tracey E. 1998. "Developing a Positive Theory of Decisionmaking on U.S. Courts of Appeals." *Ohio State Law Journal* 58:1635–96.

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

Guthrie, Chris, Jeffrey J. Rachlinksi, and Andrew J. Wistrich. 2007. "Blinking on the Bench: How Judges Decide Cases." *Cornell Law Review* 93:1–43.

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 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.

Kapardis, Andreas, and David P. Farrington. 1981. "An Experimental Study of Sentencing by Magistrates." *Law and Human Behavior* 5:107–21.

Kruskal, William H. 1958. "Ordinal Measures of Association." *American Statistical Association Journal* 53:814–851.

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.

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

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

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10:134–53.

Mashaw, Jerry L., Charles J. Goetz, Frank I. Goodman, Warren F. Schwartz, and Paul R. Verkuil. 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.

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.

Post, Robert. 2001. "The Supreme Court Opinion as Institutional Practice: Dissent, Legal Scholarship, and Decisionmaking in the Taft Court." *Minnesota Law Review* 85:1267–1390.

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. 1990. "Formal Mode of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25(3): 279–92.

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

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.

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.

Van Koppen, Peter J. and Jan Ten Kate. 1984. "Individual Differences in Judicial Behavior: Personal Characteristics and Private Law Decision-Making." *Law and Society Review* 18:225–47.

Wistrich, Andrew J., Chris Guthrie, and Jeffrey J. Rachlinksi. 2005. "Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding." *University of Pennsylvania Law Review* 153:1251–1345.

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.

Appendix

PROOF OF PROPOSITION 1: It was demonstrated in the text that equation (3) is always satisfied. It remains to be shown that both bounds can be achieved. For any rate of inconsistency $D_{ij} \in \left[\underline{D}_{ij}, \overline{D}_{ij}\right]$, let $p_{00} = 1 - \frac{1}{2}\left(r_i + r_j + D_{ij}\right), p_{01} = \frac{1}{2}\left(r_j - r_i + D_{ij}\right), p_{10} = \frac{1}{2}\left(r_i - r_j + D_{ij}\right)$, and $p_{10} = \frac{1}{2}\left(r_i + r_j - D_{ij}\right)$. Then it can be verified that $p_{01} + p_{10} = D_{ij}$, all of the joint densities are bounded between 0 and 1, and the joint densities sum to the correct marginal densities.

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 represent the set of cases in which judge j reaches a positive decision, where C_j has measure r_j. The lower bound in on average inconsistency can be derived by averaging the lower bound in equation (3) over all pairs of judges. The lower bound can be achieved by setting $C_j = \left[0, r_j\right]$.

Deriving the upper bound is more complicated, since the average of the pairwise upper bounds is not necessarily achievable. Let $R = \sum r_j$, let $Q_k = \left\{c \mid \sum_i 1_{C_i}(c) = k\right\}$ be set of cases that would garner exactly k positive decisions among the judges, and let $q_k = \mu(Q_k)$, where μ denotes Lebesgue measure.

Now $D = \frac{2}{n(n-1)} \sum_{i<j} D_{ij} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}$, since $D_{ii} = 0$. Also, $D_{ij} = \mu\left(C_i \cup C_j - C_i \cap C_j\right) = r_i + r_j - 2\mu\left(C_i \cap C_j\right)$, so

$$D = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} = \frac{2}{n(n-1)} \left[nR - \sum_{i=1}^{n} \sum_{j=1}^{n} \mu\left(C_i \cap C_j\right) \right]. \tag{7}$$

Thus, maximizing D is equivalent to minimizing $\sum_{i,j} \mu\left(C_i \cap C_j\right)$.

Now

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mu\left(C_i \cap C_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \int_C 1_{C_i}(c) 1_{C_j}(c) dc = \int_C \left(\sum_{i=1}^{n} 1_{C_i}(c) \right) \left(\sum_{j=1}^{n} 1_{C_j}(c) \right) dc$$

$$= \int_C \left(\sum_{i=1}^{n} 1_{C_i}(c) \right)^2 dc = \sum_{k=1}^{n} q_k k^2.$$

Also,

$$\sum_{k=1}^{n} q_k k = \int_C \left(\sum_{i=1}^{n} 1_{C_i}(c) \right) dc = \sum_{i=1}^{n} \int_C 1_{C_i}(c) dc = \sum_{i=1}^{n} r_i = R, \tag{8}$$

and

$$\sum_{k=1}^{n} q_k = 1, \tag{9}$$

since the Q_k's are disjoint and $\bigcup_{k=1}^{n} Q_k = C$.

This leads to the constrained optimization problem[20]

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

The Lagrangian takes the form

$$L = \sum q_k k^2 - \mu_1 \left(\sum q_k k - R \right) - \mu_2 \left(\sum q_k - 1 \right) - \sum \lambda_k q_k,$$

and the first-order conditions on the q_k's are

$$k^2 - \mu_1 k - \mu_2 = \lambda_k. \tag{10}$$

This means that $\lambda_k = 0$ for at most two distinct values of k, since the expression on the left side of (10) 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 it follows from constraints (8) and (9) that $q_i = \frac{R-j}{i-j}$ and $q_j = \frac{i-R}{i-j}$. Since $q_i, q_j > 0$ by hypothesis, it must hold that $j < R < i$.

Substituting the above expression for q_i, q_j into the minimand yields the expression $R(i+j) - ij$. It is easily verified that this expression can be reduced by substituting $(i-1)$ for i or $(j+1)$ for j. Thus, the constraints $j < R < i$ must be binding, and for non-integer R, it follows that $i = \lfloor R \rfloor + 1$ and $j = \lfloor R \rfloor$. In the case where R is an integer, either $q_i = 0$ or $q_j = 0$, so that only one of the q_k's is non-zero. It then follows that $q_R = 1$ and $q_k = 0$ for $k \neq R$. In

[20] Note that there are additional upper bounds on the q_k's, 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.

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 (7) yields the upper bound in Proposition 3.

It only remains to be shown that the minimizing values of $q_1, \cdots, q_n$ can be achieved. Let $B_j = \left[\sum_{i=1}^{j-1} r_j, \sum_{i=1}^{j} r_j\right]$, and let $C_j = f(B_j)$, where $f(x) = x - \lfloor x\rfloor$. Then $q_{\lfloor R\rfloor} = \lfloor R\rfloor + 1 - R$, $q_{\lfloor R\rfloor+1} = R - \lfloor R\rfloor$, 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 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 (5), then the lower bound will be achieved. Thus, over the subset of determinate cases, all judges must be perfectly concordant with 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 (5) 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 some 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

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

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 $\underline{E}(I) + \overline{E}(I) = 1 - I$.