Judicial Errors: Evidence from Refugee Appeals

Samuel Norris

OCTOBER 2018
Judicial Errors: Evidence from Refugee Appeals*

Samuel Norris†

October 6, 2018

Abstract

Judges with the same conviction rate might choose to convict different defendants, which violates an important standard of court quality. I show how judge disagreement can be nonparametrically bounded using information on defendant characteristics, or from other court decisions on the same cases. I implement the procedure for a Canadian refugee appeal court, and bound disagreement for the average pair of similarly-severe judges at 10% of all cases, higher than the amount of disagreement coming from cross-judge variation in leniency and large relative to the overall approval rate of only 14%. I aggregate judge-pair disagreement up into a judge-specific measure of decision quality I call consistency, and build a structural model to study the judge and institutional characteristics associated with it. Finally, I show how inconsistency implies failure of the monotonicity assumption in examiner-assignment IV designs, and adapt my bounding method into a test that is more powerful than current approaches.

*I thank my committee, Lori Beaman, Jon Guryan, Seema Jayachandran and Matt Notowidigdo for their help and encouragement over the course of this project. Arjada Bardhi, Gideon Bornstein, Michael Frakes, Ezra Friedman, Lori Hausegger, Cynthia Kinnan, Lizzie Krasner, Jens Ludwig, Justin McCrary, Laia Navarro-Sola, Aviv Nevo, Matt Pecenco, Krishna Pendakur, Will Rafey, James Rendell, Brian Rendell, Caitlin Rowe, Jesse Shapiro, Jeff Weaver, and seminar audiences at Emcon 2017, Harris, LSE, Northwestern Law, the University of Pennsylvania, Simon Fraser University and the NBER Summer Institute provided useful thoughts and comments. Aaron Dewitt introduced me to the judicial review system for refugee claims. Andrew Baumberg, Catherine Dauvergne, Lori Hausegger, Sean Rehaag and several anonymous judges and law clerks offered valuable insight into the institutional details of the Immigration and Refugee Board and the Federal Court. Paul Longley generously shared his expertise on imputing country of origin from names. I gratefully acknowledge the Social Sciences and Humanities Research Council of Canada for financial support through its Doctoral Fellowship Awards, and the Becker Friedman Institute for hosting me as a Price Theory Scholar for a very productive semester.

†Harris School of Public Policy, University of Chicago. samnorris@uchicago.edu
1 Introduction

The justice system is a major institution in all developed countries. In the US there are approximately 7 million felons and ex-felons under court supervision (Glaze and Parks, 2011), and 47 million non-traffic cases filed in state courts each year (Bureau of Justice Statistics, 2006). Other quasi-judicial institutions routinely make very costly decisions; Social Security Disability Insurance examiners who decide SSDI eligibility collectively oversee a system that spends $144 billion on annual benefits (DeHaven, 2013). Despite the importance of these decisions, there is strong evidence that judges’ rulings are affected by a wide variety of non-relevant sources of information, such as the defendants’ race or TV news about unrelated crime, meaning that some court cases would be decided differently if the case was heard by a different judge, on a different day or even at a different time.\(^1\) However, there is no existing research that attempts to assess the overall importance of these sorts of inconsistencies or measure the number of defendants affected by them.

A different literature has used the random assignment of judges and other decisionmakers as instruments for treatment (Kling, 2006; Gaulé, 2015). Interpretation of IV estimates as a local average treatment effect relies on the monotonicity assumption, which states that each judge would incarcerate a strict subset of the defendants incarcerated by her more-severe colleagues. This no-defiers assumption could be violated if different judges were more or less lenient towards particular crime types or demographic groups, which seems likely in many courts.\(^2\) Recent work has adjusted IV strategies to account for this issue; Mueller-Smith (2015) shows that constructing judge instruments using only variation within LASSO-constructed demographic groups dramatically boosts first-stage predictive power. This suggests that judge behavior is not monotonic, but cannot be used to directly estimate the number of monotonicity violations arising from naive use of judge instruments.

In this paper I show that monotonicity is intimately related to a standard of court quality called consistency, which requires that equally-severe judges would choose to incarcerate the same individual defendants and is violated by the examples above. I develop methods to nonparametrically bound the number of defiers in the Imbens and Angrist (1994) sense, and show that my approach tests a stronger implication of monotonicity than current tests. I use a very similar method to provide bounds on inconsistency, and discuss the conditions under which it can be aggregated into an individual judge—rather than court-wide—measure of quality. Finally, I build a structural model that identifies the institutional and judge characteristics that are associated with consistency.

---

\(^1\) Abrams, Bertrand, and Mullainathan (2012) show that judges are differentially harsh towards different demographic groups, implying that at least some defendants are treated differently because of their race. The TV news study is by Philippe and Ouss (2016). Other sources of extraneous information that affect decisions include upcoming elections (Canes-Wrone, Clark, and Kelly, 2014), inter-communal violence (Shayo and Zussman, 2010), the previous decision (Chen, Moskowitz, and Shue, 2016), whether the case was heard on the defendant’s birthday (Chen and Philippe, 2018) and the winner of the previous night’s football game (Eren and Mocan, 2016).

\(^2\) Indeed, the first paper to discuss the monotonicity assumption argued that examiner assignment designs were unlikely to satisfy monotonicity for this reason (Imbens and Angrist, 1994).
My test of monotonicity is built on the idea that assignment to a more severe judge must increase the average probability of treatment for all demographic subgroups. The most common test of this implication is to run subsample-specific regressions of treatment on mean judge severity calculated in a different subsample, and ask whether the linear first stages are all positive (Bhuller et al., 2016; Dobbie et al., 2018). I show that it is strictly more powerful to conduct this test for each pair of consecutively-more-severe judges, and provide an inequality test to do so. For a binary instrument, I combine the estimates of the subsample-specific first stages into a nonparametric lower bound on the $\lambda$ of Angrist, Imbens, and Rubin (1996), which allows the researcher to assess the size and importance of the monotonicity violations. I develop a multi-instrument analog to $\lambda$ suitable for evaluating monotonicity in examiner assignment contexts, and estimate lower bounds on it using the intersection bounds method of Chernozhukov, Lee, and Rosen (2013).

I then turn to understanding the determinants of court quality. My focus on consistency is motivated by the legal standard of equal treatment of equals, which requires that all decisions are made according to the same standards. Equal treatment has been recognized as a desirable characteristic in decision-making since at least the 4th century BCE (Aristotle et al., 1998). Contemporary philosophers emphasize that equal treatment is a necessary—though not sufficient—condition for substantive justice. In Rawls’ seminal A Theory of Justice (1972), he argues that equal treatment excludes significant kinds of injustices. For if it is supposed that institutions are reasonably just, then it is of great importance that the authorities should be impartial and not influenced by personal, monetary, or other irrelevant considerations in their handling of particular cases ... even where laws and institutions are unjust, it is often better that they should be consistently applied. In this way those subject to them at least know what is demanded and they can try to protect themselves accordingly.

Other prominent legal thinkers claim that equal treatment is itself a powerful bulwark against gross injustice, and that there are few examples of institutions that are substantively unjust but still satisfy equal treatment (Fuller, 1969).

Judges can differ from the ideal of equal treatment in two ways. First, there could be differential leniency among judges, and in fact in all courts that I am aware of there is substantial cross-judge variation in propensity to incarcerate randomly-assigned defendants. Characteristically, Norris, Pecenco, and Weaver (2018) find that mean incarceration rates in Ohio criminal courts vary from 29% for a 10th percentile judge to 43% for a 90th percentile judge.

Second, judges with the same overall incarceration rate could choose to incarcerate different individuals, which I refer to as pairwise inconsistency. Inconsistency is usually not observable because it is a function of the joint distribution of judge decisions—point identification requires seeing different judges make an exactly comparable decision on the exact same case, which is not feasible in real-life institutions.\(^3\)

\(^3\)The one set of institutions where multiple judges make a decision on the same case is multi-judge panels like the
The two different ways judge can violate equal treatment are both key to examiner-assignment designs. The strength of the instrument is driven by differential leniency; if all judges incarcerated the same share of defendants than there would be no first stage. On the other hand, I show that inconsistency implies violations of monotonicity. This means that IV requires violations of equal treatment along one dimension—differential leniency—but forbids them along the other, consistency.

The distinction between differential leniency and inconsistency is also of interest to legal scholars, who have approached the problem by providing judges with small numbers of identical cases and comparing their hypothetical sentences. Partridge and Eldridge (1974) find a high degree of disparity in the sentences given for the same hypothetical case, but that this disparity is not primarily caused by differential leniency. As they put it, “if there are indeed hanging judges and lenient ones—and it would appear that there are a few—their contribution to the disparity problem is minor compared to the contribution made by judges who cannot be so characterized.” However, subsequent research has not clarified whether these findings generalize to real-life court cases, or what causes judge inconsistency.

I begin with the observation that consistency, like monotonicity, implies restrictions on outcomes for pairs of judges. Suppose that a given pair with the same overall leniency is perfectly consistent, and would choose to incarcerate the same defendants. Then, when defendants are randomly assigned to the judges, the distribution of covariates among both judges' incarcerated defendants should also be the same. Differences in the distributions can be translated into bounds on the share of defendants that the judges would disagree on. The relationship is particularly intuitive for an indicator covariate: among the defendants that were incarcerated by each judge, the cross-judge difference in the covariate mean is a lower bound on the share of defendants the judges would disagree on.

Bounding is particularly powerful in contexts with many covariates, because disagreement in each covariate cell can be added together, weighted by the cell probability. Aggregation over different covariates is more difficult when—as in my setting—the researcher does not have access to covariates in the traditional sense, but subsequent outcomes that are only observed conditional on the first-round decision, and where only the marginal distribution is observed. This naturally arises in many appeal courts, where the subsequent outcome is a second judicial decision made by one of many judges. I show that in this context, bounds defined by each subsequent outcome can be aggregated up into an overall bound on judge-pair disagreement in a simple way: overall disagreement is bounded below by the largest covariate bound. As with the bound on the share of defiers, Chernozhukov, Lee, and Rosen (2013) can be used for estimation and inference.

The bounding approach is attractive because it requires very few assumptions. One cost of

---

Supreme Court. A large literature has emphasized how strategic interactions and consensus norms mean that judges' decisions are dependent and simultaneous, and not necessarily indicative of judges true opinions (Epstein, Landes, and Posner, 2013; Fischman, 2008).
this flexibility is that the estimand is a disagreement rate between two given judges, rather than a parameter associated with a single judge. This reduces precision by diluting the sample for any given parameter, and makes it difficult to learn about the judge-specific factors that drive inconsistency. For example, do judges improve with experience, or do worse when workloads are higher? Does the way the court selects judges matter for consistency? To answer these questions requires a way to determine which of two pairwise inconsistent judges is making the right decision. Here, I rely on a nice characteristic of my empirical setting: as in many appeal courts, first-round judges are explicitly told that they should make decisions their second-round colleagues are likely to agree with. Thus, if there are two first-round judges who approve the same number of first-round claimants, I say that the judge with a higher share of her claimants approved by the second-round judge is more consistent with the collective standards of the court.  

With $J$ judges, the modeling challenge is to aggregate the $J^2 - J$ measures of pairwise consistency into $J$ judge-specific measures of quality. To do so, I model judge behavior using a generalization of standard index models of choice, where judges perfectly observe the strength of each claimants case and approve them if it is larger than some judge-specific threshold. In my model, judges observe case quality with error and so rank the relative strengths of cases differently; the size of the distribution of this error (and the corresponding difference in the judges’ rankings) is inconsistency. Judicial behavior is thus summarized by a judge-specific threshold, and judge-specific inconsistency.  

I show how the same conditions used to generate a lower bound on disagreement—random assignment of judges to cases, and two separate judges making consecutive decisions on each case—play a key role in identification. With the addition of regressors that affect judge leniency but not errors, the model is nonparametrically identified and can be tractably estimated under parametric restrictions. It is applicable to any situation where two potentially fallible decision-makers are independently making a similar decision on the same case. Potential applications include judges, patent and disability examiners, and health care settings where doctors get a second opinion on a diagnosis.

I use these tools to study judicial review of refugee decisions at the Federal Court of Canada. The Federal Court is the only point of appeal for claimants who have been rejected for refugee

---

4Two recent papers on bail decisions have also exploited knowledge of judges’ objectives, since judges are instructed to grant bail only to defendants who are unlikely to re-offend before their trial. Arnold, Dobbie, and Yang (2017) show how cross-race differences in the propensity of marginal defendants to re-offend is evidence of judge discrimination. Kleinberg et al. (2017) use machine-learning tools to predict re-offending. They show that judges imperfectly grant bail to defendants who are ex ante unlikely to re-offend, and differ substantially in their ability to do so.

5Equivalently, judges convict a defendant if his guiltiness is higher than some judge-specific threshold. I use the language of approval rather than conviction to concord with my empirical application, though the concept is identical.

6My conception of decision-making has strong parallels with Abaluck et al. (2016)’s model of how doctors choose which patients to send for further tests. Since the test reveals whether the patient actually has the disease, high test yield rates (conditional on share of patients sent for a test) are an indication of good allocation of tests across patients. Their identification strategy differs somewhat from mine, owing to non-random assignment of patients to doctors, and they focus on identifying doctors’ thresholds rather than their consistency.
status by administrative decision-makers at the Immigration and Refugee Board (IRB), and is seen as a crucial backstop that ensures the fairness of the overall refugee system. The stakes are high. As noted in Rehaag (2012), “if errors in first-instance refugee determinations at the [IRB] are not caught and corrected through judicial review, refugees may be deported to countries where they face persecution, torture or death.” The judges are experts in dealing with refugee cases; about 70% of their caseload is refugee appeals. Nonetheless, I find low levels of agreement between judges on which claimants’ appeals should be granted, corresponding to a meaningful impact on decisions and outcomes. If judges were perfectly consistent, differential leniency across first-round judges means that the average pair of judges would disagree on the correct decision for 8% of cases, which is already relatively large given that the judges approve only 14% of cases. I show that this measure severely understates the level of unequal treatment. I bound the average disagreement rate for pairs of similarly-severe judges to at least 10% and reject perfect consistency for 72% of judge pairs. Under the slightly stronger assumptions of the structural model, the disagreement rate rises to 13%, implying that inconsistency is a larger contributor to unequal treatment than differential leniency.7

I provide evidence that inconsistency is caused largely by idiosyncratic observational errors like the non-relevant factors that have affected decisions in other contexts, rather than permanent cross-judge differences in racial or gender bias, judging ideologies, or statutory interpretation. The lack of consistency can also be understood as a failure of first-round judges to pick the claimants that are most likely to be successful in the second round. If all claimants automatically advanced to the second round, I estimate that 19.4% of appeals would be successful, rather than the current 6%. This difference amounts to approximately 7,700 families over my study period.

I survey refugee lawyers about judge quality, and validate the model by showing that survey responses are correlated with my measures of consistency and leniency. Model-estimated consistency improves dramatically during the first year of experience, and continues to improve at a slower rate for at least ten years. These gains from experience are a function of time on the court rather than number of cases heard, suggesting that the court could improve overall consistency by increasing judge retention. One of the ways that judges improve with experience is to get better at maintaining consistency with workload; judges with less than five years of experience are less consistent during periods of high workload, but experienced judges are unaffected by the number of cases.

In 1988, a law was passed to make it more difficult for the government to appoint unqualified judges. The reform, which gave a committee of legal experts veto power over candidates, had the intended effect of reducing the number of newly-appointed judges with ties to the party in power. I find that it dramatically improved judge consistency, implying that reforms to judicial selection processes can have meaningful effects on judicial outcomes.

The structural model uncovers wide cross-judge variation in consistency. I show how policymak-

---

7 An alternative measure of how inconsistency affects disagreement is that among all pairs of judges, the disagreement rate under perfect consistency is 8%. With inconsistency, it rises to 23%.
ers can exploit this knowledge to optimize the allocation of judges to rounds while maintaining the quality of the approved claimants. The Federal Court could reduce its workload by approximately 18% while approving the same number of similarly-qualified claimants, saving at least $4.4 million in judge salaries alone over my study period.\footnote{Alternatively, judge assignments could be reshuffled to maximize the number of successful claims while holding workloads constant and maintaining the case quality of the approved. The problem is approximately symmetric, so the same judicial resources could be used to increase the number of approvals by 19%.
}

Given the high levels of inconsistency, I unsurprisingly uncover large violations of monotonicity. These violations are not detected by current approaches. Using my structural model, I study the effect of monotonicity violations on IV and MTE estimates. Despite the violations of monotonicity, the effect on the IV estimate is small—because many individuals are compliers under one judge treatment but defiers under another, the treatment effects for the compliers and defiers are similar. I find more concerning levels of bias in the MTE estimates.

The paper proceeds in five sections. Section 2 briefly discusses the institutional background. In Section 3.1 I introduce the test for monotonicity in examiner-assignment designs. I then formally introduce the concept of equal treatment, discuss its connection to substantive justice, and bound judge-pair disagreement in Section 3.2. Section 3.3 shows how to aggregate up judge-pair disagreement into a measure of judge quality, and introduces a structural model to do so. Section 4 contains the results, and Section 5 concludes.

### 2 Institutional Background and Data

Initial refugee decisions in Canada are made by an administrative body known as the Immigration and Refugee Board (IRB). The IRB is not itself amenable to analysis because procedures to assign adjudicators to cases are opaque and non-random. My entire analysis therefore concerns the Federal Court, which hears appeals of IRB decisions.\footnote{The Court’s decision is technically a ‘judicial review,’ which refers specifically to judicial oversight of an administrative decision. For ease of language I instead use the term ‘appeal’ throughout this paper.}

However, I begin by describing the IRB in enough detail to contextualize the distribution of the denied claimants who appeal to the Federal Court. I then describe the Federal Court and the relevant institutional background.

#### 2.1 Immigration and Refugee Board

Initial decisions on inland refugee claims are made by the Members of the IRB, who are tasked with evaluating whether the claimant meets the international definition of a refugee: “a person who, by reason of a well-founded fear of persecution for reasons of race, religion, nationality, membership in a particular social group or political opinion, is outside each of their countries of nationality and is unable or, by reason of fear, unwilling to avail themselves of the protection of each of those
countries” (United Nations, 1967). Claims are non-randomly assigned to Members with expertise in either the claimant’s country of origin or the stated reason for the claim. The Members are political appointees rather than long-term, professional bureaucrats.

The IRB approves about 50% of claims, but between-Member variation in approval rates is large. Between 2006 and 2010, the 10th percentile Member approved 15.8% while the 90th percentile Member approved 82.1%. One rejected all of the 169 claims given to him over a three year period, although this was unusual enough to attract media attention (Keung, 2011). The non-random assignment of cases to IRB Members means that this difference might merely reflect cross-Member variation in strength of case rather than variation in Member severity, but the scope of the variation seems at odds with the possible extent of specialization (Rehaag, 2007). The 10th-90th percentile difference is also much larger than the same measure for judges at the Federal Court (7-24%), the Circuit Court of Cook County (roughly 31-39%, Loeffler (2013)) or Norwegian district courts (34-54%, Bhuller et al. (2016)). This is particularly important because it suggests that some claimants who reasonably meet the refugee standard may be initially denied status.

Claimants who have been rejected for refugee status may apply to the Federal Court for judicial review. Approximately 65% of denied claimants file an appeal, which allows most claimants to stay in Canada until the Federal Court makes its final decision.10

IRB procedures for refugee decisions were broadly consistent from 1995 until December 15, 2012, when an administrative appeal division partially supplanted the review work of the Federal Court (Grant and Rehaag, 2015). The only major policy change in this period concerned the composition of the IRB panel that made the decision. For refugee claims submitted before June 28, 2002, standard procedure was for the case to be heard by a two-Member panel. If either member recommended approval, refugee status would be granted. After the implementation of the Immigration and Refugee Protection Act (IRPA) in 2002, all cases were heard by a single Member. This is important because it suggests that the distribution of case quality for the rejected claimants who appeal to the Federal Court may have changed after IRPA came into affect; more high-quality claimants may have been rejected, skewing the distribution further to the right.11 To allow for this possibility, in the structural model I will allow for the distribution of case quality to vary before and after IRPA. More details are in Section 4.2.

---

10 The IRB occasionally rules that a refugee application was “without merit.” In that case, removal can occur before judicial review.

11 Before 2002, cases were sometimes heard by a single Member, but only with the consent of the claimant (Dauvergne, 2003). The claimant often knew which Member would be making the decision if they agreed to a single-Member panel. Ostensibly they would be less likely to let the decision on their refugee status be made by a Member with a low approval rate, meaning that they had some ability to pick who would decide their refugee status.
2.2 Federal Court responsibilities and protocol

The Federal Court is a national court with jurisdiction over certain issues related to the federal government. Its 33 judges hear cases related to intellectual property, maritime law, and aboriginal law, but about 70% of their caseload is devoted to refugee appeals. The scope of these appeals is limited. Judges are not tasked with determining whether the IRB Member made the right decision. Under Canadian law, judges must show deference to administrative decisions. Instead of determining whether the “correct” ruling was made, the judge must simply decide whether the government’s initial decision was “reasonable” (Rehaag, 2012).12

Success at the Federal Court requires approval by two consecutive judges. First-round judges are tasked with deciding whether a claimant has an “arguable case” to make in a second-round hearing, and second-round judges with whether the original government decision was reasonable. A natural implication of the arguable case standard is that the first-round judges goal is to approve as many claimants as possible who will be successful in the second round, conditional on her overall approval rate. Such a clear objective is in contrast to most judicial environments where the goal—to convict the guilty and acquit the innocent—has no clear empirical analog. As I discuss further in Section 3.3, this implies that a good measure of first-round judge quality is her ability to find claimants who are subsequently approved.

The first-round judge makes her decision after reviewing written records from the IRB decision and briefs written by the lawyers for the claimant (arguing for a second-round hearing) and the government (arguing against). If the judge decides against the claimant, the claim is rejected and the claimant is usually deported.13 If the petition for leave is approved, the case goes to a full judicial review hearing. The judge for the second-round hearing is also quasi-randomly assigned, so usually the second-round judge is someone different. Regardless of the first-stage outcome, the first-round judge does not provide a written explanation for her decision. This makes it relatively difficult for judges to learn how their colleagues have ruled on similar cases, and may contribute to high levels of disagreement between first-round judges.

During the second-round hearing, the justice questions the lawyers about the contents of their submissions and the IRB records, but very rarely reviews new evidence or calls witnesses. The name of the first-round judge is not immediately available. It is not difficult for the second-round judge to access this information if he wants, but my conversations with judges indicate that they rarely do.

---

12 An unreasonable decision as one where “there is no line of analysis within the given reasons that could reasonably lead the tribunal from the evidence before it to the conclusion at which it arrived.” One concrete way that this standard affects the Federal Court proceedings is how it limits the sort of evidence that can be introduced. New evidence concerning the actual merits of the case—for example, a death-threat letter implying the claimant truly is in danger in his own country—would not be considered, while evidence about how the decision was made—evidence the IRB Member had made a racially prejudiced statement during the hearing—would typically be accepted.

13 There are two legal options for claimants who have been denied leave but do not want to accept the decision, though neither is very common. Beginning the process for either does not forestall removal from Canada. For more details, see Rehaag (2012).
To reflect this, I will model the second-round decision maker as explicitly ignoring the identity of the first-round judge—the first-round judge affects the second-round decision only through her choice of which claimants to approve, not as a signal to the second-round judge. This shuts down what could be an important information channel, and eliminates the possibility of multiple equilibria.

If a claimant is successful in the judicial review stage, their case is usually returned to the IRB to be analyzed anew by a different Member. Occasionally the judge will grant refugee status to the claimant without a return to the IRB, but I will ignore this distinction in the empirical analysis.

Judge assignment works similarly in both stages. For the first stage, judges are assigned to cases using a pre-set schedule; in each office the judges rotate through “leave duty.” When enough cases have accrued the court clerk gives the leave duty judge all the outstanding files (usually on a Monday), and they are responsible for disposing of all of them. There is no review of the cases before they are given to the judge, and the leave duty schedule is not public. Previous research claims that this assignment is as good as random (Rehaag, 2007); in Section 4.3 I show that judge leniency is uncorrelated with case or claimant characteristics predictive of success. In the second stage the assignment process is similar; cases are divided between judges who are available for refugee work without review of the contents. A computer program slots hearings into the available times in the judges schedules.

2.3 Reform to selection of Federal Court justices

Federal Court justices are appointed by the Minister of Justice. For most of Canadian history the Minister has had nearly unfettered discretion over appointments and has used them to reward “active supporters of the party in power” (McKelvey, 1985).

A major reform in 1988 reduced the discretion of the government in making appointments. The reform created province-level judicial advisory councils (JACs) to pre-screen applicants before they could go to the Minister for possible selection. The committees were made up of one member of the provincial Law Society, one member of the provincial branch of the bar association, one representative of the provincial chief justice, one representative of the provincial attorney general, and three representatives of the Minister of Justice. The JACs rated each candidate as “highly recommended,” “recommended,” or “not recommended,” and the government could pick judges only from the pool of recommended and highly recommended candidates. The standards concorded well with a lay understanding of what makes a good judge: “professional competence and experience” (such as proficiency in the law, awareness of racial and gender issues); ‘personal characteristics’ (ethical standards, fairness, tolerance); and ‘potential impediments to appointment’ (drug or alcohol depen-

14 Appointments are long, lasting until the mandatory retirement age of 75. Judges sometimes elect to work past 75 by having the court classify them as a supernumerary justice, which allows them to stay on the bench but has no effect on the work they do.
dency, health, financial difficulties)” (Hausegger et al., 2010). Crucially, the direct representatives of the Minister were a minority on the committee, making it difficult to push through wholly unqualified candidates. The standards had bite—only about 40% of candidates were recommended or highly recommended—and the reform seems to have reduced the level of patronage. Russell and Ziegel (1991) report that before 1988 at least 47% of appointed judges had some involvement with the ruling Conservative party, ranging from financial contributions to running for office. Their data come from reports by surveyed respondents in the legal profession, and so estimates are likely biased downwards. Though data on post-reform connections to the ruling party are not exactly comparable, Hausegger et al. (2010) search through administrative records and find that after the reform only 30% of newly-appointed judges had donated to the party in power in the five years before their appointment. This is consistent with the new system reducing the number of unqualified party supporters being appointed, and suggests that the overall quality of the courts may have improved as a result. I will test this hypothesis in Section 4.11, and find evidence that consistency did improve after the reform.

3 Empirical strategy

I begin by describing the test of monotonicity and a method for bounding the proportion of defiers in Section 3.1. In Section 3.2 I connect monotonicity to equal treatment, and bound disagreement for pairs of similarly-severe judges. In Section 3.3 I introduce a structural model that aggregates judge-pair disagreement into a measure of judge-specific quality that I call inconsistency.

3.1 Bounding violations of monotonicity

Monotonicity is one of the assumptions that is required to interpret IV estimates as a local average treatment effect (LATE) (Imbens and Angrist, 1994). Suppose that for individual $i$ assigned to instrument $j$, the realized treatment for individual $i$ is $T_i(j) \in \{0, 1\}$. Assume that judge assignment is orthogonal to potential outcomes. Ordering the $J + 1$ discrete instruments $j \in \{0, 1, \ldots, J\}$ by $E[T_i(j)]$, monotonicity requires that for all individuals $i$, if $j' > j$ then $T_i(j') > T_i(j)$. Individuals who are induced out of treatment by being assigned instrument $j$ rather than $j - 1$ are known as $j$ defiers, so monotonicity is also referred to as the no-defier condition. This may be a strong assumption for judge instruments, and can be violated whenever judges are differentially lenient over different types of cases or defendants.

Current approaches to testing monotonicity ask whether judge mean incarceration rates in one demographic subsample are positively correlated with incarceration in other subsamples (Bhuller et al., 2016; Dobbie et al., 2018). Other approaches test the joint implications of monotonicity and the other LATE conditions (Kitagawa, 2015).
tonicity, and how to construct bounds on the share of defiers and the \( \lambda \) of Angrist, Imbens, and Rubin (1996). This provides an easily-interpretable scale to the degree of monotonicity violations, and allows the researcher to assess the potential implications. I begin by introducing this method in a binary-instrument, binary-treatment setting, then extend to multiple instruments and explain how to implement it in a judge setting.

**Binary instrument**

Monotonicity is an assumption about the joint distribution of outcomes under different treatments for a given individual. Both assignments are not observed for any single individual, so the degree of monotonicity violations cannot be directly assessed. However, a long tradition in economics has used information about marginal distributions—which are observable—to bound the joint distribution, which is not (Heckman, Smith, and Clements, 1997; Manski, 1997).

One implication of monotonicity is that the first stage coefficient should be positive for all demographic subgroups. This can be seen by using Fréchet inequalities to bound the number of defiers when assigned to instrument \( j \) as opposed to \( j - 1 \):

\[
P[i \text{ is a } j \text{ defier} = P[T_i(j) = 0, T_i(j - 1) = 1] \geq \max(0, P[T_i(j - 1) = 1] - P[T_i(j) = 1]) = \max(0, -\tilde{\alpha}_j) 
\]

(1)

where \( \tilde{\alpha}_j \) comes from a first-stage regression of treatment on the dummy \( z_j \) indicating assignment to judge \( j \), among the sample of individuals assigned to judge \( j \) or \( j - 1 \):

\[
T_{ij} = \mu + \tilde{\alpha}_j z_j + \varepsilon_{ij} 
\]

(2)

The right hand side of Equation 1 is necessarily zero, because the judges have been ordered according to propensity to assign treatment and so \( \tilde{\alpha}_j \geq 0 \). This means that the bound is uninformative. However, Equation 1 also applies conditional on covariates, and so whenever the first stage is negative for a given demographic group, the bound is informative and the share of defiers within that subgroup is larger than the negative of the first stage. This implication is directly testable. Let \( D_g \) be a dummy variable indicating that individual \( i \) belongs to demographic group \( g \). Then the first stage for individuals assigned to either judge \( j \) or \( j - 1 \) is:

\[
T_{ijg} = \sum_{g=1}^{G} \alpha^g_j z_j \times D_g + \mu_g + \varepsilon_{ijg} 
\]

(3)

The tests I discuss in this section focus only on the monotonicity assumption, at the cost of requiring access to additional covariates.
Under monotonicity, the null hypothesis is $H_0 : \alpha_j \geq 0$, where $\alpha_j = [\alpha_j^1, \ldots, \alpha_j^G]'$ and the $\geq$ is applied element-wise. This hypothesis can be tested using Wolak (1989).

More interestingly, the degree of the monotonicity violations and their implications for LATE estimates can also be assessed. Angrist, Imbens, and Rubin (1996) describe how for a single instrument $z_j$, the bias of the IV estimate is equal to $\lambda_j(\theta_C - \theta_D)$, where $\theta_C$ and $\theta_D$ represent the treatment effects for the compliers and defiers, respectively, and

$$\lambda_j = \frac{P(i \text{ is a } j \text{ defier})}{P(i \text{ is a } j \text{ complier}) - P(i \text{ is a } j \text{ defier})}$$

Applying Equation 1 within each each demographic subgroup and adding together, the overall proportion of defiers for judge $j$ (relative to judge $j-1$) is bounded below by $\sum_g -\alpha_j^g 1[\alpha_j^g < 0] w_j^g$, where $w_j^g$ is the share of the $g^{th}$ demographic group among individuals assigned to judge $j$ or $j-1$. This is equal to $\max_{\tau \in T_{[-1,0]^G}} W_j \alpha_j$, where $W_j = \text{diag}(w_j^1, \ldots, w_j^G)$ and $T_{[-1,0]^G}$ is the $2^G$-size set of all $G \times 1$ row vectors made up of $\{-1, 0\}$, so $\lambda_j$ is bounded by

$$\lambda_j \geq \sum_g -\alpha_j^g 1[\alpha_j^g < 0] w_j^g = \max_{\tau \in T_{[-1,0]^G}} \frac{\tau W_j \alpha_j}{\bar{\alpha}_j}$$

(4)

where $\bar{\alpha}_j$ is simply the first stage for instrument $z_j$ and is treated as fixed. This construction ignores sampling variation in the overall first stage, following a long tradition that treats the ordering of the instruments as fixed given the sample (Imbens and Angrist, 1994).

The advantage of the maximization form in Equation 4 is that it allows estimation and inference using tools developed for intersection bounds. These tools address one of the main problems with plug-in estimates of both forms of Equation 4, which is that they tend to overstate the true bound in finite samples due to sampling error in $\alpha_j$. I apply the method of Chernozhukov, Lee, and Rosen (2013) (hereafter CLR) to the right-hand expression, which shades the bound estimate down to account for uncertainty about whether each $\alpha_j^g$ really is smaller than zero. Their method allows the calculation of asymptotically valid half-median unbiased estimates—which means that they asymptotically fall below the true lower bound with probability at least $\frac{1}{2}$—as well as confidence intervals for the bounds. With bounds on $\lambda_j$ in hand, interpretation of the potential implications of monotonicity violations now turns on the potential difference in treatment effects between the complier and defier groups, which can be evaluated in each context.

**Multiple instruments**

Current approaches to testing monotonicity in judge designs rely on a relatively weak implication of the assumption (Bhuller et al., 2016; Dobbie et al., 2018). These tests, which I refer to as tests
of *average monotonicity*, estimate in one subsample the linear first stage of the endogenous variable on a judge-mean incarceration rate calculated in a different subsample. Monotonicity requires that the first stage should be positive no matter how those subsamples are defined—for example, over race or type of crime—and so researchers have focused on testing this implication.

In contrast, I focus on testing *pairwise monotonicity*, or monotonicity for each pair of consecutively-more-severe pair of judges. This approach has two advantages. First, I show that pairwise approaches test a strictly more powerful implication of monotonicity than average tests. Second, the tests for each judge pair can be aggregated using ordinary IV weights into a measure of the potential LATE bias analogous to the familiar $\lambda$ of Angrist, Imbens, and Rubin (1996), allowing the researcher to assess the severity of monotonicity violations.

I defer formal definitions of pairwise and average tests to Appendix A1, along with the proof that violations of average monotonicity are a strict subset of violations of pairwise monotonicity. The idea, however, is simple: if high-incarcerating judges are on average less likely to incarcerate a certain demographic group than their low-incarcerating colleagues—and so the instruments would fail a test of average monotonicity—there must be a similar monotonicity violation within at least one of the judge pairs. The converse, however, is not true, and so some violations of monotonicity are detectable by pairwise tests but not by average ones. Following is the formal statement.

**Theorem 1** (Average versus pairwise tests of monotonicity). *Violations of average monotonicity imply violations of pairwise monotonicity, but violations of pairwise monotonicity do not imply violations of average monotonicity.*

**Proof:** See Appendix Section A1.

With this theorem in hand, the first question is how to implement a pairwise test. The multi-instrument version of the subgroup-specific first stage in Equation 3 is:

$$T_{ijg} = \sum_{g=1}^{G} D_g \times \left( \sum_{\ell=1}^{J} \alpha_{\ell}^g \mathbf{1}[\ell \leq j] \right) + \mu_g + \epsilon_{ijg}$$

(5)

Since the judge indices $j$ are ordered by overall severity, each coefficient $\alpha_{j}^g$ represents the difference in approval rates for demographic group $g$ between judge $j$ and judge $j - 1$. Under monotonicity, these differences are all positive, so the null hypothesis is $H_0 : \alpha \geq 0$ for $\alpha = [\alpha_1^g, \ldots, \alpha_1^g, \ldots, \alpha_J^g, \ldots, \alpha_J^g]$. Similarly to the single instrument version, this can be tested using Wolak (1989).

As with the single instrument, an important question is how to assess the magnitude of monotonicity violations. One approach is motivated by the IV weighting of the judge-pair treatment effects. The $J + 1$ instruments define $2^{J+1}$ types in terms of their response to each consecutive
judge, $t_{\ell}$, $\ell \in \{1, \ldots, 2^{J+1}\}$. Under each treatment $j$, the type is one of an always-taker, never-taker, complier or defier with respect to the previous treatment, $j - 1$, so $t_{\ell,j} \in \{A, N, C, D\}$. Let the population share of each type be $\pi_{\ell}$, and the effect of treatment on each type be $\theta_{\ell}$. Then, the IV estimand is

$$\sum_{j=1}^{J} \varphi_j (\theta_{C,j} + \lambda_j (\theta_{C,j} - \theta_{D,j}))$$

(6)

where $\theta_{C,j} = \sum_{\ell} [t_{\ell,j} = C] \theta_{\ell} \pi_{\ell} / \sum_{\ell} [t_{\ell,j} = C] \pi_{\ell}$ is the average treatment effect for individuals who are compliers when exposed to instrument $j$ relative to $j - 1$, and $\theta_{D,j} = \sum_{\ell} [t_{\ell,j} = D] \theta_{\ell} \pi_{\ell} / \sum_{\ell} [t_{\ell,j} = D] \pi_{\ell}$ is the average treatment effect for individuals who are defiers when exposed to instrument $j$. $\varphi_j$ are weights that can be calculated from the data (Imbens and Angrist, 1994).\(^{16}\)

Analogously to Angrist et al. (1996)’s expression for the bias of a single-instrument IV estimate, Equation 6 tells us that the bias of the IV-weighted complier LATE is $\sum_{j=1}^{J} \varphi_j \lambda_j (\theta_{C,j} - \theta_{D,j})$. Therefore, the multi-instrument analog to $\lambda_j$ is $\Lambda = \sum_{j=1}^{J} \varphi_j \lambda_j$. Whenever $\Lambda$ is larger than 0, differences in treatment effects between the complier and defier groups will bias the estimate of the LATE. The lower bound of $\Lambda$, $\Lambda^l$, can be estimated by taking the Equation 5 estimates of the first stage, $\alpha_j$, and jointly maximizing the weighted absolute sum of the negative coefficients:

$$\Lambda = \sum_{j=1}^{J} \varphi_j \lambda_j \geq \Lambda^l = \max_{\tau \in \{-1,0,1\}} \sum_{j=1}^{J} \varphi_j \tau W_j \alpha_j$$

(7)

where $\alpha_j = [\alpha_{1j}, \ldots, \alpha_{Gj}]$ are the demographic-judge-specific incremental first stage (Equation 5), $\bar{\alpha}_j$ is the judge incremental first stage (Equation 2),\(^{17}\) and $W_j$ is the diagonal matrix of demographic-specific shares for judges $j$ and $j - 1$.

This expression is merely the weighted sum of the individual bounds on $\lambda_j$. Joint maximization allows the estimate of $\Lambda^l$ to account for the uncertainty in $\Lambda$ arising from the imprecision in all estimates. I implement this procedure for the refugee appeal data in Appendix A8, and decisively reject monotonicity. Use of this method in a more standard criminal court can be found in Norris, Pecenco, and Weaver (2018).

As is usual when considering the importance of monotonicity violations, the implications for LATE estimates hinge on the difference in complier and defier treatment effects. This is particularly relevant here because tests of pairwise monotonicity can uncover situations where individuals are compliers under one instrument but defiers under another, suggesting the complier and defier treatment effects might be similar. In particular, the weight on each type $\ell$ is equal to

\(^{16}\)Imbens and Angrist use $\lambda_j$ to denote the IV weights, but to avoid confusion with the bias term I substitute $\varphi_j$.

\(^{17}\)These coefficients can be estimated in one regression with $T_{ij} = \mu + \sum_{\ell=1}^{J} \bar{\alpha}_{ij}[\ell \leq j] + \epsilon_{ij}$. 

14
\[ \pi_t \sum_j \left[ \mathbb{1}[t_{t,j} = C] - \mathbb{1}[t_{t,j} = D] \right] \frac{\varphi_j}{\tilde{\alpha}_j} \]  \hspace{1cm} (8)

When types are defiers under some treatments, their weights can be negative.\(^{18}\) However, since the rescaled IV weights \( \varphi_j / \tilde{\alpha}_j \) tend to be of similar size for judges with similar approval rates, the net effect of a type being a defier and a complier for similarly-severe judges might be small.

Violations of monotonicity that can be detected by the methods introduced in this section are the result of judges’ differential severity across observable demographic groups or case types. By definition, these methods cannot detect monotonicity violations inside demographic groups, so the tests have no power against instruments constructed by interacting observable demographics with judge assignment. Using such instruments is an attractive option whenever monotonicity of the judge-assignment instruments is rejected, and can be done efficiently with a LASSO first stage (Mueller-Smith, 2015). However, failures of monotonicity along one dimension may imply failures along other, unobserved dimensions, so caution is still warranted.

### 3.2 Bounding disagreement

We observe approval decisions \( y_{is} \) for each individual \( i \) in round \( s \). The goal is to bound disagreement \( \delta_{AB} \), the share of claimants that first-round judges \( A_1 \) and \( B_1 \) would disagree on. Under a potential outcomes framework, define \( y_{is}(j) \) as an indicator for approval for individual \( i \) under judge \( j \) in round \( s \). Disagreement is defined as

\[ \delta_{AB} \equiv P[y_{i1}(A_1) \neq y_{i1}(B_1)] \]  \hspace{1cm} (9)

Since \( y_{is} \) is binary, disagreement is merely the likelihood that the claimant will be approved by one judge but not the other, or \( \delta_{AB} = P[y_{i1}(A_1) = 1, y_{i1}(B_1) = 0] + P[y_{i1}(A_1) = 0, y_{i1}(B_1) = 1] \). Comparison with the definition of monotonicity violations in Equation 1 emphasizes the connection between monotonicity and disagreement. It is simple to show that for a fixed level of disagreement between two equally-severe judges, as the number of judges grows and the incarceration rate for the next-most-severe judge moves closer to the pairs’, there must be monotonicity violations.

Like the share of individuals who violate monotonicity, disagreement is a function of the joint distribution of \((y_{i1}(A_1), y_{i1}(B_1))\) and so is only partially identified. Defining \( D_j^1 \) as a dummy variable

\(^{18}\)When there are no defiers, this expression collapses to the familiar Imbens and Angrist weights for each type.
indicating assignment to judge \( j \),

\[
\delta_{AB} = P[y_{i1}(A_1) = 1, y_{i1}(B_1) = 0] + P[y_{i1}(A_1) = 0, y_{i1}(B_1) = 1] \\
\geq \max\{0, P[y_{i1}(A_1) = 1] - P[y_{i1}(B_1) = 1]\} + \max\{0, P[y_{i1}(B_1) = 1] - P[y_{i1}(A_1) = 1]\} \\
= \max\{P[y_{i1}(A_1) = 1] - P[y_{i1}(B_1) = 1], P[y_{i1}(B_1) = 1] - P[y_{i1}(A_1) = 1]\} \\
= \max\{E[y_{11}|D^A_1] - E[y_{11}|D^B_1], E[y_{11}|D^B_1] - E[y_{11}|D^A_1]\} \\
\tag{10}
\]

where the inequality follows from Fréchet (1951), and the final equality from random assignment.

As with monotonicity violations, the Equation 10 bound on disagreement—which was first used to bound judge agreement by Fischman (2013)—is uninformative when judges \( A_1 \) and \( B_1 \) have the same approval rate. However, the bound on disagreement also applies within each demographic subgroup; if judges are in perfect agreement, they should approve the same share of cases within each demographic cell.\(^{19}\) Exploiting additional information about covariates can therefore make bounds informative even between judges with the same approval rate. For example, suppose that the claimants are equally divided between blacks and whites. If judge \( A_1 \) approved 75\% of black claimants and 25\% of white claimants, and judge \( B_1 \) approved 25\% of black claimants and 75\% of white claimants, they must disagree on at least 50\% of all claimants. This holds true for latent claimant characteristics, like whether they \textit{would} be approved in the future by second-round judge \( C_2 \). By random assignment, the share of claimants who would be approved by judge \( C_2 \), \( P[y_{i2}(C_2) = 1] \), does not depend on first-round judge assignment.\(^{20}\) This means that using the second-round judge \( C_2 \) as a latent covariate, disagreement between two first-round judges can be further bounded.

\textbf{Theorem 2} (Bounding disagreement). Suppose that \( y_{i2}(C_2) \perp D^1_1 \), or claimant characteristics as measured by the second-round judges potential decision are orthogonal to first-round judge assignment. Then, first-round disagreement is bounded by the following expression:

\[
\delta_{AB}(C) \geq \max \left\{ \begin{array}{l}
E[y_{11}|D^A_1] - E[y_{11}|D^B_1], \\
E[y_{11}|D^B_1] - E[y_{11}|D^A_1], \\
E[y_{11}|D^A_1] - E[y_{11}|D^B_1] + 2\left[ E[y_{11}|D^B_1]E[y_{12}|D^B_1, D^C_1, y_1 = 1] - E[y_{11}|D^A_1]E[y_{12}|D^A_1, D^C_1, y_1 = 1]\right], \\
E[y_{11}|D^B_1] - E[y_{11}|D^A_1] + 2\left[ E[y_{11}|D^A_1]E[y_{12}|D^A_1, D^C_1, y_1 = 1] - E[y_{11}|D^B_1]E[y_{12}|D^B_1, D^C_1, y_1 = 1]\right] \\
\end{array} \right. \\
\tag{11}
\]

\textbf{Proof:} See Appendix Section A2.

\(^{19}\)The similarities between bounding monotonicity violations and disagreement is even more clear when Equation 10 is rewritten as \( \max(\tilde{\alpha}_A - \tilde{\alpha}_A) \).

\(^{20}\)Similarly to many tests of monotonicity, this test relies on an exclusion-style assumption that first-round approval serves only a screening function, and does not independently affect the likelihood of second-round approval. This would be violated if, for example, the first-round judge gave the claimant tips on how to improve his arguments.
The comparison between Equation 10 and Equation 11 highlights the additional information gained by having access to a second-round judge $C_2$. Using Equation 11, we can do no worse than the Equation 10 bound that uses information on the difference in the overall first-round approval rates. The last two inequalities explain how differences in the share of approved claimants that are subsequently approved by the second-round judge can tighten the bound. Crucially, this means that a second-round judge allows the estimation of an informative bound even when judges $A_1$ and $B_1$ have the same first round approval rate. Indeed, the expression is particularly tractable when they do. For $\delta_{AB}^l(C)$, the lower bound of disagreement $\delta_{AB}(C)$,

$$\lim_{E[y_1 | D_i^A], E[y_1 | D_i^B] \rightarrow P_1} \delta_{AB}^l(C) = 2P_1 \max \{ \Delta_{AB}^C, -\Delta_{AB}^C \} \tag{12}$$

where $\Delta_{AB}^C$ comes from a regression of second-round approval $y_{i2}$ on the identity of the first-round judge, $D_i^A$, for all claimants approved by judges $A_1$ or $B_1$ and assigned to judge $C_2$:

$$y_{i2} = \Delta_0 + \Delta_{AB}^C D_i^A + \varepsilon_i \tag{13}$$

Finally, this relationship is true for all second-round judges $C_2$. Disagreement must be larger than the (positive) difference in approval rates for the most informative second-round judge:

$$\delta_{AB} \geq 2P_1 \max \{ \Delta_{AB}^1, -\Delta_{AB}^1, \Delta_{AB}^2, -\Delta_{AB}^2, \ldots, \Delta_{AB}^J, -\Delta_{AB}^J \} \tag{14}$$

One way to think about this result is that each second-round judge makes decisions using slightly different criteria. The difference in conditional second-round approval rates $E[y_{i2} | D_i^C, y_{i1}(A) = 1]$ and $E[y_{i2} | D_i^C, y_{i1}(B) = 1]$ speaks to how judges $A$ and $B$ differentially approve claimants that meet those criteria. Searching across all second-round judges for the largest difference is equivalent to looking for the dimensions that first-round judges $A_1$ and $B_1$ most disagree on. However, because the preferences of even the second-round judge with the largest differential in approval rates across $A_1$ and $B_1$ do not necessarily correspond exactly to the claimants that $A_1$ and $B_1$ would disagree on, $\delta_{AB}(C)$ is only bounded below by the differential approval rate $\Delta_{AB}^C$.

I conduct estimation and inference of the bounds in Equation 14 using CLR. For all pairs of first-round judges with an approval rate closer than 1 percentage point, I run a regression of second-round approval on the identity of the assigned second-round judge interacted with the first-round

---

21 The order of $A_1$ and $B_1$ is arbitrary, so alternatively regressing $y_{i2}$ on $D_i^B$ does not change estimated disagreement.

22 If the second-round judge approved only claimants would be approved by judge $A_1$ but not judge $B_1$ (or vice-versa), $\delta_{AB}(C)$ would be the exact level of disagreement.

23 Equation 14 also highlights the cost of using latent characteristics like second-round judge approval, rather than pre-existing demographic characteristics like race. It is easy to show that using the intersection of all second-round judge’s decisions as pre-existing characteristics, rather than the maximum over each judge individually, is a more powerful bounding approach.
judge who approved the case, as well as a second-round judge fixed effect.

\[ y_{i2} = \sum_{\ell_2} \Delta^{\ell_2}_{AB} D^{A_1}_{i} \times D^{\ell_2}_{i} + \theta_{\ell_2} + \varepsilon_i \] (15)

This is the single-regression version of Equation 13—putting it in a single regression allows easy recovery of a single covariance matrix for \( \Delta_{AB} \). Then, I estimate \( \hat{\delta}_{AB} \) by applying CLR to the max operator in Equation 14 and scaling by twice the mean of the empirical analogs of \( E[y_1|D^B_1] \) and \( E[y_1|D^A_1] \) for \( P_1 \).24 The CLR estimate is the plug-in analog of the max expression, shaded down to account for the sampling error of \( \hat{\Delta}_{AB} \). I also report the endpoints of one-sided 95% confidence intervals.

### 3.3 Structural model of judge choices

In this section, I introduce a structural model that relies on the known goal of first-round judges—to approve claimants who are likely to be subsequently approved in the second-round—to turn the previous section’s pairwise disagreements into judge-specific measures of consistency, which under some assumptions can be interpreted as a judge quality measure. The model more sharply connects judge decisions to the familiar index representation of decision-making, and clarifies the assumptions under which disagreement can be point-identified, rather than bounded as in the previous section.

The court receives a flow of applicants for refugee status, who are randomly assigned to judges. Strength of case for each applicant \( i \) can be represented by a scalar, \( r_i \). Strength of case is defined with respect to the decisions the judges would collectively make; the ordering of \( r_i \) across claimants is the same as the ordering of a hypothetical average approval rate taken over appearances in front of all judges.

To be approved as a refugee, a claimant must be approved by two consecutive judges. If she is denied by the first judge, her case is not seen by the second judge. Formally, in stages \( s = 1, 2 \), judges \( j = 1, \ldots, J \) approve the claimant if

\[ r_i > \varepsilon_{ij_s}(X_{ij_s}, W_{ij_s}) = \gamma_{js} + X_{ij_s}\beta_s + \tilde{\varepsilon}_{ij_s}(W_{ij_s}) \] (16)

where

---

24 This slightly understates sampling variation in \( \hat{\delta}_{AB} \) by not accounting for the variation in the empirical first-round judge approval means, but sidesteps the issue of sampling variation in determining which judge pairs have a comparable first-round approval rate. Although in this context the sampling variation in the first-stage mean approval rates is small, an alternative approach is to estimate the empirical analog of Equation 11 for all judge pairs, then take a mean of the disagreement rate for judge pairs that are close to the same first-round approval rate, weighted by distance from having the same approval rate.
Assumption 1 (Model definition).

(a) \( r_i \sim F_r \) and \( \tilde{\varepsilon}_{ijs}(W_{ijs}) \sim G_{js,W} \) are independent and have a median of zero. The variance of \( r_i \) is known and finite, and the variances of \( \tilde{\varepsilon}_{ijs}(W_{ijs}) \) are finite.

(b) For all \( j \) and \( k \), \( \tilde{\varepsilon}_{ij1} \perp \tilde{\varepsilon}_{ik2} \).

Judge leniency is captured by \( \gamma_{js} \); judges with higher values of \( \gamma_{js} \) approve fewer claimants. This threshold can be adjusted by \( X_{ijs} \), for example to allow judges to vary in leniency as a function of experience or the time of day of the hearing.

Judge consistency with respect to the court’s collective standards is defined by the distribution of \( \tilde{\varepsilon}_{js}(W_{ijs}) \). For perfectly consistent judges, \( \tilde{\varepsilon}_{js}(W_{ijs}) = 0 \). Then, the decision problem is non-stochastic for a given value of \( r_i \): 

\[
P[r_i > \varepsilon_{js}(X_{ijs}, W_{ijs})] = P[U > F_r(\gamma_{js} + X_{ijs} \beta_s)] = 1 - F_r(\gamma_{js} + X_{ijs} \beta_s),
\]

and so any two judges with the same overall approval rate would either both approve or both reject any claimant with given quality \( \tilde{r}_i \) (this is the standard model of judicial decision-making). Furthermore, both judges would approve the highest quality claimants in the sense that they are more likely than non-approved claimants to be subsequently approved in the second round. Judges become less consistent as the distribution of \( \tilde{\varepsilon}_{js}(W_{ijs}) \) widens. Across judges, more consistent judges are more likely to approve claimants with a strong case (high \( r_i \)).

3.3.1 Consistency and decision quality

Leniency and consistency are related to the level of unequal treatment in a simple way. A court that perfectly satisfied equal treatment would have all judges use the same threshold \( \gamma_{js} \), and be perfectly consistent in their decisions (e.g. \( \tilde{\varepsilon}_{js} = 0 \)). Thus, any individual would be always approved by the court, or always rejected. Cross-judge variation in \( \gamma_{js} \) increases unequal treatment by increasing the variation in individual’s outcomes generated by judge assignment. Similarly, inconsistency induces randomness into the realized outcome for each individual; unequal treatment is monotonically increasing in inconsistency.

The relationship between consistency and the quality of decisions in another abstract sense is dependent on the institution. As judges become more consistent, they are more likely to approve high-\( r_i \) claimants and less likely to approve low-\( r_i \) claimants. The relationship between consistency and quality of decisions therefore turns on the relationship between \( r_i \) and underlying claimant quality. \( r_i \) is an ordering of claimants by the index of characteristics that improve the probability of approval in both stages. This means that if the court is doing a good job on average—truly deserving claimants, however defined, are more likely to be approved than undeserving claimants—higher levels of consistency will also improve the quality of decisions, and so consistency is an
appropriate measure of judge quality. The connection between the overall quality of decisions and the desirability of consistency has been noted by the prominent political philosopher John Rawls, who argued that if “institutions are reasonably just, then it is of great importance that the authorities should be impartial and not influenced by personal, monetary, or other irrelevant considerations in their handling of particular cases” (1972).

3.3.2 Ranking judges by consistency

All judges can be ranked in terms of overall consistency by building up from two-judge comparisons of pairwise consistency. Pairwise consistency is defined for any two judges with the same approval rate in a given round. In the first round, this approval rate is

\[ P[r_i > \varepsilon_{j1}(X_{ijs}, W_{ijs})] = \int G_{j1}(r_i - \gamma_{j1} - X_{ij1} \beta_1) f_r dr \] (17)

Judge A is comparable to judge B when \( P[r_i > \varepsilon_{A1}] = P[r_i > \varepsilon_{B1}] \). The following assumption restricts the set of error distributions \( G_{js} \) so that all judges can be ordered in terms of consistency.

**Assumption 2** (Single-crossing). Suppose judge A and judge B are comparable. Then, there exists a point of single-crossing \( z \) such that:

(a) \( G_{A1}(z - \gamma_{A1}) = G_{B1}(z - \gamma_{B1}) \)

(b) Either:

(i) \( \forall w > z, G_{A1}(w - \gamma_{A1}) \geq G_{B1}(w - \gamma_{B1}) \) and \( \forall w < z, G_{A1}(w - \gamma_{A1}) \leq G_{B1}(w - \gamma_{B1}) \), or

(ii) \( \forall w > z, G_{A1}(w - \gamma_{A1}) \leq G_{B1}(w - \gamma_{B1}) \) and \( \forall w < z, G_{A1}(w - \gamma_{A1}) \geq G_{B1}(w - \gamma_{B1}) \).

Then, the more consistent judge is the one that is more likely to approve high-\( r_i \) claimants; judge A if condition (b)(i) is met, and judge B if (b)(ii).

This definition of consistency is similar to the concept of screening in Sah and Stiglitz (1986), where they define \( A \) as more discriminating than \( B \) when \( \partial G_{A1}(v - \gamma_{A1})/\partial v > \partial G_{B1}(v - \gamma_{B1})/\partial v \). In my model, the more consistent judge is weakly more discriminating at the point of single-crossing.

It is simple to see that consistency is transitive; if judge A is more consistent than judge B, and judge B is more consistent than judge C, then judge A is more consistent than judge C. Thus, under the restricted set of distributions \( G_{js} \) that satisfy single-crossing for all comparable judge pairs, we can rank all judges on a single scale. This is a key advantage over the bounding approach in Section 3.2, and in the empirical application I implement this restriction by assuming that \( G_{js} \) is normally distributed.
3.4 Identification

The model—parameters $\beta_s$, judge-round thresholds $\gamma_{js}$, judge-round error distributions $G_{js}$ and the case strength distribution $F_r$—is identified from two different sources of variation: the random assignment of cases to judges of varying severity, and regressors that shift judge thresholds. I discuss each in turn.

3.4.1 Judge-assignment identification

Take two judges with the same first-round approval rate, $A_1$ and $B_1$. Then, a weakly higher share of the more consistent judge’s claimants will be ultimately approved by a common second-round judge, $C_1$. Suppose that judge $A_1$ is more consistent. Abstracting away from covariates $X_{ij1}$ and substituting $\tilde{G}_{js}(r) = G_{js}(r - \gamma_{js})$ for clarity, this can be seen by noting that:

$$\left( P[r > \varepsilon_{C2} | r > \varepsilon_{A1}] - P[r > \varepsilon_{C2} | r > \varepsilon_{B1}] \right) P[r > \varepsilon_{B1}] = \int_{-\infty}^{z} \left[ \tilde{G}_{A1}(r) - \tilde{G}_{B1}(r) \right] \tilde{G}_{C2}(r)f_rdr + \int_{z}^{\infty} \left[ \tilde{G}_{A1}(r) - \tilde{G}_{B1}(r) \right] \tilde{G}_{C2}(z)f_rdr\geq \int_{-\infty}^{z} \left[ \tilde{G}_{A1}(r) - \tilde{G}_{B1}(r) \right] \tilde{G}_{C2}(z)f_rdr + \int_{z}^{\infty} \left[ \tilde{G}_{A1}(r) - \tilde{G}_{B1}(r) \right] \tilde{G}_{C2}(z)f_rdr = \tilde{G}_{C2}(z) \int \left[ \tilde{G}_{A1}(r) - \tilde{G}_{B1}(r) \right] f_rdr = 0$$

where $z$ is the point of single-crossing of $\tilde{G}_{A1}$ and $\tilde{G}_{B1}$. In the first line, I scale the difference in second-round conditional approval probabilities by the common probability of first-round approval to reduce the number of terms to carry around.

Key to this result is the monotonicity of $\tilde{G}_{C2}(\cdot)$, which means that the second-round judge is more likely to approve claimants with higher quality $r_i$ and his decisions are informative about which first-round judge has chosen higher-quality first-round claimants. Note in particular the contrast to the bounding exercise in Section 3.2. Under those weaker assumptions, difference in the second-round approval rates of judges’ approved claimants implies that the judges would disagree on the correct decision for some claimants, but has no normative interpretation. Here, the index representation of the judge decision in Equation 16 and Assumption 1 makes the more specific claim that the first-round judge who has more of her approvals subsequently approved is more consistent.

Identification of second-round consistency follows a similar route. Because I focus on first-round inconsistency in the empirical results, I defer discussion of judge-assignment identification in the second round to Appendix A3.

21
3.4.2 Regressors and identification

The between-judge comparisons that I discuss in the previous section are local; they measure relative consistency for judges with similar approval rates. To compare judges who approve different shares of claimants and to identify the scale of judge errors \( \tilde{\varepsilon}_{ij} \) without resorting to functional form assumptions, additional large-support continuous regressors \( X_{ijs} \) are required. These regressors affect judge severity through thresholds \( \gamma_{js} \) but do not otherwise affect errors. In a nonparametric sense, they are used as special regressors to identify the distribution of the composite error \( \tilde{\varepsilon}_{ij} - r_i \) for each round. The following assumption provides the necessary conditions.

**Assumption 3** (Support and scaling).

(a) \( X_{ij1}\beta_1|\gamma_j, W_{ij1} \) is continuous with large support, and independent of \( r_i \) and \( \tilde{\varepsilon}_{ij1} \).

(b) \( X_{ik2}\beta_2|X_{ik1}\beta_1, \gamma_j, \gamma_k, W_{ij1}, W_{ik2} \) is continuous with large support and independent of \( r_i \) and \( \tilde{\varepsilon}_{ij2} \).

(c) There exist some columns of \( X_{ij} \), referred to as \( X_{ij}^* \), that are not in \( W_{ij} \).

(d) At least one element of \( \beta_1 \) associated with \( X_{ij}^* \) is equal to the same element in \( \beta_2 \).

Assumptions (a) and (b) are standard support conditions for special regressors. Assumptions (c) and (d) pin down the relative scale of errors in each round. As I explain in the Appendix, these conditions together fit the requirements for identification of duration models first given in Chen, Heckman, and Vytlacil (2000).

**Theorem 3** (Identification). Suppose that Assumption 1 and Assumption 3 are satisfied. Then, the coefficients \( \gamma_{js} \) and \( \beta_s \), and the distributions \( F_r \) and \( G_{js,W} \) are identified.

**Proof:** See Appendix Section A4.

3.5 Interpretation

3.5.1 1st versus 2nd round inconsistency

Interpreting consistency in the second stage is complicated by the possibility that the judges gain additional information about the case in the second round (see Section 2.2 for details). I model this by making a distinction between judge errors \( e_{ij} \) and information \( I_{ij2} \), decomposing second-round errors into

\[
\tilde{\varepsilon}_{ij2} = e_{ij2} + I_{ij2}
\]  

(19)
The information shock $I_{ij2}$ represents information that if explained (e.g., written down in an opinion) would shift the cross-judge consensus on case strength. The subscript $j$ reflects that some judges may be better at finding this information, and thus have a wider distribution of $I_{ij2}$.

The potential presence of $I_{ij2}$ complicates interpretations of the size of the distribution of inconsistency, because not all of the variation in $\tilde{\varepsilon}_{ij2}$ is attributable to judge errors—some may reflect new information. A wide distribution of $\tilde{\varepsilon}_{ij2}$ could in fact reflect a particularly perceptive judge, so the relationship between judge quality and second-round consistency could go in either direction. For this reason, I focus my discussion on first-round consistency $\tilde{\varepsilon}_{ij1}$.

### 3.5.2 Interpreting the magnitude of inconsistency

The most straightforward measure of judicial inconsistency is the share of claimants that judges disagree on. This comparison is particularly sharp between judges who approve the same share of claimants, because under perfect consistency for both judges they would agree on all cases. Suppressing covariates so that $\varepsilon_{ijs} = \gamma_{js} + \tilde{\varepsilon}_{ijs}$, for judges $j$ and $k$ in the first round, disagreement can be calculated as

$$\delta_{jk} = \int \int \int \left\{ \mathbb{1}[r > \varepsilon_{ij1}] \mathbb{1}[r < \varepsilon_{ik1}] + \mathbb{1}[r < \varepsilon_{ij1}] \mathbb{1}[r > \varepsilon_{ik1}] \right\} f_r \, d\varepsilon_{ij1} \, d\varepsilon_{ik1} \, dr \quad (20)$$

This quantity, like disagreement in the bounding model, is only partially identified. The model identifies the distributions of $\tilde{\varepsilon}_{ij1}$ and $\tilde{\varepsilon}_{ik1}$, but does not identify their joint distribution. In plain language, it is possible that a particular claimant would be highly likely to be approved by all first-round judges, even though he is unlikely to be approved in the second round (e.g., low case strength $r_i$ but a high draw of the observational error $\tilde{\varepsilon}_{ij1}$ for all judges).

One way to proceed would be to bound $\delta_{jk}$. Since this is essentially the same approach I pursue in Section 3.2, I instead assume that cross-judge errors are independent in the first round.\(^{25}\) This assumption is fundamentally untestable, because we never observe different first-round judges making a decision on the same case. However, as I explain in the following section, the joint distribution of first- and second-round decisions is informative about the correlation between first- and second-round judge errors. I find little evidence of correlations between those errors for different judge pairs, suggesting that independence between same-round judges may also be reasonable.

**Assumption 4 (Point-identifying disagreement).** Cross-judge errors within the first round are independent, or $\tilde{\varepsilon}_{ij1} \perp \tilde{\varepsilon}_{ik1} \forall j \neq k$.

Under Assumption 4, the share of case on which two first-round judges would disagree is point-identified as $\delta_{jk} = \int \left[ G_{j1}(r) + G_{k1}(r) - 2G_{j1}(r)G_{k1}(r) \right] f_r \, dr$.

\(^{25}\)The structural analog to the bounding approach is to note that for each $r$, a pair of judges disagrees with at least probability $\left| G_{j1}(r - \gamma_{j1}) - G_{k1}(r - \gamma_{k1}) \right|$, and so $\delta_{jk}$ can be bounded by integrating over $r_i$. 

23
3.5.3 Taste-based versus observational errors

Under the index model of judicial decision-making and Assumption 4, more inconsistent judges are those who are worse at observing signals of quality. An alternative interpretation of the data is that different judges have different—but potentially legitimate—standards. In this alternative view, two first-round judges might always agree with each other, but disagree with the rest of their colleagues. Although this distinction is difficult to test, in this section I introduce a test for whether pairs of judges in different rounds disproportionately agree or disagree, and argue that if—as in fact it turns out—there is evidence in favor of independence of the cross-round errors for all judge pairs, that suggests that the within-round errors may also be independent.

To understand how Assumption 4 places potentially strong restrictions on how pairs of judges can share a common taste for certain kinds of claimants above and beyond that captured by scalar claimant quality $r_i$, I disaggregate judge errors into

$$\tilde{\varepsilon}_{ij1} = \lambda_j C_i + e_{ij1}$$

(21)

where $C_i$ are characteristics of the claimant conditional on quality $r_i$, $\lambda_j$ the judge-specific taste for that attribute, and $e_{ij1}$ is a pure observational error. $C_i$ could include characteristics like race or details about the case, and may be unobserved to the researcher.

Conversely, $e_{ij1}$ is an observational error, or failure to understand the merits of the case. It is a measure of test-retest consistency. If a judge was repeatedly given the same case $i$ (without memory of her previous decisions), she would observe them as having quality distributed as $r_i - \lambda_j C_i - e_{ij1}$, with $r_i - \lambda_j C_i$ fixed and variation coming only from $e_{ij1}$.

Assumption 4 assumes that $\lambda_j C_i$ is small. An indirect way to test whether this is to ask if certain pairs of judges in different rounds value the same claimants, above and beyond what would be predicted by their overall approval rates. Suppose that the composite error mostly reflected permanent differences across judges ($\lambda_j C_i$ is high variance). Then, one would expect that some pairs of judges would both value the same type of cases—for example, cases that involved an Asian claimant. By definition, the probability of approval in the second round for a claimant with quality $r_i$ and judges $j$ and $k$ (and suppressing extra regressors) is

$$P_{jk} = P[\text{Approval by } j | \text{Approval by } k \text{ in 1st}] = \frac{P[r_i > \gamma_{j2} + \tilde{\varepsilon}_{ij2} \cap r_i > \gamma_{k1} + \tilde{\varepsilon}_{ik1}]}{P[r_i > \gamma_{k1} + \tilde{\varepsilon}_{ik1}]}$$

If two judges have a similar preferences over characteristics, then they will both treat the same case either more or less positively than would be predicted by the factor refugee quality $r_i$, their $\gamma$’s and their $\sigma$’s. More formally, index $P_{jk}$ by the correlation in permanent errors $\lambda_j C_i$ and $\lambda_k C_i$, $\rho_{jk}$. It is simple to show that $P_{jk}(\rho_{jk})$ is increasing in $\rho_{jk}$.  

24
By Assumption 1, the between-round errors are uncorrelated, so $\rho = 0$ for each judge pair. However, the parametric model that I estimate summarizes judge behavior as depending on only two parameters in each round. When the number of judges is even moderately large there are more judge-pairs ($J^2 - J$, not counting the same judge in each round) than parameters determining $P_{jk}(\rho_{jk})$ (4J). This suggests a reduced-form overidentification test for residual judge-pair correlations, regressing

$$1[\text{Approval by } j|\text{Approval by } k] = \beta\hat{P}_{jk}(0) + \nu_{jk} + u_{ijk}$$

(22)

where $\hat{P}_{jk}(0)$ is calculated using the estimated model parameters. Under the null of no correlations in errors between judge pairs, the judge-pair fixed effects $\nu_{jk}$ should be jointly insignificant. This is a joint test of all the reasons pairs of judges could disproportionately agree or disagree, but failure to reject suggests that inter-rater differences $\lambda_j C_i$ are not large, and idiosyncratic observational errors $e_{ij}$ dominate. I implement this test in Section 4.8, and fail to reject the null of no judge-pair effects.

4 Data and results

4.1 Data

My main data come from Federal Court case reports available on their website. I parsed the data and verified it against a subset professionally transcribed by Rehaag (2012). I use all contested cases since 1995 that were filed at the IRB before June 28, 2012 (as discussed in Section 2.1, the institution changed considerably after this date), and appealed to the Federal Court before the end of 2012 to ensure that a final decision is observed for all cases. The dataset has information on the date the case was filed, the Federal Court office that received the application, the name of the first- and (if applicable) second-round judge, and the ultimate outcome. The office is an important covariate because it strongly predicts outcomes at the court—provinces differ in the level of free legal aid provided to claimants. I exclude offices with fewer than 200 cases, leaving Calgary, Montreal, Ottawa, Toronto, and Vancouver.

Using the first name of the claimant, I infer gender using British Columbia and Social Security Administration birth records that contain both first name and gender. To collect information on the country of origin of the claimant, I link the court records to the subset of available IRB case files. These data contain the name of the IRB Member who made the initial determination, and in some cases the country of origin and gender of the claimant. I also use the commercial service Onomap to predict country of origin for each claimant, which I collapse to continent dummies.

For each judge, I collected information on the date and party of appointment. Appendix Table

---

26http://cas-cdc-ww02.cas-satj.gc.ca/IndexingQueries/infp_queries_e.php
A1 shows that 25% are female, and their dates of appointment range from 1982 to 2010. Since the Liberals held power for most of this time period, 72% of judges are Liberal appointees. The average judge has 6.5 years of experience with a maximum of 28.

4.2 Estimation

Fully nonparametric identification of the structural model as outlined in Section 3.4 requires special regressors with large support conditional on judge assignment. This is a very high bar, and one that is not met in my empirical application. Instead, I parameterize the distributions of case strength $r_i$ and judge errors $\varepsilon_{ij1}$ and $\varepsilon_{ij2}$, generating tractable analytic expressions for approval probabilities and allowing rapid estimation of the entire model by maximum likelihood (the derivations are available in Appendix Section A5). I begin by assuming that $\varepsilon_{ij}$ is mean-zero and normally distributed with standard deviation $\sigma_{js}$ to be estimated as the measure of judge inconsistency. Larger $\sigma_{js}$ corresponds to more inconsistency, i.e., a wider distribution of judge errors $\varepsilon_{ij}$. I additionally allow regressors $W_{ij}$ to affect errors, so

$$\varepsilon_{ij}(W_{ij}) \sim \mathcal{N}(0, \sigma_{js}(W_{ij})^2), \quad \sigma_{js}(W_{ij}) = e^{\overline{\sigma}_{js} + W_{ij}\psi_s} \tag{23}$$

Since occasionally the same judge is assigned to the first- and second-round decision, I allow $\varepsilon_{ij1}$ and $\varepsilon_{ik2}$ to be correlated (with the correlation estimated as an additional parameter) whenever $j = k$.

As discussed in Section 2.2, the distribution of unobserved case strength for claimants at the Federal Court is likely right-skewed, since it is the distribution of individuals who were denied refugee status by government decision-makers. A relatively small number of high-case strength refugees are likely not initially granted status by the government. To reflect this intuition I therefore assume that the distribution is single-tailed; $r_i$ is exponential-Pareto distributed (I show the distribution of $r_i$ relative to the estimated parameters in Figure 4). I allow flexibility in the distribution of $r_i$ across two dimensions. First, cases filed at different offices vary considerably in strength. This is partially because office is correlated with country of origin, but more closely related to varying levels of legal aid funding. Second, the government made changes to the decision process for initial refugee applications in 2002 (see Section 2.1 for more details). Combining these, I fix the distribution of $r_i$ to have a scale and shape parameter of 1 for the largest office (Toronto) before 2002, and then separately estimate the scale and shape parameters for each office before and after 2002. In practice, I find almost no difference in the distribution of case quality between these time periods, suggesting

---

27The two main political parties in Canada are much closer ideologically than the major parties in the United States, as are the judges they appoint. There is less dissent within the legal community about the correct approach to statutory interpretation, although the Conservative party is generally more skeptical of refugee claims than the Liberal party.
that the institutional changes had little affect on which claimants were approved. However, the 
between-office variation is considerable.

As discussed in Assumption 3, nonparametric identification requires a subset of regressors $X_{ij}^*$ 
that shift judge thresholds $\gamma_{js}$ but do not affect judge errors $\bar{\varepsilon}_{js}$. One immediate implication of this 
is that $X_{ij}^*$ is not in $W_{ij}$, so that there is variation in $X_{ij}^*$ conditional on $W_{ij}$.

For $X_{ij}^*$ I use the timing of the decisions during the week, and of the second-round hearing 
during the day, as regressors. Danziger, Levav, and Avnaim-Pesso (2011) argue that when decision-
makers make many decisions in a row, they become fatigued and are more likely to pick the default 
option. They study parole decisions in Israel and find that rejections become more likely just before 
lunch and revert to baseline levels immediately after the break. I follow them and use a dummy for 
the noon hearing as a regressor.28

Second, I exploit the fact that judges make decisions on refugee cases only irregularly (non-
refugee cases occupy much of their time). Judges who are making decisions in the first stage are 
given a tranche of case files at the start of the week and work on them until they have made all 
the decisions. One might expect that judges would become more fatigued as the week goes on. 
Ideally, I would define a decision as having been made late if it was made on Tuesday or later, 
after the first full day of decisions. However, judges may endogeneously change the order in which 
they make decisions as a function of case characteristics (concretely, they may delay the difficult 
decisions), which would violate the exclusion restriction. I therefore include a dummy for the last 
submission before the judge’s decision happening on Wednesday or later as a regressor—empirically, 
this predicts the leave decision is likely to be made Tuesday or later. Similarly, for second-round 
cases the scheduling is done by court staff without knowledge of the case characteristics. I define the 
end-of-week regressor in the second round as the case being heard on Wednesday or later (Monday 
hearings are rare, so Wednesday is usually the second day of hearings). I assume that the effect of a 
late-week hearing is the same in both the first and second round, satisfying part (d) of Assumption 3.

In Appendix Section A6.1 I show that these regressors are highly predictive of approval, sug-
uggesting that they lend identifying power in the spirit of the support conditions of the Assumption 3. 
However, they are also uncorrelated with the likelihood of approval as predicted by fixed demo-
graphic characteristics of the claimants, suggesting they are orthogonal to judge errors $\bar{\varepsilon}_{ij}$($W_{ij}$). 
To further test this assumption, I estimate versions of the model where $X_{ij}^*$, the excluded com-
ponents of $X_{ij}$, are added in turn to $W_{ij}$—in other words, I allow the regressors to directly affect 
the level of inconsistency. Since there are two regressors in the second round (dummies for the end 
of the week and a lunchtime hearing), I then test whether $\sigma_{j2}^2(W_{ij2})$ varies with the regressors and 
find that the effect is small and statistically insignificant, in line with the assumption.

28There are two main reasons why one would expect denial to be the default option. First, most appeals are 
rejected. Second, because Canadian law requires judges show deference to governmental decision-makers, judges tend 
to see overruling IRB decisions as the exception rather than the rule.
As a final robustness check, I estimate the model without relying on regressors. Although identification then leans more heavily on functional form assumptions, it necessarily avoids making identifying assumptions on how the regressors affect judge errors. In Appendix Section A6.2, I present estimates for this model and show that all the results are qualitatively similar.

Estimation throughout is by maximum likelihood. Standard errors are clustered at the level of the first-round judge.

4.3 Randomization tests

In Table 1 I explore whether the cases are assigned quasi-randomly to judges. One implication of quasi-random assignment is that judge characteristics should be unrelated to case characteristics. To test this, for each round I regress claimant characteristics on judge-level mean approval rates in that round, controlling for office X pre-2002 fixed effects to account for office and time variation analogously to the controls in the structural model. I also regress the characteristics on judge fixed effects and report the F-stat and p-value for the joint test of the judge fixed effects.

The predictive power of judge assignment is low in the subsample of IRB-linked case files where I observe claimant characteristics. The coefficients from the regression of covariates on judge-level approval rates are all insignificant. Similarly to other examiner-effect contexts with random or quasi-random assignment, the F-statistics are small (about 1) but the test is sensitive enough to reject slightly more than half the time (Mueller-Smith, 2015).

Columns 1-4 are relatively straightforward outcomes: gender and region of origin. Column 5 is the mean approval rate of the IRB Member that denied the claimants’ initial application for refugee status. Because it is not obvious how to weigh the different columns, I predict round-specific approval using the demographic variables in columns 1 through 5. Then, I use this predicted value as the regressor in Column 6. In this omnibus test, the coefficient on judge approval is small and insignificant. Finally, in Column 7 I regress the 1st-round judges mean approval rate on the 2nd-round judge’s. The 2nd-round judges’ approval rates do not predict the 1st-round judge’s, suggesting that assignment between rounds is quasi-random.

Claimant demographics come from the IRB case files, which are only available for about 85% of cases. In Appendix Table A2, I display similar regressions for the entire sample, substituting gender and continent of origin imputed from claimant name as dependent variables. Judge leniency has some predictive power for imputed continent of origin, but not in a way that is correlated with predicted approval.
4.4 Reduced form judge behavior

Federal Court judges are obliged to show deference to the government’s initial determination of refugee status. Perhaps because of this, approval rates in the first round are low, at only 14%. There is a large amount of cross-judge heterogeneity: the histogram in Panel A of Figure 1 shows that four judges approved less than 5% of cases, while one judge approved 70% (after this judge, the next highest rate is 28%).

The second stage approval rate is much higher, at 44%. Similarly to the first round, there is a large amount of dispersion in approval rates, from 13% to 87%. The dramatic improvement in the success rate in the second round suggests that first-round judges are successful to some extent in terms of choosing claimants who can make an “arguable case” in the second round. In terms of the structural model, this implies that variation in refugee quality \( r_i \) is substantial. More evidence in favor of a latent factor \( r_i \) that judges have systematic preferences over can be seen in Panel C of Figure 1, which shows that there is a high correlation (0.56) between the first- and second-round approval rates for the same judge.

To the extent that there is a common case strength factor observed by judges, claimants approved in the first round by strict judges should fare better in the second round than those approved by lenient judges. Table 2 conducts this analysis, regressing second-round approval on the exclusive mean approval rate for the first-round judge. Moving across the columns, I include no other controls, the mean approval rate of the second-round judge, and second-round judge FE s. The results are similar; having been approved by a 10 percentage point more lenient first-round judge gives claimants a 2.6-3.2 percentage point lower chance of being approved in the second round. Individuals who were approved by a more lenient judge in the first stage have, on average, a weaker case in the eyes of the second-round judges. This again suggests that there is a refugee quality factor \( r_i \) that is commonly observed by the judges up to some observational error.

For evidence on variation in \( \sigma_{js} \), judges’ ability to pick the highest-quality claimants, I turn to Figure 2. One source of identifying variation in the structural model comes from how often a first-stage judge’s approved claimants are approved by a different judge in the next round (difference in success rates also implies disagreement in the Section 3.2 sense). First-round judges with more ultimately successful claimants, the model says, are better at picking high-\( r_i \) cases. Figure 2 displays reduced form evidence on the size of the variation in this ability. For each first-round judge, I take the mean approval rate of her approved claimants in the second round, residualizing out the approval rate of the second round judges and shrinking the estimates towards the grand mean via Empirical Bayes to account for small cell sizes. The Figure shows a scatter of this measure of judge ability against the same judges first-round approval rate. As expected, the relationship between a first round judges’ approval rate in the first round and the second-round approval rate of her first-round approvals is negative. However, there is a large degree of cross-judge dispersion in second-round
approval for each first-round approval average—the subsequent approval rate for judges approving about 15% of first-round applicants ranges from 38 to 48%. In other words, there is a lot of cross-judge variation in ability to pick claimants who will be approved in the second round, suggesting a high level of inconsistency.

4.5 Bounding judge disagreement

In Section 3.2, I describe how multiple rounds of decision-making allow the construction of nonparametric bounds on the share of cases for which a pair of first-round judges would make a different decision. In Figure 3, I calculate these bounds for each pair of first-round judges with a similar overall approval rate. I restrict to judges with an approval rate within 1% after partialling out controls for a late-week decision and the office of origination interacted with a dummy for the case being filed before and after 2002, analogously to the structural model. For each first-round judge pair \( j \) and \( k \), I then implement the bound analysis by estimating Equation 15, a regression of second-round approval on the identity of an arbitrary first-round judge interacted with the identity of the second-round judge. As explained in Section 3.2, the plug-in estimator of the lower bound for disagreement is

\[
\delta_{jk}^\ell = 2P_1 \max(\Delta^1_{jk}, -\Delta^1_{jk}, \Delta^2_{jk}, -\Delta^2_{jk}, \ldots, \Delta^J_{jk}, -\Delta^J_{jk})
\]

where \( P_1 \) is the limiting value of the two judges first-round approval rates, and \( \Delta^\ell_{jk} \) is the difference between judge \( j \) and \( k \)'s subsequent approval rates under second-round judge \( \ell \). I substitute in the mean judge value for \( P_1 \), and instead of the plug-in estimates for the maximum, I calculate half-median-unbiased estimates of the bound and the endpoints of 95% one-sided confidence intervals (Chernozhukov, Lee, and Rosen, 2013). Figure 3 reports the estimates \( \hat{\delta}_{AB}^\ell \) and the lower endpoints of the CI. To account for the estimated first-round approval rates not being exactly the same, I adjust the bounds by subtracting the absolute difference in approval rates.

I individually test each of the 170 pairs of first-round judges, and reject perfect agreement for 123. This rejection rate is too high given multiple testing, so I also adjust the critical values using the step-down method of Romano and Shaikh (2006), controlling the false discovery proportion to 10%. Under no assumptions on the relationship between the test statistics, I reject perfect consistency for 91 pairs, implying that at least 82 pairs of judges with the same approval rate would disagree on the correct decision for some claimants. Since inconsistency implies violations of monotonicity, this is also a rejection of monotonicity for first-round judges. I directly test monotonicity and bound the share of defiers in Appendix A8, and discuss the implications in Section 4.13.

\(^{29}\)Tests that control false discovery proportion (FDP, the ratio of false rejections to total rejections) to a given level \( \gamma \) ensure that \( P\{FDP > \gamma\} \leq \alpha \).
The average lower bound on disagreement, taken across the judge pairs, is 10%, with a maximum of 35.6% (the corresponding endpoints of the CIs are 6% and 31.7%). This level of disagreement is particularly high given that the overall first-round approval rate is only 14.4%.

Also striking is the extent to which having multiple judges tightens the bounds. It is possible to conduct a comparable analysis without granular information on the identity of the second-round judge. Instead of estimating Equation 15, I estimate

\[ y_{i2} = \Delta_0 + \Delta_{AB}^* D_{iA1} + \epsilon_i \]

so \( \Delta_{jk}^* \) measures the overall difference in second-round success between first-round judges \( j \) and \( k \) analogously to Figure 2, and \( \delta_{jk}^1 = 2P_1 \max(\Delta_{jk}^*, -\Delta_{jk}^*) \). Under this specification, the bounds are almost entirely uninformative: I cannot reject perfect agreement for 150 of the 170 pairs, and the average disagreement is only 5%. I take this as evidence that observing many different second-round judges substantially improves precision.

To explore the degree and causes of inconsistency further, for the rest of the paper I turn to the fully structural model of Section 3.3. This model has the advantage of aggregating data up into a single, judge-specific measure of consistency, as opposed to a judge-pair-specific measure like the one presented here. That allows me to more accurately measure how consistency—and by implication, disagreement—varies with judge covariates and other factors.

### 4.6 Structural results

The structural model estimates the parameters of the judges decision about whether to approve a claimant of quality \( r_i \) displayed in Equation 16 and reproduced here:

\[ r_i > \varepsilon_{ijs}(X_{ij}, W_{ij}) = \gamma_{js} + X_{ij} \beta_s + \tilde{\varepsilon}_{ijs}W_{ij} \]

\[ \tilde{\varepsilon}_{ijs}(W_{ij}) \sim \mathcal{N}(0, \sigma_{js}(W_{ij})^2), \quad \sigma_{js}(W_{ij}) = e^{\tilde{\sigma}_{js} + W_{ij} \psi_s} \]

My baseline model includes as regressors \( X_{ij} \), a dummy for a late-week decision and a dummy for whether the second-round hearing was heard at lunch. I assume that \( \beta_1 = \beta_2 \). Thresholds \( \gamma_{js} \) vary by judge and round, and inconsistency \( \sigma_{js} \) varies by judge-round but not by any covariates (ie, \( W_{ij} \) is empty). Case strength \( r_i \) is distributed as an exponential-Pareto, with different parameters in each office of case origination and before and after 2002 (and the largest office-time period normalized to have a location parameter of 0 and a scale parameter of 1), when there were changes to how the government made initial refugee determinations. I discuss these modeling decisions in detail in Section 4.2, and defer discussion of robustness and specification checks to Appendix A6.

Figure 4 plots the distribution of judge-round thresholds \( \gamma_{js} \) and inconsistency \( \sigma_{js} \). The red dot-
ted lines are the raw coefficients. Because of estimation error the distribution of the raw coefficients is slightly too wide; in blue I plot the distribution of the underlying coefficients after deconvolving out the measurement error (Delaigle, Hall, and Meister, 2008). The coefficients are precisely estimated, so for most of the estimates the distributions are similar.

In Panel A, the distribution of $\gamma_{j1}$ is large relative to the distribution of refugee strength $r_i$, plotted in black. This tells us that not all variation in approval rates across judges is coming from differences in ability to pick the highest-quality claimants—there are real differences in standards across judges.

In Panel C, the distribution of second-round thresholds $\gamma_{j2}$ is narrower and slightly smaller on average than in the first round. The second-round distribution of $\gamma_j$ does not quite second-order stochastically dominate the distribution of $\gamma_{j1}$, but the standard deviation is substantially smaller (0.39 vs. 0.74). I interpret this as reflecting the larger amount of precedent available to judges in the second round relative to the first round. In the first round, no decisions are written up, making it difficult for judges to learn about the decisions their colleagues have made in similar situations. Conversely, nearly all second-round decisions are published as precedent, so the distribution of standards is tighter.

Panel B of Figure 4 shows the round-specific distributions of $\sigma_{js}$. The standard deviation of the first-round judge errors for the median judge is 1.08, which is large relative to the standard deviation of 1 for underlying case quality $r_i$. To contextualize the size of the inconsistency, I match pairs of first-round judges with approval rates within one percentage point. I calculate that the average pair of judges with the same first-round approval rate disagrees on 13.2% of all cases, slightly more than the nonparametric lower bound estimate of 10%. This is large relative to the overall first-round approval rate of only 14%. Interestingly, it is also remarkably close to the model-estimated disagreement rate of 14.6% for first- and second-round judges if second-round judges saw all cases rather than only approved ones.

How large is a disagreement rate of 13.2% for judges with the same first-round approval rate? If all judges were perfectly consistent, there would be no disagreement. Alternatively, if the judges were perfectly inconsistent—all picks were made by tossing a fair coin that indicated approval at the judges' overall approval rate—their average disagreement rate would be 23.2%. In other words, judges are slightly worse than halfway between flipping coins and making perfectly consistent decisions.

Another relevant comparison is to the level of disagreement engendered by cross-judge variation in leniency. For all pairs of first-round judges (rather than just those with the same approval rate), the average disagreement rate is 23.1%, suggesting that differential leniency and inconsistency among similarly-severe judges contribute about equally to the unequal treatment. Similarly, if all judges were perfectly consistent, cross-judge variation in approval rates implies that the average pair of
judges disagrees on 8% of cases, meaning that moving to a world where judges are inconsistent (and so the disagreement rate is 23.1%) nearly triples the level of unequal treatment.

The distribution of $\sigma_{j1}$ in Panel B is wide—some judges are much more consistent than others—so replacing judges with more consistent alternatives dramatically reduces disagreement. If the least-consistent half of judges were made as consistent as the median judge (adjusting thresholds $\gamma_{j1}$ to keep approval rates the same), disagreement among similarly-severe judges would fall from 13.2 to 8%. Policies that replaced the least consistent judges with average replacements would therefore have an important effect on the overall consistency of the justice system.

In Panel D of Figure 4, the size of the distribution of second-round errors $\sigma_2$ is harder to interpret. Recall from Section 3.5.1 that the second-round error may contain an informational component $I_{ij2}$. If this is the case, then a larger $\sigma_{j2}$ may reflect better information-gathering abilities in the second-round hearing. Some suggestive evidence that information-gathering is an important part of second-round errors is in Appendix Figure A1, a scatter of judge-specific first- and second-round $\sigma_{js}$. The correlation is small and slightly negative (-0.043). It is reasonable to assume that the size of the non-information components of first- and second-round errors are positively correlated, given that the other measurable aspects of judge behavior (thresholds $\gamma_j$, overall approval rates) are correlated across rounds. This suggests there is a non-trivial information shock in the second round. The size of this shock is negatively correlated with the magnitude of first-round inconsistency—consistent first-round judges are also better at uncovering relevant information in the second-round hearing.\(^{31}\) I take this as a further reason to focus on first-round consistency for the remainder of the paper.

First-round judges do a poor job of predicting which claimants will be successful in the second round; for most claimants the probability of second-round approval is higher than their probability of first-round approval. Figure 5 shows how this works. For each $r_i$, I calculate the first-round approval probability and the second-round approval probability conditional on first-round approval. I then plot them against each other. The figure shows that the median claimant has a 6% chance of first-round approval, but conditional on approval has a 12% chance of approval in the second round. This does not reflect selection, which is accounted for by conditioning on $r$. Instead, it shows that second-round decisions are unpredictable from the perspective of the first round, and even claimants with a relatively low quality factor $r_i$ are sometimes approved in the second round. Furthermore, no one has a very high chance of approval: the 95\(^{th}\) percentile claimant has only a 62% chance of first-round approval and a 66% chance of second-round approval—a 41% chance of overall success at the Federal Court. The empirical first-round approval rate is 14%; however, if the highest-$r_i$ 14% was selected in the first round the overall approval rate would only climb to 8.5%.

\(^{30}\)Because I adjust the thresholds $\gamma_{j1}$ to keep approval rates the same, these changes reflect only changes in the pair-specific disagreement rates, not changing pair composition.

\(^{31}\)Decomposing the first and second round residuals, we know that $\text{corr}(\text{var}(e_{ij1}), \text{var}(e_{ij2}) + \text{var}(I_{ij2})) \approx 0$, so if $\text{var}(e_{ij1})$ and $\text{var}(e_{ij2})$ are positively correlated that implies $\text{var}(e_{ij1})$ and $\text{var}(I_{ij2})$ are negatively correlated.
Integrating over the entire distribution of \( r \), I find that on average 19.4\% of claimants would be approved in a second-round hearing if they were approved in the first round. This is in contrast to the 6\% of claimants who are approved under the current system. Although it may be surprising that the number of refugee appeals granted by the Federal Court would triple under this alternative decision-making process, the result is foreshadowed by the Figure 2 scatter of first-round approval rates by judge against the approval rates for that judges approved claimants in the second round. After partialling out controls, the most lenient judge approved 66.7\% of claimants in the first round, and of those 27.4\% were approved in the second round, bounding the overall approval rate in the absence of a first round at 18.3\% (0.667 × 0.274). In terms of the policy implication, an important caveat is that judges might change their second-round behavior in unpredictable ways if all cases were automatically approved in the first round—one first-order effect would be that each judge would have 7 times more second-round cases to hear.

4.7 Relationship between structural parameters and reduced form moments

In this section, I show how reduced-form moments translate into the structural parameters in the first round (I defer similar discussion for second-round inconsistency to Appendix Section A6.3). In the model, the principal determinant of first-round approval rates is judge thresholds \( \gamma_{j1} \). In Panel A of Figure 6, I vary the \( \gamma_{j1} \)'s from their estimated values by adding a common shifter to each \( \gamma_{j1} \). As they change, the estimated approval rate moves away from the observed value of 14\%. Reassuringly, the change is monotonic and steep.

As discussed in Section 3.4.1, a main source of identification of the overall size of first-round judge inconsistency \( \sigma_{j1} \) is whether the approved claimants are subsequently approved in the second round. In Panel B, I adjust the \( \sigma_{j1} \)'s away from their estimated values by multiplying each coefficient by a common factor. As inconsistency increases (\( \sigma_{j1} \) gets larger), this dramatically reduces the second-round approval rate.

Panels C hews closer to the judge-pair comparison intuition. Comparing two judges with the same first-round approval rate, Equation 18 indicates that the more consistent judge has the higher approval rate in the second round for her approved claimants. In Panel C, I match judges with first-round approval rates within 1 percentage point of each other, then plot the difference in second-round approval rates against the difference in estimated \( \sigma_{j1} \). As expected, judges with a higher second-round approval rate than their matched colleague have a lower estimated \( \sigma_{j1} \), or are more consistent.
4.8 Taste-based versus observational errors

In Section 3.5.3, I discussed how judicial inconsistency can be decomposed into permanent differences in preferences between judges conditional on scalar quality r_i (inter-rater inconsistency) and pure observational errors (test-retest inconsistency). Point-identification of disagreement requires that the errors be all observational, rather than permanent disagreements (Assumption 4). One test of the assumption is to ask whether particular pairs of judges in different rounds disproportionately agree with each other. If the correlation is caused by permanent inter-rater differences, it is likely that certain pairs of judges share the same weighting over different aspects of cases. Then, in a regression of second-round approval on model-predicted likelihood of approval \( P[\text{Approval by } j_2|\text{Approval by } k_1] \) and judge-pair fixed effects \( \nu_{jk} \), the fixed effects should add meaningful predictive power.

I implement the test in Table 3. The left two columns take order into account when constructing the judge pairs (ie, judge A then judge B is different from judge B then judge A), while the rightmost two columns ignore ordering. I drop cases where the same judge made both the first- and second-round decisions.\(^{32}\)

Across columns, the coefficient on \( P[\text{Approval by } j_2|\text{Approval by } k_1] \) is very close to one. However, in all specifications the F-stat for the joint test of judge-pair fixed effects is about 1, corresponding to \( p \)-values in the range of 0.30-0.60.\(^{33}\) In other words, the judge-pair effects do not predict second-round approval beyond the model estimates. This suggests that there is in fact no correlation between errors \( \tilde{\epsilon}_{ij} \) for cross-round judge pairs, and that taste-based errors are small relative to observational errors. As a descriptive analysis, I calculate the Empirical Bayes judge-pair means of the residuals of a regression of second-round approval on \( P[\text{Approval by } j_2|\text{Approval by } k_1] \).\(^{34}\) This confirms the F-stat result: the standard deviation of the judge-pair means is only 0.004 relative to a mean approval rate of 0.44.

One potential issue with the judge-pair test is that it may be underpowered because there are so many estimated coefficients. Monte Carlo simulations indicate that the power of the test reaches 0.8 only when the true SD of the judge-pair effects is 0.055, which is relatively large. An alternative is to replace the judge-pair fixed effects in Equation 22 with interactions of judge demographic characteristics. Now, instead of testing whether there is residual explanatory power in knowing the identity of the judge pair that made the decision, this test asks whether observable groups of judges share preferences for certain types of claimants. I conduct this test in Table A10 for judge experience (more than 1 and more than 5 years of experience), gender, political party of appointment (Liberal

\(^{32}\)Correlation in same-judge errors could be linked to strong judge-specific tastes for characteristics, or to a misreading of the facts of the case that is common to both rounds. The baseline model estimates this correlation for same-judge pairs and finds it is 0.32 (SE=0.04), suggesting that one of these factors is large.

\(^{33}\)I report asymptotic \( p \)-values. Since the F-test can over-reject when the judge-pair cells are small, I also follow Abrams et al. (2012) and bootstrap the distribution of the null. In this case, however, the results are similar.

\(^{34}\)The Empirical Bayes means are the judge-pair residuals, shrunk towards the grand mean to account for measurement error.
or Conservative), and native language (French or English, though most are bilingual). With one exception—less than 1 year of experience, which I discuss in the following section—I find similar results to the judge-pair test. Estimated $p$-values range from 0.2 to 0.8, and the F-stats range from 0.39 to 1.48.

I take these two tests as evidence that the cross-round judge errors $\tilde{\epsilon}_{ijs}$ are mostly composed of observational errors, rather than differential weighting of different aspects of the case. 35 This is consistent with a large body of evidence discussed in the introduction, where factors as disparate as the upcoming judicial elections, the previous decision the judge made, the weather, and the winner of a sporting event the night before have been shown to influence judges decisions. Importantly, it implies that modeling errors as uncorrelated for judges in the same round (Assumption 4) is reasonable, and that disagreement is likely to be closer to the 13.2% estimated under that assumption than the 10% lower bound.

4.9 Judicial inconsistency, experience and workload

Judging is difficult. Particularly in this environment, where there are no published first-round decisions that allow judges to learn before they start work, an important question for understanding how to optimize court functioning is how quickly judges learn from experience. If judges learn slowly, that suggests that judicial churn is costly and should be avoided.

Table 4 presents models where I allow experience to enter the judge thresholds $\gamma_{j}$ (ie, in $X_{ijs}$) and the variance of the judge error, $\sigma_{j}$ (ie, in $W_{ijs}$) as in Equation 23. I parameterize experience with an indicator for more than one year of experience and with indicators for more than 1, 5, and 10 years of experience. Column 1 shows that first-round inconsistency $\sigma_{j1}$ shrinks dramatically after the first year (0.77 log points), followed by smaller but still substantial decreases of 0.29 log points after five years and 0.54 log points after ten years. To put these numbers into context, I estimate that if all judges had less than one year of experience, pairs of judges with the same approval rate would disagree 22% of all cases. After one year of experience for all judges that declines to 17%; after a total of 5 years, 14%; and after 10 years 10%. 36

This pattern of front-loaded gains to experience is similar to that observed in teachers, who see the most dramatic gains after the first year (Rivkin, Hanushek, and Kain, 2005). In contrast to teachers, I see further gains even after 10 years of experience, perhaps reflecting the more complicated nature of judging.

An important question is exactly what judges are learning. One likely possibility is that new

35 In the Appendix, I show results for the same test using the model that is identified without the use of regressors. I find almost identical results, alleviating the concern that the types of errors implied by the end-of-week and noon-time regressors are more likely to be idiosyncratic rather than ideological errors.

36 In each counterfactual I adjust the thresholds $\gamma_{j1}$ to keep overall approval rates the same. This means that the same judges are matched to each other in all counterfactuals.
judges slowly learn the relevant standards from their older colleagues. If that were true, then you could expect that claimants approved in the first round by new judges would be disproportionately rejected by their more experienced colleagues in the second round, even accounting for overall consistency and leniency of all judges. That is exactly what I find in Table A10, where Column 5 says that claimants are 9.6 percentage points less likely to be approved when the first-round judge has less than one year of experience and the second-round judge has more than one year of experience, but when both judges are at the same level of experience there is no difference. In other words, first-year and more experienced judges are operating on slightly different standards, and one of the ways that first-year judges improve is to learn those standards. In contrast to the overall improvement in consistency over at least the first ten years of experience, this particular type of learning is relatively short-lived. Analogously comparing judges with more and less than 5 years of experience, there is no evidence that judges in these groups are disproportionately likely to agree with each other than with the other group.

A potential implication of these high returns to experience is that more cases should be given to experienced judges. This will be particularly true if judges improve as a function of years on the job rather than directly from the number of cases they have seen.\textsuperscript{37} Improvement strictly over time is perhaps more likely at the Federal Court than elsewhere. Recall from Section 2.2 that first-round decisions are not written up as precedent. What’s more, the second-round decision is not made for months (the median wait is 89 days), making it difficult for first-round judges to follow up on which specific cases have been approved in the second round. The main way that judges learn how their colleagues make decisions in the first round are indirect: inferring them from the types of cases they observe themselves in the second round, and discussing cases with colleagues. Both of these factors suggest that years of experience will be more relevant than number of cases. I take this distinction to the data in Column 2 of Table 4, allowing years of experience and career number of cases to affect inconsistency. Analogously to experience, I implement the control for number of career cases with dummies, setting the thresholds at the 95\textsuperscript{th} percentile of the number of cases seen for judges with fewer than 1, 6, and 11 years of experience. Interestingly, the results are nearly the same: inconsistency declines by 0.69, 0.24, and 0.44 log points after 1, 5, and 10 years of experience, respectively. In contrast, the effect of number of cases on inconsistency is small and statistically insignificant, at effects (in log points) of 0.16 (SE=0.53), 0.066 (0.11), and 0.18 (0.67) after the career number-of-case thresholds corresponding to 1, 5, and 10 years of experience. This suggests that the Federal Court could indeed improve overall consistency by shifting some cases from inexperienced to experienced judges.

Another factor that may affect judicial consistency is workload. Higher caseloads reduce the amount of time judges can spend on each case. Chen, Moskowitz, and Shue (2016) show that refugee

\textsuperscript{37}If they improve as a function of number of cases, then the cost in terms of the number of inconsistent decisions does not depend on when the cases are given to the judge.
judges in the US make more gambler’s fallacy errors—judges being less likely to approve a claimant when they approved their last claimant—when caseloads are high, so one might expect to observe the same phenomenon here. To test this, I calculate monthly log workload as the number of leave cases a judge is assigned in a given month and add it to the vector of variables that affect judge thresholds ($X_{ij}$) and errors ($W_{ij}$). Judges are also responsible for non-refugee cases, so this is an imperfect measure of workload. The Court does not assign judges to cases or case types as a function of performance, fearing that this would result in challenges to the assignment procedure, so there is no reason to think that judge ability would be related to caseload. Nonetheless, the model accounts for judge-specific consistency in $\sigma_{j1}$, guaranteeing these estimates do not reflect time-invariant selection of more- or less-consistent judges into refugee work. However, to the extent that judges with higher refugee caseloads have lower non-refugee caseloads, these estimates are likely biased towards zero.

Column 3 shows that a 10% higher workload reduces consistency by about 2%. In Column 4, I show the workload effect is unchanged by the addition of experience controls. In Column 5 I interact workload with indicators for more or less than 5 years experience, and find that the effect comes entirely from judges with less than 5 years of experience (the $p$-value of a test of equality is 0.11). This finding—that more experienced judges are better able to maintain decision quality as workload increases—is another reason why it might be optimal to shift more cases to experienced judges.

4.10 Judge inconsistency and expert opinion

The judge inconsistency parameters are related to readily-observable reduced form moments in the data, as well as with experience and workload in sensible ways. In this section, I explore whether they are also related to lawyer perceptions of judge ability. Higher degrees of correlation between model-based measures and expert opinion serve as a validation of the model, and suggest that surveys contain useful information about judge quality.

To measure expert opinion, I conducted an email survey of refugee lawyers who have appeared in refugee hearings at the Federal Court. I asked respondents to rate the judges with whom they had personal experience along dimensions analogous to the parameters of the model: how lenient is the judge to claimants (corresponding to judge threshold $\gamma_{js}$), and how consistent and predictable is the judge (I reverse this scale so it corresponds to judge inconsistency $\sigma_{js}$). More details about the survey are in Appendix Section A7.

Each response is on a five-point likert scale, which I normalize by the mean and standard deviation. Table 5 describes the relationship between model coefficients and the survey results. I model the relationship as
\[
\bar{C}_{j\ell} = \beta_0 + \beta_1 \text{Favorability}_{j\ell} + \beta_2 \text{Inconsistency}_{j\ell} + \eta_{\ell} + u_{j\ell},
\]  

where \(\ell\) indexes lawyers and \(\bar{C}_j = \{\hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}_1\}\). I use model estimates that account for experience (which is highly predictive of behavior), and for each judge-respondent pair use the coefficient combinations reflecting experience at the time of their modal interaction. To account for estimation error in the model coefficients I use Hanushek’s (1974) efficient estimator. In Panel A, the dependent variable is the first-round \(\hat{\gamma}_1\). As expected, higher lawyer-reported favorability is associated with a lower threshold. The correlation is large but imprecise in the first round; in the right-most preferred specification one SD higher favorability corresponds to a 0.19 lower \(\gamma_1\), which is significant at the 10% level and about 0.21 SD of the cross-judge distribution of \(\gamma_1\). Panel B displays the relationship between \(\gamma_2\) and the survey measures. The relationship is stronger than with \(\gamma_1\); adding one SD of predicted favorability decreases \(\gamma_2\) by 0.42 SDs of the judge distribution. This may be because second-round judge behavior is more salient than first-round behavior for lawyers, since they appear in front of the judge only in the second round.

Finally, Panel C shows the relationship between surveyed judge characteristics and \(\hat{\sigma}_1\). Survey inconsistency is positively related with the model estimate of inconsistency \(\hat{\sigma}_1\); across specifications one extra SD of surveyed inconsistency translates to between 0.16 to 0.22 higher \(\hat{\sigma}_1\), or about 0.1 SD of the cross-judge distribution. In other words, the model and the judge survey describe the same judges as being consistent, suggesting that the survey is picking up true variation in judge ability to assess case strength.

4.11 Judge selection reform and judge consistency

In 1988, the government enacted an important reform to how it selects judges. The goal of the reform was to make it more difficult for the party in power to appoint unqualified party supporters by requiring that all candidates be approved by an independent Judicial Advisory Council (JAC). As I detail in Section 2.3, the limited evidence available suggests that the policy change reduced the number of new judges with ties to the ruling party. In this section, I show that the reform was also successful in reducing judge inconsistency \(\sigma_{j1}\).

Table 6 presents a regression of \(\hat{\sigma}_{j1}\) on a dummy for whether the judge was appointed before

---

38 Hanushek’s two-step method exploits knowledge of the standard error of the dependent variable \(C_j\) (i.e., the model coefficients) to construct observation-level estimates of the variance of the residual \(u_j\). The second step reweights observations by the inverse standard deviation of the residual.

39 For comparability to Panels A and B, Panel C includes the same set of controls. In Appendix Table A12 I show that this result is robust to the inclusion of controls for first-round approval rates, which is closer to the nonparametric identification intuition of Section 3.4.

40 Because second-round inconsistency \(\sigma_{j2}\) should not be interpreted as reflecting only judge errors, I do not include it in the table. However, consistent with it being partially correlated with the level of observational errors, I find that \(\sigma_{j2}\) is positively but insignificantly correlated with surveyed inconsistency.
the reform. I weight the regressions to account for estimation error in the dependent variable (Hanushek, 1974), and in my preferred specifications control for judge gender and party of appointment. Because the reform took place seven years before the start of my sample, the pre-reform judges are mechanically more experienced. As I show in Section 4.9, more experienced judges are more consistent (have lower $\sigma_{j1}$), so this likely works against finding that the reform improved judge consistency.\footnote{Alternatively, if high-consistency judges are more likely to be promoted to a higher court, then pre-reform consistent judges might not be observed in my data. This would mechanically make the pre-reform judges look less consistent. Although about 20\% of the justices are promoted in seven or fewer years, there is not a strong or statistically significant correlation between promotion and estimated inconsistency—judges who are eventually promoted have an estimated $\sigma_{j1}$ 0.15 lower (standard error 0.207) than non-promoted judges.}

I show results for a baseline model that does not control for experience, and controlling for experience with categorical variables for more than 1, 5 and 10 years of experience (for approximate comparability, I adjust all coefficients to the median experience of 6 years). The first three columns show that average consistency improved by 0.34 after the reform, a 26\% reduction ($p$-value is 0.12). This does not appear to be related to the change in party that occurred just after the reform—Column 3 shows that party of appointment does not have a large or statistically significant effect on consistency. In Columns 4-6, the effect is larger once the model properly accounts for differing experience among pre- and post-reform judges, with $\sigma_{j1}$ dropping by a statistically significant 1.75 points (relative to a pre-reform mean of 2.2) for judges appointed after the reform.\footnote{In Table A15, I estimate an identical table using second-round inconsistency $\sigma_2$ as the dependent variable. Because $\sigma_2$ may partially reflect skill at gathering information in hearings, the predicted effect is ambiguous. In all but one specification, I find statistically insignificant results. This is consistent with information acquisition being an important part of second-round judge errors.}

Both of these affects are large. The estimates from the preferred right-most model imply a pre-reform disagreement rate of 18.3\% for pairs of judges with the same approval rate, compared to the baseline estimate of 13.2\%. If all judges were appointed post-reform, that drops to 6.6\%.\footnote{An alternative explanation for this finding would be that pre- and post-reform judges value different things, but because there are more post-reform judges the model describes them as more consistent. I can test this using the characteristics test in Table A10, I regress second-round approval on model-estimated approval rates plus interactions of whether the first- and second-round judges were appointed before and after the reform. I do not reject the hypothesis that these two groups are operating under the same standards ($p$-value= 0.85).} The strength of the effect speaks to the effectiveness of the reform, which materially restricted the minister’s options. Government data shows that the JACs approve only 40\% of applicants; ostensibly some of the rejected candidates would otherwise have been appointed. Interestingly, the effect is not driven by changes in judge leniency. In Appendix Table A13, I show that judges appointed after the reform approved a statistically indistinguishable share of first-round claimants. More directly, in Appendix Table A14 I control for judge approval rates and find almost identical results, with an average $\sigma_{j1}$ 1.72 smaller for judges appointed after the reform (SE=0.65).

The table also shows that there is no significant or substantial difference in consistency between judges appointed by the two parties, which is in line with the relative similarity in judicial
philosophies between liberal and conservative judges in Canada. The Federal Court is a prestigious appointment, and so governments are unlikely to be constrained by supply limitations. Male judges are slightly (and marginally statistically significantly) more consistent than female judges, though this difference is dwarfed by both the gains to experience and the post-reform effect.

4.12 Optimal judge allocation

The Court assigns cases to judges taking into account only their availability, not their behavior in previous cases. In this section, I show how the Court could optimize judge allocation to minimize effort while maintaining the same standards.44

Second-round decisions are much more costly to the court than first-round decisions. Instead of reading documents from the IRB’s initial decision, a second-round decision entails a full hearing in front of the opposing lawyers, time to prepare for the hearing and time to write the decision—about ten times as long as a first-round decision. The Court could minimize workload while approving the same number of total claimants by reducing the number of first-round acceptances and approving all second-round claimants. To some extent, they are already pursuing this strategy—as I discuss in Section 4.4, the approval rate jumps from 14% to 45% between the first and second round.

However, it is unclear from the reduced form evidence what further reducing first-round approvals (and thus costly second-round decisions) would do to the distribution of case quality for approved claimants. A natural requirement is that any acceptable counterfactual judge assignment mechanism approves at least the same number of claimants, and that the distribution of posterior case strength \( r_i \) of the approved claimants first-order stochastically dominates the baseline distribution. I also require that no judge works more than she currently does.

Under this problem, there are three ways to minimize caseload. First, judges should be reallocated to rounds where they make more consistent decisions. Second, first-round judges should be made more strict to improve the posterior case quality of claimants approved in the first round and decrease the number of second-round decisions. Third, second-round judges should approve a higher share of cases so that the overall approval rate remains the same given lower first-round approval rates.

In Figure 7, I conduct exactly this maximization. I find that overall workload would be reduced by 17.5% (or 28,000 hours), amounting to savings of approximately $4.4 million in judge salaries alone over the study period. This counterfactual policy would also save staff time and allow claimants to receive their decision faster. The figure demonstrates the second two kinds of savings, but not the first. To summarize how the re-assignment procedure works, I present histograms of the baseline

44 Alternatively, I could maximize acceptance rates while maintaining the same standards and keeping workload no higher than in baseline. This problem is almost symmetric, and for given model estimates the potential cost savings holding acceptance fixed are always close to the percentage increase in refugees holding costs fixed.
judge coefficients by round as well as histograms that have been reweighted to reflect the assignment of judges to rounds after optimization. The average first-round threshold $\gamma_{j1}$ for optimally-assigned judges is higher (Panel A), meaning that fewer cases will be approved in the first round but case strength conditional on first-round approval will be higher. Conversely, judges in the second round are much more lenient, as evinced by the lower thresholds $\gamma_{j2}$ (Panel C). However, in Panel B and D the change in the overall distribution of consistency is much less dramatic. Average inconsistency is almost unchanged, from 1.11 to 1.13 in the first round, though the most inconsistent judge-rounds are eliminated. I interpret this to mean that judge thresholds $\gamma_{js}$ are a stronger driver of who is selected, but that knowledge of judge-specific consistency has an important role to play in allowing the researcher to discipline the allocation process.

The problem as I’ve described it takes the posterior distribution of $r_i$ as the relevant measure of quality. Implicitly, this assumes that the second-round error is all judge error $e_{ij2}$ rather than information $I_{ij2}$. As I discuss in Section 3.5.3, it may also reflect additional information gained in the second-round hearing. In that case, the relevant measure of quality is $r_i + I_{ij2}$. I do not directly estimate the distribution of of the information shock $I_{ij2}$, so cannot perfectly condition on the posterior (I estimate the distribution of inconsistency plus information shock, $e_{ij2} + I_{ij2}$). However, under the assumption that the second-round error is all information, I can minimize workload so that the posterior of $r_i + I_{ij2}$ first-order stochastically dominates the baseline distribution. Under this specification, the allocation of judges is highly correlated with the baseline optimization (0.58), and workload is reduced by 16% rather than 17.5%. It also satisfies the constraints of the baseline allocation, so a cautious planner could implement the second design and enjoy most of the gains of judge reallocation.

4.13 Implications for monotonicity in examiner-assignment designs

The presence of inconsistency has important implications for examiner-assignment research designs. This identification strategy uses random or quasi-random assignment of decision-makers to cases to generate random variation in a treatment, then studies the effect of treatment on outcomes. Prominent examples of the treatments studied include incarceration, patent receipt, and being placed in foster care (Bluhler et al., 2016; Gaulé, 2015; Doyle, 2008).

As I discuss in Section 3.1, the monotonicity assumption in this context requires that all individuals are weakly more likely to be approved by a high-approval judge, and less likely to be approved by a low-approval judge (individuals for whom this does not hold are defiers in the Angrist, Imbens, and Rubin (1996) sense). The presence of inconsistency implies violations of monotonicity, and in the section I provide tools to test for monotonicity in other contexts. I implement these tools in Appendix Section A8 and uncover large violations of monotonicity that are not detected by current approaches.
The bias in LATE estimates generated by monotonicity violations depends on both the size of the violation, and the difference in the treatment effects between compliers and defiers. To answer whether these sort of monotonicity violations are likely to be harmful, in Appendix Section A9 I simulate the implications of inconsistency for IV and MTE estimates of the effect of first-round approval on second-round approval. Although the high level of inconsistency means that there are a high number of monotonicity violations, I find that the bias in IV estimates is only 5%. My analysis in Section 3.1 suggests that this is due to a small average difference in treatment effects for compliers and defiers, in large part because individuals are often compliers under one treatment and defiers under another. More worryingly, I find that inconsistency causes the estimated MTE to be considerably flatter than the true MTE. I conclude that there are reasons for caution in interpreting MTE estimates in judge designs, particularly when the researcher rejects monotonicity using the tests of Section 3.1.

5 Conclusion

Much research has focused on non-relevant factors that affect judge behavior: among others, the decision in the previous case, the outcome of a college football game, or the timing of the hearing relative to lunch. The existence of these phenomena suggests that a convicted defendant might have been acquitted had his case been heard the following day, or even at a different time. In a similar vein, judges and juries apply different standards to different racial groups, implying that some claimants would be incarcerated under one judge and acquitted under another (Abrams et al., 2012; Anwar et al., 2012).

All of this research suggests that courts do not always satisfy the fundamental judicial principle of equal treatment of equals (Fuller, 1969; Rawls, 1972). In this paper, I study the extent to which judges contribute to this problem, and estimate the share of cases that would be decided differently if they were heard by a different judge or under different circumstances. I differentiate between two sources of unequal treatment: cross-judge variation in leniency, and inconsistency. The former part is well-studied, and easily observable when judges are randomly assigned to cases. The second part, inconsistency, is any situation when judges with the same incarceration rate would make a different decision, or a judge would make a different decision at a different time.

Understanding the relative size of differential leniency and inconsistency is key to interpreting examiner-assignment LATE estimates. Power in these designs depends on high levels of differential leniency—so that random judge assignment affects the share of treated individuals—but requires monotonicity in treatment assignment across judges—in other words, each judge must incarcerate a subset of the defendants his more-severe colleagues would incarcerate. I show that monotonicity violations are implied by inconsistency, and so understanding the sources of unequal treatment is
informative about the reliability of examiner-assignment estimates.

I begin by introducing a simple test of one implication of monotonicity: that each consecutively-more-severe judge must be more likely to incarcerate defendants in each observable demographic group. This approach tests a stronger implication of monotonicity than current approaches, and can be adapted to generate a bound on the proportion of defiers for each pair of judges, as well as a bound on the multi-instrument analog to Angrist, Imbens, and Rubin (1996)’s $\lambda$. This allows the researcher to assess the potential effect of monotonicity violations.

I then show how to use a similar bounding approach to bound the level of pairwise inconsistency, or the share of cases that judges with the same incarceration rate would disagree on. To study the judge-specific (as opposed to court-level) determinants of decision quality, I rely on the fact that in many judicial systems with appeals, the explicit goal of judges is to write decisions that their colleagues will agree with. A good measure of judge quality is therefore the share of decisions that are not subsequently overturned, and I show how this quality measure aggregates up the bounds on judge-pair disagreement. In my model, judges approve all candidates with a case strength larger than a judge-specific threshold. Judges observe case strength with some error (more inconsistent judges have larger errors), which generates differences across judges in which claimants they approve, even for judges who approve the same share of cases. I show that this model is nonparametrically identified in two-stage judicial processes by a combination of cross-judge comparisons and regressors that shift judge thresholds without affecting errors.

I implement the model using data on judicial review of initially-denied refugee claims at the Federal Court of Canada. Although the justices of the Federal Court are experts in refugee cases, I uncover relatively high levels of inconsistency. I compare first-round judges who approve the same share of cases (on average, 14%), meaning that they would never disagree if they were perfectly consistent. I nonparametrically bound disagreement to at least 10% of all cases and under the extra assumptions of the structural model find that they disagree on 13.2% of cases, about halfway between perfect consistency and the 23.2% disagreement rate implied by judges who arrived at decisions by flipping coins. Even more strikingly, inconsistency contributes more to unequal treatment than cross-judge variation in leniency. If all judges were perfectly consistent they would disagree on 8% of cases, but with the addition of inconsistency the disagreement rate jumps to 23%. Overidentification tests suggest that most disagreement arises from idiosyncratic observational errors, rather than permanent differences between judges in racial or gender bias, ideology, or statutory interpretation.

Cross-judge variation in inconsistency is large. Replacing the least consistent half of judges with median-consistency judges would reduce the disagreement rate between similarly-severe judges from 13.2% to 8%. I validate the measured variation in consistency against a survey that solicited estimates of judicial characteristics from refugee lawyers who had appeared in front of the judges. Judicial consistency improves dramatically after the first year, and continues to improve (albeit at
a slower rate) for at least the first ten years of experience. This result is true even after controlling for total number of refugee cases by the judge, suggesting that shifting cases from inexperienced to experienced judges would improve overall, long-term consistency. Inexperienced judges—but not experienced ones—are more consistent when they have a smaller workload. A reform in the late 1980s designed to stop the government from appointing unqualified party supporters dramatically improved judicial consistency, suggesting that well-designed judge selection processes can indeed improve court quality.

I study how reallocating judges to rounds where they are more consistent could improve the functioning of the court. Using the structural model, I estimate the posterior distribution of case quality for approved claimants. I then find the counterfactual allocation of judges to cases that most reduces workload while approving claimants whose posterior quality first-order stochastically dominates the baseline distribution. The optimal policy would lead to a quicker resolution of cases and reduce judge hours by 18%, saving at least $4.4 million over the study period.

Given the tight relationship between inconsistency and monotonicity, I unsurprisingly uncover evidence of large violations of monotonicity. My test decisively rejects the assumption, even though other popular tests do not. Although the bias in IV estimates engendered by monotonicity violations depends crucially on the context, I find that in my setting the bias is relatively small, at only about 5%. However, the biases in MTE estimates are much larger, suggesting that researchers should be cautious about examiner-assignment identification of the MTE. Researchers should be particularly concerned when using data from less professional courts, where judges are less likely to behave consistently.

My research has strong implications for the assessment of the Federal Court, and suggests that many cases hinge on the identity of the assigned judge, rather than the merits of the case. Under current policy, 14% of all claimants proceed to the second stage and 6% of the total are eventually successful in having the Court return their case to the government for redetermination. I find that first-round judges reject many claimants who might be successful in the second stage—if first-stage approval became automatic, 19.4% of all claimants would be granted redetermination. Over the 17 years from 1995 that comprise my study period, that difference amounts to approximately 7,700 families.
References


PHILIPPE, A. AND A. OUSS (2016): “‘No Hatred or Malice, Fear or Affection’: Media and Sentencing,” .


6 Figures

Figure 1: Approval rates by judge

(a) First round

(b) Second round

(c) Second vs. first round approval rates

Panel A and B contain histograms of approval rates by judge for the first and second round, respectively. Both are weighted by the number of observations per judge. Panel C contains the scatter plot of judge-level first- and second-round approval rates. The correlation is 0.57, and 0.40 without the outlier.
Figure 2: Share of approved claimants subsequently approved versus first-round approval rates

The Figure shows second round approval rates for the claimants approved by each first round judge plotted against the judge’s first-round approval rates, with second-round judge approval rates residualized out and means shrunk towards the grand mean via Empirical Bayes to account for measurement error.

Regression coefficient is $-0.28 (0.042)$. 
This figure displays half-median-unbiased estimates of disagreement rates for all first-round judges with an approval rate within 1% of each other calculated using the method of Chernozhukov, Lee and Rosen (2013), adjusted downwards to account for differences in approval rates. Lower endpoints of 95% one-sided confidence intervals displayed as whiskers.
Figure 4: Distribution of judge parameters

(a) Threshold $\gamma_1$, first round

(b) Observational error $\sigma_1$, first round

(c) Threshold $\gamma_2$, second round

(d) Observational error $\sigma_2$, second round

This figure presents coefficient estimates for the decision model $1[r_i > \gamma_{js} + X_{ijs}\beta_s + \tilde{\epsilon}_{ijs}]$, $\tilde{\epsilon}_{ijs} \sim \mathcal{N}(0, \sigma^2_{js})$. All models include controls for time/date of decision in $X_{ijs}$, and allow the parameters of the Pareto distribution of $r_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Each panel contains the density of the raw and shrunken estimates of the judge-round specific thresholds $\gamma_1$ and $\gamma_2$, and inconsistency $\sigma_1$ and $\sigma_2$. Black line in Panels A and C is density of case quality $r_i$. Shrunken estimates recovered via deconvolution of estimates accounting for coefficient-specific standard errors, clustered at the level of the first round judge (Delaigle, Hall, and Meister, 2008).
Figure 5: Model estimates of first- versus second-round approval

Figure plots first-round approval probability against second-round approval probability conditional on first-round approval for each value of case strength \( r_i \). Secondary graph displays cumulative density of first-round approval. Black dotted comparison line marks out 45°.
This figure relates to the estimated decision model \( \mathbb{1}[r_i > \gamma_{ij} + X_{ij}\beta_\delta + \tilde{\epsilon}_{ij}] \), \( \tilde{\epsilon}_{ij} \sim N(0, \sigma_{ij}^2) \). Panel A contains model estimates of the first-round approval probability as a function of the deviation from estimated judge thresholds \( \gamma_{ij} \). Panel B contains model estimates of second-round approval as a function of mean first-round error relative to the estimated value—higher errors make second-round approval less likely. In Panel C, I match pairs of judges with similar first-round approval rates (within 1 percentage point), then display difference in model-estimated judge errors \( \hat{\sigma}_1 \). See Section 3.4.1 for more details.
Figure 7: Distribution of judge parameters under optimal allocation

(a) Threshold $\gamma_1$, first round

(b) Observational error $\sigma_1$, first round

(c) Threshold $\gamma_2$, second round

(d) Observational error $\sigma_2$, second round

I minimize judge workload requiring that a) no judge works more than she does in the baseline, b) at least as many claimants are approved, and c) the posterior distribution of case strength $r_i$ for approved claimants under the counterfactual assignment first-order stochastically dominates the baseline distribution. Each panel contains a histogram of the baseline distribution of coefficients, as well as the distribution after maximization. The overall reduction in workload is 17.5%.
7 Tables

Table 1: Randomization tests of claimant characteristics on judge leniency

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Africa</th>
<th>Asia</th>
<th>South America</th>
<th>IRB mean approval</th>
<th>Predicted approval</th>
<th>1st-round mean approval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>First-round approval rate</td>
<td>0.011</td>
<td>-0.031</td>
<td>-0.089</td>
<td>-0.006</td>
<td>0.010</td>
<td>-0.002</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.030)</td>
<td>(0.071)</td>
<td>(0.023)</td>
<td>(0.014)</td>
<td>(0.002)</td>
<td></td>
</tr>
<tr>
<td>F-stat</td>
<td>0.88</td>
<td>2.82</td>
<td>9.54</td>
<td>2.47</td>
<td>4.23</td>
<td>2.99</td>
<td></td>
</tr>
<tr>
<td>Prob</td>
<td>0.73</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>50,435</td>
<td>50,435</td>
<td>50,435</td>
<td>50,435</td>
<td>50,435</td>
<td>50,435</td>
<td></td>
</tr>
</tbody>
</table>

Panel A: First round judges

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Africa</th>
<th>Asia</th>
<th>South America</th>
<th>IRB mean approval</th>
<th>Predicted approval</th>
<th>1st-round mean approval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>Second-round approval rate</td>
<td>-0.048</td>
<td>-0.012</td>
<td>0.000</td>
<td>0.032</td>
<td>-0.005</td>
<td>0.002</td>
<td>-0.024</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.037)</td>
<td>(0.042)</td>
<td>(0.042)</td>
<td>(0.019)</td>
<td>(0.002)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>F-stat</td>
<td>1.01</td>
<td>1.53</td>
<td>1.71</td>
<td>1.22</td>
<td>1.54</td>
<td>2.11</td>
<td>3.07</td>
</tr>
<tr>
<td>Prob</td>
<td>0.45</td>
<td>0.01</td>
<td>0.00</td>
<td>0.12</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Observations</td>
<td>7,143</td>
<td>7,143</td>
<td>7,143</td>
<td>7,143</td>
<td>7,143</td>
<td>7,143</td>
<td>7,143</td>
</tr>
</tbody>
</table>

Panel B: Second round judges

All regressions include office X pre-2002 fixed effects to account for cross-office differences in case strength and changes in government policy in 2002. IRB mean approval is the approval rate of the IRB Member who initially denied refugee status to the claimant. Predicted approval comes from a regression of approval on gender, continent of origin and IRB Member approval rate. Judge approval rates on right side partial out office X pre/post-2002. F-stats come from separate regression of outcome on judge fixed effects. Standard errors clustered at the judge level in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.
Table 2: Second-round approval on mean approval rate of first-round judge

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean first round approval, exclusive</td>
<td>-0.264***</td>
<td>-0.312***</td>
<td>-0.324***</td>
</tr>
<tr>
<td></td>
<td>(0.0521)</td>
<td>(0.0423)</td>
<td>(0.0437)</td>
</tr>
<tr>
<td>Mean second round approval, exclusive</td>
<td>0.958***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0239)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second-round judge FE</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>8,446</td>
<td>8,446</td>
<td>8,446</td>
</tr>
</tbody>
</table>

Standard errors clustered by second-round judge. * p < 0.10, ** p < 0.05, *** p < 0.01.
Table 3: Second-round approval on model approval probability and judge-pair FEs

<table>
<thead>
<tr>
<th></th>
<th>Judge-pair round FEs</th>
<th>Judge-pair FEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Model approval probability</td>
<td>0.945***</td>
<td>0.938***</td>
</tr>
<tr>
<td></td>
<td>(0.146)</td>
<td>(0.169)</td>
</tr>
<tr>
<td>Mean approval</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>F-stat for judge pairs</td>
<td>1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>P-value</td>
<td>0.350</td>
<td>0.344</td>
</tr>
<tr>
<td>Bootstrap p-value</td>
<td>0.694</td>
<td>0.693</td>
</tr>
<tr>
<td>SD of judge-pair EB means</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Observations</td>
<td>8,196</td>
<td>8,196</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Model approval probability</td>
<td>0.967***</td>
<td>0.962***</td>
</tr>
<tr>
<td></td>
<td>(0.0463)</td>
<td>(0.0470)</td>
</tr>
<tr>
<td>Mean approval</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>F-stat for judge pairs</td>
<td>1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>P-value</td>
<td>0.350</td>
<td>0.344</td>
</tr>
<tr>
<td>Bootstrap p-value</td>
<td>0.694</td>
<td>0.693</td>
</tr>
<tr>
<td>SD of judge-pair EB means</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Observations</td>
<td>8,196</td>
<td>8,196</td>
</tr>
</tbody>
</table>

Regresses second-round approval on model-predicted likelihood of approval and judge-pair fixed effects. Left two columns construct judge-pair FEs accounting for order of assignment; right two columns ignore this distinction. Model controls are analogous to structural model and include office of origination X pre-post 2002, and an end-of-week and noon hearing dummy for the second-round hearing. Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.
Table 4: First-round judge consistency by experience and workload

<table>
<thead>
<tr>
<th>Experience &gt; 1 year</th>
<th>Experience &gt; 5 years</th>
<th>Experience &gt; 10 years</th>
<th>Log monthly caseload</th>
<th>Log caseload (≤ 5 yrs exp)</th>
<th>Log caseload (&gt; 5 yrs exp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients $\psi_1$ affecting judge inconsistency $\sigma_1$</td>
<td>Coefficients $\psi_1$ affecting judge inconsistency $\sigma_1$</td>
<td>Coefficients $\psi_1$ affecting judge inconsistency $\sigma_1$</td>
<td>Coefficients $\psi_1$ affecting judge inconsistency $\sigma_1$</td>
<td>Coefficients $\psi_1$ affecting judge inconsistency $\sigma_1$</td>
<td>Coefficients $\psi_1$ affecting judge inconsistency $\sigma_1$</td>
</tr>
<tr>
<td>Experience &gt; 1 year</td>
<td>-0.774***</td>
<td>-0.688***</td>
<td>-0.337***</td>
<td>-0.460***</td>
<td>(0.096)</td>
</tr>
<tr>
<td>Experience &gt; 5 years</td>
<td>-0.290***</td>
<td>-0.237***</td>
<td>0.047</td>
<td>-0.053</td>
<td>(0.046)</td>
</tr>
<tr>
<td>Experience &gt; 10 years</td>
<td>-0.536***</td>
<td>-0.443**</td>
<td>-0.476***</td>
<td>-0.509***</td>
<td>(0.175)</td>
</tr>
<tr>
<td>Log monthly caseload</td>
<td>0.197***</td>
<td>0.239***</td>
<td>(0.009)</td>
<td>(0.023)</td>
<td> </td>
</tr>
</tbody>
</table>

Second-round experience control Yes Yes No Yes Yes
Career number of cases No Yes No No No

Reports coefficients for decision model $\mathbb{I}[r_i > \gamma_{js} + X_{ijs}\beta_s + \bar{\epsilon}_{ijs}(W_{ijs})]$, $\bar{\epsilon}_{ijs}(W_{ijs}) \sim N(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\bar{\sigma}_{js} + W_{ijs}\psi_s)$. All models include controls for time/day of decision in $\beta$, and allow the parameters of the Pareto distribution of $r_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Table 5: Model coefficients on survey responses

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Threshold</strong> $\gamma_1$ <em>(mean=2.26, SD=.87)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>-0.144***</td>
<td>-0.237***</td>
<td>-0.055</td>
<td>-0.186*</td>
<td>(0.104)</td>
</tr>
<tr>
<td>Inconsistency, SD</td>
<td>0.316***</td>
<td>0.212***</td>
<td>0.304***</td>
<td>0.134*</td>
<td>(0.046)</td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
</tr>
</tbody>
</table>

| **Panel B: Threshold** $\gamma_2$ *(mean=2.19, SD=.99)* | | | | | |
| Favorability, SD | -0.280** | -0.402*** | -0.264*** | -0.420*** | (0.107) | (0.080) | (0.097) | (0.090) |
| Inconsistency, SD | 0.121 | 0.104 | 0.058 | -0.053 | (0.078) | (0.093) | (0.069) | (0.079) |
| Respondent FE | No | Yes | No | Yes | No | Yes |
| Observations | 182 | 182 | 182 | 182 | 182 | 182 |

| **Panel C: Inconsistency** $\sigma_1$ *(mean=1.89, SD=2.17)* | | | | | |
| Favorability, SD | 0.082 | 0.020 | 0.152*** | 0.092 | (0.060) | (0.099) | (0.044) | (0.094) |
| Inconsistency, SD | 0.184*** | 0.163*** | 0.224*** | 0.194*** | (0.047) | (0.056) | (0.049) | (0.061) |
| Respondent FE | No | Yes | No | Yes | No | Yes |
| Observations | 182 | 182 | 182 | 182 | 182 | 182 |

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) weights for estimated dependent variables. Decision model is $I[r_i > \gamma_{js} + X_{ij} \beta_s + \bar{\epsilon}_{ij}W_{ij,s}], \bar{\epsilon}_{ij}W_{ij,s} \sim N(0, \sigma_{js}(W_{ij,s})^2), \sigma_{js}(W_{ij,s}) = \exp(\tilde{\sigma}_{js} + W_{ij,s}\psi_s)$. All models include controls for time/day of decision in $\beta_s$, and allow the parameters of the Pareto distribution of $r_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. I adjust for judge experience in $\beta_s$ and $\psi_s$. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 

62
Table 6: Inconsistency before and after judge selection reform

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Experience control in $\sigma_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>After reform (=1)</td>
<td>-0.215</td>
<td>-0.332</td>
</tr>
<tr>
<td></td>
<td>(0.151)</td>
<td>(0.207)</td>
</tr>
<tr>
<td>Liberal appointee (=1)</td>
<td>-0.0133</td>
<td>-0.108</td>
</tr>
<tr>
<td></td>
<td>(0.106)</td>
<td>(0.109)</td>
</tr>
<tr>
<td>Male judge (=1)</td>
<td>-0.0873</td>
<td>-0.210*</td>
</tr>
<tr>
<td></td>
<td>(0.101)</td>
<td>(0.122)</td>
</tr>
<tr>
<td>Year appointed</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pre-reform mean</td>
<td>1.29</td>
<td>1.29</td>
</tr>
<tr>
<td>N judges</td>
<td>53</td>
<td>53</td>
</tr>
</tbody>
</table>

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency $\sigma_1$, which is estimated from decision model $1[r_i > X_{ijs} \beta_s + \gamma_{js} + \tilde{\epsilon}_{ijs}(W_{ijs})]$, $\tilde{\epsilon}_{ijs}(W_{ijs}) \sim N(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. All models include controls for time/day of decision in $\beta$, and allow the parameters of the Pareto distribution of $r_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. In the right-hand panel, $\beta_s$ and $\psi_s$ include dummies for more than 1, 5, and 10 years of experience. Standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Appendix for *Judicial Errors: Evidence from Refugee Appeals*

A1  Proof of Theorem 1

In this section, I prove that monotonicity violations that are detectable by tests of average monotonicity are a strict subset of the violations detectable by pairwise tests. I begin by defining each type of test, then present the proof.

Tests of pairwise monotonicity ask whether the binary-instrument first stage is positive within each demographic subgroup. For endogenous variable $x_{ijg}$, dummy variable $z_j$ indicating assignment to judge $j$, and dummy variable $D_g$ indicating membership in group $g$, the first stage regression is

$$x_{ijg} = \sum_{g=1}^{G} \alpha_j^g z_j \times D_g + \mu_g + \varepsilon_{ijg} \tag{A1}$$

estimated for each sample of individuals $i$ assigned to judge $j$ or $j-1$, where the judge ordering is with respect to overall severity. The null hypothesis is that $\alpha \geq 0$, where the $\geq$ is applied elementwise and $\alpha$ consists of all judge-demographic combinations of $\alpha_{j,g}$. This null hypothesis can be tested with Wolak (1989), or indirectly with Equation 7.

On the other hand, tests of average monotonicity test the sign of the linear first stage estimated on a particular demographic $g$. Define the mean incarceration rate of judge $j$ as $\bar{x}_j$, where the average can be taken over all cases (as in Dobbie et al. (2018)), or only over another demographic group (Bhuller et al., 2016). Then, the first stage is

$$x_{ijg} = \mu_0 + \bar{\alpha}_g \bar{x}_j + \varepsilon_{ijg} \tag{A2}$$

This test can be repeated for all demographics, and then the null hypothesis under monotonicity is that $\bar{\alpha} \geq 0$, for $\bar{\alpha} = [\bar{\alpha}_1, \ldots, \bar{\alpha}_G]$.

I prove Theorem 1 by establishing that $\bar{\alpha}_g$ is a weighted average of $\alpha_j^g$, where all the weights are positive. This means that whenever average monotonicity is violated—that is, when $\bar{\alpha}_g < 0$—at least one $\alpha_j^g$ must be negative, and so pairwise monotonicity is also violated. However, some $\alpha_j^g$ can be negative and $\bar{\alpha}_g$ can still be positive, meaning that the converse is not true.

Define $x_{jg} = E[x_{ijg}|j,g]$, and note that $x_{jg} = x_{0g} + \sum_{k=1}^{j} \alpha_j^g$. Conditioning on $g$, $\bar{\alpha}_g =$
\[ E[x_{ijg} \bar{x}_j | g] / E[\bar{x}_j^2 | g] \] where \( \bar{x}_j \) is demeaned. The numerator of this expression is equal to:

\[
E[x_{ijg} \bar{x}_j | g] = \sum_{j=0}^{J} \pi_j \bar{x}_j E[x_{ijg} | j, g]
\]
\[
= \sum_{j=0}^{J} \pi_j \bar{x}_j \left[ x_{0g} + \sum_{k=1}^{j} \alpha_k^q \right]
\]
\[
= \sum_{j=1}^{J} \pi_j \bar{x}_j \left( \sum_{k=1}^{j} \alpha_k^q \right)
\]
\[
= \sum_{k=1}^{J} \alpha_k^q \sum_{j=k}^{J} \pi_j \bar{x}_j
\]

where the third equality follows from the demeaning of \( \bar{x}_j \). Since \( E[\bar{x}_j^2 | g] = \sum_{j=1}^{J} \pi_j \bar{x}_j^2 \), we can rewrite the coefficient from the linear first stage:

\[
\bar{\alpha}_g = \sum_{j=1}^{J} w_j \alpha_j^q, \quad w_j = \frac{\sum_{k=1}^{J} \pi_k \bar{x}_k \bar{x}_j}{\sum_{k=1}^{J} \pi_k \bar{x}_k^2}
\] (A3)

where the weights are all positive since \( \bar{x}_j \) is demeaned and \( \bar{x}_0 < \bar{x}_k \) for all \( k \).
A2 Proof of Theorem 2

Bounds on disagreement can be tightened from the ordinary Fréchet bounds using additional information on decisions by subsequent judges. Beginning with the definition of disagreement and expanding using the law of total probability,

$$\delta_{AB}(C) = P[y_1(A) \neq y_1(B)|y_2(C) = 1]P[y_2(C) = 1] + P[y_1(A) \neq y_1(B)|y_2(C) = 0]P[y_2(C) = 0] \quad (A4)$$

Taking the probability of disagreement from the first term on the left side of the expression, $P[y_1(A) \neq y_1(B)|y_2(C) = 1]$, Fréchet inequalities imply that

$$P[y_1(A) \neq y_1(B)|y_2(C) = 1] = P[y_1(A) = 1, y_1(B) = 0|y_2(C) = 1] + P[y_1(A) = 0, y_1(B) = 1|y_2(C) = 1]$$

$$\geq \max\{0, P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1]\} + \max\{0, P[y_1|D_1=0, y_2(C) = 1] - P[y_1|D_1=1, y_2(C) = 1]\}$$

$$= \max\{P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1], P[y_1|D_1=0, y_2(C) = 1] - P[y_1|D_1=1, y_2(C) = 1]\}$$

where the inequality follows from the Fréchet bound and random assignment of judges to cases. The analogous argument can be made for $P[y_1(A) \neq y_1(B)|y_2(C) = 0]$. Plugging into Equation A4 and expanding the max operators,

$$\delta_{AB}(C) \geq \max\left\{\left[P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1]\right]P[y_2(C) = 1] + \left[P[y_1|D_1=0, y_2(C) = 0] - P[y_1|D_1=0, y_2(C) = 0]\right]P[y_2(C) = 0],
\left[P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1]\right]P[y_2(C) = 1] + \left[P[y_1|D_1=0, y_2(C) = 0] - P[y_1|D_1=1, y_2(C) = 0]\right]P[y_2(C) = 0],
\left[P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1]\right]P[y_2(C) = 1] + \left[P[y_1|D_1=0, y_2(C) = 0] - P[y_1|D_1=1, y_2(C) = 0]\right]P[y_2(C) = 0],
\left[P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1]\right]P[y_2(C) = 1] + \left[P[y_1|D_1=0, y_2(C) = 0] - P[y_1|D_1=1, y_2(C) = 0]\right]P[y_2(C) = 0]\right\} \quad (A5)$$

It is clear that the first argument of the max function in Equation A5 is equal to $P[y_1|D_1=y_1] - P[y_1|D_1=0]$. The second argument can be expanded to

$$\left[P[y_1|D_1=y_1, y_2(C) = 1] - P[y_1|D_1=0, y_2(C) = 1]\right]P[y_2(C) = 1] + \left[P[y_1|D_1=0, y_2(C) = 0] - P[y_1|D_1=0, y_2(C) = 0]\right]P[y_2(C) = 0]$$

$$= P[y_1|D_1=y_1] - P[y_1|D_1=0] + 2P[y_2(C) = 1]P[y_1|D_1=y_1, y_2(C) = 1] - 2P[y_2(C) = 0]P[y_1|D_1=0, y_2(C) = 1]$$

$$= P[y_1|D_1=y_1] - P[y_1|D_1=0] + 2P[y_2(C)|D_1=y_1, y_1 = 1]P[y_2(C) = 1] - 2P[y_2(C)|D_1=0, y_1 = 1]P[y_2(C) = 1]$$

$$= P[y_1|D_1=y_1] - P[y_1|D_1=0] + 2P[y_2(C)|D_1=y_1, y_1 = 1]P[y_1|D_1=y_1, D_2=0]P[y_2(C) = 1] - 2P[y_2(C)|D_1=0, y_1 = 1]P[y_1|D_1=0, D_2=0]P[y_2(C) = 1]$$

where the third line follows from Bayes rule and the fourth from the independence of $y_2(C)$ and first-round judge assignment. The third and fourth lines of Equation A5 follow analogously, completing the proof.
A3 Judge-assignment identification of second-round inconsistency

Identification of second-round inconsistency follows a similar principle to first-round inconsistency, but since there is no third round to use as a check, identification requires what I call a *non-limiting* judge. Suppose we are trying to determine which second-round judge, $A_2$ or $B_2$, is more consistent. I assume that there is a known first-round judge $D_1$ that approves nearly anyone, and a known first-round comparison judge $C_1$. Formally, I require that $\tilde{G}_{C1}(\cdot)/\tilde{G}_{D1}(\cdot)$ is monotonically increasing wherever $\tilde{G}_{A2}(\cdot) \neq \tilde{G}_{B2}(\cdot)$. This is trivially satisfied when $G_{D1}(\cdot) = 1$ (judge $D_1$ literally approves everyone), and can be satisfied when judge $D_1$ is fairly consistent and has a very low threshold $\gamma$ relative to judge $C_1$.\footnote{The requirement is also satisfied by assuming that $\tilde{G}_{C1}(z)/\tilde{G}_{D1}(z) > \tilde{G}_{C1}(w)/\tilde{G}_{D1}(w)$ whenever $z > w$, and $\tilde{G}_{C1}(z)/\tilde{G}_{D1}(z) < \tilde{G}_{C1}(w)/\tilde{G}_{D1}(w)$ when $z < w$, which is considerably weaker but is harder to understand because it is defined in reference to the point of single-crossing.}

Define judge $A_2$ and $B_2$ as second-round comparable if they have the same second-round approval rate conditional on first-round approval by judge $D_1$: $\int \tilde{G}_{D1}(r)\tilde{G}_{A2}(r)f_r\,dr = \int \tilde{G}_{D1}(r)\tilde{G}_{B2}(r)f_r\,dr$. Then, if judge $A_2$’s second-round approval rate increases more than judge $B_2$’s for decisions conditional on judge $C_1$’s first-round approval (vs. judge $D_1$’s), judge $A_2$ is more consistent than judge $B_2$. This can be seen by the following derivation,

$$(P[r > \varepsilon_{A2}|r > \varepsilon_{C1}] - P[r > \varepsilon_{B2}|r > \varepsilon_{C1}])P[r > \varepsilon_{C1}]$$

$$= P[r > \varepsilon_{C1} \cap r > \varepsilon_{A2}] - P[r > \varepsilon_{C1} \cap r > \varepsilon_{B2}]$$

$$= \int \tilde{G}_{C1}(r) \left[ \tilde{G}_{A2}(r) - \tilde{G}_{B2}(r) \right] f_r\,dr$$

$$= \int_{-\infty}^{z} \frac{\tilde{G}_{C1}(r)}{\tilde{G}_{D1}(r)} \tilde{G}_{D1}(r) \left[ \tilde{G}_{A2}(r) - \tilde{G}_{B2}(r) \right] f_r\,dr + \int_{z}^{\infty} \frac{\tilde{G}_{C1}(r)}{\tilde{G}_{D1}(r)} \tilde{G}_{D1}(r) \left[ \tilde{G}_{A2}(r) - \tilde{G}_{B2}(r) \right] f_r\,dr$$

$$> \int_{-\infty}^{z} \frac{\tilde{G}_{C1}(z)}{\tilde{G}_{D1}(z)} \tilde{G}_{D1}(r) \left[ \tilde{G}_{A2}(r) - \tilde{G}_{B2}(r) \right] f_r\,dr + \int_{z}^{\infty} \frac{\tilde{G}_{C1}(z)}{\tilde{G}_{D1}(z)} \tilde{G}_{D1}(r) \left[ \tilde{G}_{A2}(r) - \tilde{G}_{B2}(r) \right] f_r\,dr$$

$$= 0$$

where monotonicity of $\tilde{G}_{C1}(\cdot)/\tilde{G}_{D1}(\cdot)$ takes the place of monotonicity of second-round approval in the identification of first-round consistency.
A4 Proof of Theorem 3

With a small change of notation, the main model of Section 3 can be recast as a single-spell duration model (Chen, Heckman, and Vytlacil, 1999), where the duration is the number of rounds until a judge rejects the applicant’s case (duration is capped at 2). Equation 16 from the main text sets out the problem as identifying the parameters of the choice model where approval in each stage $s$ occurs if

$$r_i > \varepsilon_{ijs}(X_{ijs}, W_{ijs}) = \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}(W_{ijs})$$  \hspace{1cm} (A7)

We want to identify $G_{js,W}$, the distributions of the errors $\tilde{\varepsilon}_{ijs}(W_{ijs})$, the distribution of $r_i$, $F_r$, as well as the coefficients $\gamma_{js}$ and $\beta_s$. Nonparametric identification requires Assumption 1 and Assumption 3.

Assumption 1 and Assumption 3(a) are familiar from the standard literature on nonparametric binary choice models. In Assumption 3(a), note that identification requires variation in $X_{ijs}$ conditional on regressors $W_{ijs}$ that affect the distribution of errors. Assumption 3(b) guarantees that there is variation in the second-round regressors conditional on the first-round regressors and judge identities in both rounds.

Rewriting Equation A7, in each stage an individual is approved if

$$1[-X_{ijs}\beta_s - \gamma_{js} > \tilde{\varepsilon}_{ijs}(W_{ijs}) - r_i] = H_{js,W}(-X_{ijs}\beta_s - \gamma_{js})$$  \hspace{1cm} (A8)

where $H_{js,W}$ is the distribution of $\eta_{ks} = \tilde{\varepsilon}_{ijs}(W_{ijs}) - r_i$, the composite error of the refugee-level equality variable $r_i$ and the case-judge idiosyncratic error $\tilde{\varepsilon}_{ijs}(W_{ijs})$. As in Manski (1975), the assumption of median-zero errors allows nonparametric identification of $\beta_1$ and $H_{k1}$ up to scale. However, the identity of judge $j$ and the regressors $W_{ijs}$ enter the distribution $H_{js,W}$, and thus neither $W_{ijs}$ or the judge effect $\gamma_j$ can be used for identification. Instead, $X_{ij1}$ traces out the distribution of $H_{js,W}$, which is why Assumption 3(a) calls for large support conditional on judge assignment and $W_{ijs}$.

In the second round, the second and third conditions imply that
\[
\lim_{X_{ij1, \beta_1 \to -\infty}} 1 \left(-X_{ik2, \beta_2 - \gamma k_2} > \bar{\varepsilon}_{ik2} (W_{ik2}) - r_i - X_{ij1, \beta_1 - \gamma j_1} > \bar{\varepsilon}_{ij1} (W_{ij2}) - r_i \right) = H_{k2, W} (-X_{ij2, \beta_2 - \gamma j_2})
\]

so \(\beta_2\) and \(H_{k2}\) are similarly identified to scale after conditioning on values of \(X_{ij1}\) so that all first-round claimants are accepted. Note the parallels between the use of \(X_{ij1}\) to conditionally approve all claimants and the use of the non-limiting judge in Section 3.4 to similarly approve all claimants.

By Assumption 3(d), the scale in the two rounds is the same. Then, as in Chen, Heckman, and Vytlacil (1999), the variances of \(\bar{\varepsilon}_{ij1}\) and \(\bar{\varepsilon}_{ij2}\) are identified relative to \(r_i\) from the variance of the first and second round residuals and their covariance. Finally, the result of Rao (1971) recovers the distributions \(G_{j1, W}, G_{j2, W}\), and \(F_r\).
A5 Estimation details for structural model

For notational simplicity, I collapse all coefficients and regressors into the distribution of the observational error $\varepsilon_s$, which I denote with mean $\mu_s$ and standard deviation $\sigma_s$. I first explain the derivation of first-round approval probabilities, then the second-round probabilities.

A5.1 First round approval

$$P(r - \varepsilon_1 > 0) = \int_0^\infty P(r > \tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1)$$

$$= \int_0^{x_m} P(r > \tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1) + \int_{x_m}^\infty P(r > \tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1)$$ (A10)

The first term in Equation A10 can be shown to be equal to $\Phi[\frac{\ln(x_m) - \mu_1}{\sigma_1}]$. Then,

$$\int_{x_m}^\infty P(r > \tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1) = \int_{x_m}^\infty x_m^{\alpha} \phi \left( \frac{\ln(\tilde{\varepsilon}_1) - \mu_1}{\sigma_1} \right) \frac{1}{\sigma_1} d\tilde{\varepsilon}_1$$

$$= x_m^{\alpha} \int_{\ln(x_m) - \mu_1}^{\infty} e^{-\alpha(\sigma_1 y + \mu_1)} \phi(y)dy$$

$$= x_m^{\alpha} e^{-\alpha \mu_1} \int_{\ln(x_m) - \mu_1}^{\infty} e^{-\alpha \sigma_1 y - \frac{y^2}{2}} dy$$

$$= x_m^{\alpha} e^{-\alpha \mu_1 + \frac{\alpha^2 \sigma_1^2}{2}} \int_{\ln(x_m) - \mu_1}^{\infty} e^{-\frac{1}{2}(y + \alpha \sigma_1)^2} dy$$

$$= x_m^{\alpha} e^{-\alpha \mu_1 + \frac{\alpha^2 \sigma_1^2}{2}} \int_{\ln(x_m) - \mu_1 + \alpha \sigma_1}^{\infty} e^{-\frac{1}{2}y^2} dy$$

$$= x_m^{\alpha} e^{-\alpha \mu_1 + \frac{\alpha^2 \sigma_1^2}{2}} \left[ 1 - \Phi \left( \frac{\ln(x_m) - \mu_1 + \alpha \sigma_1}{\sigma_1} \right) \right]$$ (A11)

where the second equality follows from substituting $y = \frac{\ln(x_m) - \mu_1}{\sigma_1}$ and $\tilde{\varepsilon}_1^{-\alpha} = e^{-\alpha \ln(\tilde{\varepsilon}_1)}$. The fourth equality follows from completing the square: $-\frac{1}{2}(y^2 + 2\alpha \sigma_1 y) = -\frac{1}{2}(y^2 + 2\alpha \sigma_1 y + \alpha^2 \sigma_1^2) + \frac{\alpha^2 \sigma_1^2}{2} = -\frac{1}{2}(y + \alpha \sigma_1)^2 + \frac{\alpha^2 \sigma_1^2}{2}$.

A5.2 Approval in both rounds

In the model I estimate, occasionally the same judge is assigned to make the first and second round decision for a defendant. I model this by allowing between-round errors to be correlated whenever
it is the same judge and estimate the correlation as an additional parameter. Below, I present the full derivations for the no-correlation case (which is more intuitive), then explain how the model works with correlations.

The likelihood of approval in the first round is

\[ P(r > \varepsilon_2 \cap r > \varepsilon_1) = \int_0^\infty \int_0^\infty P(r > \varepsilon_2 \mid r > \varepsilon_1)P(r > \varepsilon_1)dF(\varepsilon_1)dF(\varepsilon_2) \] (A12)

The terms inside the integrals can be rewritten

\[ P(r > \varepsilon_1) = 1[\varepsilon_1 < x_m] + 1[\varepsilon_1 \geq x_m] \frac{x_m^{\alpha}}{\varepsilon_1^{\alpha}} \] (A13)

and

\[ P(r > \varepsilon_2 \mid r > \varepsilon_1) = 1[\varepsilon_1 < x_m] \left[ 1[\varepsilon_2 < x_m] + 1[\varepsilon_2 \geq x_m] \frac{x_m^{\alpha}}{\varepsilon_2^{\alpha}} \right] + \right. \\
\left. 1[\varepsilon_1 \geq x_m] \left[ 1[\varepsilon_2 < \varepsilon_1] + 1[\varepsilon_2 \geq \varepsilon_1] \frac{\varepsilon_1^{\alpha}}{\varepsilon_2^{\alpha}} \right] \] (A14)

Substituting into Equation A12 and expanding the integrals,

\[ P(r > \varepsilon_2 \cap r > \varepsilon_1) = \int_0^\infty \int_0^{x_m} 1[\varepsilon_2 < x_m] + 1[\varepsilon_2 \geq x_m] \frac{x_m^{\alpha}}{\varepsilon_2^{\alpha}} dF(\varepsilon_1)dF(\varepsilon_2) \]

\[ + \int_0^\infty \int_{x_m}^{\infty} 1[\varepsilon_2 < \varepsilon_1] + 1[\varepsilon_2 \geq \varepsilon_1] \frac{\varepsilon_1^{\alpha}}{\varepsilon_2^{\alpha}} dF(\varepsilon_1)dF(\varepsilon_2) \]

Further separate the integrals into four components:

\[ \int_0^{x_m} \int_0^{x_m} dF(\varepsilon_1)dF(\varepsilon_2) \] (A15)

\[ \int_{x_m}^{\infty} \int_0^{x_m} \frac{x_m^{\alpha}}{\varepsilon_2^{\alpha}} dF(\varepsilon_1)dF(\varepsilon_2) \] (A16)
\[
\int_{x_m}^{\infty} \int_{0}^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2)dF(\tilde{\varepsilon}_1) = (A17)
\]

\[
\int_{x_m}^{\infty} \int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_2)dF(\tilde{\varepsilon}_1) = (A18)
\]

These four equations (A15-A18) are all simple to evaluate because the distribution of a Pareto-distributed random variable conditional on being larger than a given threshold is itself Pareto. I solve them in turn:

\[
\int_{0}^{x_m} \int_{0}^{x_m} dF(\tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_2) = \Phi \left( \frac{x_m - \mu_1}{\sigma_1} \right) \Phi \left( \frac{x_m - \mu_2}{\sigma_2} \right)
\]

\[
\int_{x_m}^{\infty} \int_{0}^{x_m} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_2) = \Phi \left( \frac{x_m - \mu_1}{\sigma_1} \right) \int_{x_m}^{\infty} e^{-\alpha \ln \tilde{\varepsilon}_2} dF(\tilde{\varepsilon}_2)
= \Phi \left( \frac{x_m - \mu_1}{\sigma_1} \right) e^{-\alpha \mu_2 + \frac{\alpha^2 \sigma_2^2}{2}} \left[ 1 - \Phi \left( \frac{\ln(x_m) - \mu_2}{\sigma_2} + \alpha \sigma_2 \right) \right]
\]

The last two make use of the additional fact that

\[
\int_{z}^{\infty} \phi(x) \Phi \left( \frac{x - b}{a} \right) dx = P[Y < \frac{X - b}{a}, X > z]
= P[aY - X < -b, X < -z]
= BvN \left( \frac{-b}{\sqrt{a^2 + 1}}, -z, \frac{1}{\sqrt{a^2 + 1}} \right)
\]

where \(BvN\) is the CDF of the standard bivariate normal. This is important because bivariate normals can be cheaply evaluated using Gauss-Legendre quadrature.

\[
\int_{x_m}^{\infty} \int_{0}^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2)dF(\tilde{\varepsilon}_1) = \int_{x_m}^{\infty} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \Phi \left( \frac{\ln(\tilde{\varepsilon}_1) - \mu_2}{\sigma_2} \right) dF(\tilde{\varepsilon}_1)
\]

\[
= x_m^\alpha \int_{x_m}^{\infty} e^{-\alpha \ln(\tilde{\varepsilon}_1)} \Phi \left( \frac{\ln(\tilde{\varepsilon}_1) - \mu_2}{\sigma_2} \right) \phi \left( \frac{\ln(\tilde{\varepsilon}_1) - \mu_1}{\sigma_1} \right) \frac{1}{\sigma_1 \tilde{\varepsilon}_1} d\tilde{\varepsilon}_1
\]
Equation A18 in detail; the same method works for all the joint first- and second-round probabilities. This is used when the same judge sees the case in both rounds. I describe the version for A5.3 Approval in both rounds with error correlation

\[ \int_{x_m}^{\infty} x_m^\alpha e^{-\alpha \mu_2} \frac{\alpha^2 \sigma_2^2}{2} \left[ 1 - \Phi \left( \frac{\ln(x_m) - \mu_1}{\sigma_1} \right) \right] dF_1(x_m) \]

where \( \tilde{B} = x_m^\alpha e^{-\alpha \mu_2} \frac{\alpha^2 \sigma_2^2}{2} \).

A5.3 Approval in both rounds with error correlation

In this section I describe how the probabilities can be modified to allow for correlation between rounds. This is used when the same judge sees the case in both rounds. I describe the version for Equation A18 in detail; the same method works for all the joint first- and second-round probabilities.

\[ \int_{x_m}^{\infty} \int_{\tilde{\epsilon}_1}^{\infty} x_m^\alpha e^{-\alpha \mu_2} \frac{\alpha^2 \sigma_2^2}{2} \left[ 1 - \Phi \left( \frac{\ln(x_m) - \mu_1}{\sigma_1} \right) \right] dF_1(x_m) d\tilde{\epsilon}_1 \]

\[ = \tilde{B} \left\{ \left[ 1 - \Phi \left( \frac{\ln(x_m) - \mu_1}{\sigma_1} \right) \right] - \int_{\ln(x_m) - \mu_1}^{\infty} \Phi \left( \frac{y + (\mu_1 - \mu_2 + \alpha \sigma_2^2)/\sigma_1}{\sigma_2/\sigma_1} \right) \phi(y) dy \right\} \]

Complete the square in the exponentiated part, then substitute into the above equation. This allows you to take the integral with respect to \( x \), leaving
\[ x_m^\alpha e^{-\alpha \mu_2 + \frac{\alpha^2 \sigma_2^2}{2}} \int_{\ln(x_m) - \mu_1}^{\infty} \left( 1 - \Phi \left( \frac{\frac{\sigma_1 y + \mu_1 - \mu_2}{\sigma_2} - (\rho y - \alpha \sigma_2 (1 - \rho^2))}{\sqrt{1 - \rho^2}} \right) \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (y + \alpha \sigma_2 \rho)^2} dy \]

Rearrange the term in the normal:

\[ \frac{\sigma_1 y + \mu_1 - \mu_2}{\sigma_2} - (\rho y - \alpha \sigma_2 (1 - \rho^2)) = y + (\mu_1 - \mu_2 + \alpha \sigma_2^2 (1 - \rho^2))/(\sigma_1 - \rho \sigma_2) \]

Substitute back in, then change of variables the constant term in the normal. This puts the expression in a form where the probability can be expressed as a bivariate normal, and hence cheaply evaluated.

\[ \overline{B} \left\{ 1 - \Phi \left( \frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha \sigma_2 \rho \right) \right\} - \overline{BvN} \left( \frac{-b}{\sqrt{a^2 + 1}}, \frac{-\ln(x_m) - \mu_1}{\sigma_1} - \alpha \sigma_2 \rho, \frac{1}{\sqrt{a^2 + 1}} \right) \]

\[ \overline{B} = x_m^\alpha e^{-\alpha \mu_2 + \frac{\alpha^2 \sigma_2^2}{2}} \]

\[ b = \alpha \sigma_2 \rho - (\mu_1 - \mu_2 + \alpha \sigma_2^2 (1 - \rho^2))/(\sigma_1 - \rho \sigma_2) \]

\[ a = \sigma_2 \sqrt{1 - \rho^2}/(\sigma_1 - \rho \sigma_2) \]
A6 Model assumption checks and robustness

A6.1 Decision timing as regressors

Identification requires that the case timing regressors affect judge thresholds $\gamma_{js}$ but are not correlated with judge errors $\tilde{\varepsilon}_{ij}$ or case strength $r_i$ (Assumption 3). In this section I test some of the implications of this assumption.

I explore whether the regressors are uncorrelated with case strength in Table A3. Because case strength is unobserved, I predict first- and second-round approval from country of origin and gender of the claimant, then test whether this omnibus measure of $r_i$ can be predicted by the regressors. The coefficients in Columns 1 and 2 are small and insignificant, suggesting that the timing of the cases is uncorrelated with case strength. In Columns 3 and 4, I show that the timing of the decision has a significant effect on both first- and second-round approval. This is important because it suggests that $X_{ij}$ sizably affects judge thresholds $\gamma_{js}$, and that the regressors make a substantive contribution to identification.

Another fear is that decision timing affects approval through changing judge errors $\tilde{\varepsilon}_{ij}$ rather than case strength $r_i$ or judge thresholds $\gamma_{js}$. If this were true, one possible implication is that the regressors would affect the distribution of errors. I test this directly in Table A4, where I include in turn the two decision timing regressors in $W_{ij}$. In a nonparametric sense, the model is identified by an excluded regressor in each round that affects thresholds but not errors; since the noon-hearing regressor affects errors only second-round outcomes these should be interpreted as tests on second-round identification. In line with the relevance tests in Table A3, both regressors have strong and statistically significant effects on $\gamma_{js}$. However, neither has a statistically significant affect on $\sigma_{j2}$—the log-log end-of-week coefficient is only 0.04 (SE=0.06) and the noon-hearing coefficient a very imprecisely estimated 0.37 (SE=1.29).

A6.2 Model parameters without additional regressors

The baseline model uses dummies for a late-week decision and whether the second-round hearing was made over lunch to aid in identification. In this section I present estimates from a model identified without regressors, as well as the main results. Identification now leans more strongly on functional form, though judge randomization still identifies relative consistency for judges with
similar approval rates. In Figure A2 I find that the results are qualitatively unchanged but slightly less precise, suggesting that $X_{ijS}$ indeed affects outcomes only through judge thresholds, not judge errors, as required by Assumption 3.

In Section 4.8 I presented evidence that judicial inconsistency is due more to idiosyncratic observational errors than permanent taste-based differences between judges in statutory interpretation or bias. One fear with this approach is that errors arising from late-week and noon-time decisions may be precisely idiosyncratic errors rather than taste-based ones. In other words, the choice of regressors is determining the result. Table A5 tests the additional explanatory power of judge identity analogously to Table 3, but uses the no-regressor model probabilities. The results are comparable to the baseline specification: the model predicts second-stage approval well, but conditional on the model probabilities there is little additional predictive power from knowing the exact judge pairs. The distribution of the EB means of the judge pairs is very similar—in my preferred, rightmost specification, the standard deviation of the judge pair effects is 0.003 in the both models—and the F-test similarly does not reject that the judge pair effects are jointly zero.

In Table A6, I test the effect of experience and workload on inconsistency. Similarly to Table 4, I find that judges become dramatically more consistent after one year of experience, but continue to make gains through at least the first ten years on the job. Higher caseloads decrease consistency (Columns 3 and 4), though only for judges with fewer than 6 years of experience (Column 5).

Table A7 contains estimates of the effect of judicial selection reform on judge consistency. As I describe in Section 2.3, changes to the laws governing judicial selections made it much more difficult for governments to grant judgeships to unqualified party supporters after 1988. In Section 4.11 I show that this reduced baseline estimates of consistency by approximately 75%. In the model estimated without regressors, the results are large but not quite so dramatic—in the baseline specification inconsistency declines by 0.7 from a pre-reform mean of 1.7.

Finally, Table A8 mirrors the results of the baseline model: estimates of judge thresholds $\gamma_{jS}$ are negatively correlated with survey measures of judge favorability to claimants, and model-estimated inconsistency $\sigma_{j1}$ is negatively correlated with surveyed consistency. Again, this suggests that the correlation between model and survey results are not driven by the use of regressors in identification.
A6.3 Reduced-form and structural parameters for second-round inconsistency

For second-round judges, identification relies on matching pairs of second-round judges with similar approval rates conditional on first-round approval by a very lenient first-round judge. Equation A6 shows that second-round approval rates conditional on first-round approval by a different, less lenient judge will be higher for the more consistent second-round judge. Figure A3 shows how the estimated model reflects this logic. Fortunately, my data contain one judge who approves 70% of first-round claimants, while the next-most-lenient judge approves only 28%. I match second-round judges by approval rates conditional on first-round approval by the outlier judge, taking all pairs with approval rates within 5 percentage points. The figure displays the binned scatter plot of the within-pair difference in estimated second-round inconsistency $\sigma_{j2}$ and the difference in second-round approval rates conditional on first-round approval by all other judges, and find that the larger the difference in approval rates, the larger the difference in estimated $\sigma_{j2}$. Higher approval rates under the comparison judges correspond to higher consistency (lower $\sigma_{j2}$).
A7 Survey questions

As I discuss in Section 4.10, I fielded a survey of lawyers who had appeared in front of the Federal Court justices in my sample. The goal of the survey was to generate expert measures of the same parameters that are identified by my structural model.

From the court records, I located the names of 931 lawyers who had appeared in front of one of the judges in my sample. I was able to find online contact information for 551 of them. In April 2017, I contacted the lawyers and requested that they fill out an online survey on their experience with Federal Court judges. After one reminder email, 64 lawyers responded for an overall response rate of 14%. Table A9 compares responders to non-responders and lawyers for whom I couldn’t find contact information. The main differences are that responders are more successful, with a first-round approval rate of 27% versus 19% for non-responders (the contacted sample is mostly lawyers for the claimants; government lawyers were included in the sample but their names are recorded much less frequently in the court documents). Respondents are slightly younger, with their first recorded case coming about one year later.

Each survey asked three questions on up to four judges, personalized to reflect the justices they had actually appeared in front of (there was also an option to fill out a non-personalized, anonymous survey on my academic website if they were concerned about privacy). The questions were:

1. On a scale from 1 to 5, how would you rate the listed judges in terms of favourableness towards claimants? Do they rule for the claimant more or less often than other judges? Given the facts of the case, are they more likely to either grant leave or rule for the claimant during judicial review?

Each question concerns one judge only, and your answer should reflect your holistic understanding of the judge’s behavior across both leave and judicial review stages, not the outcome of a specific case or what you feel the decisions ought to be.

---

2 The main source of contact information was www.canadianlawlist.com, where I found 370 emails. Another 140 were on lawyers’ own websites. The rest of the contact information was in the form of online form submissions on lawyer-directory websites like www.lawyer.com, although the response rate from these forms was almost zero.

3 This response rate compares favorably to telephone political polls, where response rates are below 10% (Pew, 2012). However, it is significantly lower than the 20% response rate for an email poll conducted by Card et al. (2012) surveying UC Berkeley staff about job satisfaction. The difference in response rates is likely due to declining survey rates over time (Card et. al surveyed in 2008), a pecuniary incentive, and that they had the advantage of being able to present themselves as in-group members (other University of California employees).
2. On a scale from 1 to 5, how would you rate the listed judges in terms of **consistency**? Are their decisions predictable compared to other judges with similar grant rates? Do they decide cases on similar grounds as other justices? Can you predict what grounds the case will be decided on?

Each question concerns one judge only, and your answer should reflect your **holistic understanding** of the judge’s behavior across both leave and judicial review stages, not the outcome of a specific case or what you feel the decisions ought to be. This can include information you’ve heard from colleagues.

3. On a scale from 1 to 5, how would you rate the listed judge in terms of **accuracy**? Do they make the right legal decisions?

Each question concerns one judge only, and should be answered relative to other judges. Your answer should reflect your **holistic understanding** of the judge’s behavior across both leave and judicial review stages, not only the specific cases you have been involved with. Unlike the previous questions, it can reflect your personal opinion on how cases should be decided.

I expected that the first question would be related to the judge-specific threshold $\gamma_j$, and the second question with the variance of the observational error $\sigma_j$. By design, the second question encompasses the two distinct aspects of $\sigma_j$ detailed in Section 3.5.3. First, asking about predictability concerns test-retest reliability—will the judge understand the merits of the case? On the other hand, asking whether the judge decides cases on similar grounds as other judges is trying to unearth information about how judges consistently value different aspects of the case (inter-rater reliability), such as the relative weight they place on procedural versus substantive merits.

Each response was on a five-point likert scale (I reverse the ordering of the consistency response so it is analogous to the estimated inconsistency coefficients). I normalize responses by the mean and standard deviation, but it is worth noting that the likert responses were centered at 3 (“average”) for both consistency and accuracy. For favorability, the median lawyer response was a 4 (“slightly more favourable to claimants than average”).

The main results are in Table 5, where I include only the first two questions. I discuss these in Section 4.10. The final question of the survey, which asked about how accurate the judge is, I did not discuss in the main text. This question does not have as clear an interpretation as the other
two. There is no direct mapping of accuracy into the model, since accuracy implies a normative judgement about the correct outcome of the case. Reported accuracy is correlated positively with favorability and negatively with inconsistency, but more strongly with the former ($\rho = 0.7$ versus $-0.46$). Anecdotally, many of the lawyers that I corresponded with about the survey were involved in refugee-rights non-profits, so it is likely that they believe the claimants should win more cases than they currently do. Table A11 adds accuracy to the regression of model coefficients on survey responses; with no other regressors higher accuracy predicts lower second-stage thresholds $\gamma_{j2}$, but this disappears when favorability and inconsistency are added. The relationship between favorability and $\gamma_2$, and inconsistency and $\sigma_1$ is almost unchanged.
A8  Testing monotonicity and bounding $\Lambda$

In Section 3.1 I show how to test one implication of monotonicity: that for each pair of consecutively-more-severe judges, the more lenient judge should approve fewer of each demographic group. Although the number of demographic variables are limited in this context, I tested that assumption using demographic groups defined by the intersection of gender, the claimant being African, and the case coming from Montreal (which provides the lowest level of legal aid, and so might have lower-quality cases). I exclude the high-approving judge for comparability with more standard judge contexts (when including him, he accounts for about 40% of the IV weight), and partial out year and office fixed effects.

I estimate Equation 5 using the demographic groups defined above, and then test the null that all the $\alpha$ are weakly larger than 0. The histogram of t-stats is displayed in Figure A4; although the median coefficient is larger than zero there are many precisely estimated negative coefficients. Unsurprisingly, a formal Wolak (1989) test definitely rejects the null of monotonicity, with a $p$-value of $1 \times 10^{-10}$.

To understand the possible impacts on IV estimates, I also estimate bounds on $\Lambda = \sum_{j=1}^{J} \varphi_j \lambda_j$ using Equation 7, where $\varphi_j$ are the IV weights and $\lambda_j$ are the pairwise $\lambda$ for the instrument of assignment to judge $j$ rather than judge $j-1$. With $J$ judges and $G$ demographic groups, calculating the bound requires checking each of $2^{JG}$ possible combinations of coefficients in Equation 5, which is infeasible for even modest numbers of judges and demographic groups. However, the number of calculations can be reduced to a handful whenever the coefficients $\alpha^g_j$ (which are main ingredients for the bound) are mutually independent.\footnote{Estimation of the CLR bound in this case requires knowing the highest-variance linear combination of the coefficients $\tau \alpha$ where $\tau = \in \{ -1, 0 \}^{JG}$ is the set of $JG$ length vectors combining -1 and 0. Under independence of $\alpha$ this is the 1 $\times$ $JG$ length -1 vector. It also requires knowing the maximum value of the lower confidence interval of all linear combinations for a wide variety of p-values, which under independence can be found by taking the vector with a -1 entry whenever the corresponding coefficient that can be rejected at the $p$ level, and a zero elsewhere.} This is true whenever Equation 5 does not contain controls, which is why I chose to pre-residualize out the year and office effects rather than include them as controls.

As expected given the results of the Wolak test, I reject that $\Lambda$ is equal to zero, with a median-unbiased estimate of 1.14 and a 95% confidence interval lower endpoint of 0.91. Figure A5 displays the IV weights, as well as the bound on the IV-weighted $\lambda_j$'s for each histogram bin.\footnote{To be precise, in each block of judges I estimate the lower bound on $\sum_{j \in \text{block}} \varphi_j \lambda_j / (\sum_{k \in \text{block}} \varphi_k)$.} Here I see...
more evidence of possible biases from the IV estimate; the median-unbiased bound is larger than zero for most of the support of the instrument. For most values of the instrument the bound on mean $\lambda_j$ is larger than 0.5, and zero can be rejected.

Despite the unequivocal rejections of monotonicity by both of these tests, traditional methods of testing monotonicity in judge settings would not have detected the issue. In Table A16 I conduct the whole-sample and reverse-sample tests of Dobbie et al. (2018) and Bhuller et al. (2016). They consist of separate first stage regressions in different demographic subsamples, asking whether higher judge mean approval calculated either overall or in a different subsample predicts higher approval in the subsample. I show in Section 3.1 that these tests of average monotonicity are weaker than tests of pairwise monotonicity. The improvement in sensitivity is important in my application. In Table A16 I find that the overall and reverse-sample judge severity instruments have a very strong, very positive first stage in all subsamples, with t-stats never smaller than 8. If the researcher conducted only this test, it would mistakenly lead them to conclude that there are no monotonicity violations.
A9 Simulated effect of monotonicity violations on IV estimates

As I show in Appendix A8, in my context there are large violations of monotonicity caused by differential harshness across claimant demographic types. This immediate problem could be addressed by interacting the instrument with the charge type, or using Mueller-Smith (2015)’s LASSO instrument construction approach to efficiently predict incarceration rates from judge assignment flexibly interacted with covariates. However, neither of these methods will address monotonicity violations along unobserved dimensions—for example, if some judges were differentially harsh towards defendants accused of violence against different-race victims, but this was not recorded in the case file—potentially leading to bias in judge-assignment designs. In this section, I use the estimated levels of inconsistency in this data to explore the bias engendered by inconsistency on IV and MTE estimates of the effect of first round approval on second round approval.

There is not a single way to quantify the effect of inconsistency on the estimated MTE. There are many potential non-degenerate joint distributions of first-round judge errors \( \tilde{\varepsilon}_{ij1} \) that do not generate violations of monotonicity—for example, if all judges had the same error \( \tilde{\varepsilon}_{ij1} \) for each claimant. This is important because although the estimated parameters guarantee violations of monotonicity (recall from Section 4.5 that I bound the share of cases judge pairs disagree on above zero), different assumptions on the cross-judge joint distribution of \( \tilde{\varepsilon}_{ij1} \) allow for different counterfactuals that satisfy the monotonicity assumption. For simplicity I choose the most straightforward alternative: judges are perfectly consistent \( \sigma_{j1} = 0 \) for all judges), guaranteeing monotonicity is satisfied. I adjust thresholds \( \gamma_{j1} \) to keep the approval rate the same for all judges, then calculate the MTE under both the baseline coefficients and the counterfactual. Klein (2010) emphasizes that the direction of the bias is \textit{a priori} unknown and depends on the slope and curvature of the underlying MTE, as well as the variance of mean inconsistency over the limit points (e.g., judges) of the incarceration probability. Because there are fewer judges in the tails of the distribution, we might expect larger levels of bias there, since the variance of mean judge inconsistency with respect to incarceration probability increases as the number of judges decreases. This is precisely what we see in Figure A6; as consistency declines, the estimated MTE becomes shallower. This results in a medium-size negative bias for defendants induced into treatment by severe (low-probability of approval) judges of about 11%, and a relatively large positive bias of about 25% from the perfect-consistency estimate for high-approval judges. This reflects how inconsistency generates a distribution of marginal claimants for
each judge, each with a different treatment effect. Since it is difficult to predict which judges will be inconsistent—and thus what range of claimants are marginal when assigned to them—it is difficult to sign the direction of the bias.

In Section 3.1, I discuss how the LATE estimate is a weighted average of types defined by complier and defier status under different judges. I show how the net effect of an individual being a complier and a defier under different judges may be small because the effects (at least partially) offset each other, suggesting that the bias may be less severe under IV than under MTE. When I estimate the LATE for the simulated data, I find that this is indeed the case. The two types of bias observed in the MTE work in opposite directions, and so the simulated linear IV estimate is barely affected by the change from the perfect-consistency to the baseline estimate, moving from 0.251 to 0.265, or 5%. This is particularly striking because the tests of monotonicity estimated in Appendix A8 decisively reject monotonicity, and bound the mean $\lambda_j$ of Imbens and Angrist (1994) above zero for parts of the distribution of judge strength. More broadly, it suggests that inconsistency in judge-assignment designs may more strongly affect the MTE than the IV estimate.
A10 Appendix Figures

Figure A1: Scatter plot of first- and second-round consistency $\sigma_{j1}$ and $\sigma_{j2}$.
Figure A2: Distribution of judge coefficients, model identified without regressors

(a) Threshold $\gamma_1$, first round

(b) Observational error $\sigma_1$, first round

(c) Threshold $\gamma_2$, second round

(d) Observational error $\sigma_2$, second round

Figure displays coefficients for decision model $1 [r_i > \gamma_{js} + \tilde{\epsilon}_{ij}]$, $\tilde{\epsilon}_{ij} \sim N(0, \sigma_{js}^2)$. All models allow the parameters of the Pareto distribution of $r_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Each panel contains the density of the raw and shrunken estimates of the judge thresholds $\gamma_1$ and $\gamma_2$, and judge inconsistency $\sigma_1$ and $\sigma_2$. Black line is density of case quality $r_i$. Shrunken estimates recovered via deconvolution of estimates accounting for coefficient-specific standard errors (Delaigle, Hall, and Meister, 2008).
This figure relates to the estimated decision model $I[r_i > \gamma_{js} + X_{ij} \beta_s + \tilde{\varepsilon}_{ij}], \tilde{\varepsilon}_{ij} \sim \mathcal{N}(0, \sigma^2_{js})$. I match pairs of second-round judges with similar approval rates conditional on first-round approval by a high-approving (non-limiting) first-round judge. I then compare the difference in second-round observational error $\sigma_2$ as a function of within-pair differences in second-round approval rates conditional on first-round approval by all other judges. See Appendix A3 for more details.
The figure displays a histogram of the t-stats for $\alpha_j^g$ from Equation 5, where the demographic groups are defined using the intersection of the claimant being African, the claimant being male, and the case being decided in Montreal. Under monotonicity, all coefficients should be positive.
Figure A5: Bounds on IV-weighted judge-pair $\lambda_j$

On the left axis, the figure displays a histogram of the IV weights. On the right axis, the solid line displays the median-unbiased estimate of the bound on the IV-weighted $\lambda_j$’s for judges of that severity. The dashed line marks the one-sided 95% confidence interval.
Figure A6: Estimated MTE at baseline and under consistency

Figure demonstrates how inconsistency affects the estimated MTE. I plot the baseline MTE in solid blue, then use model estimates to construct an estimate of the MTE if all judges were perfectly consistent (ie, $\sigma_{ij} = 0$). MTE estimated by parametric regression of simulated outcome on a cubic polynomial of judge probability of approval, then taking the derivative.
### Table A1: Judge summary statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male judge (=1)</td>
<td>0.75</td>
<td>0.44</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Liberal appointee (=1)</td>
<td>0.72</td>
<td>0.45</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Experience (years)</td>
<td>6.51</td>
<td>5.63</td>
<td>0.00</td>
<td>28.00</td>
</tr>
<tr>
<td>Workload</td>
<td>-0.07</td>
<td>0.80</td>
<td>-3.45</td>
<td>1.53</td>
</tr>
<tr>
<td>Male (=1)</td>
<td>0.63</td>
<td>0.43</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>African (=1)</td>
<td>0.19</td>
<td>0.39</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Asia (=1)</td>
<td>0.10</td>
<td>0.31</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>South American (=1)</td>
<td>0.35</td>
<td>0.48</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Calgary (=1)</td>
<td>0.02</td>
<td>0.14</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Montreal (=1)</td>
<td>0.42</td>
<td>0.49</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Ottawa (=1)</td>
<td>0.02</td>
<td>0.13</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Vancouver (=1)</td>
<td>0.03</td>
<td>0.18</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Observations</td>
<td>58,604</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table A2: Randomization tests using name-imputed continent of origin

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Africa</th>
<th>Asia</th>
<th>South America</th>
<th>Predicted approval</th>
<th>1st-round mean approval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
</tbody>
</table>

**Panel A: First round judges**

<table>
<thead>
<tr>
<th>First-round approval rate</th>
<th>-0.002</th>
<th>-0.068***</th>
<th>-0.011</th>
<th>0.098</th>
<th>-0.002</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.023)</td>
<td>(0.040)</td>
<td>(0.071)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>F-stat</td>
<td>0.87</td>
<td>2.90</td>
<td>3.04</td>
<td>6.65</td>
<td>3.51</td>
</tr>
<tr>
<td>Prob</td>
<td>0.75</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Observations</td>
<td>58,604</td>
<td>58,604</td>
<td>58,604</td>
<td>58,604</td>
<td>58,604</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second-round approval rate</th>
<th>-0.042</th>
<th>0.032</th>
<th>-0.022</th>
<th>0.027</th>
<th>0.002</th>
<th>-0.027</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.039)</td>
<td>(0.042)</td>
<td>(0.041)</td>
<td>(0.003)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>F-stat</td>
<td>1.07</td>
<td>1.83</td>
<td>1.19</td>
<td>1.61</td>
<td>1.54</td>
<td>4.02</td>
</tr>
<tr>
<td>Prob</td>
<td>0.33</td>
<td>0.00</td>
<td>0.15</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Observations</td>
<td>8,446</td>
<td>8,446</td>
<td>8,446</td>
<td>8,446</td>
<td>8,446</td>
<td>8,446</td>
</tr>
</tbody>
</table>

All regressions include office X pre-2002 fixed effects to account for cross-office differences in case strength and changes in government policy in 2002. Standard errors clustered at the judge level in parentheses. Gender and continent of origin predicted from claimant name. IRB mean approval is the approval rate of the IRB Member who initially denied refugee status to the claimant. Predicted approval comes from a regression of approval on gender, continent of origin and IRB Member approval rate. Judge approval rates on right side partial out office X pre/post-2002. F-stats come from separate regression of outcome on judge fixed effects. \( * p < 0.10, \* * p < 0.05, \* * * p < 0.01 \).
Table A3: Placebo tests and relevance for regressors

<table>
<thead>
<tr>
<th>Predicted approval</th>
<th>Actual approval</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Panel A: First round</td>
<td></td>
</tr>
<tr>
<td>End of week</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Observations</td>
<td>58,604</td>
</tr>
<tr>
<td>Noon hearing</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
</tr>
<tr>
<td>Judge fixed effects</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>8,446</td>
</tr>
</tbody>
</table>

Predicted approval from regression of approval in each round on ethnicity and gender. Controls include year filed and office. All specifications include judge fixed effects. End of week regressor in first panel is dummy for final pre-decision filing taking place on Thursday, Friday, Saturday or Sunday (which predicts the decision will be made after Monday). Standard errors clustered at the judge level in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.
Table A4: Testing effect of regressors on distribution of judge errors

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coefficients $\beta$ affecting judge threshold $\gamma_1$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>End-of-week decision</td>
<td>0.057***</td>
<td>0.087***</td>
<td>0.051**</td>
</tr>
<tr>
<td>(0.004)</td>
<td>(0.029)</td>
<td>(0.021)</td>
<td></td>
</tr>
<tr>
<td>Hearing schedule over lunch</td>
<td>0.411***</td>
<td>0.381**</td>
<td>0.510***</td>
</tr>
<tr>
<td>(0.077)</td>
<td>(0.193)</td>
<td>(0.165)</td>
<td></td>
</tr>
<tr>
<td><strong>Coefficients $\psi$ affecting judge inconsistency $\sigma_1$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>End-of-week decision</td>
<td>0.040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.058)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hearing schedule over lunch</td>
<td>0.371</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1.287)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD of $\gamma_1$</td>
<td>0.836</td>
<td>0.833</td>
<td>0.840</td>
</tr>
<tr>
<td>SD of $\sigma_1$</td>
<td>0.485</td>
<td>0.467</td>
<td>0.475</td>
</tr>
</tbody>
</table>

Reports coefficients for choice model \( 1[\gamma_i > \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}(W_{ijs})] \), \( \tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0,\sigma_{js}(W_{ijs})^2) \), \( \sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s) \). All models include controls for time/date of decision in $\beta$, and allow the parameters of the Pareto distribution of $\gamma_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Table A5: Second-round approval on model approval probability and judge-pair FEs

<table>
<thead>
<tr>
<th>Judge-pair round FEs</th>
<th>Judge-pair FEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Model approval probability</td>
<td>0.949***</td>
</tr>
<tr>
<td></td>
<td>(0.166)</td>
</tr>
<tr>
<td>Model controls</td>
<td>No</td>
</tr>
<tr>
<td>Mean approval</td>
<td>0.44</td>
</tr>
<tr>
<td>F-stat for judge pairs</td>
<td>1.02</td>
</tr>
<tr>
<td>P-value</td>
<td>0.318</td>
</tr>
<tr>
<td>Bootstrap p-value</td>
<td>0.674</td>
</tr>
<tr>
<td>SD of judge-pair EB means</td>
<td>0.006</td>
</tr>
<tr>
<td>Observations</td>
<td>8,196</td>
</tr>
</tbody>
</table>

Regresses second-round approval on model-predicted likelihood of approval and judge-pair fixed effects. In contrast to Table 3, model is estimated without using regressors for identification. Left two columns construct judge-pair FEs accounting for order of assignment; right two columns ignore this distinction. Model controls include office of origination, pre-post 2002, and an end-of-week and noon hearing dummy for the second-round hearing. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Table A6: First-round judge consistency by experience and workload

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coefficients ( \psi ) affecting judge inconsistency ( \sigma_1 )</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experience &gt; 1 year</td>
<td>(-0.797^{**})</td>
<td>(-0.600^{***})</td>
<td>(-0.557^{**})</td>
<td>(-0.462^{**})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.365)</td>
<td>(0.212)</td>
<td>(0.267)</td>
<td>(0.203)</td>
<td></td>
</tr>
<tr>
<td>Experience &gt; 5 years</td>
<td>(-0.351^{***})</td>
<td>(-0.345^{*})</td>
<td>(-0.049)</td>
<td>(-0.054)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td>(0.199)</td>
<td>(0.561)</td>
<td>(0.380)</td>
<td></td>
</tr>
<tr>
<td>Experience &gt; 10 years</td>
<td>(-0.530)</td>
<td>(-0.452^{***})</td>
<td>(-0.447^{***})</td>
<td>(-0.501^{***})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.524)</td>
<td>(0.101)</td>
<td>(0.170)</td>
<td>(0.012)</td>
<td></td>
</tr>
<tr>
<td>Log caseload</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.204^{***})</td>
<td>0.139^{***})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.038)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log caseload (≤ 5 yrs exp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.179^{***})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.025)</td>
</tr>
<tr>
<td>Log caseload (&gt; 5 yrs exp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.146)</td>
</tr>
<tr>
<td>SD of ( \gamma_1 )</td>
<td>0.817</td>
<td>0.715</td>
<td>1.417</td>
<td>1.122</td>
<td>1.044</td>
</tr>
<tr>
<td>SD of ( \sigma_1 )</td>
<td>1.562</td>
<td>1.486</td>
<td>0.537</td>
<td>0.991</td>
<td>1.084</td>
</tr>
<tr>
<td>Second-round experience control</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Career number of cases</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Reports coefficients for decision model \( \mathbb{1}[r_i > \gamma_{js} + X_{ij} \beta_{js} + \tilde{\varepsilon}_{ij}(W_{ij})] \), \( \tilde{\varepsilon}_{ij} \sim N(0, \sigma_{js}^2) \), \( \sigma_{js}(W_{ij}) = e^{\tilde{\beta}_{js} \cdot 5} \). In contrast to the baseline model, the reported models do not use timing regressors for identification. All models allow the parameters of the Pareto distribution of \( r_i \) to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Standard errors clustered at the level of the first stage judge. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).
Table A7: Inconsistency $\tilde{\sigma}_1$ before and after judge selection reform

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Experience control in $\sigma_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) (2) (3)</td>
<td>(4) (5) (6)</td>
</tr>
<tr>
<td>After reform (=1)</td>
<td>-0.154 -0.703** -0.820*** -1.096*** -0.745** -0.698*</td>
<td>(0.150) (0.287) (0.283) (0.307) (0.339) (0.376)</td>
</tr>
<tr>
<td>Liberal appointee (=1)</td>
<td>0.0549</td>
<td>-0.0788</td>
</tr>
<tr>
<td>Male judge (=1)</td>
<td>-0.409**</td>
<td>0.00181</td>
</tr>
<tr>
<td>Year appointed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-reform mean</td>
<td>1.49</td>
<td>1.49</td>
</tr>
<tr>
<td>N judges</td>
<td>53</td>
<td>53</td>
</tr>
</tbody>
</table>

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency $\sigma_{j1}$, which is estimated from decision model $\mathbb{1}[r_{i} > \gamma_{js} + X_{ij}s + \bar{\varepsilon}_{ij}s(W_{ij}s)]$, $\bar{\varepsilon}_{ij}s(W_{ij}s) \sim \mathcal{N}(0, \sigma_{js}(W_{ij}s)^2)$, $\sigma_{js}(W_{ij}s) = \exp(\tilde{\sigma}_{js} + W_{ij}s\psi_{s})$. In contrast to the baseline model, the reported models do not use timing regressors for identification. All models allow the parameters of the Pareto distribution of $r_i$ to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. In the right-hand panel, $\beta_{s}$ and $\psi_{s}$ include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 


### Table A8: Model coefficients on survey responses

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: ( \gamma_1 ) (mean=2.65, SD=1.11)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>-0.314</td>
<td>-0.328**</td>
<td>-0.267</td>
<td>-0.321*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.265)</td>
<td>(0.163)</td>
<td>(0.260)</td>
<td>(0.181)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency, SD</td>
<td>-0.247**-0.135**-0.180***-0.022</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.100)</td>
<td>(0.060)</td>
<td>(0.057)</td>
<td>(0.083)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel B: ( \gamma_2 ) (mean=2.24, SD=1.22)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>-0.315***-0.435***-0.329***-0.487***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
<td>(0.151)</td>
<td>(0.118)</td>
<td>(0.156)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency, SD</td>
<td>-0.046</td>
<td>-0.071</td>
<td>0.045</td>
<td>0.123*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td>(0.087)</td>
<td>(0.064)</td>
<td>(0.073)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel C: ( \sigma_1 ) (mean=2.28, SD=2.19)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>0.130</td>
<td>0.065</td>
<td>0.187**</td>
<td>0.138</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td>(0.105)</td>
<td>(0.077)</td>
<td>(0.105)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency, SD</td>
<td>-0.110</td>
<td>-0.126**-0.167**-0.185***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.059)</td>
<td>(0.065)</td>
<td>(0.063)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
<td>182</td>
</tr>
</tbody>
</table>

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) correction for estimated dependent variable. Decision model is \( r_i > \gamma_j + \tilde{\epsilon}_{ij} \), \( \tilde{\epsilon}_{ij} \sim \mathcal{N}(0, \sigma_{js}(W_{ij})^2) \), \( \sigma_{js}(W_{ij}) = \exp(\bar{\sigma}_{js} + W_{ij}\psi_{js}) \). In contrast to the baseline model, the reported model does not use timing regressors for identification. The parameters of the Pareto distribution of \( r_i \) vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. I adjust for judge experience in \( \beta_s \) and \( \psi_s \). Model standard errors clustered at the level of the first stage judge, linear standard errors at the judge level. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).
Table A9: Lawyer characteristics, survey respondents vs lawyer population

<table>
<thead>
<tr>
<th></th>
<th>Respondents</th>
<th>NR/NC</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success rate (first round)</td>
<td>0.27</td>
<td>0.19</td>
<td>0.078***</td>
</tr>
<tr>
<td></td>
<td>[0.22]</td>
<td>[0.21]</td>
<td>(0.027)</td>
</tr>
<tr>
<td>Success rate (second round)</td>
<td>0.13</td>
<td>0.08</td>
<td>0.049***</td>
</tr>
<tr>
<td></td>
<td>[0.16]</td>
<td>[0.15]</td>
<td>(0.019)</td>
</tr>
<tr>
<td>First case (year)</td>
<td>2002.55</td>
<td>2001.37</td>
<td>1.179*</td>
</tr>
<tr>
<td></td>
<td>[5.36]</td>
<td>[5.39]</td>
<td>(0.698)</td>
</tr>
<tr>
<td>Number of cases (total)</td>
<td>141.77</td>
<td>101.62</td>
<td>40.149</td>
</tr>
<tr>
<td></td>
<td>[225.93]</td>
<td>[221.69]</td>
<td>(28.752)</td>
</tr>
<tr>
<td>Male (=1)</td>
<td>0.67</td>
<td>0.60</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>[0.47]</td>
<td>[0.48]</td>
<td>(0.067)</td>
</tr>
<tr>
<td>Observations</td>
<td>64</td>
<td>867</td>
<td></td>
</tr>
</tbody>
</table>

Sample is all lawyers who appeared before the Federal Court. NR/NC = no response or no contact information. Standard deviations in square brackets and standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 


Table A10: Second-round approval on model approval probability and judge characteristics

<table>
<thead>
<tr>
<th>Characteristic =</th>
<th>Outcome = approved in second round</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>French (1)</td>
</tr>
<tr>
<td>Model approval probability</td>
<td>0.993***</td>
</tr>
<tr>
<td>(0.0321)</td>
<td>(0.0297)</td>
</tr>
<tr>
<td>Characteristic, first-round judge (=1)</td>
<td>-0.0128</td>
</tr>
<tr>
<td>(0.0146)</td>
<td>(0.0217)</td>
</tr>
<tr>
<td>Characteristic, second-round judge (=1)</td>
<td>-0.0115</td>
</tr>
<tr>
<td>(0.0157)</td>
<td>(0.0217)</td>
</tr>
<tr>
<td>Characteristic, both judges (=1)</td>
<td>0.0122</td>
</tr>
<tr>
<td>(0.0215)</td>
<td>(0.0258)</td>
</tr>
<tr>
<td>Model controls</td>
<td>Yes</td>
</tr>
<tr>
<td>Mean approval</td>
<td>0.44</td>
</tr>
<tr>
<td>F-stat for characteristic pairs</td>
<td>0.39</td>
</tr>
<tr>
<td>Prob</td>
<td>0.759</td>
</tr>
<tr>
<td>BS prob</td>
<td>0.767</td>
</tr>
<tr>
<td>SD of judge-pair EB means</td>
<td>0.077</td>
</tr>
<tr>
<td>Observations</td>
<td>8,196</td>
</tr>
</tbody>
</table>

Regresses second-round approval on model-predicted likelihood of approval and characteristics of the judges first and second round judges. Model controls include office of origination, pre-post 2002, and an end-of-week and noon hearing dummy for the second-round hearing. Standard errors in parentheses. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).
Table A11: Model coefficients on survey responses including accuracy response

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Threshold ( \gamma_1 ) (mean=2.49, SD=1.05)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, SD</td>
<td>-0.130</td>
<td>-0.136*</td>
<td>0.111</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>(0.096)</td>
<td>(0.076)</td>
<td>(0.172)</td>
<td>(0.112)</td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>-0.221</td>
<td>-0.315*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.260)</td>
<td>(0.184)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inconsistency, SD</td>
<td>0.153*</td>
<td>0.075</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.087)</td>
<td>(0.069)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>174</td>
<td>174</td>
<td>174</td>
<td>174</td>
</tr>
<tr>
<td><strong>Panel B: Threshold ( \gamma_2 ) (mean=2.16, SD=1.46)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, SD</td>
<td>-0.195***</td>
<td>-0.308***</td>
<td>-0.016</td>
<td>-0.167</td>
</tr>
<tr>
<td></td>
<td>(0.059)</td>
<td>(0.079)</td>
<td>(0.078)</td>
<td>(0.104)</td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>-0.225***</td>
<td>-0.309***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td>(0.107)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inconsistency, SD</td>
<td>0.039</td>
<td>-0.133*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.055)</td>
<td>(0.072)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>174</td>
<td>174</td>
<td>174</td>
<td>174</td>
</tr>
<tr>
<td><strong>Panel C: Inconsistency ( \sigma_1 ) (mean=2.06, SD=2.15)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, SD</td>
<td>0.029</td>
<td>0.092</td>
<td>0.011</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td>(0.073)</td>
<td>(0.083)</td>
<td>(0.099)</td>
<td>(0.113)</td>
</tr>
<tr>
<td>Favorability, SD</td>
<td>0.123</td>
<td>0.052</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.124)</td>
<td>(0.111)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inconsistency, SD</td>
<td>0.126*</td>
<td>0.168**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.078)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>174</td>
<td>174</td>
<td>174</td>
<td>174</td>
</tr>
</tbody>
</table>

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) correction for estimated dependent variable. Decision model is \( 1[r_i > \gamma_{js} + X_{ij}s\beta_s + \bar{\epsilon}_{ij}W_{ij}s], \bar{\epsilon}_{ij}s(W_{ij}s) \sim N(0, \sigma_{js}(W_{ij}s)^2), \sigma_{js}(W_{ij}s) = \exp(\bar{\sigma}_{js} + W_{ij}s\psi_s). \) All models include controls for time/date of decision in \( \beta_s, \) and allow the parameters of the Pareto distribution of \( r_i \) to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. I adjust for judge experience in \( \beta_s \) and \( \psi_s. \) Model standard errors clustered at the level of the first stage judge, linear standard errors at the judge level. * \( p < 0.10, \)** \( p < 0.05, \)**\( p < 0.01. \)
Table A12: Inconsistency $\sigma_1$ on survey responses with approval controls

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorability, SD</td>
<td>0.0408</td>
<td>-0.108</td>
<td></td>
<td>0.125**</td>
<td>-0.0430</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0694)</td>
<td>(0.105)</td>
<td></td>
<td>(0.0539)</td>
<td>(0.0981)</td>
<td></td>
</tr>
<tr>
<td>Inconsistency, SD</td>
<td></td>
<td></td>
<td>0.166***</td>
<td>0.151**</td>
<td>0.210***</td>
<td>0.131***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0563)</td>
<td>(0.0583)</td>
<td>(0.0505)</td>
<td>(0.0437)</td>
</tr>
<tr>
<td>Approval rate control</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Respondent FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
</tr>
</tbody>
</table>

Estimated with Hanushek (1974) weights for estimated dependent variable. All models include respondent fixed effects and a control for the judge-experience level approval rate. Standard errors clustered at the judge level in parentheses. ** $p < 0.05$, *** $p < 0.01$. 
Table A13: Approval rate for judges before and after reform

<table>
<thead>
<tr>
<th></th>
<th>Approval rate</th>
<th>Approval, year residualized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>After 1988 reform (=1)</td>
<td>0.0199</td>
<td>0.0490</td>
</tr>
<tr>
<td></td>
<td>(0.0272)</td>
<td>(0.0546)</td>
</tr>
<tr>
<td>Liberal appointee (=1)</td>
<td>-0.00621</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0211)</td>
<td></td>
</tr>
<tr>
<td>Year appointed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Table A14: Inconsistency $\hat{\sigma}_1$ before and after reform, with control for judge approval rate

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Experience control in $\sigma_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>After 1988 reform (=1)</td>
<td>-0.211</td>
<td>-0.362</td>
</tr>
<tr>
<td></td>
<td>(0.139)</td>
<td>(0.251)</td>
</tr>
<tr>
<td>Liberal appointee (=1)</td>
<td>-0.0429</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td></td>
</tr>
<tr>
<td>Male judge (=1)</td>
<td>-0.201*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.113)</td>
<td></td>
</tr>
<tr>
<td>Year appointed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-reform mean</td>
<td>1.26</td>
<td>1.26</td>
</tr>
<tr>
<td>N judges</td>
<td>53</td>
<td>53</td>
</tr>
</tbody>
</table>

Estimated with Hanushek (1974) weights for estimated dependent variable and a control for judge approval rate. Dependent variable is consistency $\sigma_j$, which is estimated from decision model $I[r_i > X_{ijs}\beta_s + \gamma_{js} + \tilde{\epsilon}_{ijs}(W_{ijs})]$, $\tilde{\epsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. In the right-hand panel, $\beta_s$ and $\psi_s$ include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Table A15: Second-round inconsistency $\hat{\sigma}_2$ for judges before and after reform

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Experience control in $\sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>After reform (=1)</td>
<td>-0.0448</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>(0.290)</td>
<td>(0.377)</td>
</tr>
<tr>
<td>Liberal appointee (=1)</td>
<td></td>
<td>0.0973</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.175)</td>
</tr>
<tr>
<td>Male judge (=1)</td>
<td>-0.453**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.198)</td>
<td></td>
</tr>
</tbody>
</table>

Year appointed

<table>
<thead>
<tr>
<th>Pre-reform mean</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-reform mean</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>3.31</td>
<td>3.31</td>
<td>3.31</td>
</tr>
<tr>
<td>N judges</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
</tbody>
</table>

Estimated with Hanushek (1974) weights for estimated dependent variable. Dependent variable is consistency $\sigma_{j2}$, which is estimated from decision model $\mathbb{1}[r_{ij} > X_{ij,s}\beta_s + \gamma_{js} + \tilde{\epsilon}_{ij,s}(W_{ij,s})]$, $\tilde{\epsilon}_{ij,s}(W_{ij,s}) \sim \mathcal{N}(0, \sigma_{js}(W_{ij,s})^2)$, $\sigma_{js}(W_{ij,s}) = \exp(\tilde{\sigma}_{js} + W_{ij,s}\psi_s)$. In the right-hand panel, $\beta_s$ and $\psi_s$ include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 
Table A16: Subsample first stages using reverse-sample and whole-sample instruments

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Reverse-sample judge approval rate as instrument</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reverse-sample instrument</td>
<td>0.926***</td>
<td>0.732***</td>
<td>0.681***</td>
<td>0.756***</td>
<td>0.733***</td>
<td>0.791***</td>
<td>0.256</td>
<td>0.481***</td>
</tr>
<tr>
<td>(0.112)</td>
<td>(0.058)</td>
<td>(0.091)</td>
<td>(0.119)</td>
<td>(0.092)</td>
<td>(0.094)</td>
<td>(0.252)</td>
<td>(0.032)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>20274</td>
<td>37549</td>
<td>37730</td>
<td>20093</td>
<td>48868</td>
<td>8955</td>
<td>32111</td>
<td>24831</td>
</tr>
<tr>
<td><strong>Panel B: Whole-sample judge approval rate as instrument</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole-sample instrument</td>
<td>0.992***</td>
<td>0.889***</td>
<td>0.937***</td>
<td>0.916***</td>
<td>0.937***</td>
<td>0.872***</td>
<td>1.078***</td>
<td>0.769***</td>
</tr>
<tr>
<td>(0.083)</td>
<td>(0.043)</td>
<td>(0.077)</td>
<td>(0.100)</td>
<td>(0.049)</td>
<td>(0.084)</td>
<td>(0.091)</td>
<td>(0.033)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>20274</td>
<td>37549</td>
<td>37730</td>
<td>20093</td>
<td>48868</td>
<td>8955</td>
<td>32992</td>
<td>24831</td>
</tr>
</tbody>
</table>

Includes controls for year and office. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 

43